ANÁLISIS DE DATOS MULTIVARIANTES

Daniel Peña

23 de enero de 2002

Índice General

0.1 Prefacio 10

1 INTRODUCCIÓN **13**

1.1 EL ANÁLISIS DE DATOS MULTIVARIANTES 13

1.2 ESTRUCTURA DEL LIBRO 15

1.3 PROGRAMAS DE ORDENADOR 18

1.4 UN POCO DE HISTORIA 18

1.5 LECTURAS COMPLEMENTARIAS 21

2 ÁLGEBRA MATRICIAL **23**

2.1 INTRODUCCIÓN 23

2.2 VECTORES 24

2.2.1 Definiciones básicas 25

2.2.2 Dependencia Lineal 29

2.3 MATRICES 31

2.3.1 Definiciones básicas 32

2.3.2 Productos entre matrices 33

2.3.3 Rango de una matriz 35

2.3.4 Matrices Cuadradas 36

2.3.5 Matrices Particionadas 45

2.4 VECTORES Y VALORES PROPIOS 46

2.4.1 Definición 47

2.4.2 Valores y vectores propios de matrices simétricas 49

2.4.3 Diagonalización de Matrices Simétricas 52

2.4.4 Raiz cuadrada de una matriz semidefinida positiva 54

2.4.5 Descomposición en valores singulares 56

2.4.6 (*)Diagonalización de Matrices generales 56

2.4.7 (*)Inversas Generalizadas 57

2.5 (*)PROYECCIÓN ORTOGONAL 58

2.5.1 Matrices Idempotentes 58

2.5.2 Proyección Ortogonal 59

2.6 (*)DERIVADAS MATRICIALES 64

3 DESCRIPCIÓN DE DATOS MULTIVARIANTES **67**
3.1 INTRODUCCIÓN . . . 67
3.2 DATOS MULTIVARIANTES . . . 67
3.2.1 Tipos de variables . . . 67
3.2.2 La matriz de datos . . . 68
3.2.3 Análisis univariante . . . 70
3.3 MEDIDAS DE CENTRALIZACIÓN: EL VECTOR DE MEDIAS . . . 72
3.4 LA MATRIZ DE VARIANZAS Y COVARIANZAS . . . 74
3.4.1 Cálculo a partir de la matriz de datos centrados . . . 75
3.4.2 Propiedades . . . 79
3.4.3 Variables redundantes: El caso con Matriz **S** singular . . . 80
3.5 MEDIDAS GLOBALES DE VARIABILIDAD . . . 83
3.5.1 La variabilidad total y la varianza promedio . . . 83
3.5.2 La Varianza Generalizada . . . 83
3.5.3 La variabilidad promedio . . . 85
3.6 VARIABILIDAD Y DISTANCIAS . . . 86
3.6.1 El concepto de distancia . . . 86
3.6.2 La Distancia de Mahalanobis . . . 88
3.6.3 La distancia promedio . . . 89
3.7 MEDIDAS DE DEPENDENCIA LINEAL . . . 91
3.7.1 Dependencia por pares: La matriz de correlación . . . 91
3.7.2 Dependencia de cada variable y el resto: Regresión Múltiple . . . 92
3.7.3 Dependencia directa entre pares: Correlaciones parciales . . . 95
3.7.4 El coeficiente de Dependencia . . . 96
3.8 La matriz de precisión . . . 98
3.9 COEFICIENTES DE ASIMETRÍA Y KURTOSIS . . . 99

4 ANALISIS GRAFICO Y DATOS ATIPICOS **107**
4.1 INTRODUCCIÓN . . . 107
4.2 REPRESENTACIONES GRÁFICAS . . . 107
4.2.1 Histogramas y diagramas de dispersión . . . 107
4.2.2 Representación mediante figuras . . . 111
4.2.3 (*)Representación de Proyecciones . . . 112
4.3 TRANSFORMACIONES LINEALES . . . 114
4.3.1 Consecuencias . . . 114
4.3.2 Estandarización univariante . . . 115
4.3.3 (*)Estandarización multivariante . . . 115
4.4 TRANSFORMACIONES NO LINEALES . . . 117
4.4.1 Simplicidad en las distribuciones . . . 117
4.4.2 Simplicidad en las relaciones . . . 119
4.5 DATOS ATÍPICOS . . . 120
4.5.1 Definición . . . 120
4.5.2 Los efectos de los atípicos . . . 121
4.5.3 (*)Identificación de grupos de atípicos . . . 122

4.6 Lecturas complementarias 125

5 COMPONENTES PRINCIPALES 137
5.1 INTRODUCCIÓN 137
5.2 PLANTEAMIENTO DEL PROBLEMA 138
5.3 CALCULO DE LOS COMPONENTES 141
5.3.1 Cálculo del primer componente 141
5.3.2 Cálculo del segundo componente 144
5.3.3 Generalización 145
5.4 PROPIEDADES DE LOS COMPONENTES 149
5.5 ANÁLISIS NORMADO O CON CORRELACIONES 151
5.6 INTERPRETACIÓN DE LOS COMPONENTES 155
5.6.1 Selección del número de componentes 158
5.6.2 Representación gráfica 159
5.6.3 Datos atípicos 162
5.6.4 Distribución de los componentes 163
5.7 Generalizaciones 171
5.8 Lecturas complementarias 171

6 ESCALADO MULTIDIMENSIONAL 179
6.1 INTRODUCCIÓN 179
6.2 ESCALADOS MÉTRICOS: COORDENADAS PRINCIPALES 180
6.2.1 Construcción de variables a partir de las distancias 180
6.3 Matrices compatibles con métricas euclídeas 183
6.3.1 Construcción de las Coordenadas Principales 186
6.4 RELACIÓN ENTRE COORDENADAS Y COMPONENTES PRINCIPALES 189
6.5 BIPLOTS 190
6.6 ESCALADO NO MÉTRICO 193
6.7 Lecturas complementarias 198

7 ANÁLISIS DE CORRESPONDENCIAS 201
7.1 INTRODUCCIÓN 201
7.2 BÚSQUEDA DE LA MEJOR PROYECCIÓN 202
7.2.1 Proyección de las Filas 203
7.2.2 Proyección de las columnas 210
7.2.3 Análisis Conjunto 211
7.3 LA DISTANCIA JI-CUADRADO 214
7.4 ASIGNACIÓN DE PUNTUACIONES 220
7.5 Lecturas complementarias 225

8 ANÁLISIS DE CONGLOMERADOS 227
8.1 FUNDAMENTOS 227
8.2 MÉTODOS CLÁSICOS DE PARTICIÓN 228
8.2.1 Fundamentos del algoritmo de k-medias 228
8.2.2 Implementación del algoritmo 228

8.2.3 Número de grupos 230
8.3 MÉTODOS JERÁRQUICOS 240
8.3.1 Distancias y Similaridades 240
8.3.2 Algoritmos Jerárquicos 244
8.3.3 Métodos Aglomerativos 244
8.4 CONGLOMERADOS POR VARIABLES 252
8.4.1 Medidas de distancia y similitud entre variables 252
8.5 Lecturas complementarias 253

9 DISTRIBUCIONES MULTIVARIANTES 257
9.1 CONCEPTOS BÁSICOS. 257
9.1.1 Variables aleatorias vectoriales. 257
9.1.2 Distribución conjunta 258
9.1.3 Distribuciones marginales y condicionadas 259
9.1.4 Independencia 262
9.1.5 La maldición de la dimensión 262
9.2 PROPIEDADES DE VARIABLES VECTORIALES 263
9.2.1 Vector de medias 263
9.2.2 Esperanza de una función 264
9.2.3 Matriz de varianzas y covarianzas 264
9.2.4 Transformaciones de vectores aleatorios. 265
9.2.5 Esperanzas de transformaciones lineales 266
9.3 Dependencia entre variables aleatorias 267
9.3.1 Esperanzas condicionadas 267
9.3.2 Varianzas condicionadas 268
9.3.3 Matriz de correlación 269
9.3.4 Correlaciones Múltiples 270
9.3.5 Correlaciones Parciales 270
9.4 LA DISTRIBUCIÓN MULTINOMIAL 271
9.5 LA DISTRIBUCIÓN DE DIRICHLET 273
9.6 LA NORMAL k-DIMENSIONAL 274
9.6.1 Distribuciones condicionadas 277
9.7 DISTRIBUCIONES ELÍPTICAS 281
9.7.1 Distribuciones esféricas 281
9.7.2 Distribuciones elípticas 282
9.8 (*)LA DISTRIBUCIÓN DE WISHART 283
9.8.1 Concepto 283
9.8.2 Propiedades de la distribución 285
9.9 LA T^2 DE HOTELLING 286
9.10 DISTRIBUCIONES MEZCLADAS 288
9.11 Lecturas complementarias 290

10 INFERENCIA CON DATOS MULTIVARIANTES 295
10.1 INTRODUCCIÓN 295
10.2 Fundamentos de la Estimación Máximo Verosimil 295
10.3 Estimación de los parámetros de variables normales p-dimensionales. 297
10.4 El método de la razón de verosimilitudes 299
10.5 Contraste sobre la media de una población normal 301
10.6 Contrastes sobre la matriz de varianzas de una población normal 303
10.6.1 Contraste de un valor particular 304
10.6.2 Contraste de independencia 305
10.6.3 Contraste de esfericidad 305
10.6.4 (*)Contraste de esfericidad parcial 306
10.6.5 Ajustes en la distribución 307
10.7 Contraste de igualdad de varias medias: el Análisis de la Varianza Multivariante 307
10.8 Contrastes de datos atípicos 312
10.9 Contrastes de Normalidad 313
10.9.1 Transformaciones 314
10.10Lecturas recomendadas 316

11 METODOS DE INFERENCIA AVANZADA MULTIVARIANTE 321
11.1 INTRODUCCIÓN 321
11.2 ESTIMACIÓN MV CON DATOS FALTANTES 322
11.2.1 Estimación MV con el algoritmo EM 323
11.2.2 Estimación MV de mezclas 325
11.2.3 Estimación de poblaciones normales con datos ausentes 331
11.3 ESTIMACIÓN ROBUSTA 334
11.4 ESTIMACIÓN BAYESIANA 337
11.4.1 Concepto 337
11.4.2 Distribuciones a priori 339
11.4.3 Cálculo de la Posterior 340
11.4.4 Estimación Bayesiana de referencia en el modelo normal 341
11.4.5 Estimación con información a priori 342
11.5 CONTRASTES BAYESIANOS 344
11.5.1 Conceptos básicos 344
11.5.2 Comparación entre los contraste bayesianos y los clásicos 346
11.6 Selección de Modelos 346
11.6.1 El Criterio de Akaike 346
11.6.2 El criterio BIC 348
11.6.3 Relación entre el BIC y EL AIC 350
11.7 Lecturas complementarias 350

12 ANÁLISIS FACTORIAL 355
12.1 INTRODUCCIÓN 355
12.2 EL MODELO FACTORIAL 356
12.2.1 Hipótesis básicas 356

12.2.2 Propiedades . . . 357
12.2.3 Unicidad del modelo . . . 358
12.2.4 Normalización del modelo factorial . . . 360
12.2.5 Número máximo de factores . . . 361
12.3 EL MÉTODO DEL FACTOR PRINCIPAL . . . 362
12.3.1 Estimación de las comunalidades . . . 363
12.3.2 Generalizaciones . . . 368
12.4 ESTIMACIÓN MÁXIMO VEROSÍMIL . . . 370
12.4.1 Estimación MV de los parámetros . . . 370
12.4.2 Otros métodos de estimación . . . 372
12.5 DETERMINACIÓN DEL NÚMERO DE FACTORES . . . 374
12.5.1 Contraste de verosimilitud . . . 374
12.5.2 Criterios de selección . . . 377
12.6 ROTACIÓN DE LOS FACTORES . . . 379
12.7 ESTIMACIÓN DE LOS FACTORES . . . 381
12.7.1 Los factores como parámetros . . . 381
12.7.2 Los factores como variables aleatorias . . . 382
12.8 DIAGNOSIS DEL MODELO . . . 383
12.9 Análisis Factorial Confirmatorio . . . 386
12.10Relación con componentes principales . . . 388
12.11Lecturas recomendadas . . . 389

13 ANÁLISIS DISCRIMINANTE 397
13.1 INTRODUCCIÓN . . . 397
13.2 CLASIFICACIÓN ENTRE DOS POBLACIONES . . . 398
13.2.1 Planteamiento del Problema . . . 398
13.2.2 Poblaciones Normales: Función lineal discriminante . . . 401
13.2.3 Interpretación Geométrica . . . 402
13.2.4 Cálculo de Probabilidades de error . . . 405
13.2.5 Probabilidades a posteriori . . . 406
13.3 GENERALIZACIÓN PARA VARIAS POBLACIONES NORMALES . . . 407
13.3.1 Planteamiento General . . . 407
13.3.2 Procedimiento operativo . . . 409
13.4 POBLACIONES DESCONOCIDAS. CASO GENERAL . . . 412
13.4.1 Regla estimada de clasificación . . . 412
13.4.2 Cálculo de Probabilidades de error . . . 414
13.5 VARIABLES CANÓNICAS DISCRIMINANTES . . . 415
13.5.1 El caso de dos grupos . . . 415
13.5.2 Varios Grupos . . . 417
13.5.3 Variables canónicas discriminantes . . . 420
13.6 DISCRIMINACIÓN CUADRÁTICA. DISCRIMINACIÓN DE POBLACIONES NO NORMALES . . . 424
13.7 DISCRIMINACIÓN BAYESIANA . . . 427
13.8 Lecturas complementarias . . . 428

14 DISCRIMINACIÓN LOGÍSTICA Y OTROS MÉTODOS DE CLASIFICACIÓN 433
14.1 INTRODUCCIÓN . . . 433
14.2 EL MODELO LOGIT . . . 434
14.2.1 Modelos con respuesta cualitativa . . . 434
14.2.2 El modelo logit con datos normales . . . 436
14.2.3 Interpretación del Modelo Logístico . . . 437
14.3 LA ESTIMACIÓN DEL MODELO LOGIT . . . 438
14.3.1 Estimación MV . . . 438
14.3.2 Contrastes . . . 442
14.3.3 Diagnosis . . . 445
14.4 EL MODELO MULTILOGIT . . . 445
14.5 OTROS MÉTODOS DE CLASIFICACIÓN . . . 446
14.5.1 Árboles de Clasificación . . . 446
14.5.2 Redes Neuronales . . . 449
14.5.3 Métodos no Paramétricos . . . 452
14.5.4 Otros Métodos . . . 454
14.6 Lecturas complementarias . . . 455

15 CLASIFICACIÓN MEDIANTE MEZCLAS DE DISTRIBUCIONES 457
15.1 FUNDAMENTOS . . . 457
15.2 EL METODO de K-MEDIAS para mezclas . . . 458
15.2.1 Número de grupos . . . 460
15.3 ESTIMACIÓN DE MEZCLAS DE NORMALES . . . 464
15.3.1 Las ecuaciones de máxima verosimilitud para la mezcla . . . 464
15.3.2 Resolución mediante el algoritmo EM . . . 466
15.3.3 Aplicación al análisis de conglomerados . . . 468
15.4 MÉTODOS BAYESIANOS . . . 470
15.4.1 Estimación Bayesiana de Mezclas de Normales . . . 470
15.5 MÉTODOS DE PROYECCIÓN . . . 472
15.6 Lecturas complementarias . . . 475

16 CORRELACIÓN CANÓNICA 477
16.1 INTRODUCCIÓN . . . 477
16.2 Construcción de las variables canónicas . . . 478
16.2.1 La primera variable canónica . . . 478
16.3 Las r variables canónicas . . . 481
16.3.1 Propiedades de las variables y correlaciones canónicas . . . 482
16.4 ANÁLISIS MUESTRAL . . . 483
16.5 INTERPRETACIÓN GEOMÉTRICA . . . 487
16.6 CONTRASTES . . . 488
16.7 EXTENSIONES A MÁS DE DOS GRUPOS . . . 490
16.8 RELACIÓN CON OTRAS TÉCNICAS ESTUDIADAS . . . 491
16.9 ANÁLISIS CANÓNICO ASIMÉTRICO . . . 492

16.9.1 Coeficientes de redundancia 493
16.9.2 Análisis canónico asimétrico 494
16.10Lecturas complementarias 495

A Datos 505

0.1 Prefacio

El crecimiento de los ordenadores y la facilidad de adquirir y procesar grandes bancos de datos en todas las ciencias ha estimulado el desarrollo y utilización del análisis estadístico multivariante en muchas disciplinas. En las Ciencias Económicas y empresariales los métodos estadísticos multivariantes se utilizan para cuantificar el desarrollo de un país, determinar las dimensiones existentes entre ingresos y gastos familiares, comprender el comportamiento de los consumidores y medir la calidad de productos y servicios. En Ingeniería para diseñar máquinas inteligentes que reconozcan formas o caracteres, para construir clasificadores que aprendan interactivamente con el entorno y para establecer sistemas de control de procesos. En Ciencias de la computación para desarrollar sistemas de inteligencia artificial. En Medicina para construir procedimientos automáticos de ayuda al diagnóstico. En Psicología para interpretar los resultados de pruebas de aptitudes. En Sociología y Ciencia Política para el análisis de encuestas de actitudes y opiniones sociales y políticas.

Este libro presenta las técnicas actuales más utilizadas del Análisis multivariante. Su contenido se ha seleccionado para que pueda ser útil a distintos tipos de audiencias, pero esta especialmente orientado como texto en un curso orientado a las aplicaciones pero donde se desee proporcionar al estudiante los fundamentos de las herramientas presentadas de manera que se facilite su utilización inteligente conociendo sus posibilidades y limitaciones. Para conseguir este objetivo, el libro incluye numerosos ejemplos de aplicación de la técnicas, pero también presenta con cierto detalle los fundamentos estadísticos de las técnicas expuestas. En la exposición se ha procurado prescindir de los detalles técnicos que tienen más interés para especialistas, y este material se ha presentado en los apéndices y en los ejercicios al final de cada capítulo. Por otro lado, se recomienda que los estudiantes realizen un proyecto donde apliquen los métodos estudiados a sus propios datos, para que adquieran la experiencia práctica que les permitirá utilizarlos después con éxito en su trabajo profesional.

Este libro ha tenido una largo período de gestación. Mi interés por el Análisis Multivariante se lo debo a Rafael Romero, Catedrático en la Universidad Politécnica de Valencia y excelente profesor, de quien aprendí, a finales de los años 70, la potencia de estos métodos como herramientas de investigación empírica y su inmenso campo de aplicación. La primera versión de este libro tenía la mitad del tamaño actual y se redactó por primera vez a finales de los años 80 para un curso de Doctorado en la Universidad Politécnica de Madrid. Desde entonces, cada año el manuscrito ha ido sufrido revisiones y ampliaciones, fruto de su uso como notas de clase en varias universidades, y especialmente en la Universidad Carlos III de Madrid. Estoy agradecido a mis estudiantes del curso de doctorado sobre análisis multivariante que han sugerido muchas mejoras y detectado errores y erratas de versiones anteriores. En esa labor estoy especialmente en deuda con Ana Justel, Juan Antonio Gil, Juan Carlos

Ibañez, Mónica Benito, Pilar Barrios, Pedro Galeano y Rebeca Albacete, por sus numerosas sugerencias y cuidadosa lectura de versiones anteriores de estos capítulos. He tenido también la fortuna de contar con excelentes comentarios de mis colegas Carlos Cuadras, Javier Girón, Jorge Martinez, Alberto Muñoz, Rosario Romera, Juan Romo, Santiago Velilla, George Tiao, Victor Yohai y Rubén Zamar, que me han ayudado a mejorar el texto en muchos aspectos. El libro incorpora resultados recientes, fruto de investigaciones conjuntas con Javier Prieto y Julio Rodriguez, con los que ha sido un placer trabajar y de los que ha aprendido mucho. Además, Julio Rodriguez, me ha ayudado en la preparación de muchos de los ejemplos y ha leído y comentado sucesivas versiones del manuscrito encontrando siempre formas de mejorarlo.

Capítulo 1

INTRODUCCIÓN

1.1 EL ANÁLISIS DE DATOS MULTIVARIANTES

Describir cualquier situación real, por ejemplo, las características físicas de una persona, la situación política en un país, las propiedades de una imagen, el rendimiento de un proceso, la calidad de una obra de arte o las motivaciones del comprador de un producto, requiere tener en cuenta simultáneamente varias variables. Para describir las características físicas de una persona podemos utilizar variables como su estatura, su peso, la longitud de sus brazos y de sus piernas, etc. Para describir la situación política de un país, variables como la existencia o no de un régimen democrático, el grado de participación política de los ciudadanos, el número de partidos y sus afiliados, etc. El análisis de datos multivariantes tienen por objeto el estudio estadístico de varias variables medidas en elementos de una población. Pretende los siguientes objetivos.

1. Resumir el conjunto de variables en una pocas nuevas variables, construidas como transformaciones de las originales, con la mínima pérdida de información.
2. Encontrar grupos en los datos si existen.
3. Clasificar nuevas observaciones en grupos definidos.
4. Relacionar dos conjuntos de variables.

Vamos a explicar estos objetivos. El lector habrá encontrado que la descripción de una realidad compleja donde existen muchas variables se simplifica mediante la construccíon de uno o varios índices o indicadores que la resumen. Por ejemplo, el crecimiento de los precios en una economía se resume en un índice de precios, la calidad de una universidad o de un departamento se resume en unos pocos indicadores y las dimensiones del cuerpo humano se resumen en la ropa de confección en unas pocas variables indicadoras del conjunto. Disponer de estas indicadores tiene varias ventajas: (1) si son pocas podemos representarlas gráficamente y comparar distintos conjuntos de datos o instantes en el tiempo; (2) simplifican el análisis al permitir trabajar con un número menor de variables; (3) si las variables indicadoras pueden interpretarse, podemos mejorar nuestro conocimiento de la realidad estudiada. El análisis multivariante de datos proporciona métodos objetivos para conocer cuántas variables indicadoras, que a veces se denomina factores, son necesarias para describir una realidad compleja y determinar su estructura.

El segundo objetivo es identificar grupos si existen. Si observamos un conjunto de vari-

ables en empresas, esperamos los datos indiquen una división de las empresas en grupos en función de su rentabilidad, su eficacia comercial o su estructura productiva. En muchas situaciones los grupos son desconocidos a priori y queremos disponer de un procedimiento objetivo para obtener los grupos existentes y clasificar las observaciones. Por ejemplo, deseamos construir una tipología de clientes, de votantes o de procesos productivos.

Un tercer objetivo relacionado con el anterior aparece cuando los grupos están bien definidos a priori y queremos clasificar nuevas observaciones. Por ejemplo, queremos clasificar a clientes que solicitan créditos como fiables o no, personas como enfermas o no, o diseñar una máquina que clasifique monedas o billetes en clases prefijadas.

Para alcanzar estos tres objetivos una herramienta importante es entender la estructura de dependencia entre las variables, ya que las relaciones entre las variables son las que permiten resumirlas en variables indicadoras, encontrar grupos no aparentes por las variables individuales o clasificar en casos complejos. Un problema distinto es relacionar dos conjuntos de variables. Por ejemplo, podemos disponer de un conjunto de variables de capacidad intelectual y otros de resultados profesionales y queremos relacionar ambos conjuntos de variables. En particular, los dos grupos de variables pueden corresponder a las mismas variables medidas en dos momentos distintos en el tiempo o en el espacio y queremos ver la relación entre ambos conjuntos.

Las técnicas de análisis multivariante tienen aplicaciones en todos los campos científicos y comenzaron desarrollándose para resolver problemas de clasificación en Biología, se extendieron para encontrar variables indicadoras y factores en Psicometría, Marketing y las Ciencias sociales y han alcanzado una gran aplicación en Ingeniería y Ciencias de la computación como herramientas para resumir la información y diseñar sistemas de clasificación automática y de reconocimiento de patrones. Algunos ejemplos indicativos de sus aplicaciones en distintas disciplinas, muchos de los cuales serán objeto de análisis detallado en este libro, son:

Administración de Empresas: Construir tipologías de clientes.

Agricultura: Clasificar terrenos de cultivo por fotos aéreas.

Arqueología: Clasificar restos arqueológicos.

Biometría: Identicar los factores que determinan la forma de un organismo vivo.

Ciencias de la Computación: Diseñar algoritmos de clasificación automática.

Ciencias de la Educación: Investigar la efectividad del aprendizaje a distancia.

Ciencias del medio ambiente: Investigar las dimensiones de la contaminación ambiental.

Documentación: Clasificar revistas por sus artículos y construir indicadores bibliométricos.

Economía: Identificar las dimensiones del desarrollo económico.

Geología: Clasificar sedimentos.

Historia: Determinar la importancia relativa de los factores que caracterizan los periodos prerevolucionarios.

Ingeniería: Transmitir óptimamente señales por canales digitales.

Lingüística: Encontrar patrones de asociación de palabras.

Medicina: Identificar tumores mediante imágenes digitales.

Psicología: Determinar los factores que componen la inteligencia humana

Sociología y Ciencia Política: Consstruir tipologías de los votantes de un partido.

Algunas de esas aplicaciones han tenido una repercusión importante en la evolución del análisis mutivariante, como veremos en la sección 1.4.

1.2 ESTRUCTURA DEL LIBRO

Los datos de partida para un análisis multivariante están habitualmente en una tabla de dos o mas dimensiones y para trabajar con ellos es muy conveniente considerar la tabla como una o varias matrices. El capítulo 2 presenta los fundamentos matemáticos de álgebra lineal que son necesarios para trabajar con matrices y entender sus propiedades. Este capítulo está diseñado de forma instrumental para proporcionar los conceptos básicos y las técnicas necesarias para los análisis estadísticos presentados en los capítulos posteriores.

El análisis multivariante pude plantearse a dos niveles. En el primero, el objetivo es utilizar sólo los datos disponibles y extraer la información que contienen. Los métodos encaminados a este objetivo se conocen como métodos de **exploración de datos,** y se presentan en la primera parte del libro que cubre los capítulos 3 al 8. A un nivel más avanzado, se pretende obtener conclusiones sobre la población que ha generado los datos, lo que requiere la construcción de un modelo que explique su generación y permita prever lo datos futuros. En este segundo nivel hemos generado conocimiento sobre el problema que va más allá del análisis particular de los datos disponibles. Los métodos encaminados a este objetivo se conocen como métodos de **inferencia,** y se presentan en la segunda parte del libro, capítulos 9 al 16 .

El primer paso en la descripción de datos multivariantes es describir cada variable y comprender la estructura de dependencia que existe entre ellas. Este análisis se presenta en el capítulo 3. Siempre que sea posible conviene utilizar técnicas gráficas para resumir y representar la información contenida en los datos, y analizar la forma de medir las variables para obtener una representación lo más simple posible. Estos problemas se estudian en el capítulo 4. Estos dos capítulos extienden al caso multivariante la descripción habitual de datos estadísticos estudiada en los textos básicos de Estadística.

El problema de resumir o condensar la información de un conjunto de variables se aborda, desde el punto de vista descriptivo, construyendo una nuevas variables indicadoras que sinteticen la información contenida en las variables originales. Existen distintos métodos exploratorios para conseguir este objetivo. Con variables continuas, el método más utilizado se conoce como *componentes principales*, y se estudia en el capítulo 5. Los componentes principales nos indican las dimensiones necesarias para representar adecuadamente los datos. Con ellos podemos hacer gráficos de los datos en pocas dimensiones, con mínima pérdida de información, para entender su estructura subyacente.

El análisis de componentes principales puede generalizarse en dos direcciones: la primera cuando los datos disponibles no corresponden a variables sino a similitudes o semejanzas entre elementos. Interesa entonces investigar cuantas dimensiones tienen estas similitudes, este es el objetivo de las *escalas multidimensionales,* que se estudian en el capítulo 6. La segunda generalización de componentes principales es para datos cualitativos, que se presentan en una tabla de contingencia, y esto conduce al *análisis de correspondencias,* que se presenta en el capítulo 7. Esta técnica permite además cuantificar de forma objetiva atributos cualitativos.

El problema descriptivo de investigar si los elementos de nuestra muestra forman un grupo homogéneo o no, y, en caso de que existan varios grupos de datos identicar que elementos pertenecen a cada uno, se aborda con las herramientas de *métodos de agrupamiento* (cluster methods en inglés). Por ejemplo, supongamos que tenemos una encuesta de los gastos en los hogares españoles. Podemos encontrar que, de manera natural, la estructura de gastos es distinta en los hogares unipersonales que en los que conviven adultos con niños pequeños, y la relación entre distintas variables del gasto puede ser distinta en ambos. Conviene, en estos casos, dividir la muestra en grupos de observaciones homogéneas y estudiarlos separadamente. En otros casos el análisis de la homogeneidad de los datos tiene por objeto construir tipologías: de consumidores, de artistas por su utilización del color, de productos, o del tipo de consultas en una página web. Estos métodos se estudian en el capítulo 8.

Las técnicas descriptivas para resumir, condensar y clasificar datos y relacionar variables se conocen a veces como técnicas de **exploración de datos multivariantes**, y se han popularizado en los últimos años en ingeniería y ciencias de la computación con el nombre de **minería de datos,** nombre que indica la capacidad de estas técnicas para extraer información a partir de la matería prima datos. Los capítulos 3 al 8 forman pues un curso básico de minería de datos. Sin embargo estas herramientas no permiten directamente obtener conclusiones generales respecto al proceso o sistema que genera los datos. Para ello necesitamos los métodos presentados en la segunda parte del libro, que comprende los capítulos, 9 al 16, y allí se aborda el objetivo más ambicioso de crear conocimiento respecto al problema mediante un modelo estadístico.

La construcción de un modelo estadístico requiere el concepto de probabilidad y las herramientas básicas para la construcción de modelos para varias variables se exponen en el capítulo 9. La construcción del modelo requiere estimar los parámetros del modelo a partir de los datos disponibles, y contrastar hipótesis respecto a su estructura. Los fundamentos de la inferencia multivariante se estudian en el capítulo 10. Algunos problemas de estimación multivariante pueden formularse como estimación con valores ausentes, y un método eficiente para llevar a cabo esta estimación, el algoritmo EM, se presenta en el capítulo 11. Este capítulo aborda también la estimación (1) permitiendo la posibilidad de que una pequeña fracción de los datos incluyan errores de medida o datos heterogeneos; (2) incorporando además información a priori respecto a los parámetros. En el primer caso tenemos la estimación robusta y en el segundo la Bayesiana (que, como la clásica, puede además hacerse robusta). Este capítulo considera además el problema de seleccionar el mejor modelo explicativo entre varios posibles.

Los modelos para generar conocimiento mediante una reducción del número de variables se conocen como modelos de *análisis factorial,* y pueden verse como una generalización de los componentes principales. Si podemos reemplazar un conjunto amplio de variables por unos pocos factores o variables latentes, no observables, que permiten prever las variables originales hemos aumentado nuestro conocimiento del problema. En primer lugar, descubrimos el mecanismo generador de los datos, en segundo podemos realizar predicciones respecto a datos no observados pero generados por el mismo sistema. Este es el objeto del análisis factorial que se presenta en el capítulo 12.

El problema de la homogeneidad se aborda desde el punto de vista inferencial según dos puntos de vista principales. El primero es el problema de *clasificación o discriminación*:

Objetivos	Enfoque descriptivo (información)	Enfoque inferencial (conocimiento
Resumir los datos	Descripción de datos (Cap. 3 y 4)	Constr. de modelos (Cap 9, 10 y
Obtener indicadores	Componentes principales (Cap. 5)	Análisis Factorial (Cap. 12)
	Escalas multid. (Cap.6)	
	Análisis de Correspon.(Cap. 7)	
Clasificar	Análisis de Conglomerados (Cap. 8)	A. Discriminante (Cap.13 y 14)
Construir grupos	Análisis de Conglomerados (Cap. 8)	Clas. con mezclas (Cap 15)
Relacionar Conj. variab.	Regres. mul.(Cap 3) y Mult. (Cap 16)	Correlación canónica (Cap. 16)

Tabla 1.1: Clasificación de los métodos multivariantes estudiados en este libro

conocemos que los datos pueden provenir de una de dos (o más) poblaciones conocidas y se desea clasificar un nuevo dato en una de las poblaciones consideradas. Por ejemplo, se desea clasificar declaraciones de impuestos como correctas o fraudulentas, personas como enfermos o sanos, empresas como solventes o no, billetes por su valor en una maquina automática, cartas escritas a mano por su código postal en un máquina clasificadora, preguntas a un contestador telefónico por su contenido. Este es el objetivo de los métodos de análisis discriminante presentados en los capítulos 13 y 14.

El segundo punto de vista es investigar si los datos disponibles han sido generados por una sola o varias poblaciones desconocidas. Los métodos para clasificar las observaciones se conocen como *métodos de clasificación* mediante mezclas, y se estudian en el capítulo 15. Estos métodos generalizan los métodos de agrupamiento estudiados desde el punto de vista descriptivo.

El problema inferencial multivariante de relacionar variables aparece cuando estas se separan en dos conjuntos, y tiene varias variantes:

(1) *Análisis de la varianza multivariante:* el primero incluye variables cualitativas y el segundo variables continuas, y queremos ver el efecto de las cualitativas sobre las del segundo grupo. Por ejemplo, observamos en distintas clases de una universidad, definidas por variables cualitativas como titulación, curso etc, varias variables que miden los resultados de las encuestas de evaluación docente y se desea investigar como los resultados varían en las distintos clases. Este tema se estudia en el capítulo 10 como una aplicación directa de los contrastes estadísticos multivariantes

(2) *Regresión multivariante:* el primer conjunto incluye variables continuas o discretas y queremos utilizarlas para explicar las variables continuas del segundo grupo. Por ejemplo, queremos relacionar un conjunto de variables de inversión con un grupo de variables resultado en distintas empresas. Estos métodos se presentan brevemente en el capítulo 16.

(3) *Correlación canónica:* queremos encontrar indicadores del primer grupo que explique lo más posible a los indicadores de las variables del segundo grupo. El número de relaciones independientes entre los indicadores nos informa respecto a la dimensión de la relación. Por ejemplo, queremos buscar indicadores de la inversión en empresas, medida por un conjunto de variables, que explicen indicadores de resultado, construidos también como resumen de un conjunto de variables de resultados económicos. Estos métodos se desarrollan en el capítulo 16.

La tabla1.1 resume la clasificación de métodos multivariantes estudiados en el libro.

omo

1.3 PROGRAMAS DE ORDENADOR

Es impensable aplicar el análisis multivariante sin utilizar el ordenador y recomendamos al lector que reproduzca los ejemplos y realice los ejercicios del libro con cualquiera de los programas estadísticos disponibles. En el libro se han utilizado, por orden de dificultad,los siguientes:

(1) STATGRAPHICS que permite aplicar las herramientas básicas contenidas en el libro, teniendo buenas capacidades gráficas fáciles de usar.

(2) MINITAB es un programa más completo, también de fácil manejo. Es más completo que el anterior y más comodo para la manipulación de datos y la lectura de ficheros en distintos formatos.

(3) SPSS es un programa más avanzado y con mejores capacidades para el manejo de datos. Está dirigido a investigadores sociales que desean analizar grandes encuestas con variables de distintos tipos y permite mucha flexibilidad en la entrada de los datos y en su manipulación, así como en la presentación de los resultados. Además este programa tiene algoritmos de cálculo bastante fiables y muy contrastados con distintas aplicaciones.

(4) S-PLUS está dirigido a un usuario con buena formación estadística, e incluye muchas rutinas que el lector puede combinar para hacer análisis de datos más a su medida. Puede programarse también fácilmente para implantar nuevos desarrollos, y contiene los métodos más modernos que todavía no se han implantado en SPSS. El programa R es similar a S-PLUS y tiene la ventaja de ser de distribución gratuita.

(5) MATLAB y GAUSS son programas con excelentes capacidades para la manipulación matricial, por lo que son muy recomendables para los lectores que quieran escribir sus propios programas y probar análisis nuevos, no incluidos en los paquetes tradicionales. Tienen la ventaja de la flexibilidad y el inconveniente de que son menos automáticos para análisis tradicionales.

Además de estos programas existen otros muchos paquetes estadísticos, como SAS, BMDP, STATA, etc, que estan también bien adaptados para aplicar las técnicas multivariantes descritas en este libro, pero sobre los que el autor tiene menos experiencia directa.

1.4 UN POCO DE HISTORIA

El primer método para medir la relación estadística entre dos variables es debido a Francis Galton (1822-1911), que introduce el concepto de recta de regresión y la idea de correlación entre variables en su libro *Natural Inheritance*, publicado en 1889 cuando Galton tenía 67 años. Estos descubrimientos surgen en sus investigaciones sobre la transmisión de los rasgos hereditarios, motivadas por su interés en contrastar empíricamente la teoría de la evolución de las especies, propuesta por su primo Charles Darwin en 1859. El concepto de correlación es aplicado en las ciencias sociales por Francis Edgeworth (1845-1926), que estudia la normal multivariante y la matriz de correlación. Karl Pearson (1857-1936), un distinguido estadístico británico creador del famoso contraste ji-cuadrado que lleva su nombre, obtuvo el estimador

del coeficiente de correlación en muestras, y se enfrentó al problema de determinar si dos grupos de personas, de los que se conocen su medidas físicas, pertenen a la misma raza. Este problema intrigó a Harold Hotelling (1885-1973), un joven matemático y economista americano, que, atraído por la Estadística, entonces una joven disciplina emergente, viaja en 1929 a la estación de investigación agrícola de Rothamsted en el Reino Unido para trabajar con el ya célebre científico y figura destacada de la estadística, R. A. Fisher (1890-1962). Hotelling se interesó por el problema de comparar tratamientos agrícolas en función de varias variables, y descubrió las semejanzas entre este problema y el planteado por Pearson. Debemos a Hotelling (1931) el contraste que lleva su nombre, que permite comparar si dos muestras multivariantes vienen de la misma población. A su regreso a la Universidad de Columbia en Nueva York, Truman Kelley, profesor de pedadogía en Harvard, planteó a Hotelling el problema de encontrar los factores capaces de explicar los resultados obtenidos por un grupo de personas en test de inteligencia. Hotelling (1933) inventó los componentes principales, que son indicadores capaces de resumir de forma óptima un conjunto amplio de variables y que dan lugar posteriormente al análisis factorial. El problema de obtener el mejor indicador resumen de un conjunto de variables había sido abordado y resuelto desde otro punto de vista por Karl Pearson en 1921, en su trabajo para encontrar el plano de mejor ajuste a un conjunto de observaciones astronómicas. Posteriormente, Hotelling generaliza la idea de componentes principales introduciendo el análisis de correlaciones canónicas, que permiten resumir simultaneamente dos conjuntos de variables.

El problema de encontrar factores que expliquen los datos fué planteado por primera vez por Charles Spearman (1863-1945), que observó que los niños que obtenían buenas puntuaciones en un test de habilidad mental también las obtenían en otros, lo que le llevó a postular queeran debidas a un factor general de inteligencia, el factor g (Spearman, 1904). L. Thurstone (1887-1955) estudió el modelo con varios factores y escribió uno de los primeros textos de análisis factorial (Thurstone, 1947). El análisis factorial fué considerado hasta los años 60 como una técnica psicométrica con poca base estadística, hasta que los trabajos de Lawley y Maxwell (1971) establecieron formalmente la estimación y el contraste del modelo factorial bajo la hipótesis de normalidad. Desde entonces, las aplicaciones del modelo factorial se han extendido a todas las ciencias sociales. La generalización del modelo factorial cuando tenemos dos conjuntos de variables y unas explican la evolución de las otras es el modelo LISREL, que ha sido ampliamente estudiado por Joreskov (1973), entre otros.

La primera solución al problema de la clasificación es debida a Fisher en 1933. Fisher inventa un método general, basado en el análisis de la varianza, para resolver un problema de discriminación de cráneos en antropología. El problema era clasificar un cráneo encontrado en una excavación arqueológica como perteneciente a un homínido o no. La idea de Fisher es encontrar una variable indicadora, combinación lineal de las variables originales de las medidas del cráneo, que consiga máxima separación entre las dos poblaciones en consideración. En 1937 Fisher visita la India invitado por P. C. Mahalanobis (19***), que había inventado la medida de distancia que lleva su nombre, para investigar las diferentes razas en la India. Fisher percibe enseguida la relación entre la medida de Mahalanobis y sus resultados en análisis discriminante y ambos consiguen unificar estas ideas y relacionarlas con los resultados de Hotelling sobre el contraste de medias de poblaciones multivariantes. Unos años después, un estudiante de Mahalanobis, C. R. Rao, va a extender el análisis de

Fisher para clasificar un elemento en más de dos poblaciones.

Las ideas anteriores se obtienen para variables cuantitativas, pero se aplican poco después a variables cualitativas o atributos. Karl Pearson había introducido el estadístico que lleva su nombre para contrastar la independencia en una tabla de contingencia y Fisher, en 1940, aplica sus ideas de análisis discriminante a estas tablas. Paralelamente, Guttman, en psicometría, presenta un procedimiento para asignar valores numéricos (construir escalas) a variables cualitativas que está muy relacionado con el método de Fisher. Como éste último trabaja en Biometría, mientras Guttman lo hace en psicometría, la conexión entre sus ideas ha tardado más de dos décadas en establecerse. En Ecología, Hill (1973) introduce un método para cuantificar variables cualitativas que esta muy relacionado con los enfoques anteriores. En los años 60 en Francia un grupos de estadísticos y lingüistas estudian tablas de asociación entre textos literarios y J. P. Benzecri inventa el análisis de correspondencias con un enfoque geométrico que generaliza, y establece en un marco común, muchos de los resultados anteriores. Benzecri visita la Universidad de Princeton y los laboratorios Bell donde Carroll y Shepard están desarrollando los métodos de escalado multidimensional para analizar datos cualitativos, que habían sido iniciados en el campo de la psicometría por Torgeson (1958). A su vuelta a Francia, Benzecri funda en 1965 el Departamento de Estadística de la Universidad de París y publica en 1972 sus métodos de análisis de datos cualitativos mediante análisis de correspondencias.

La aparición del ordenador transforma radicalmente los métodos de análisis multivariante que experimentan un gran crecimiento desde los años 70. En el campo descriptivo los ordenadores hacen posible la aplicación de métodos de clasificación de observaciones (análisis de conglomerados o análisis cluster) que se basan cada vez más en un uso extensivo del ordenador. MacQueen (1967) introduce el algoritmo de de k-medias. El primer ajuste de una distribución mezclada fue realizado por el método de los momentos por K. Pearson y el primer algortimo de estimación multivariante es debido a Wolfe (1970). Por otro lado, en el campo de la inferencia, el ordenador permite la estimación de modelos sofisticados de mezclas de distribuciones para clasificación, tanto desde el punto de vista clásico, mediante nuevos algoritmos de estimación de variables ausentes, como el algoritmo EM, debido a Dempster, Laird y Rubin (1977), como desde el punto de vista Bayesiano, con los métodos modernos de simulación de cadenas de Markov, o métodos MC^2(Markov Chain Monte Carlo).

En los últimos años los métodos multivariantes están sufriendo una transformación en dos direcciones: en primer lugar, las grandes masas de datos disponibles en algunas aplicaciones están conduciendo al desarrollo de métodos de aproximación local, que no requieren hipótesis generales sobre el conjunto de observaciones. Este enfoque permite construir indicadores no lineales, que resumen la información por tramos en lugar de intentar una aproximación general. En el análisis de grupos, este enfoque local esta obteniendo también ventajas apreciables. La segunda dirección prescinde de las hipótesis sobre las distribuciones de los datos y cuantifica la incertidumbre mediente métodos de computación intensiva. Es esperable que las crecientes posibilidades de cálculo proporcionadas por los ordenadores actuales amplie el campo de aplicación de estos métodos a problemas más complejos y generales.

1.5 LECTURAS COMPLEMENTARIAS

Existe una excelente colección de textos de análisis multivariante en lengua inglesa. Entre ellos destacaremos Flury (1997), Johnson and Wichern (1998), Mardia, Kent y Bibby (1979), Gnandesikan (1997) y Seber (1984). Estos libros combinan la presentación de resultados teóricos y ejemplos y cubren un material similar al aquí expuesto. Textos más aplicados son Dillon y Goldstein (1984), Flury y Riedwyl (1988) y Hair et al (1995). En español, Cuadras (1991), es un excelente texto. Otras referencias de interés son Escudero (1977), Lebart et al (1985) y Batista y Martinez (1989). Hand et al (2000) es una buena referencia para la relación entre minería de datos y estadística.

El libro de Krzanowski y Marriot (1994, 1995) contiene numerosas referencias históricas del desarrollo de estos métodos. Otros textos más específicos que comentan sobre los orígenes históricos de una técnica y presentan abundantes regerencias son Jackson (1991) para componentes principales, Gower and Hand (1996), para los escalogramas multidimensionales, Greenacre (1984) para el análisis de correspondencias, Hand (1997) para los métodos de clasificación, Harman (1980) y Bartholomew (1995) para el análisis factorial, Bollen (1989) para el modelo LISREL, McLachlan y Basford (1988) para los métodos de clasificación mediante mezclas y Schafer (1997) para el algoritmo EM y los nuevos métodos MC^2 de cálculo intensivo. Efron y Tibshirani (1993) presentan interesantes ejemplos de las posibilidades del bootstrap para el análisis multivariante.

Capítulo 2

ÁLGEBRA MATRICIAL

2.1 INTRODUCCIÓN

La información de partida en el análisis multivariante es una tabla de datos correspondiente a distintas variables medidas en los elementos de un conjunto. La manipulación de estos datos se simplifica mucho utilizando el concepto de matriz y su propiedades, que se presentan en este capítulo. La descripción de datos parte de las posiciones de las observaciones como puntos en el espacio y las ideas que aquí se presentan pueden ayudar al lector a desarrollar la intuición geométrica, de gran ayuda para visualizar y comprender la estructura de los procedimientos del análisis multivariante. Por esta razón, recomendamos al lector dedicar el tiempo necesario para comprender los conceptos básicos presentados en este capítulo. Su estudio puede abordarse con dos objetivos distintos. Para los lectores interesados en las aplicaciones y sin formación previa en álgebra lineal, recomendamos concentrarse en las secciones 2.1, 2.2 y 2.3 y la introducción a la sección 2.4. Para los lectores que hayan seguido ya un curso de álgebra, este capítulo puede servir de repaso de los conceptos básicos y de profundización en el uso de valores y vectores propios y proyecciones ortogonales, que forman la base de muchas de las técnicas estudiadas en este libro.

El concepto principal de este capítulo es el concepto de vector. Un conjunto de n datos numéricos de una variable puede representarse geométricamente asociando cada valor de la variable a una dimensión del espacio n dimensional, obteniendo un punto en ese espacio, y también el vector que une el origen con dicho punto. Esta analogía entre variables y vectores es útil, porque los métodos de descripción estadística de una variable tienen una correspondencia clara con las operaciones básicas que realizamos con vectores.

Cuando en lugar de medir una variable en n elementos observamos en cada elemento los valores de p variables, los datos pueden disponerse en una tabla rectangular con p columnas y n filas, de manera que cada columna tenga los valores de una variable y cada fila los valores de las p variables en cada elemento. Si consideramos cada columna como un vector n dimensional, este conjunto de p vectores se denomina matriz. Así como la descripción univariante se asocia a operar con el vector de datos, la descripción de datos multivariantes se asocia geométricamente a operar con la matriz de datos. En particular, el estudio de la variabilidad y la dependencia lineal entre las p variables conduce al concepto de matrices cuadradas, que son aquellas que contienen el mismo número de filas que de columnas. Las

matrices cuadradas son útiles para representar, por ejemplo, las varianzas y covarianzas o correlaciones entre las p variables, y sobre ellas se definen ciertas funciones escalares, como el determinante y la traza, que veremos tienen una clara interpretación estadística: el determinante es una medida de la dependencia lineal y la traza de la variabilidad, del conjunto de las variables. Además, las matrices cuadradas tienen ciertas propiedades básicas, asociadas al tamaño y la dirección de los vectores que la forman. El tamaño de una matriz está relacionada con sus valores propios, y las direcciones con los vectores propios.

La estimación de parámetros mediante una muestra en modelos lineales puede verse geométricamente como la proyección ortogonal del vector (o vectores) que representa la muestra sobre un subespacio. Por esta razón se presentan con detalle algunos resultados de proyecciones ortogonales que no suelen incluirse en textos introductorios de álgebra lineal. Finalmente, este capítulo incluye algunos resultados básicos de cálculo diferencial con vectores y matrices.

Para favorecer el aprendizaje del material de este capítulo al estudiante que se enfrenta a él por primera vez hemos incluido ejercicios después de cada sección, y recomendamos al lector que intente resolverlos. Las secciones marcadas con un asterístico son algo más avanzadas y pueden saltarse en una primera lectura sin pérdida de continuidad. El lector puede encontrar una explicación más detallada de los conceptos aquí expuestos en cualquier texto de álgebra matricial. Un libro claro en español es Arvesú, Alvarez y Marcellán (1999), y en inglés Hadi (1996) presenta una introducción muy sencilla y fácil de seguir con pocos conocimientos básicos. Searle (1982) y Basilevsky (1983) están especialmente orientados a las aplicaciones estadísticas. Noble y Daniel (1977) es una buena introducción de carácter general.

2.2 VECTORES

Geométricamente un dato numérico puede representarse como un punto en un espacio de dimensión uno. Si elegimos una recta con origen y dirección (positiva o negativa) definidos, podemos asociar a cada punto de la recta la magnitud del segmento que une el origen con el punto. Un conjunto de n datos numéricos puede representarse como n puntos sobre una recta pero también, y esta representación es muy útil, como un punto en el espacio de n dimensiones. En dicho espacio podemos también asociar al conjunto de datos el vector que une el origen de coordenadas con dicho punto. La longitud de un vector se denomina norma.

Existe una correspondencia entre las propiedades del conjunto de datos y las propiedades del vector asociado. La media de los datos es proporcional a la proyección del vector de datos sobre la dirección del vector constante (que se define como el que tiene todas las coordenadas iguales). La desviación típica es la distancia promedio entre el vector de datos y el vector constante. La dependencia lineal entre dos variables se mide por la covarianza. El concepto análogo vectorial es el de producto escalar, que es la herramienta principal para estudiar la posición en el espacio de dos vectores. Con variables estandarizadas la covarianza se reduce al coeficiente de correlación, que es equivalente al producto escalar de dos vectores de norma unitaria.

Cuando consideramos varios vectores, por ejemplo p variables medidas sobre n elementos de una población, el concepto principal es la noción de dependencia lineal. La dependencia

lineal establece cuantas variables realmente distintas tenemos. Por ejemplo, si en un conjunto de variables una de ellas representa salarios en euros y otra los mismos salarios pero medidos en miles de euros, aunque ambas variables no sean idénticas (la primera es siempre mil veces mas grande que la segunda), es claro que ambas miden la misma característica y contienen la misma información: las dos variables son linealmente dependientes, ya que conocida una podemos determinar el valor de la otra. Generalizando esta idea, diremos que p variables son linealmente dependientes si podemos obtener los valores de una cualquiera de ellas mediante una combinación lineal del resto. Por ejemplo, las tres variables, número de hombres, número de mujeres y número de personas (que es la suma de las anteriores), son linealmente dependientes, ya que podemos calcular el valor de cualquiera conocidos los valores de las otras dos.

2.2.1 Definiciones básicas

Un conjunto de n números reales $\mathbf{x}$ puede representarse como un punto en el espacio de n dimensiones, $\Re^n$. Definiremos el vector $\mathbf{x}$ como el segmento orientado que une el origen de coordenadas con el punto $\mathbf{x}$. La dirección es importante, porque no es lo mismo el vector $\mathbf{x}$ que el $-\mathbf{x}$. Con esta correspondencia, a cada punto del espacio en $\Re^n$ le asociamos un vector. Por ejemplo, en la figura 2.1 se representa dos vectores en el plano ($\Re^2$): el vector $\mathbf{x} = \binom{3}{2}$, y el vector $\mathbf{y} = \binom{-1}{0}$. En adelante, representaremos un vector mediante $\mathbf{x}$, para diferenciarlo del escalar x, y llamaremos $\Re^n$ al espacio de todos los vectores de n coordenadas o componentes. En particular, un conjunto de números con todos los valores iguales se representará por un vector **constante,** que es aquel con todas sus coordenadas iguales. Un vector constante es de la forma $c\mathbf{1}$, donde c es cualquier constante y $\mathbf{1}$ el vector con todas sus coordenadas iguales a la unidad.

En Estadística podemos asociar a los valores de una variable en n elementos un vector en $\Re^n$, cuyo componente iésimo es el valor de la variable en el elemento i. Por ejemplo, si medimos las edades de tres personas en una clase y obtenemos los valores 20, 19 y 21 años, esta muestra se representa por el vector tridimensional

$$\mathbf{x} = \begin{bmatrix} 20 \\ 19 \\ 21 \end{bmatrix}$$

La suma (o diferencia) de dos vectores $\mathbf{x}, \mathbf{y}$, ambos en $\Re^n$, se define como un nuevo vector con componentes iguales a la suma (diferencia) de los componentes de los sumandos:

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix}.$$

Es inmediato comprobar que la suma de vectores es asociativa $(\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z})$ y conmutativa $(\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x})$.

La suma de dos vectores corresponde a la idea intuitiva de trasladar un vector al extremo del otro y construir la línea que va desde el origen del primero al extremo del segundo. Por ejemplo, la suma de los vectores $\mathbf{x} = \binom{3}{2}$ e $\mathbf{y} = \binom{-1}{0}$,en la figura 2.1, es el vector $\mathbf{z} = \binom{2}{2}$.

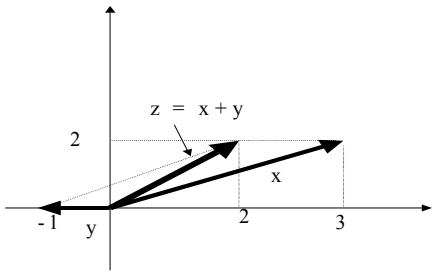


Figura 2.1. Suma de dos vectores

La operación suma (resta) de dos vectores da lugar a otro vector y estadísticamente corresponde a generar una nueva variable como suma (resta) de otras dos anteriores. Por ejemplo, si $\mathbf{x}$ representa el número de trabajadores varones en un conjunto de empresas e $\mathbf{y}$ el número de trabajadoras, la variable $\mathbf{x}+\mathbf{y}$ representa el número total de trabajadores y la variable $\mathbf{x}-\mathbf{y}$ la diferencia entre hombres y mujeres en cada empresa.

El producto de una constante por un vector, es un nuevo vector cuyos componentes son los del vector inicial multiplicados por la constante.

$$\mathbf{z} = k\mathbf{x} = \begin{bmatrix} kx_1 \\ \vdots \\ kx_n \end{bmatrix}.$$

Multiplicar por una constante equivale a un cambio en las unidades de medición. Por ejemplo, si en lugar de medir el número de trabajadores en unidades (variable $\mathbf{x}$) lo hacemos en centenas (variable $\mathbf{z}$) entonces la variable $\mathbf{z}$ es igual a $\mathbf{x}/100$.

Llamaremos **vector transpuesto** $\mathbf{x}'$, de otro $\mathbf{x}$, a un vector con los mismos componentes, pero escritos ahora en fila:

$$\mathbf{x}' = (x_1, ..., x_n).$$

Al transponer un vector columna se obtiene un vector fila. Generalmente los vectores fila se utilizan para describir los valores de p variables distintas en un mismo elemento de una población.

El **producto escalar o interno** de dos vectores $\mathbf{x}, \mathbf{y}$, ambos en $\Re^n$, que escribiremos $\mathbf{x}'\mathbf{y}$ o $\mathbf{y}'\mathbf{x}$, es el escalar obtenido al sumar los productos de sus componentes.

$$\mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x} = \sum_{i=1}^{n} x_i y_i.$$

Se llamará **norma** o longitud de un vector $\mathbf{x}$, a la raíz cuadrada del producto escalar $\mathbf{x}'\mathbf{x}$. Se escribe $\|\mathbf{x}\|$:

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{x_1^2 + \ldots + x_n^2}.$$

La norma es la longitud del segmento que une el origen con el punto $\mathbf{x}$. Por ejemplo, la norma del vector $\mathbf{x}$ en la figura 2.1 es

$$\|\mathbf{x}\| = \sqrt{3^2 + 2^2} = \sqrt{13}$$

que corresponde a la longitud de la hipotenusa en el triángulo rectángulo formado por el vector y sus proyecciones sobre los ejes.

El producto escalar de dos vectores puede calcularse también como el producto de las normas de los vectores por el coseno del ángulo que forman. Para ilustrar este concepto consideremos la figura 2.1 donde se representan los vectores $\mathbf{x} = \binom{a}{0}$ y $\mathbf{y} = \binom{a}{c}$. Observemos que el producto escalar es $\mathbf{x}'\mathbf{y} = a^2$ y que este mismo resultado se obtiene multiplicando la norma de ambos vectores, $\|\mathbf{x}\| = a$ y $\|\mathbf{y}\| = \sqrt{a^2 + c^2}$ por el coseno del ángulo θ que forma, dado por $a/\sqrt{a^2 + c^2}$. Observemos que el producto escalar puede también expresarse como el producto de la norma de un vector por la proyección del otro sobre él. Si uno de los vectores tiene norma uno, el producto escalar es directamente la proyección sobre él del otro vector.

Generalizando esta idea, se define el **ángulo** entre dos vectores $\mathbf{x}$, $\mathbf{y}$ por la relación:

$$\cos\theta = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}.$$

Si dos variables tiene media cero, el coseno del ángulo que forman dos vectores es su coeficiente de correlación. Como $\cos\theta \leq 1$, se demuestra en general que:

$$|\mathbf{x}'\mathbf{y}| \leq \|\mathbf{x}\|\,\|\mathbf{y}\|\,.$$

que se conoce como la **desigualdad de Schwarz**.

Dos vectores son **ortogonales**, o perpendiculares, si y sólo si su producto escalar es cero. Por la definición de ángulo

$$\mathbf{x}'\mathbf{y} = \|\mathbf{x}\|\,\|\mathbf{y}\|\cos\theta,$$

siendo θ el ángulo que forman los vectores. Si $\theta = 90°$ el coseno es cero y también lo será el producto escalar.

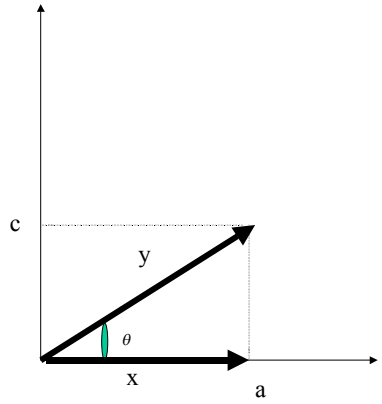


Figura 2.1: Coseno del ángulo entre dos vectores

El producto escalar tiene una clara interpretación estadística. Para describir una variable tomamos su media. Para describir un vector podemos tomar su proyección sobre el vector constante. El vector constante de modulo unitario en dimensión n es $\frac{1}{\sqrt{n}}\mathbf{1}$, y la proyección de $\mathbf{x}$ sobre este vector $\frac{1}{\sqrt{n}}\mathbf{1}'\mathbf{x} = \sum x_i/\sqrt{n} = \overline{x}\sqrt{n}$. El vector constante resultante de esta proyección es $\frac{1}{\sqrt{n}}\mathbf{1}(\overline{x}\sqrt{n}) = \overline{x}\mathbf{1}$. Por tanto, la media es el escalar que define el vector obtenido al proyectar el vector de datos sobre la dirección constante. También puede interpretarse como la norma estandarizada del vector obtenido al proyectar los datos en la dirección del vector constante, donde para estandarizar la norma de un vector dividiremos siempre por $\sqrt{n}$, siendo n la dimensión del espacio.

La variabilidad de los datos se mide por la desviación típica, que es la distancia entre el vector de datos y el vector constante. La proyección del vector de datos sobre la dirección del vector constante produce el vector $\overline{x}\mathbf{1}$, y la norma del vector diferencia, $\mathbf{x}-\overline{x}\mathbf{1}$, mide la distancia entre el vector de datos y el vector constante. Tomando la norma estandarizada, dividiendo por la raíz de la dimensión del espacio

$$\frac{1}{\sqrt{n}}\left\|\mathbf{x}-\overline{x}\mathbf{1}\right\| = \sqrt{\frac{\sum(x_i-\overline{x})^2}{n}}$$

La medida de dependencia lineal entre dos variables, $\mathbf{x}, \mathbf{y}$, es la covarianza. La covarianza es el producto escalar estandarizado de los dos vectores medidos en desviaciones a la media, o tomando sus diferencias respecto al vector constante. Si promediamos el producto escalar de estos vectores

$$\frac{1}{n}(\mathbf{x}-\overline{x}\mathbf{1})'(\mathbf{y}-\overline{y}\mathbf{1}) = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{n}$$

se obtiene directamente la covarianza. Para variables con media cero, el producto escalar promedio de los dos vectores que representan sus valores normalizado es directamente la

covarianza. Para variables estandarizadas, de media cero y desviación típica unidad, la covarianza es el coeficiente de correlación. Para vectores de norma unitaria, el producto escalar es el coseno del ángulo que forman, que es la interpretación geométrica del coeficiente de correlación. La implicación estadística de ortogonalidad es incorrelación. Si dos variables son ortogonales, es decir los vectores que las caracterizan forman un ángulo de 90 grados, llamando r al coeficiente de correlación como $r = \cos\theta = 0$, las variables están incorreladas.

2.2.2 Dependencia Lineal

Un conjunto de vectores $\mathbf{x}_1, ..., \mathbf{x}_p$ es **linealmente dependiente** si existen escalares $c_1, ..., c_p$, no todos nulos, tales que:

$$c_1\mathbf{x}_1 + ... + c_p\mathbf{x}_p = \mathbf{0}$$

donde $\mathbf{0}$ representa el vector nulo que tiene todos los componentes iguales a cero. En particular el vector de ceros, $\mathbf{0}$, es siempre linealmente dependiente de cualquier otro vector $\mathbf{x}$ no nulo. En efecto, aplicando la definición siempre podemos escribir para cualquier escalar c no nulo

$$0\mathbf{x}+c\mathbf{0} = \mathbf{0}$$

Intuitivamente, si los vectores son linealmente dependientes podemos expresar alguno de ellos como combinación lineal de los demás. Por ejemplo, supuesto $c_1 \neq 0$ y llamando $a_i = c_i/c_1$, tenemos

$$\mathbf{x}_1 = a_2\mathbf{x}_2 + \ldots + a_p\mathbf{x}_p.$$

Si un conjunto de vectores no es linealmente dependiente diremos que los vectores son **linealmente independientes**. En el espacio $\Re^p$ el número máximo de vectores linealmente independientes es p. En efecto, si tenemos un conjunto de $p + h$ vectores donde existen, al menos, p linealmente independientes $(\mathbf{x}_i, i = 1, ..., p)$ podemos expresar cualquier otro vector del conjunto, $\mathbf{x}_{p+1}$, como

$$\mathbf{x}_{p+1} = \sum_{i=1}^{p} a_i\mathbf{x}_i,$$

y resolviendo este sistema de p ecuaciones y p incógnitas obtendremos los coeficientes a_i. Por tanto, el máximo número de vectores linealmente independientes es p.

En Estadística un conjunto de vectores linealmente independientes corresponde a un conjunto de variables que no están relacionadas linealmente de forma exacta. Por ejemplo, si dos variables miden la misma magnitud pero en unidades distintas serán linealmente dependientes. También serán linealmente dependientes si el conjunto de variables incluye una que se ha generado como una combinación lineal de las otras (por ejemplo, tenemos p variables que representan los precios en euros de p productos en n países de Europa $(n > p)$ y se incluye también como variable $p+1$ el precio ponderado de estos productos en los mismos países).

Dado un conjunto de p vectores linealmente independientes $(\mathbf{x}_1, ..., \mathbf{x}_p)$, en $\Re^n$ $(p \leq n)$, llamaremos **espacio generado** por este conjunto de vectores al espacio que contiene todos los vectores $\mathbf{z}$, en $\Re^n$, que pueden expresarse como combinación lineal de éstos. El conjunto $(\mathbf{x}_1, ..., \mathbf{x}_p)$ se llama **base generadora del espacio**, o simplemente **base** del espacio. Si $\mathbf{z}$ pertenece a este espacio:

$$\mathbf{z} = c_1\mathbf{x}_1 + ... + c_p\mathbf{x}_p.$$

Es fácil comprobar que $\mathbf{z}$ estará en un espacio de dimensión p: en efecto, podemos tomar las primeras p coordenadas de $\mathbf{z}$ y obtener los coeficientes $c_1, ..., c_p$ del sistema de p ecuaciones y p incógnitas resultante. Las $n - p$ coordenadas siguientes de $\mathbf{z}$ quedan determinadas, al estarlo los c_i, por lo que, obviamente, $\mathbf{z}$ sólo tiene p componentes independientes, estando, por lo tanto, en un espacio de dimensión p. El espacio generado por un conjunto de variables incluye a todas las variables que pueden generarse como índices o combinaciones lineales de las originales.

La **dimensión** de un espacio E_p se define como el número de vectores linealmente independientes que lo generan.

Diremos que un vector $\mathbf{x}$ es **ortogonal a un subespacio** E_p si $\mathbf{x}$ es ortogonal a todo vector de E_p, es decir, si $\mathbf{y}$ pertenece al subespacio E_p, que escribiremos $\mathbf{y} \in E_p$, entonces:

$$\mathbf{y}'\mathbf{x} = 0.$$

Llamaremos **complemento ortogonal** de un subespacio E_p, de dimensión p, y lo denotaremos por $C(E_p)$, al espacio que contiene todos los vectores ortogonales a E_p. Entonces, si $\mathbf{x} \in E_p$, $\mathbf{y} \in C(E_p)$ se verifica $\mathbf{x}'\mathbf{y} = 0$. La dimensión de $C(E_p)$ será $n - p$. En particular el complemento ortogonal del espacio generado por un vector que contiene todos los vectores ortogonales a él se denomina **espacio nulo** del vector.

Ejercicios 2.2

2.2.1 Dados los tres vectores
$\mathbf{a} = \binom{1}{2}$, $\mathbf{b} = \binom{2}{1}, \mathbf{c} = \binom{-2}{1}$

a) Representarlos en el plano $\Re^2$.

b) Calcular los vectores suma y diferencia de $\mathbf{a}$ y $\mathbf{b}, \mathbf{a} \pm \mathbf{b}$.

c) Calcular la norma de los tres vectores.

d) Calcular los productos escalares,$\mathbf{ab}$, $\mathbf{bc}, \mathbf{ac}$.¿ Qué podemos deducir de estos productos?

e) Calcular la proyección del vector $\mathbf{a}$ sobre el $\mathbf{b}$.

f) Justificar si los tres vectores son linealmente independientes. Si no lo son, expresar uno cualquiera como combinación lineal de los otros dos.

2.2.2 En $\Re^3$ se denomina base canónica a la formada por los vectores $\mathbf{a} = (1, 0, 0)', \mathbf{b} = (0, 1, 0)'$, y $\mathbf{c} = (0, 0, 1)$. Se pide

a) Expresar el vector $\mathbf{d} = (1, 1, 2)'$, como suma de los vectores de la base canónica.

b) Calcular la proyección del vector $\mathbf{d}$ sobre cada uno de los vectores de la base canónica.

c) Calcular el coseno del ángulo entre el vector $\mathbf{d}$ y los vectores de la base canónica.

d) Indicar la dimensión del espacio generado por el vector $\mathbf{d}$ y obtener una base del complemento ortogonal a ese espacio.

2.2.3 Dados los vectores en $\Re^3, \mathbf{a} = (1, 0, 2)'$, $\mathbf{b} = (1, 1, 2)'$, $\mathbf{c} = (2, 1, 6)'$.

a) Calcular los vectores $-\mathbf{b}$, $\mathbf{a} + \mathbf{c}$, y $\mathbf{b} + \mathbf{c}$

b) Calcular la norma de los vectores, 4**a** y -2**c**.

c) Calcular el producto escalar, **ab** y **bc**.

d) Calcular la proyección del vector **a** sobre el **b**.

2.2.4 Calcular la dimensión del espacio generado por los tres vectores del ejercicio anterior

a) ¿Pertenece el vector $\mathbf{d} = (-2, 0, -8)'$ al espacio generado por estos tres vectores? Si es así expresarlo como suma de una base del espacio.

b) Indicar la dimensión del espacio complemento ortogonal al generado por estos tres vectores.

c) Encontrar una base del complemento ortogonal.

d) Calcular el coseno de ángulo entre los vectores **d** y **a**.

2.2.5 Dados los tres vectores $\mathbf{a} = (1, 0, 0, 0, 1)', \mathbf{b} = (1, 1, 0, 0, 0)'$, y $\mathbf{c} = (0, 0, 0, 1, 1)$, en $\Re^5$.

a) Indicar la dimensión del espacio generado por estos vectores y obtener un nuevo vector miembro de ese espacio.

b) Calcular la dimensión del espacio complemento ortogonal al generado por esos vectores.

c) Calcular una base del espacio complemento ortogonal.

d) Demostrar que los vectores $\mathbf{a} + \mathbf{b}, \mathbf{a} + \mathbf{c}$, y $\mathbf{b} + \mathbf{c}$ también son linealmente independientes.

2.2.6 Considerar las 9 variables que definen los productos alimenticios en los datos EUROALI del apéndice de datos como 9 vectores en un espacio de dimensión 25. Se pide:

a) Calcular el vector proyección de cada vector sobre el vector de constantes.

b) Calcular la distancia entre cada vector y el vector de constantes.

c) Calcular el producto escalar de los vectores correspondientes a las variables CR y CB.

d)Calcular el coseno del ángulo que forman los vectores CR y CB.

2.2.7 Considerar cada país de los datos EUROALI del apéndice de datos como un vectores en un espacio de dimensión 9. Se pide:

a) Indicar si estos vectores son linealmente independientes

b) Justificar que el número máximo de vectores linealmente independientes es ocho.

c) Calcular e interpretar el producto escalar entre Austria y Bélgica.

d) Determinar el ángulo que forman Austria y Bélgica.

e) Calcular la distancia de cada país al vector de constantes. Interpretar el resultado.

2.3 MATRICES

Para trabajar conjuntamente con p variables o vectores definimos el concepto de matriz. Una matriz es un conjunto de números dispuestos en filas y columnas y puede verse como un conjunto de vectores columna o un conjunto de vectores fila. Diremos que una matriz tiene dimensiones $n \times p$ si tiene n filas y p columnas. Si en una matriz intercambiamos las filas por las columnas, se obtiene una nueva matriz que se denomina la traspuesta de la primera. En particular, un vector columna de orden n es una matriz de dimensiones $n \times 1$(su traspuesta es un vector fila), y un escalar es una matriz de dimensiones 1×1 (e igual a su traspuesta).

La generalización del concepto de producto escalar entre dos vectores es el producto matricial, que se define como una nueva matriz que contiene todos los productos escalares entre los vectores fila de la primera matriz y los vectores columna de la segunda. Para que este producto sea posible la primera matriz tiene que tener tantas columnas como filas la segunda. Por la propia definición de deduce que este producto no es conmutativo. Diremos que premultiplicamos la matriz $\mathbf{A}$ por la $\mathbf{B}$ cuando realizamos el producto $\mathbf{BA}$ y que postmultiplicamos la $\mathbf{A}$ por la $\mathbf{B}$ si realizamos el producto $\mathbf{AB}$. Un producto matricial que puede siempre aplicarse entre dos matrices cualesquiera es el producto de Kronecker.

Una propiedad básica de una matriz es el rango, que indica el número máximo de vectores fila o columna linealmente independientes que la forman. En una matriz de n filas y p columnas ($n > p$), sus p columnas pueden ser vectores linealmente independientes en $\Re^n$, pero sus n filas no, ya los vectores fila pertenecen a $\Re^p$ donde sólo pueden existir $p < n$ vectores fila linealmente independientes. El rango máximo de la matriz es p y cuando esto ocurre decimos que la matriz tienen rango completo. El rango de una matriz es igual al de su traspuesta.

Las matrices cuadradas son aquellas que tienen el mismo número de filas que de columnas. Las matrices cuadradas tienen ciertas propiedades similares a los escalares. Podemos definir la matriz inversa, y existen distintas formas de obtener una medida escalar de una matriz cuadrada. La primera es la traza, la segunda el determinante y la tercera construir una forma cuadrática a partir de la matriz. Veremos en el capítulo siguiente que todas estas propiedades tienen una interpretación estadística en el análisis de datos multivariantes.

2.3.1 Definiciones básicas

Llamaremos matriz, $\mathbf{A}$, de dimensiones $(n \times p)$ a un conjunto de $n \times p$ números reales, ordenados en n filas y p columnas. Por ejemplo, si medimos p variables en n individuos de una población podemos representar cada variable por un vector columna de dimensión n y el conjunto de datos muestrales será una matriz $n \times p$. En particular, cada vector columna es pues una matriz $(n \times 1)$. Una matriz $(n \times p)$, puede verse como un conjunto de p vectores columna en $\Re^n$, o como un conjunto de n vectores fila en $\Re^p$. Llamaremos **matriz traspuesta** $\mathbf{A}'$ a la matriz obtenida a partir de $\mathbf{A}$ intercambiando filas por columnas. Si $\mathbf{A}$ es $n \times p$, $\mathbf{A}'$ será $p \times n$. Se verifica:

$$(\mathbf{A}')' = \mathbf{A}.$$

La suma de dos matrices se define sólo cuando ambas tienen las mismas dimensiones. Cada elemento de la matriz suma se obtiene sumando los elementos correspondientes de los sumandos

$$\mathbf{A} + \mathbf{B} = \mathbf{C} \Rightarrow \begin{bmatrix} a_{11} & \dots & a_{1p} \\ ... & & ... \\ a_{n1} & \dots & a_{np} \end{bmatrix} + \begin{bmatrix} b_{11} & \dots & b_{1p} \\ ... & & ... \\ b_{n1} & \dots & b_{np} \end{bmatrix} = \begin{bmatrix} c_{11} & \dots & c_{1p} \\ ... & & ... \\ c_{n1} & \dots & c_{np} \end{bmatrix}$$

con $c_{ij} = a_{ij} + b_{ij}$. Se verifica:

(a) $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$

(b) $(\mathbf{A}+\mathbf{B})' = \mathbf{A}'+\mathbf{B}'$.

Sumar dos matrices equivale en términos estadísticos a sumar los valores de las variables correspondientes a las columnas de las matrices. Por ejemplo, si la matriz $\mathbf{A}$ representa el número de incidencias leves de p clases distintas en una empresa en n semanas y la $\mathbf{B}$ el número de incidencias graves en las mismas semanas, la suma representa el número total de incidencias.

2.3.2 Productos entre matrices

Vamos a estudiar dos tipos de productos entre matrices. El primero y más importante es el **producto matricial,** lo representaremos por $\mathbf{AB}$ y sólo es posible cuando el número de columnas de $\mathbf{A}$ es igual al número de filas de $\mathbf{B}$. Entonces, si $\mathbf{A}(n \times p)$ y $\mathbf{B}(p \times h)$, el producto es una matriz $\mathbf{C}(n \times h)$ con términos:

$$c_{ij} = \sum_{m=1}^{p} a_{im} b_{mj}$$

Es decir, el término c_{ij} representa el producto escalar del vector $\mathbf{a}'_i$, definido por la i-ésima fila de $\mathbf{A}$, por el vector $\mathbf{b}_j$, de la j-ésima columna de $\mathbf{B}$. Si escribimos:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}'_1 \\ \vdots \\ \mathbf{a}'_n \end{bmatrix} \qquad \mathbf{B} = [\mathbf{b}_1 ... \mathbf{b}_h]$$

donde todos los vectores tienen dimensiones p, el producto matricial de estas dos matrices es:

$$\mathbf{AB} = \mathbf{C} = \begin{bmatrix} \mathbf{a}'_1\mathbf{b}_1 & \dots & \mathbf{a}'_1\mathbf{b}_h \\ \vdots & & \vdots \\ \mathbf{a}'_n\mathbf{b}_1 & \dots & \mathbf{a}'_n\mathbf{b}_h \end{bmatrix}_{(n \times h)} .$$

Observemos que el producto de dos matrices no es en general conmutativo, ya que si $\mathbf{AB}$ existe (el número de columnas de $\mathbf{A}$ es igual al número de filas de $\mathbf{B}$), el producto $\mathbf{BA}$ puede no existir. Además, cuando existe, el producto $\mathbf{AB}$ es, en general, distinto de $\mathbf{BA}$.

En particular, el producto de una matriz $(n \times p)$ por un vector $(p \times 1)$, $\mathbf{Ax}$, será un nuevo vector de dimensión $(n \times 1)$ cuyos componentes se obtienen por el producto escalar de las filas de $\mathbf{A}$ por el vector $\mathbf{x}$. Si

$$\mathbf{y} = \mathbf{Ax},$$

la matriz $\mathbf{A}$ transforma un vector $\mathbf{x}$ en $\Re^p$ en otro vector $\mathbf{y}$ en $\Re^n$. Como veremos más adelante, los movimientos y deformaciones de vectores en el espacio son el resultado de multiplicar el vector por una matriz.

Definimos la **matriz identidad** de dimensión n, $\mathbf{I}_n$, como la matriz de dimensiones $n \times n$ que tiene unos en las posiciones ii y ceros fuera de ella. En general la dimensión está clara

por el contexto y utilizaremos la letra $\mathbf{I}$ para representar la matriz identidad de cualquier dimensión:

$$\mathbf{I} = \begin{bmatrix} 1 & \dots & 0 \\ \vdots & 1 & \vdots \\ 0 & \dots & 1 \end{bmatrix}.$$

El producto matricial tiene, entre otras, las propiedades siguientes, donde suponemos que las matrices tienen las dimensiones adecuadas para que los productos están definidos:

(a) $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$

(b) $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}$

(c) $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$

(*)Producto de Kronecker

El **producto de Kronecker** nos resuelve el problema de construir matrices grandes cuyos elementos son matrices dadas más pequeñas y se define para matrices cualesquiera. Dadas dos matrices $\mathbf{A}_{k\times n}$ y $\mathbf{B}_{p\times q}$, su producto de Kronecker, que representaremos con el símbolo $\otimes$, se efectúa multiplicando cada elemento de la primera por todos los elementos de la segunda, de manera que la matriz resultante tiene un número de filas igual al producto de las filas, kp, y un número de columnas igual al producto de las columnas, nq. Este producto existe siempre sean cual sean las dimensiones de las matrices, y se representa por :

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{k1}\mathbf{B} & a_{k2}\mathbf{B} & \dots & a_{kn}\mathbf{B} \end{bmatrix}.$$

donde la matriz producto es de orden $kp \times nq$. Por ejemplo,

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 3 \\ 2 & 0 & 6 \end{bmatrix}$$

Las propiedades siguientes son resultado directo de la definición:

(a) si c es un escalar $c\otimes\mathbf{A} = \mathbf{A}\otimes c = c\mathbf{A}$.

(b) si $\mathbf{x}$ e $\mathbf{y}$ son vectores:

$$\mathbf{x}\otimes\mathbf{y}' = \mathbf{y}'\otimes\mathbf{x}$$

(c) $(\mathbf{A}\otimes\mathbf{B})' = \mathbf{A}'\otimes\mathbf{B}'$

(d) $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = \mathbf{AC}\otimes\mathbf{BD}$, supuesto que los productos $\mathbf{AC}$ y $\mathbf{BD}$ existen.

En estadística el producto de Kronecker se utiliza para construir matrices cuyos elementos son a su vez matrices, con frecuencia repetidas. Por ejemplo, si queremos construir una matriz que tenga como elementos diagonales la matriz $\mathbf{A}$, definimos el producto

$$\mathbf{I}_3 \otimes \mathbf{A} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A} \end{bmatrix}$$

donde si $\mathbf{I}_3$ es la matriz identidad y $\mathbf{0}$ es una matriz de ceros, ambas de dimensiónes 3×3.

2.3.3 Rango de una matriz

Una propiedad básica de una matriz es el **rango**, que indica el número máximo de vectores fila o columna linealmente independientes que contiene la matriz. En una matriz de $n \times p$, suponiendo $n > p$, el máximo número de vectores linealmente independientes es p. En efecto, si consideramos los vectores formados por las p columnas, tenemos p vectores en $\Re^n$, que pueden ser linealmente independientes. Sin embargo, si consideramos los n vectores fila, estos son vectores de $\Re^p$, y el máximo número de vectores independientes en este espacio es p. Por tanto, el rango máximo de la matriz es p, y cuando esto ocurre decimos que la matriz es de rango completo. Por la definición es inmediato que el rango de una matriz y de su transpuesta es el mismo.

En general, si llamamos $rg(\mathbf{A})$ al rango de la matriz $\mathbf{A}$ se verifica:

1. $rg(\mathbf{A}_{n\times p}) \leq \min(n, p)$. El rango es igual o menor que el menor de n y p.
2. Si $rg(\mathbf{A}_{n\times p}) = n < p$ o $rg(\mathbf{A}_{n\times p}) = p < n$, se dice que $\mathbf{A}$ es de rango completo.
3. $rg(\mathbf{A} + \mathbf{B}) \leq rg(\mathbf{A}) + rg(\mathbf{B})$.
4. $rg(\mathbf{AB}) \leq$ mínimo$(rg(\mathbf{A}), rg(\mathbf{B}))$
5. $rg(\mathbf{A}'\mathbf{A}) = rg(\mathbf{A}\mathbf{A}') = rg(\mathbf{A})$.

Las dos primeras propiedades resultan de la definición. Es fácil comprobar que el rango de la suma no puede ser mayor que la suma de rangos. Por ejemplo en la suma

$$\begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

la primera matriz tiene rango dos, (los dos vectores columna no nulos son linealmente independientes), la segunda rango uno (solo un vector es linealmente independiente) y la suma tiene rango uno.

Si multiplicamos dos matrices, el rango de la matriz resultante no puede exceder a la de menor rango. Por ejemplo, en el producto

$$\begin{bmatrix} 1 & -1 & 1 \\ 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 3 & 2 \end{bmatrix}$$

cada una de las matrices que se multiplican tiene rango dos, pero el producto tiene solo rango uno. Finalmente, si multiplicamos una matriz por su transpuesta el producto tiene el mismo rango que la matriz original. En Estadística el rango de una matriz de datos nos indica la dimensión real necesaria para representar el conjunto de datos, o el número real de variables distintas de que disponemos. Analizar el rango de una matriz de datos es la clave para reducir el número de variables sin pérdida de información.

2.3.4 Matrices Cuadradas

Una matriz es **cuadrada** si $n = p$. Dentro de las matrices cuadradas se llaman simétricas a las que tienen cada fila igual a la correspondiente columna, es decir $a_{ij} = a_{ji}$. Una matriz simétrica es, por tanto, idéntica a su traspuesta, y diremos que $\mathbf{A}$ es **simétrica** si

$$\mathbf{A}' = \mathbf{A}.$$

Una clase de matrices cuadradas y simétricas muy importante son las **matrices diagonales,** que tienen únicamente términos no nulos en la diagonal principal. Un caso particular importante de matriz diagonal es la **matriz identidad o unidad, I**, ya estudiada.

En particular, los productos $\mathbf{AA}'$ y $\mathbf{A}'\mathbf{A}$ conducen a matrices simétricas. Las matrices cuadradas aparecen de manera natural cuando consideramos estos productos en matrices de datos. Si $\mathbf{A}$ es $(n \times p)$ y representa los valores de p variables de media cero en n individuos de una población, la matriz cuadrada de orden p, $\mathbf{A}'\mathbf{A}/n$, va a contener, como veremos en el capítulo siguiente, las varianzas y covarianzas entre las variables. Otra matriz cuadrada y simétrica de amplio uso en estadística es la matriz de correlación, que contiene unos en la diagonal y fuera de ella los coeficientes de correlación entre las variables.

Sobre las matrices cuadradas podemos definir dos medidas escalares que resumen su tamaño global : el determinante y la traza. Ambas son medidas relativas, ya que se modifican si multiplicamos los elementos de la matriz por constantes, como veremos a continuación.

Determinante de una matriz

Dada una matriz $\mathbf{A}$ cuadrada y diagonal con términos a_{ii} se denomina determinante de la matriz, y lo representaremos por $|\mathbf{A}|$, al escalar resultante de multiplicar todos los términos diagonales de la matriz. Supongamos inicialmente una matriz de orden dos como

$$\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}$$

si consideramos las columnas de esta matriz como vectores, cada vector está situado en uno de los ejes coordenados. La figura 2.2 ilustra esta situación. El determinante de esta matriz es 2×4=8, igual al área del rectángulo determinado por ambos vectores.

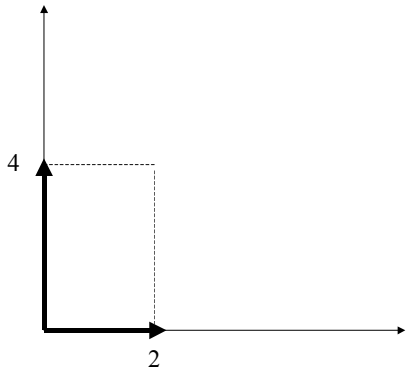


Figura 2.2: El determinante como área encerrada por los vectores columna de la matriz

Generalizando esta idea, dada una matriz $\mathbf{A}$ cuadrada de orden n con términos a_{ij}, se denomina **determinante** de la matriz, y lo representaremos por $|\mathbf{A}|$, al escalar obtenido mediante la suma de todos los productos de n elementos de la matriz, $a_{1i_1}a_{2i_2}, ..., a_{ni_n}$, que podemos formar de manera que en cada producto aparezca una vez un elemento de cada fila y uno de cada columna. Cada término tiene además un signo, que depende del número de cambios entre dos subíndices consecutivos, que es necesario para poner los subíndices $i_1, \ldots, i_n$ de ese término en el orden natural $1, 2, \ldots, n$. Escribiremos :

$$|\mathbf{A}| = \sum (-1)^r \, a_{1i_1}a_{2i_2}, ..., a_{ni_n}$$

donde el sumatorio está extendido a las $n!$ permutaciones de los segundos índices. Los índices $i_1, \ldots i_n$ son una permutación de los números $1, 2, \ldots, n$ y r es el número de cambios entre dos subíndices necesario para ponerlos en el orden $1, 2, \ldots, n$.

Por ejemplo, en la matriz 2×2 el número de permutaciones de los números 1 y 2 es dos ((1,2) y (2,1)). La primera permutación está en el orden natural luego el número de cambios es $r = 0$ y el término $a_{11}a_{22}$ será positivo. La segunda requiere permutar el uno y el dos, con lo que $r = 1$ y el término $a_{12}a_{21}$ será negativo. El determinante será:

$$|\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}.$$

y, como demostraremos más adelante, puede interpretarse de nuevo como el área del paralelogramo determinado por los vectores columna. La situación se ilustra en la figura **??**. Esta interpretación sugiere que si una columna es proporcional a la otra, los dos vectores estarán en la misma dirección y el área encerrada por ambos, que es el determinante de la matriz, será cero. La comprobación de esta propiedad es inmediata: si la primera columna es $(a, b)\prime$ y la segunda$(\lambda a, \lambda b)\prime$ el determinante será $a\lambda b - b\lambda a = 0$.

En una matriz 3×3 el determinante tiene $3! = 6$ términos que se obtiene de las 6 posibles permutaciones:

$$\begin{array}{ccc} 1 & 2 & 3 \\ 1 & 3 & 2 \\ 2 & 1 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \\ 3 & 2 & 1 \end{array}$$

la primera permutación va en el orden natural, luego $r = 0$. Las dos siguientes podemos ponerlos en orden natural con un solo cambio entre índices consecutivos, luego $r = 1$. Las dos siguientes requieren dos cambios (por ejemplo, en la cuarta primero pasamos a 2,1,3 y luego a 1,2,3). Finalmente, en la última son necesarios 3 cambios, con lo que tendrá signo menos. En consecuencia:

$$\begin{aligned} |\mathbf{A}| &= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + \\ &\quad + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}, \end{aligned}$$

y puede demostrarse que ahora el determinante es el volumen del paralepípedo generado por las columnas de la matriz.

Para matrices mayores de 3, la interpretación del determinante como hipervolumen es la misma, pero su cálculo es tedioso. Para obtenerlo utilizaremos el concepto de **menor**. Llamaremos menor del elemento a_{ij} de una matriz cuadrada de orden n, m_{ij}, al determinante de la matriz de orden $n-1$ que resulta al eliminar de la matriz original $\mathbf{A}$ la fila i y la columna j. Se denomina **adjunto** del elemento a_{ij} al escalar $(-1)^{i+j} m_{ij}$. Se demuestra que el determinante de una matriz puede calcularse multiplicando cada elemento de una fila por sus adjuntos. Entonces:

$$|\mathbf{A}| = \sum_{j=1}^{n} a_{ij} (-1)^{i+j} m_{ij}$$

para cualquier fila i. Por ejemplo, en una matriz 3×3, desarrollando por los elementos de la primera fila

$$|\mathbf{A}| = a_{11} (a_{22}a_{33} - a_{23}a_{32}) - a_{12} (a_{21}a_{33} - a_{23}a_{31}) + a_{13} (a_{21}a_{32} - a_{22}a_{31}),$$

que coincide con el resultado anterior. Aplicando sucesivamente esta idea es posible calcular el determinante de matrices grandes.

El determinante se calcula muy fácilmente cuando una matriz es diagonal, ya que entonces, como hemos visto, el determinante es el producto de los términos diagonales de la matriz. El mismo resultado se obtiene si la matriz es **triangular,** que es aquella que tiene todos los elementos por encima o por debajo de la diagonal principal nulos. Por ejemplo, una matriz diagonal de orden tres es

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 1 & 4 & 2 \end{bmatrix}$$

Para calcular el determinante desarrollamos por la primera fila, con lo que obtenemos el producto del primer término diagonal, 1, por su adjunto, que es otra matriz triangular ahora de orden dos. Desarrollando de nuevo este matriz por su primera fila tenemos el producto del segundo término diagonal, 3, por un escalar, 2. Aplicando esta misma idea a matrices de cualquier tamaño comprobamos que el determinante es el producto de los términos diagonales.

Los determinantes tienen las propiedades siguientes:

(a) $|\lambda \mathbf{A}| = \lambda^n |\mathbf{A}|$

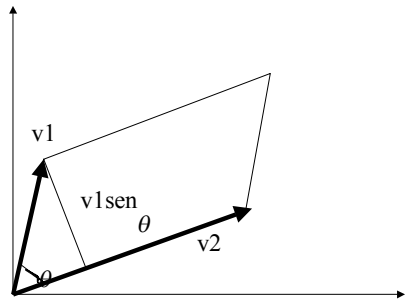
(b) $|\mathbf{A}'| = |\mathbf{A}|$

(c) Si $\mathbf{A}$ y $\mathbf{B}$ son matrices cuadradas, $|\mathbf{AB}| = |\mathbf{A}|\,|\mathbf{B}|$.

(d) Si permutamos dos filas o dos columnas entre sí, el determinante cambia sólo su signo.

(e) Si una fila (o columna) de una matriz es una combinación lineal de las restantes filas (o columnas), lo que supone que su rango es menor que n, la matriz es **singular** y el determinante de la matriz es cero.

θ



El determinante como area del paralelogramo formado por los dos vectores

La figura **??** ilustra la interpretación del determinante como area del paralelogramo definido por los dos vectores. Este area es el producto de la base, $\|v_2\|$, por la altura, $\|v_1\|\text{sen}\theta$. Consideremos el determinante de la matriz

$$|\mathbf{A}| = |\mathbf{C}'\mathbf{C}| = \left| \begin{bmatrix} \mathbf{v}_1' \\ \mathbf{v}_2' \end{bmatrix} [\mathbf{v}_1 \mathbf{v}_2] \right| = \left| \begin{bmatrix} \mathbf{v}_1'\mathbf{v}_1 & \mathbf{v}_1'\mathbf{v}_2 \\ \mathbf{v}_2'\mathbf{v}_1 & \mathbf{v}_2'\mathbf{v}_2 \end{bmatrix} \right|$$

entonces:

$$|\mathbf{A}| = |\mathbf{C}|^2 = \|v_2\|^2 \|v_1\|^2 (\text{sen}\theta)^2$$

y el determinante de la matriz formada por los dos vectores es $\|v_1\| \, \|v_2\| \text{sen}\theta$, el área del paralelogramo que forman. Observemos se obtiene el mismo resultado cuando los vectores $\mathbf{v}_1$ y $\mathbf{v}_2$ son vectores de R^n, ya que la matriz $\mathbf{C}'\mathbf{C}$ será cuadrada, y su determinante es el cuadrado del área encerrada por los vectores. Si interpretamos los vectores $\mathbf{v}_1$ y $\mathbf{v}_2$ como variables, veremos en el capítulo 3 que el producto $\mathbf{C}'\mathbf{C}$ es su matriz de varianzas y covarianzas, y su determinante, que es el área que forman, es una medida global de la independencia entre las variables, como veremos en la sección 3.5. Por ejemplo, en el caso general de p variables, si una variable es combinación lineal de las demás, las variables son linealmente dependientes, la columna correspondiente a esa variable en la matriz de varianzas y covarianzas será también combinación lineal de las demás columnas y el determinante de la matriz de covarianzas será nulo. Por otro lado, si las variables estan incorreladas su matriz de covarianzas es diagonal y el determinante será, en términos relativos, máximo. Por tanto podemos concluir que cuanto mayor sea el determinante mayor es la independencia entre los vectores.

Traza de una matriz

Se denomina diagonal principal de una matriz cuadrada $\mathbf{C}$ de orden n con elementos c_{ij} al conjunto de elementos c_{ii}, $i = 1, ..., n$. La **traza** de una matriz cuadrada es la suma de los elementos de la diagonal principal de la matriz, escribiremos:

$$tr(\mathbf{C}) = \sum_{i=1}^{n} c_{ii}$$

La traza es un operador lineal. En efecto, de la definición se obtiene:

(a) $tr(\mathbf{A} + \mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$.

(b) $tr(\lambda\mathbf{A}) = \lambda tr(\mathbf{A})$, donde λ es un escalar.

(c) Se demuestra que: $tr(\mathbf{ABC}) = tr(\mathbf{BCA}) = tr(\mathbf{CAB})$, en el supuesto de que todos los productos estén definidos.

(d) Si la matriz $\mathbf{C}$ es simétrica, $tr(\mathbf{C}^2) = tr(\mathbf{CC}) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^2$.

La traza es una medida global de tamaño de la matriz que se obtiene sumando sus elementos diagonales. Por ejemplo, la traza de una matriz de varianzas y covarianzas es la suma de todas las varianzas de las variables. Al sumar los elementos diagonales es una medida global de variabilidad, pero, a diferencia del determinante, no tiene en cuenta las relaciones entre las variables.

Rango de una matriz cuadrada

El rango máximo de una matriz cuadrada de orden n es n. Cuando el rango es menor que n una fila o columna es combinación lineal de las demás y decimos que la matriz es **singular**. Por otro lado, se comprueba que

1. Para matrices cuadradas del mismo orden, $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$, donde $\mathbf{B}$ y $\mathbf{C}$ son no singulares, $rg(\mathbf{CAB}) = rg(\mathbf{A})$.

2. Si $\mathbf{A}$ y $\mathbf{B}$ son cuadradas de orden n y $\mathbf{AB} = 0$, entonces $rg(\mathbf{A}) + rg(\mathbf{B}) \leq n$.

Formas cuadráticas

Si transformamos un vector $\mathbf{x}$ mediante una transformación lineal, $\mathbf{y} = \mathbf{Bx}$, la norma al cuadrado del nuevo vector será

$$\mathbf{y}'\mathbf{y} = \mathbf{x}'\mathbf{B}'\mathbf{Bx} = \mathbf{x}'\mathbf{Ax}$$

donde $\mathbf{A} = \mathbf{B}'\mathbf{B}$ es una matriz cuadrada y simétrica. En función del vector original la forma resultante se denomina forma cuadrática. Llamaremos **forma cuadrática** a una expresión escalar del tipo:

$$\mathbf{x}'\mathbf{Ax}$$

donde $\mathbf{x}$ es un vector, $\mathbf{x}'$ su transpuesto, y $\mathbf{A}$ una matriz cuadrada y simétrica. La forma cuadrática es siempre un escalar. Su expresión general es:

$$\sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} a_{ij}x_i x_j.$$

Diremos que una matriz $\mathbf{A}$ **es semidefinida positiva** si cualquier forma cuadrática formada a partir de ella es un número no negativo, para cualquier vector $\mathbf{x} \neq 0$. Si la forma cuadrática es siempre un numero positivo diremos que la matriz $\mathbf{A}$ es definida positiva. Se demuestra que las formas escalares, como el determinante y la traza, que pueden obtenerse a partir de matrices semidefinidas positivas son números no negativos. Una matriz semidefinida positiva tiene pues propiedades similares a los números no negativos y una matriz definida positiva a los números positivos.

Matriz Inversa

Dada una matriz $\mathbf{A}$ cuadrada $n \times n$, no singular, definimos su inversa, $\mathbf{A}^{-1}$, como una matriz $n \times n$ tal que:

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

donde $\mathbf{I}$ es la matriz identidad, que tiene unos en la diagonal y ceros fuera de ella. Es decir, escribiendo $\mathbf{A}$ con vector fila $\mathbf{a}_i'$, la matriz $\mathbf{A}^{-1}$ tendrá vectores columna $\mathbf{b}_i$ tales que:

$$\begin{bmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_n' \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 & \dots & \mathbf{b}_n \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1'\mathbf{b}_1 & \dots & \mathbf{a}_1'\mathbf{b}_n \\ \vdots & & \vdots \\ \mathbf{a}_n'\mathbf{b}_1 & \dots & \mathbf{a}_n'\mathbf{b}_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & \dots & 1 \end{bmatrix}.$$

En consecuencia la matriz $\mathbf{A}^{-1}$ debe tener por columnas vectores $\mathbf{b}$ tales que: (1) $\mathbf{b}_i$ es ortogonal a $\mathbf{a}_j$,es decir el producto escalar $\mathbf{b}_i'\mathbf{a}_j$ es cero $\forall j \neq i$; (2) el producto escalar de los vectores $\mathbf{b}_i'\mathbf{a}_i = \mathbf{a}_i'\mathbf{b}_i$ es uno.

Observemos que el cálculo de la matriz inversa resuelve el problema de calcular vectores ortogonales a uno dado (o variables incorreladas con una dada). Por ejemplo, el espacio ortogonal al vector $\mathbf{a}_1$ puede calcularse construyendo una matriz que tenga a este vector como primera fila y calculando la inversa de la matriz. Si llamamos $\mathbf{b}_2, ...\mathbf{b}_n$ a los vectores columna de la matriz inversa, estos vectores forman el espacio nulo del vector $\mathbf{a}_1$. Como ilustración, dada la matriz

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 0 & 4 \end{bmatrix},$$

es fácil comprobar que la inversa es

$$\mathbf{A}^{-1} = \begin{bmatrix} .5 & -.125 \\ 0 & .25 \end{bmatrix}$$

y el primer (segundo) vector columna de la inversa define el espacio ortogonal al segundo (primer) vector fila de la matriz original.

La necesidad de calcular la inversa de una matriz aparece de manera natural al resolver sistemas de ecuaciones lineales

$$\mathbf{Ax} = \mathbf{b}$$

donde $\mathbf{A}$ es una matriz conocida cuadrada de orden n, $\mathbf{b}$ un vector de constantes y $\mathbf{x}$ un vector de n incógnitas. Para que este sistema tenga solución única las n ecuaciones deben de ser distintas, lo que supone que no existe una fila de $\mathbf{A}$ que sea combinación lineal de las demás. Entonces $\mathbf{A}$ es no singular y la solución se obtiene mediante:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

El cálculo de la matriz inversa de una matriz dada es engorroso y debe realizarse mediante un ordenador si la dimensión de $\mathbf{A}$ es alta. Se demuestra que la inversa de una matriz puede calcularse por las tres operaciones siguientes:

1. Se sustituye cada elemento por su adjunto.
2. Se transpone la matriz resultante. Se obtiene una matriz que llamaremos **adjunta** de la matriz $\mathbf{A}$.
3. Se divide cada término de la matriz adjunta por el determinante de la matriz original.

Como ejemplo, calcularemos la inversa de la matriz

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix}$$

comenzaremos sustituyendo cada elemento por su adjunto. Por ejemplo, para el elemento $(1,1)$ su adjunto es $(-1)^2 [2 \times 3 - 1 \times 0] = 6$. Para el $(1,2)$, $(-1^3) [-1 \times 3 - 1 \times 0] = 3$, etc. Así obtenemos la matriz

$$\begin{bmatrix} 6 & 3 & 0 \\ -3 & 3 & 0 \\ -1 & -1 & 3 \end{bmatrix},$$

y al transponerla resulta la matriz adjunta :

$$Adj\,(\mathbf{A}) = \begin{bmatrix} 6 & -3 & -1 \\ 3 & 3 & -1 \\ 0 & 0 & 3 \end{bmatrix}.$$

Si dividimos ahora por el determinante de la matriz $\mathbf{A}$

$$|\mathbf{A}| = 6 + 3 = 9,$$

se obtiene la expresión de la inversa

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{9} \\ \frac{1}{3} & \frac{1}{3} & -\frac{1}{9} \\ 0 & 0 & \frac{1}{3} \end{bmatrix}$$

y podemos comprobar que $\mathbf{A}\ \mathbf{A}^{-1} = \mathbf{I}$.

La inversa de una matriz $\mathbf{A}$ tiene las propiedades siguientes:

1. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ para matrices cuadradas no singulares.
2. $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$
3. $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$
4. si $\mathbf{A}$ es simétrica también lo es $\mathbf{A}^{-1}$.

La matriz inversa de una matriz de varianzas y covarianzas tiene una interesante interpretación en Estadística, como veremos en el siguiente capítulo. La matriz inversa recoge la información de la dependencia conjunta de todas las variables de manera más completa que la matriz de varianzas y covarianzas.

Inversas de sumas de matrices

Es muy útil poder calcular la inversa de una suma de matrices en función de las inversas de los sumandos. La forma general es la siguiente: supongamos que las matrices $\mathbf{A}$ y $\mathbf{C}$ son matrices cuadradas no singulares de orden n y p respectivamente, y $\mathbf{B}$ y $\mathbf{D}$ son matrices rectangulares $(n \times p)$ y $(p \times n)$, se comprueba por multiplicación directa que

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A^{-1}B}(\mathbf{DA^{-1}B} + \mathbf{C}^{-1})^{-1}\mathbf{DA^{-1}}. \tag{2.1}$$

Si tomamos en esta expresión $\mathbf{C}$ =1 y las matrices $\mathbf{B}$ y $\mathbf{D}$ son vectores, que llamaremos $\mathbf{b}$ y $\mathbf{d}'$, se obtiene que

$$(\mathbf{A} + \mathbf{b}\mathbf{d}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{b}(\mathbf{d}'\mathbf{A}^{-1}\mathbf{b}+1)^{-1}\mathbf{d}'\mathbf{A}^{-1}$$

Cuando $\mathbf{A}$ y $\mathbf{C}$ tienen el mismo orden, se comprueba que la expresión de la inversa puede escribirse como:

$$(\mathbf{A} + \mathbf{C})^{-1} = \mathbf{C}^{-1}(\mathbf{A}^{-1} + \mathbf{C}^{-1})^{-1}\mathbf{A}^{-1}. \tag{2.2}$$

Veremos que estas fórmulas son muy útiles para estudiar el cambio de la matriz de varianzas y covarianzas, y otros estadísticos relevantes, al eliminar observaciones o variables.

Matrices ortogonales

Llamaremos matriz ortogonal, $\mathbf{C}$, a una matriz cuadrada, que representa un giro en el espacio. Para caracterizar estas matrices, supongamos que dado un vector $\mathbf{x}$ le aplicamos una matriz no singular $\mathbf{C}$ y obtenemos un nuevo vector $\mathbf{y} = \mathbf{C}\mathbf{x}$. Si esta operación es un giro, la norma de $\mathbf{y}$ debe ser idéntica a la de $\mathbf{x}$, lo que implica la condición :

$$\mathbf{y}'\mathbf{y} = \mathbf{x}'\mathbf{C}'\mathbf{C}\mathbf{x} = \mathbf{x}'\mathbf{x},$$

es decir, deberá verificarse que :

$$\mathbf{C}'\mathbf{C} = \mathbf{I}.$$

De la definición $\mathbf{y} = \mathbf{C}\mathbf{x}$ deducimos que $\mathbf{x} = \mathbf{C}^{-1}\mathbf{y}$. Por otro lado, multiplicando por $\mathbf{C}'$ tenemos que $\mathbf{C}'\mathbf{y} = \mathbf{C}'\mathbf{C}\mathbf{x} = \mathbf{x}$. De estas dos condiciones concluimos que la matriz inversa debe ser igual a su traspuesta. Esta es la condición de ortogonalidad:

$$\mathbf{C}' = \mathbf{C}^{-1}.$$

Una matriz ortogonal debe tener filas (o columnas) que son vectores ortogonales entre sí y de longitud unidad, ya que:

$$\begin{bmatrix} \mathbf{c}_1' \\ \vdots \\ \mathbf{c}_n' \end{bmatrix} [\mathbf{c}_1 \dots \mathbf{c}_n] = \begin{bmatrix} \mathbf{c}_1'\mathbf{c}_1 & \dots & \mathbf{c}_1'\mathbf{c}_n \\ \vdots & & \vdots \\ \mathbf{c}_n'\mathbf{c}_1 & \dots & \mathbf{c}_n'\mathbf{c}_n \end{bmatrix} = \begin{bmatrix} 1 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 1 \end{bmatrix}$$

además: $|\mathbf{C}| = |\mathbf{C}'|$ =1, donde $|\mathbf{C}|$ es el determinante de $\mathbf{C}$.

Por ejemplo, en $\Re^2$, la matriz

$$\mathbf{C} = \begin{pmatrix} \cos\alpha & -sen\ \alpha \\ sen\ \alpha & \cos\alpha \end{pmatrix}$$

es ortogonal, ya que $\mathbf{C}\mathbf{C}' = \mathbf{I}$.

Los vectores de una matriz ortogonal de orden n forman una **base ortonormal** de $\Re^n$ ya que son ortogonales y de norma uno.

2.3.5 Matrices Particionadas

Una matriz puede subdividirse en elementos que sean a su vez matrices y a los que se aplican las reglas anteriores. Esta operación es importante cuando queremos dividir las variables en bloques distintos. Por ejemplo, la matriz

$$\mathbf{A} = \left[\begin{array}{c|cc} 2 & 3 & 4 \\ 5 & 6 & 1 \\ \hline 0 & 2 & 3 \end{array}\right],$$

puede escribirse también como una matriz 2×2 particionada:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \tag{2.3}$$

donde:

$$\mathbf{A}_{11} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}, \quad \mathbf{A}_{12} = \begin{bmatrix} 3 & 4 \\ 6 & 1 \end{bmatrix}, \quad \mathbf{A}_{21} = 0, \quad \mathbf{A}_{22} = [2\ 3]\,.$$

Podemos obtener la inversa y el determinante de una matriz particionada en otra 2×2 de manera que los términos diagonales $\mathbf{A}_{11}$ y $\mathbf{A}_{22}$ sean matrices cuadradas no singulares. La inversa de la matriz $\mathbf{A}$ dada por (2.3) se calcula mediante:

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{B}^{-1} & -\mathbf{B}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix}$$

donde

$$\mathbf{B} = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})$$

como puede comprobarse por multiplicación directa.

El determinante se obtiene mediante:

$$|\mathbf{A}| = |\mathbf{A}_{22}||\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}| = |\mathbf{A}_{11}||\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}| = |\mathbf{A}_{22}|\,|\mathbf{B}|$$

Observemos que si la matriz es diagonal por bloques y $\mathbf{A}_{12} = \mathbf{0}$, $\mathbf{A}_{21} = \mathbf{0}$, entonces $\mathbf{A}^{-1}$ se obtiene simplemente como $\begin{bmatrix} \mathbf{A}_n^{-1} & 0 \\ 0 & \mathbf{A}_{22}^{-1} \end{bmatrix}$ y $|\mathbf{A}| = |\mathbf{A}_{11}|\,|\mathbf{A}_{12}|\,.$

Ejercicios 2.3

2.3.1 Calcular el determinante de la matriz formada por los tres vectores del ejercicio 2.2.2, $\mathbf{a} = (1, 0, 2)'$, $\mathbf{b} = (1, 1, 2)'$, $\mathbf{c} = (2, 1, 6)'$.¿Qué conclusiones podemos extraer de este resultado respecto a la independencia lineal de estos vectores?

2.3.2 Dada la matriz rectangular $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 2 & 2 \end{bmatrix}$, calcular la matriz $\mathbf{A}'\mathbf{A}$ y su determinante y traza. Hacer lo mismo para la matriz $\mathbf{A}\mathbf{A}'$.

2.3.3 Calcular la inversa de la matriz $\mathbf{A}'\mathbf{A}$ del ejercicio anterior. Dibujar en el plano los vectores que forman esta matriz y su inversa y comentar sobre el resultado obtenido.

2.3.4 Demostrar que la matriz $\begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix}$ es ortogonal. Aplicarla al vector $\binom{1}{1}$ y dibujar el resultado. ¿Qué giro produce esta matriz?

2.3.5 Se miden tres dimensiones físicas en un grupo de 10 personas y estos datos se disponen en una matriz rectangular $\mathbf{A}$, de dimensiones (10×3), justificar las siguientes afirmaciones:

a) El rango máximo de esta matriz es 3.

b) La operación $\mathbf{A}'\mathbf{1}_{10}$, donde $\mathbf{1}_{10}$ es un vector (10×1) con todas sus componentes iguales a uno proporciona un vector (3×1) cuyos componentes son la suma de los valores de cada variable.

c) La operación $\frac{1}{10}\mathbf{A}'\mathbf{1}_{10}$, proporciona un vector cuyos componentes son las medias de las variables.

d) La operación $\mathbf{1}_{10}(\frac{1}{10}\mathbf{A}'\mathbf{1}_{10})' = \frac{1}{10}\mathbf{1}_{10}\mathbf{1}'_{10}\mathbf{A}$, proporciona una matriz rectangular de dimensiones (10×3), cuyas columnas contienen la media de cada variable.

e) La matriz $\widetilde{\mathbf{A}}= \mathbf{A}-\frac{1}{10}\mathbf{1}_{10}\mathbf{1}'_{10}\mathbf{A}$ proporciona una matriz rectangular de dimensiones (10×3), cuyas columnas contienen las desviaciones de cada variable con respecto a su media.

f) La matriz $\widetilde{\mathbf{A}}'\widetilde{\mathbf{A}}$ proporciona una matriz cuadrada de dimensiones (3×3), cuyos términos diagonales son las sumas de las desviaciones a la media de cada variable al cuadrado.

2.3.6 Con la matriz de datos de EUROALI del apéndice de datos calcular las varianzas y covarianzas de las variables y colocarlas en una matriz cuadrada y simétrica de orden nueve, donde en la diagonal estén las varianzas y fuera de la diagonal las covarianzas. Calcular la traza y el determinante y pensar en su interpretación. Repetirlo para las variables estandarizadas. ¿Qué análisis le parece más informativo?

2.3.7 Calcule una base del espacio ortogonal al vector $\mathbf{a}'_1$ =(1 0 0 0 -1) de la forma siguiente: (1) construya una matriz arbitraria cuadrada de dimension 5 que tenga como primera fila el vector $\mathbf{a}'_1$; (2) calcule la inversa de la matriz y tome el espacio generado por las columnas 2 a la 5. Justifique el resultado obtenido.

2.3.8 Demuestre por multiplicación directa la fórmula (2.2). (Nota, utilice que $(\mathbf{A^{-1}} + \mathbf{C}^{-1})^{-1}\mathbf{A^{-1}}$ puede escribirse como $(\mathbf{I} + \mathbf{AC}^{-1})^{-1}$).

2.3.9 Demuestre por multiplicación directa que $(\mathbf{I} + \mathbf{C})^{-1}= \mathbf{I} - (\mathbf{I} + \mathbf{C}^{-1})^{-1}$.

2.3.10 Demuestre por multiplicación directa la fórmula (2.1). (Nota, al sacar factor común utilice que $(\mathbf{DA^{-1}B} + \mathbf{C}^{-1})^{-1}$ puede escribirse como $\mathbf{C}(\mathbf{I} + \mathbf{DA^{-1}BC})^{-1}$.

2.4 VECTORES Y VALORES PROPIOS

Dada una matriz cuadrada hay determinadas propiedades que esperamos sean invariantes ante ciertas transformaciones lineales que preservan la información existente en la matriz. Por ejemplo, si transponemos la matriz las propiedades básicas de los vectores que la forman no varían, y hemos visto que ni la traza ni el determinante se modifican. Si giramos los vectores que la forman, es decir multiplicamos la matriz por una ortogonal, no se alteran ni sus magnitudes ni sus posiciones relativas, por lo que esperamos que las propiedades básicas

de la matriz se mantengan. Por ejemplo, si en lugar de trabajar con los ingresos y los costes decidimos trabajar con los beneficios, construidos como ingresos-costes, y el volumen de actividad, definido como ingresos más costes, hemos aplicado una transformación ortogonal. Aunque la matriz cuadrada que representa las varianzas y covarianzas de las nuevas variables sea distinta de la original, la esencia del problema es la misma, y esperamos que la matriz de las nuevas variables tenga características idénticas a las de las variables originales. Para precisar esta idea aparece el concepto de valores y vectores propios de una matriz cuadrada.

Los valores propios son las medidas básicas de tamaño de una matriz, que no se ven alteradas si hacemos un cambio de coordenadas que equivale a una rotación de los ejes. Se demuestra que las medidas globales de tamaño de la matriz, como la traza o el determinante, son sólo función de los valores propios y, en consecuencia, serán también invariantes ante las transformaciones que preservan los valores propios.

Los vectores propios representan las direcciones características de la matriz y no son invariantes. Al aplicar una matriz cuadrada de orden n a un vector de dimensión n este se transforma en dirección y magnitud. Sin embargo, para cada matriz cuadrada existen ciertos vectores que al transformarlos por la matriz sólo se modifica su longitud (norma) y no su posición en el espacio. Estos vectores se denominan vectores propios de la matriz.

2.4.1 Definición

Llamaremos **vectores propios** de una matriz cuadrada de orden n a aquellos vectores cuya dirección no se modifica al transformarlos mediante la matriz. Por tanto $\mathbf{u}$ es un vector propio de la matriz $\mathbf{A}$ si verifica que :

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}. \tag{2.4}$$

donde λ es un escalar, que se denomina valor propio de la matriz. En esta relación suponemos $\mathbf{u} \neq 0$, ya que si no es trivialmente cierta. Si $\mathbf{u}$ es un vector propio de $\mathbf{A}$ y multiplicamos (2.4) por cualquier $a \neq 0$, resulta que $a\mathbf{u}$ será también un vector propio de $\mathbf{A}$. Para evitar esta indeterminación suponemos que los vectores propios están normalizados de manera que $\|\mathbf{u}\| = 1$. Sin embargo, el signo queda indeterminado: si $\mathbf{u}$ es un vector propio también lo es $-\mathbf{u}$.

Para calcular el vector propio podemos escribir la ecuación anterior como:

$$(\mathbf{A}-\lambda\mathbf{I})\mathbf{u} = \mathbf{0},$$

y este es un sistema homogéneo de ecuaciones que tendrá solución no nula si y solo si la matriz del sistema, $(\mathbf{A}-\lambda\mathbf{I})$, es singular. En efecto, si esta matriz fuese invertible multiplicando por la inversa tendríamos que la única solución es $\mathbf{u} = \mathbf{0}$. Por tanto, este sistema tiene solución no nula si se verifica que

$$|\mathbf{A} - \lambda\mathbf{I}| = 0.$$

Esta ecuación se denomina la **ecuación característica** de la matriz. Es una ecuación polinómica en λ de orden n y sus n raíces se denominan **valores propios** de la matriz. Es

inmediato de la definición que si una matriz es diagonal los valores propios son los elementos de la diagonal principal. En efecto, tendremos:

$$|\mathbf{A}-\lambda\mathbf{I}| = \left|\begin{bmatrix} a_1 & \dots & 0 \\ \vdots & a_2 & \vdots \\ 0 & ... & a_n \end{bmatrix} - \begin{bmatrix} \lambda & \dots & 0 \\ \vdots & \lambda & \vdots \\ 0 & \dots & \lambda \end{bmatrix}\right| = \left|\begin{bmatrix} a_1-\lambda & \dots & 0 \\ \vdots & a_2-\lambda & \vdots \\ 0 & \dots & a_n-\lambda \end{bmatrix}\right|$$

$$|\mathbf{A}-\lambda\mathbf{I}| = (a_1-\lambda)...(a_n-\lambda),$$

y las soluciones de esta ecuación polinómica son $a_1, ..., a_n$.

Aunque una matriz de orden n tiene siempre n valores propios, estos pueden aparecer repetidos. En general, una matriz tiene $h \leq n$ valores propios distintos. Si un valor propio aparece repetido r veces se dice que tiene multiplicidad r. Por ejemplo, la matriz diagonal:

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

tiene como valores propios 2, 3 y 0, este último valor con multiplicidad dos (aparece dos veces).

A cada valor propio distinto de una matriz cuadrada podemos asociarle un único vector propio que satisface (2.4). En efecto, dado λ podemos resolver el sistema y obtener $\mathbf{u}$. Como la matriz del sistema es singular, existen infinitas soluciones, ya que si $\mathbf{u}$ es una solución también lo es $a\mathbf{u}$, lo que resolvemos tomando el vector de norma uno. Si un valor propio es múltiple, es decir, la matriz **no** tiene n valores propios distintos, los vectores propios asociados a valores propios con multiplicidad mayor de uno no están definidos en general de manera única. Para ilustrar esta idea, consideremos la matriz

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

que tiene el valor propio 1 con multiplicidad 2. Los vectores $\mathbf{u}_1 = (1,\, 0,\, 0)'$ y $\mathbf{u}_2 = (0,\, 1,\, 0)'$ son vectores propios asociados al valor 1, pero también lo es $\mathbf{u}_3 = \gamma\mathbf{u}_1 + (1-\gamma)\,\mathbf{u}_2$, para cualquier valor de γ. Los vectores propios están en un espacio igual a la multiplicidad del valor propio, 2, y cualquier vector normalizado de este espacio de dimensión 2 es un vector propio de $\mathbf{A}$.

Cuando la matriz tiene n valores propios **distintos,** a cada valor propio le podemos asociar un vector propio bien definido y se demuestra que el conjunto de los n vectores propios es linealmente independiente.

Los valores propios de una matriz tienen las propiedades siguientes:

1. Si λ es un valor propio de $\mathbf{A}$, λ^r es un valor propio de $\mathbf{A}^r$. En particular , si $\mathbf{A}^{-1}$ existe, λ^{-1} es un valor propio de $\mathbf{A}^{-1}$.

2. Los valores propios de una matriz y su transpuesta son los mismos.

3. La suma de los valores propios de $\mathbf{A}$ es igual a la traza.

$$tr(\mathbf{A}) = \sum \lambda_i.$$

4. El producto de los valores propios de $\mathbf{A}$ es igual al determinante

$$|\mathbf{A}| = \prod \lambda_i.$$

5. Las matrices $\mathbf{A}$ y $\mathbf{P}^{-1}\mathbf{AP}$ tiene los mismos valores propios.

6. Las matrices $\mathbf{A}$ y $\mathbf{A} \pm \mathbf{I}$ tienen los mismos vectores propios y si λ es un valor propio de $\mathbf{A}$, $\lambda \pm 1$ es un valor propio de $\mathbf{A} \pm \mathbf{I}$ y la

7. Las matrices cuadradas $\mathbf{ABC}$, $\mathbf{BCA}$ y $\mathbf{CAB}$, donde las matrices $\mathbf{A}, \mathbf{B}$, y $\mathbf{C}$ son generales con la condición de que los productos existan, tienen los mismos valores propios no nulos.

8. Si $\mathbf{A}$ es triangular los valores propios son los elementos diagonales.

9. Si $\mathbf{A}$ y $\mathbf{B}$ son cuadradas de órdenes n y p los np vectores propios de su producto de Kronecker, $\mathbf{A} \oplus \mathbf{B}$, son el producto de Kronecker de los vectores propios de $\mathbf{A}$ y $\mathbf{B}$.

La propiedad 1 se demuestra fácilmente ya que si $\mathbf{Au} = \lambda\mathbf{u}$, multiplicando esta ecuación por $\mathbf{A}^{-1}$, resulta $\mathbf{u} = \lambda\mathbf{A}^{-1}\mathbf{u}$, es decir $\mathbf{A}^{-1}\mathbf{u} = \lambda^{-1}\mathbf{u}$. Para comprobar la segunda escribiendo $\mathbf{Au} = \lambda\mathbf{u}$ y $\mathbf{A}'\mathbf{v} = \mu\mathbf{v}$ y multiplicando la primera por $\mathbf{v}'$ y la segunda por $\mathbf{u}'$ se tiene $\mathbf{v}'\mathbf{Au} = \lambda\mathbf{v}'\mathbf{u}$ y $\mathbf{u}'\mathbf{A}'\mathbf{v} = \mu\mathbf{u}'\mathbf{v}$ y como el primer miembro de ambas es el mismo (un escalar es igual a su transpuesto) el segundo lo será y λ=μ. Las propiedades 3 y 4 son consecuencia de las propiedades de diagonalización de matrices que comentamos a continuación. La 5 se comprueba fácilmente ya que si $\mathbf{Au} = \lambda\mathbf{u}$, multiplicando ambos miembros por $\mathbf{P}^{-1}$ por la derecha y $\mathbf{P}$ por la izquierda, se obtiene que $\mathbf{P}^{-1}\mathbf{A}\,\mathbf{P}\,\mathbf{u} = \lambda\,\mathbf{u}$ y las matrices tienen los mismos valores propios. Los vectores propios de la matriz $\mathbf{P}^{-1}\mathbf{AP}$ son $\mathbf{P}^{-1}\mathbf{u}$, siendo $\mathbf{u}$ un vector propio de la matriz $\mathbf{A}$. La propiedad 6 es consecuencia de que si $\mathbf{Au} = \lambda\mathbf{u}$, entonces $\mathbf{Au} + \mathbf{Iu} = \lambda\mathbf{u} + \mathbf{u}$, es decir, $(\mathbf{A} + \mathbf{I})\mathbf{u} = (1 + \lambda)\mathbf{u}$. Por otro lado si $|\mathbf{A} - \lambda\mathbf{I}| = 0$, entonces también $|\mathbf{A} + \mathbf{I} - \mathbf{I} - \lambda\mathbf{I}| = |\mathbf{A} + \mathbf{I} - (\mathbf{1} + \boldsymbol{\lambda})\mathbf{I}| = 0$. La 9 resulta de la definición de producto de Kronecker.

2.4.2 Valores y vectores propios de matrices simétricas

En este libro vamos a obtener vectores y valores propios principalmente de matrices simétricas. En estas matrices:
(1) los valores propios son siempre reales;
(2) los vectores propios son ortogonales.

Para comprobar esta segunda propiedad observemos que si $\mathbf{Au}_i = \lambda_i\mathbf{u}_i$ y $\mathbf{Au}_j = \lambda_j\mathbf{u}_j$ son dos valores y vectores propios distintos, multiplicando la primera ecuación por $\mathbf{u}_j'$ y la

segunda por $\mathbf{u}_j'$ los primeros miembros son iguales y los segundos como $\lambda_i \neq \lambda_j$ sólo serán iguales si $\mathbf{u}_j'\mathbf{u}_i = 0$.

Para interpretar el significado de los valores y vectores propios de estas matrices consideremos matrices simétricas de orden 2 cuyos vectores pueden dibujarse en un plano. Por ejemplo la matriz simétrica

$$\mathbf{A}_1 = \begin{bmatrix} a & b \\ b & a \end{bmatrix}$$

es fácil comprobar que sus valores propios se obtienen de $(a-\lambda)^2 = b^2$ y que sus vectores propios están en las direcciones (1,1) y (1,-1), y normalizados a norma uno son los vectores (0.7071, 0.7071)' y (0.7071,- 0.7071)' . Por ejemplo, si $a = 3$ y $b = 1$, de manera que la matriz está formada por los dos vectores columna $(3,1)\prime$ y $(1,3)'$, los valores propios son (4 y 2). Supongamos que construimos una elipse con centro en el origen y que pase por los extremos de los dos vectores que forman la matriz, como indica la figura 2.3. Entonces los valores propios representan la distancia del extremo de cada eje de la elipse al origen. Por ejemplo el valor 4 indica que el eje principal de la elipse mide 4 unidades desde el origen, o 8 en total. Analogamente, el valor 2 indica la longitud del otro semieje de la elipse. Los vectores propios asociados a estos valores propios representan las direcciones de los ejes: el asociado al mayor valor propio es un vector unitario en la dirección de la diagonal principal y el segundo es perpendicular a el, como indica la figura 2.3. Si modificamos los valores de a y b los vectores propios no se modifican pero sí los valores propios. Si aumentamos a manteniendo fijo b alejamos los extremos de los vectores y la elipse tiene cada vez los ejes más similares. Por ejemplo, la matriz formada por los vectores columna $(100,1)\prime$ y $(1,100)'$ tiene valores propios (101 y 99) y los mismos vectores propios. Por el contrario si aumentamos b manteniendo fijo a acercamos los extremos de los vectores y apuntamos más la elipse, lo que aumentará la diferencia entre sus ejes. Por ejemplo, la matriz formada por los vectores columna $(1.2,1)\prime$ y $(1,1.2)'$ tiene valores propios (2.2, 0,2) y los mismos vectores propios.

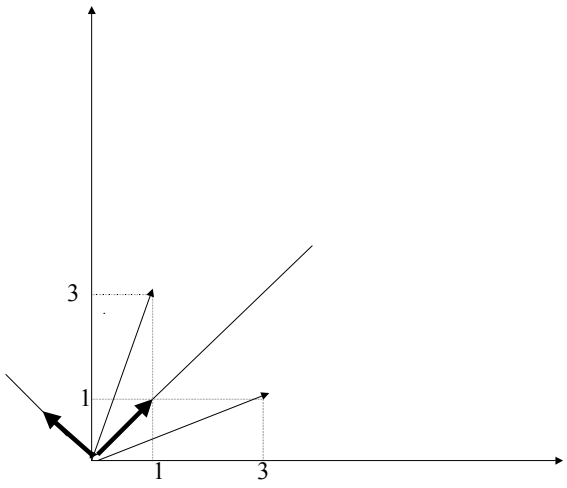


Figura 2.3: Representación de los valores y vectores propios de una matriz simétrica $\mathbf{A}_1$

En la matriz anterior al ser los elementos diagonales idénticos la orientación de la elipse era según las bisectrices de los ejes. Esto no ocurrirá si los elementos diagonales son distintos. Por ejemplo, la matriz

$$\mathbf{A}_2 = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

se encuentre representada en la figura 2.4. Ahora el eje mayor de la elipse esta mucho más cerca del vector de módulo mayor y puede comprobarse que los vectores propios son (0.9239 0.3827) y (-0.3827 0.9239), y los valores propios (4.41, 1.59).

Generalizando este ejemplo, los valores propios de una matriz simétrica representan las magnitudes de los ejes del elipsoide con centro el origen y determinado por los extremos de los vectores. Los vectores propios indican las direcciones de estos ejes principales.

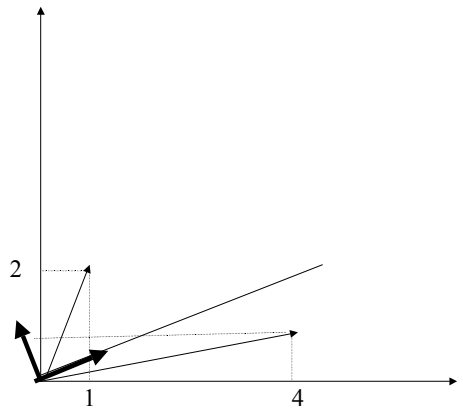


Figura 2.4: Representación de los valores y vectores propios de una matriz simétrica $\mathbf{A}_2$

2.4.3 Diagonalización de Matrices Simétricas

Una propiedad muy importante de las matrices simétricas es que pueden convertirse en una matriz diagonal mediante una transformación ortogonal. Sea $\mathbf{A}$ una matriz cuadrada y simétrica de orden n. Hemos visto que esta matriz tiene valores propios reales y vectores propios ortogonales. Entonces los vectores propios, $\mathbf{u}_1, \ldots, \mathbf{u}_n$, son linealmente independientes y forman una base en $\Re^n$. Podemos escribir

$$\mathbf{A}\left[\mathbf{u}_1, \ldots, \mathbf{u}_n\right] = \left[\lambda_1\mathbf{u}_1, \ldots, \lambda_n\mathbf{u}_n\right].$$

donde $\lambda_1, \ldots \lambda_n$ son los valores propios que son números reales y que pueden no ser todos distintos. En particular, algunos de estos valores propios pueden ser nulos. Esta ecuación puede escribirse, llamando $\mathbf{D}$ a la matriz diagonal con términos λ_i, como

$$\mathbf{AU} = \mathbf{UD}$$

donde la matriz $\mathbf{U}$ es ortogonal. Multiplicando por $\mathbf{U}' = \mathbf{U}^{-1}$, tenemos que

$$\mathbf{U}'\mathbf{AU} = \mathbf{D} \tag{2.5}$$

y hemos transformado la matriz original en una matriz diagonal, $\mathbf{D}$, mediante una matriz $\mathbf{U}$ ortogonal. La ecuación (2.5) tiene una interesante interpretacion geométrica. Observemos que $\mathbf{U}'\mathbf{A}$ es una rotación de los vectores que forman la matriz, y esta ecuación nos dice que estos vectores rotados son iguales a $\mathbf{DU}'$, que es el resultado de multiplicar por los términos de $\mathbf{D}$ a una base de vectores ortonormales. En otros términos, como $\mathbf{A} = \mathbf{UDU}'$ vemos como se genera siempre una matriz simétrica: se parte de una base ortonormal de vectores, $\mathbf{U}'$,se modifica la norma de cada vector de esta base, multiplicandolo por una matriz diagonal, y luego se rotan de nuevo los vectores así obtenidos. Diagonalizar una matriz simétrica

consiste en recuperar esta operación y los valores propios representan las constantes por las que se han multiplicado los vectores ortonormales iniciales y los vectores propios indican el giro realizado.

Si tomamos determinantes en (2.5):

$$|\mathbf{U}'||\mathbf{A}||\mathbf{U}| = |\mathbf{D}|,$$

y como $|\mathbf{U}| = |\mathbf{U}'| = 1$, el determinante de $\mathbf{A}$ será el producto de sus raíces características. Por lo tanto, si una de las raíces características es nula, el determinante será 0 y la matriz singular.

Por otro lado, como en (2.5) las matrices $\mathbf{U}'$ y $\mathbf{U}$ son no singulares, el rango de $\mathbf{A}$ será igual al de $\mathbf{D}$, que al ser diagonal será igual al número de términos diagonales no nulos, que son los valores propios de $\mathbf{A}$. Por tanto: *El rango de una matriz simétrica es igual al número de raíces características distintas de cero.*

Al diagonalizar una matriz simétrica obtenemos su rango, observando el número de elementos no nulos en la diagonal principal de la matriz transformada $\mathbf{D}$.

Descomposición espectral

Es interesante poder descomponer una matriz cuadrada simétrica en sus fuentes de variación intrínsecas, es decir en las direcciones de los vectores propios con coeficientes que dependen de los valores propios. Esto es lo que consigue la descomposión espectral. Premultiplicando (2.5) por $\mathbf{U}$ y postmultiplicando por $\mathbf{U}'$ se obtiene

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}'$$

que, como hemos comentado en la sección anterior indica cómo se genera una matriz simétrica a partir de una base ortonormal. Esta descomposición puede escribirse:

$$\mathbf{A} = [\mathbf{u}_{1,\dots,}\mathbf{u}_n] \begin{bmatrix} \lambda_1\mathbf{u}_1' \\ \vdots \\ \lambda_n\mathbf{u}_n' \end{bmatrix}$$

de donde resulta:

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i' \tag{2.6}$$

que descompone la matriz $\mathbf{A}$ como suma de n matrices de rango uno $\mathbf{u}_i\mathbf{u}_i'$ con coeficientes λ_i.

Si la matriz $\mathbf{A}$ tiene rango r la descomposición espectral (2.6) indica que puede expresarse como suma de r matrices de rango unidad. La importancia de esta descomposición es que si algunos valores propios son muy pequeños, podemos reconstruir aproximadamente $\mathbf{A}$ utilizando los restantes valores y valores propios.

Observemos que la descomposición espectral de $\mathbf{A}^{-1}$ es

$$\mathbf{A}^{-1} = \sum_{i=1}^{n} \lambda_i^{-1} \mathbf{u}_i \mathbf{u}_i'$$

ya que $\mathbf{A}^{-1}$ tiene los mismos vectores propios que $\mathbf{A}$ y valores propios λ_i^{-1}.

2.4.4 Raiz cuadrada de una matriz semidefinida positiva

Una matriz cuadrada, simétrica y semidefinida positiva puede siempre descomponerse como producto de una matriz por su transpuesta:

$$\mathbf{A} = \mathbf{H}\,\mathbf{H}',$$

en efecto, por la descomposición espectral de una matriz simétrica

$$\mathbf{A} = \left(\mathbf{U}\,\mathbf{D}^{1/2}\right)\left(\mathbf{D}^{1/2}\mathbf{U}'\right)$$

y tomando $\mathbf{H} = \mathbf{U}\,\mathbf{D}^{1/2}$ se obtiene la descomposición. A la matriz $\mathbf{H}$ se la denomina una **raiz cuadrada** de la matriz $\mathbf{A}$. La raíz cuadrada de una matriz no es única, ya que si $\mathbf{A} = \mathbf{H}\,\mathbf{H}'$ también $\mathbf{A} = \mathbf{H}^*\,\mathbf{H}'^*$ donde $\mathbf{H}^* = \mathbf{H}\mathbf{C}$ para cualquier matriz ortogonal $\mathbf{C}$. Una forma de definir la raíz de manera única es exigir que la matriz $\mathbf{H}$ sea simétrica, con lo que $\mathbf{A} = \mathbf{H}\,\mathbf{H}$. Esto puede hacerse tomando

$$\mathbf{H} = \mathbf{U}\,\mathbf{D}^{1/2}\mathbf{U}'$$

Otra forma de hacer la descomposición de manera única es la descomposición de Cholesky que estudiamos a continuación.

Descomposición de Cholesky (*)

Puede demostrarse que la raíz cuadrada de una matriz cuadrada, simétrica y definida positiva puede obtenerse de manera que $\mathbf{H} = \mathbf{T}$ sea triangular ($\mathbf{T}'$ será también triangular) con términos diagonales positivos. Entonces la descomposición es única y se denomina descomposición de **Cholesky**. Tenemos

$$\mathbf{A} = \mathbf{T}\mathbf{T}'$$

Demostraremos la existencia de esta matriz por inducción, que tiene la ventaja de proporcionar además un método para su cálculo. Si la matriz es un escalar a trivialmente $\mathbf{T} = \sqrt{a}$. Supongamos que hemos encontrado esta descomposición para dimensión p y veamos como obtenerla para dimensión $p+1$. Sea

$$\mathbf{A}_p = \mathbf{T}_p\mathbf{T}_p' \tag{2.7}$$

y vamos a obtener la descomposición para

$$\mathbf{A}_{p+1} = \begin{bmatrix} \mathbf{A}_p & \mathbf{a}_{12} \\ \mathbf{a}_{12}' & a_{22} \end{bmatrix}$$

donde $\mathbf{a}_{12}$ es un vector $p \times 1$ y a_{22} un escalar. Vamos a demostrar que esta matriz puede escribirse como $\mathbf{T}_{p+1}\mathbf{T}_{p+1}'$ donde, tomando $\mathbf{T}_{p+1}$ como triangular inferior:

$$\mathbf{T}_{p+1} = \begin{bmatrix} \mathbf{T}_p & \mathbf{0} \\ \mathbf{t} & t_{p+1} \end{bmatrix}.$$

Entonces, la condición $\mathbf{A}_{p+1}= \mathbf{T}_{p+1}\mathbf{T}'_{p+1}$ equivale a las condiciones:

$$\mathbf{a}_{12} = \mathbf{T}_p\mathbf{t},$$

y

$$a_{22} = \mathbf{t}'\mathbf{t}+t^2_{p+1},$$

conjuntamente con (2.7). Como $\mathbf{T}_p$ es no singular, podemos obtener

$$\mathbf{t} = \mathbf{T}_p^{-1}\mathbf{a}_{12}$$

y utilizando (2.7) podemos escribir

$$t_{p+1} = \sqrt{a_{22} - \mathbf{a}'_{12}\mathbf{A}_p^{-1}\mathbf{a}_{12}},$$

que debe ser positivo si la matriz es definida positiva. Esta descomposición se utiliza mucho en análisis numérico ya que puede calcularse iterativamente con el método propuesto. Por ejemplo, supongamos que $\mathbf{A}$ es una matriz de la forma

$$\mathbf{A} = \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix}$$

con las varianzas y covarianzas de dos variables. Entonces $A_p = a_1 = s_1^2$, $T_p = s_1$; $\mathbf{t} = s_{12}/s_1$

$$\mathbf{T} = \begin{bmatrix} s_1 & 0 \\ s_{12}/s_1 & \sqrt{s_2^2 - s_{12}^2/s_1^2} \end{bmatrix}$$

y contiene en la diagonal las desviaciones típicas de la primera variable y de la regresión de la segunda dada la primera. Esta propiedad es general.

La descomposición de Cholesky proporciona un método eficiente de calcular el determinante de una matriz ya que si $\mathbf{A} = \mathbf{T}\mathbf{T}'$ entonces $|\mathbf{A}| = |\mathbf{T}|\,|\mathbf{T}'| = \sum t_{ii}^2$, siendo t_{ii} los elementos diagonales de $\mathbf{T}$ o $\mathbf{T}'$.

Diagonalización de dos matrices simétricas (*)

Supongamos que $\mathbf{A}$ y $\mathbf{B}$ son dos matrices simétricas de la misma dimensión y $\mathbf{A}$ es además definida positiva. Entonces la matriz $\mathbf{H} = \mathbf{A}^{-1/2}\mathbf{C}$, donde $\mathbf{C}$ contiene los vectores propios de la matriz simétrica $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$ verifica

$$\mathbf{H}'\mathbf{A}\mathbf{H} = \mathbf{I}$$

y

$$\mathbf{H}'\mathbf{B}\mathbf{H} = \mathbf{D}$$

donde la matriz $\mathbf{D}$ es diagonal.

Para comprobar esta propiedad observemos que como la matriz $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$ es simétrica la matriz $\mathbf{C}$ es ortogonal. Por tanto

$$\mathbf{H}'\mathbf{A}\mathbf{H} = \mathbf{C}'\mathbf{A}^{-1/2}\mathbf{A}\mathbf{A}^{-1/2}\mathbf{C} = \mathbf{I}$$

y

$$\mathbf{H}'\mathbf{B}\mathbf{H} = \mathbf{C}'\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}\mathbf{C} = \mathbf{D}$$

donde la matriz $\mathbf{D}$ diagonal contiene los valores propios de la matriz $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$.

2.4.5 Descomposición en valores singulares

Para matrices rectangulares generales puede conseguirse una descomposición similar a la descomposición espectral de una matriz simétrica. Como en el caso de matrices cuadradas y simétricas, toda matriz rectangular $\mathbf{A}$ de dimensiones $(n \times p)$ y de rango r puede expresarse como producto de tres matrices, dos con vectores ortogonales y una diagonal. La descomposición es

$$\mathbf{A} = \mathbf{U}_1 \mathbf{D}^{1/2} \mathbf{V}_1'$$

donde $\mathbf{U}_1$ es $(n \times r)$, $\mathbf{D}$ es $(r \times r)$ y $\mathbf{V}_1'$ es $(r \times p)$. La matriz diagonal $\mathbf{D}^{1/2}$ contiene las raíces cuadradas de los valores propios no nulos de las matrices $\mathbf{A}\,\mathbf{A}'$ o $\mathbf{A}'\,\mathbf{A}$, que son positivos. Estos términos diagonales de $\mathbf{D}$ se denominan los **valores singulares** de la matriz $\mathbf{A}$. La matriz $\mathbf{U}_1$ contiene en columnas los vectores propios unidos a valores propios no nulos de $\mathbf{A}\,\mathbf{A}'$ y $\mathbf{V}_1$ contiene en columnas los vectores propios unidos a valores propios no nulos de $\mathbf{A}'\mathbf{A}$. Las columnas de $\mathbf{U}_1$ son ortogonales entre sí y también lo serán las de $\mathbf{V}_1$. Los elementos diagonales de $\mathbf{D}^{1/2}$ se denominan los valores singulares de la matriz $\mathbf{A}$.

2.4.6 (*)Diagonalización de Matrices generales

Sea $\mathbf{A}$ una matriz cuadrada de orden n. Esta matriz es diagonalizable si, y sólo si, sus vectores propios son linealmente independientes. En efecto, supongamos que los vectores propios, $\mathbf{u}_1, \ldots, \mathbf{u}_n$, son linealmente independientes y forman una base en $\Re^n$. Podemos escribir

$$\mathbf{A}\left[\mathbf{u}_1, \ldots, \mathbf{u}_n\right] = \left[\lambda_1 \mathbf{u}_1, \ldots, \lambda_n \mathbf{u}_n\right].$$

donde $\lambda_1, \ldots \lambda_n$ son los valores propios que pueden no ser distintos. En particular, algunos de estos valores propios pueden ser nulos. Esta ecuación puede escribirse, llamando $\mathbf{D}$ a la matriz diagonal con términos λ_i, como

$$\mathbf{AU} = \mathbf{UD}$$

Como la matriz $\mathbf{U}$ es no singular si los vectores propios son linealmente independientes, multiplicando por la inversa se obtiene

$$\mathbf{U}^{-1}\mathbf{A}\,\mathbf{U} = \mathbf{D}$$

y hemos diagonalizado la matriz $\mathbf{A}$. Podemos también escribir

$$\mathbf{A} = \mathbf{U}\,\mathbf{D}\,\mathbf{U}^{-1}. \tag{2.8}$$

Hemos comprobado que una matriz es diagonalizable si tiene n vectores propios linealmente independientes. Entonces puede escribirse como (2.8), donde $\mathbf{U}$ contienen los vectores propios y la matriz diagonal, $\mathbf{D}$, los valores propios.

Se demuestra que una condición suficiente para que una matriz sea diagonalizable es que tenga valores propios distintos.

Consideremos ahora el caso general de una matriz cuadrada de orden n con p valores propios $\lambda_1, \ldots \lambda_p$, con multiplicidad m_i, $\sum_{i=1}^{p} m_i = n$. Puede demostrarse que la condición para que $\mathbf{A}$ tenga n vectores propios linealmente independientes es que el rango de la matriz $(\mathbf{A} - \boldsymbol{\lambda}_i\mathbf{I}) = n - m_i$, y que esta condición se cumple si la matriz tiene valores propios distintos. En efecto, los valores propios se obtienen de $|\mathbf{A} - \lambda\mathbf{I}| = 0$, lo que implica que, si todos son distintos, el rango de la matriz $(\mathbf{A} - \boldsymbol{\lambda}_i\mathbf{I})$ es $n - 1$.

2.4.7 (*)Inversas Generalizadas

Se denomina matriz inversa generalizada de una matriz rectangular $\mathbf{A}_{n\times p}$ a una matriz $\mathbf{A}^-$ de dimensiones $p \times n$ que verifica:

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}.$$

En general existen muchas matrices que verifican esta condición. Si además imponemos las condiciones:

$$\begin{aligned} \mathbf{A}^-\mathbf{A}\mathbf{A}^- &= \text{simétrica} \\ \mathbf{A}^-\mathbf{A} &= \text{simétrica} \\ \mathbf{A}\mathbf{A}^- &= \text{simétrica} \end{aligned}$$

entonces $\mathbf{A}^-$ es única y se denomina la matriz inversa generalizada Moore-Penrose (MP) de $\mathbf{A}$. Si $n > p$ y $\mathbf{A}$ tiene rango completo, $rg(\mathbf{A}) = p$, la matriz inversa MP es:

$$\mathbf{A}^- = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'. \tag{2.9}$$

El lector puede comprobar que esta matriz verifica las propiedades anteriores. Si $p > n$ y $rg(\mathbf{A}) = n$, esta matriz es:

$$\mathbf{A}^- = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}.$$

Si $\mathbf{A}$ no tiene rango completo esta expresión no es válida ya que ni $(\mathbf{A}'\mathbf{A})^{-1}$ ni $(\mathbf{A}\mathbf{A}')^{-1}$ existen. La inversa MP se construye a partir de la descomposición espectral de la matriz $\mathbf{A}'\mathbf{A}$ (supuesto $n > p$). Si $\lambda_1, \ldots \lambda_r$, $r < p$, son los valores propios no nulos de $\mathbf{A}'\mathbf{A}$ y $\mathbf{u}_1, \ldots \mathbf{u}_r$ sus vectores propios asociados podemos escribir:

$$\mathbf{A}'\mathbf{A} = \mathbf{U}_r\mathbf{D}_r\mathbf{U}_r',$$

donde $\mathbf{U}_r$ es rectangular $p \times r$ con los vectores $\mathbf{u}_i$ en columnas y $\mathbf{D}_r$ es diagonal $r \times r$ e incluye los valores propios no nulos. Entonces es fácil comprobar que

$$\mathbf{A}^- = \mathbf{U}_r\mathbf{D}_r^{-1}\mathbf{U}_r'\mathbf{A}'$$

que es la generalización de (2.9) para matrices de rango no completo.

Ejercicios 2.4

2.4.1 Calcular los vectores y valores propios de la matriz $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ y representarlos gráficamente.

2.4.2 Escribir la representación espectral de la matriz $\mathbf{A}$ de 2.4.1

2.4.3 Calcular los vectores y valores propios de la matriz $\mathbf{A}^{-1}$ y su representación espectral.

2.4.4 Demostrar que 0 es un valor propio de una matriz $\mathbf{A}$ si y solo si esta matriz es singular.

2.4.5 Demostrar que los valores propios de una matriz son iguales a los de su transpuesta.

2.4.6 Dada la matriz $\mathbf{A} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & 2 \\ 1 & 1 & 2 \end{bmatrix}$ calcular la matriz inversa generalizada.

2.4.7 Calcular la descomposición en valores singulares de la matriz $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$.

2.4.8 Demostrar que $|\mathbf{A} + \mathbf{v}\mathbf{v}'| = |\mathbf{A}|\,(1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{v}')$, donde $\mathbf{A}$ es una matriz cuadrada no singular y $\mathbf{v}$ un vector. Para ello utilizar que si llamamos λ_1 al valor propio no nulo de la matriz de rango uno $\mathbf{A}^{-1}\mathbf{v}\mathbf{v}'$, $|(I + \mathbf{A}^{-1}\mathbf{v}\mathbf{v}')| = \prod(1 + \lambda_i) = 1 + \lambda_1 = 1 + tr(\mathbf{v}'\mathbf{A}^{-1}\mathbf{v})$.

2.4.5 Calcular la decomposición de Cholesky de la matriz definida positiva $\mathbf{A}'\mathbf{A}$, donde $\mathbf{A}$ es la matriz del ejercicio 2.4.6

2.5 (*)PROYECCIÓN ORTOGONAL

2.5.1 Matrices Idempotentes

En un modelo lineal la estimación por mínimos cuadrados equivale a la proyección ortogonal del vector de datos sobre el espacio generado por las variables explicativas. La proyección ortogonal tiene una importancia capital en los métodos de estimación lineal y se realiza multiplicando el vector que se desea proyectar por una matriz idempotente. Vamos a definir formalmente estas matrices.

Llamaremos matriz idempotente[1] a una matriz cuadrada, simétrica, y que verifica la propiedad:

$$\mathbf{A}\mathbf{A} = \mathbf{A} = \mathbf{A}'\mathbf{A}.$$

Es inmediato comprobar que una matriz idempotente o bien es singular ($|\mathbf{A}| = 0$), con rango r menor que el orden n de la matriz, o bien es la matriz identidad. En efecto, como $\mathbf{A}$ es idempotente:

$$\mathbf{A}\mathbf{A} = \mathbf{A}$$

si $|\mathbf{A}| \neq 0$, existirá la matriz inversa $\mathbf{A}^{-1}$, y multiplicando por $\mathbf{A}^{-1}$

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{A} = \mathbf{A} = \mathbf{I}.$$

[1] Una matriz idempotente puede no ser simétrica, pero todas las matrices idempotentes que utilicemos lo serán; por lo tanto, en adelante idempotente será simétrica e idempotente, sin que detallemos que es simétrica.

Por tanto, una matriz idempotente que no es la matriz $\mathbf{I}$ será singular. Comprobaremos que las raíces características de una matriz idempotente son cero o la unidad. Llamemos λ a sus raíces características y $\mathbf{u}$ a sus vectores característicos. Entonces:

$$\mathbf{Au} = \lambda\mathbf{u},$$

multiplicando por $\mathbf{A}$, el primer miembro es:

$$\mathbf{AAu} = \mathbf{Au} = \lambda\mathbf{u}$$

y el segundo:

$$\lambda\mathbf{Au} = \lambda^2\mathbf{u},$$

es decir,

$$\lambda\mathbf{u} = \lambda^2\mathbf{u}$$

de donde resulta:

$$\left(\lambda^2 - \lambda\right)\mathbf{u} = 0.$$

Para que λ sea una raíz característica el vector $\mathbf{u}$ debe ser distinto de cero, entonces:

$$\lambda^2 - \lambda = \lambda\left(\lambda - 1\right) = 0$$

que tiene como soluciones $\lambda = 1$ ó $\lambda = 0$. Por lo tanto, si se diagonaliza una matriz idempotente —lo que siempre puede hacerse al ser simétrica— obtendremos en la diagonal principal un número de unos igual al rango de la matriz y el resto de los elementos serán cero.

Una conclusión inmediata de este resultado es que una matriz idempotente $\mathbf{A}$ es siempre semidefinida positiva. En efecto:

$$\mathbf{x}'\mathbf{Ax} = \mathbf{x}'\mathbf{A}'\mathbf{Ax} = (\mathbf{Ax})'\mathbf{Ax} \geq 0.$$

Finalmente, si $\mathbf{A}$ es idempotente también lo es $\mathbf{I} - \mathbf{A}$ ya que:

$$(\mathbf{I} - \mathbf{A})(\mathbf{I} - \mathbf{A}) = \mathbf{I} - \mathbf{A} - \mathbf{A} + \mathbf{AA} = \mathbf{I} - \mathbf{A}$$

De las propiedades anteriores se deduce que si $\mathbf{A}$ es una matriz idempotente simétrica, su rango es igual a su traza.

2.5.2 Proyección Ortogonal

Dado un vector $\mathbf{y}$ de n componentes diremos que $\mathbf{v}$ es la proyección ortogonal de $\mathbf{y}$ sobre un subespacio E_p contenido en $\Re^n$ y de dimensión p, $p < n$ si:

1. $\mathbf{y} = \mathbf{v} + \mathbf{w}$ con $\mathbf{v} \in E_p$

2. $\mathbf{v}'\mathbf{w} = 0$ para todo $\mathbf{v} \in E_p$.

Esta definición indica que $\mathbf{y}$ puede descomponerse como suma de dos vectores perpendiculares: el primero, $\mathbf{v}$, es la proyección ortogonal de $\mathbf{y}$ sobre E_p y pertenece, por tanto, a E_p; el segundo, $\mathbf{w}$, es ortogonal a todos los vectores de E_p (y por tanto a E_p), y pertenece, en consecuencia, al espacio E_{n-p}, complemento ortogonal al E_p. Es fácil demostrar que esta descomposición es única. La figura 2.5 ilustra esta situación.

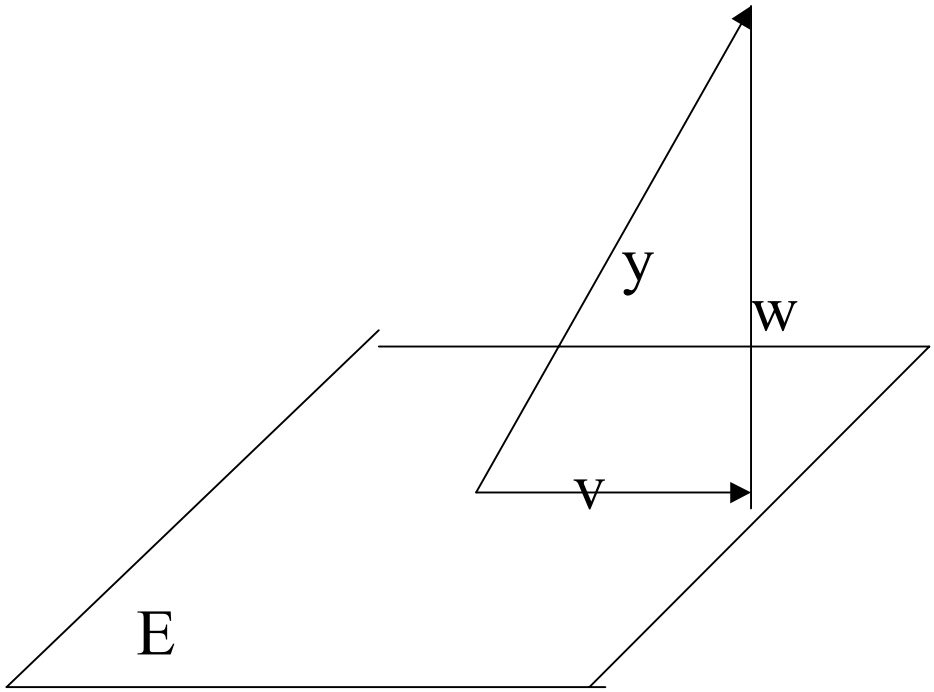


Figura 2.5: Proyección ortogonal del vector y sobre el plano E

Como ilustración, sea E_p un espacio de dimensión uno engendrado por el vector $\mathbf{x}$. Entonces la proyección del vector $\mathbf{y}$ sobre la dirección del vector $\mathbf{x}$ será:

$$\mathbf{v} = c\mathbf{x}$$

donde c es un escalar. Para determinar c, impondremos la condición de que la diferencia $\mathbf{w} = \mathbf{y} - \mathbf{v}$ debe ser ortogonal a $\mathbf{v}$, y por tanto a $\mathbf{x}$:

$$\mathbf{x}'(\mathbf{y} - \mathbf{v}) = 0,$$

es decir, $\mathbf{x}'\mathbf{y} = \mathbf{x}'\mathbf{x}c$, que implica:

$$c = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}.$$

Sustituyendo este valor de c en la expresión de $\mathbf{v}$, la proyección será:

$$\mathbf{v} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \mathbf{A}\mathbf{y}$$

es decir, la proyección de un vector $\mathbf{y}$ sobre otro $\mathbf{x}$ se obtiene multiplicando el vector por la matriz $\mathbf{A} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$. Esta matriz $\mathbf{A}$, es cuadrada $(n \times n)$, idempotente y de rango igual a la dimensión del espacio sobre el que proyectamos, que es, en este caso, uno. Comprobemos que es idempotente:

$$\left(\mathbf{x}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\right)\left(\mathbf{x}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\right) = \mathbf{x}\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}',$$

y que es de rango uno:

$$rg(\mathbf{A}) = tr(\mathbf{A}) = tr\left(\left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\mathbf{x}\right) = tr(1) = 1$$

Observemos que en el caso particular en que el vector $\mathbf{x}$ tiene norma unitaria, $(\mathbf{x}'\mathbf{x}) = 1$, y la expresión del vector proyección es

$$\mathbf{v} = \mathbf{x}\mathbf{x}'\mathbf{y}$$

que tiene una interpretación inmediata: el vector proyección estará en la dirección de $\mathbf{x}$ (lo que implica es de la forma $c\mathbf{x}$) y su norma viene dada por la longitud de la proyección que es $\mathbf{x}'\mathbf{y}$, (ya que $\mathbf{x}$ tiene norma unitaria).

A continuación generalizamos estos resultados a proyecciones más generales.

Teorema 2.1 *Sea* $\mathbf{y} \in \Re^n$ *y sea* $\mathbf{X}$ *una matriz* $(n \times p)$ *cuyas columnas son una base de un cierto subespacio* E_p*. Entonces la proyección del vector* $\mathbf{y}$ *sobre el espacio* E_p *es* $\mathbf{A}\mathbf{y}$*, donde la matriz cuadrada* $\mathbf{A}$ *es simétrica, idempotente, de rango* p*, y tal que* $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$*.*

Demostración La proyección de $\mathbf{y}$ sobre un subespacio debe ser siempre del tipo $\mathbf{v} = \mathbf{A}\mathbf{y}$, donde $\mathbf{A}$ es idempotente. En efecto la proyección de $\mathbf{v}$ sobre dicho espacio, dada por $\mathbf{A}\mathbf{v}$, tendrá que ser igual a $\mathbf{v}$, ya que $\mathbf{v}$ pertenece al subespacio. Por tanto, si $\mathbf{A}\mathbf{v} = \mathbf{v}$, resuta que:

$$\mathbf{A}(\mathbf{A}\mathbf{y}) = \mathbf{A}\mathbf{y}$$

para todo vector $\mathbf{y}$, lo que requiere $\mathbf{A} = \mathbf{A}^2$, es decir, la matriz proyección debe ser idempotente. Demostraremos ahora que la matriz idempotente $\mathbf{A}$ que proyecta sobre el espacio generado por las columnas de una matriz $\mathbf{X}$, E_p, viene dada por:

$$\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Probemos primero que $\mathbf{A}$ depende del subespacio E_p, pero no de la base elegida. En efecto, si consideramos otra base $\mathbf{B}$ generadora del subespacio dada por:

$$\mathbf{B} = \mathbf{X}\mathbf{C}$$

donde $\mathbf{C}$ es $(p \times p)$ y no singular, como $(\mathbf{G}\mathbf{P})^{-1} = \mathbf{P}^{-1}\mathbf{G}^{-1}$ para $\mathbf{G}$ y $\mathbf{P}$ matrices cuadradas no singulares, tendremos que:

$$\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}' = \mathbf{X}\mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}'\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{A}$$

por tanto, $\mathbf{A}$ no depende de la base escogida. A continuación veremos que el vector $\mathbf{v}$ definido por:

$$\mathbf{v} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

verifica las condiciones de una proyección. Demostraremos, en primer lugar, que $\mathbf{v}$ está contenido en E_p. Llamemos

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

a los coeficientes de la proyección de $\mathbf{y}$ sobre el espacio de las columnas de $\mathbf{X}$, que representaremos por $\mathbf{x}_1, ..., \mathbf{x}_p$. Entonces

$$\mathbf{v} = \mathbf{X}\boldsymbol{\beta} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + ... + \beta_p\mathbf{x}_p,$$

y al ser $\mathbf{v}$ una combinación lineal de las columnas de $\mathbf{X}$ pertenece a E_p. Demostraremos ahora que $\mathbf{y} - \mathbf{v}$ es ortogonal a E_p. Todo vector de E_p puede expresarse como:

$$\mathbf{u} = \alpha_1\mathbf{x}_1 + ... + \alpha_p\mathbf{x}_p = \mathbf{X}\boldsymbol{\alpha}$$

y por tanto,

$$\mathbf{u}'(\mathbf{y} - \mathbf{v}) = \mathbf{u}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \boldsymbol{\alpha}'(\mathbf{X}' - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{0}$$

es decir, $\mathbf{y} - \mathbf{v}$ es ortogonal a cualquier vector de E_p, lo que demuestra el teorema. ■

Teorema 2.2 *La condición necesaria y suficiente para que* $\mathbf{v} = \mathbf{A}\mathbf{y}$*, donde* $\mathbf{A}$ *es una matriz cuadrada, sea la proyección ortogonal de* $\mathbf{y} \in \Re^n$ *sobre un cierto espacio* E_p*, es que* $\mathbf{A}$ *sea idempotente* $(\mathbf{A} = \mathbf{A}', \mathbf{A}^2 = \mathbf{A})$ *de rango p.*

Demostración La condición es necesaria: si $\mathbf{A}$ define una proyección, según el teorema anterior puede expresarse como $\mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, siendo $\mathbf{X}$ una matriz que contiene, en columnas, una base del espacio, por lo que $\mathbf{A}$ es simétrica e idempotente.

La condición es suficiente: supongamos que $\mathbf{A}$ es idempotente y hagamos

$$\mathbf{y} = \mathbf{A}\mathbf{y} + (\mathbf{I} - \mathbf{A})\mathbf{y}$$

Vamos a demostrar que el vector $(\mathbf{I} - \mathbf{A})\mathbf{y}$ es ortogonal a todo vector que pertenezca a E_p. Sea $\mathbf{A}\mathbf{c}$ un vector cualquiera que pertenece a E_p.

$$(\mathbf{A}\mathbf{c})'(\mathbf{I} - \mathbf{A})\mathbf{y} = \mathbf{c}'(\mathbf{A}' - \mathbf{A})\mathbf{y} = 0$$

por tanto, si $\mathbf{A}$ es idempotente, $\mathbf{A}\mathbf{y}$ es la proyección de $\mathbf{y}$ sobre el espacio generado por las columnas de $\mathbf{A}$. ■

Teorema 2.3 *Si* $\mathbf{y} \in \Re^n$*,* $\mathbf{v}$ *es su proyección sobre* E_p *y* $\mathbf{z}$ *es cualquier otro vector de* E_p*, se verifica, llamando* $\|\mathbf{y}\|$ *a la norma del vector* $\mathbf{y}$*:*

$$\|\mathbf{y}\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{y} - \mathbf{v}\|^2$$

$$\|\mathbf{y} - \mathbf{z}\|^2 = \|\mathbf{v} - \mathbf{z}\|^2 + \|\mathbf{y} - \mathbf{v}\|^2.$$

Demostración Estas expresiones representan el teorema de Pitágoras en un espacio general. Como, por definición de proyección, $\mathbf{v}'(\mathbf{y}-\mathbf{v})=0$, entonces $\mathbf{v}'\mathbf{y}=\mathbf{v}'\mathbf{v}$. Por otro lado:

$$(\mathbf{y}-\mathbf{v})'(\mathbf{y}-\mathbf{v})=\mathbf{y}'\mathbf{y}-\mathbf{v}'\mathbf{y}-\mathbf{y}'\mathbf{v}+\mathbf{v}'\mathbf{v}=\mathbf{y}'\mathbf{y}-\mathbf{v}'\mathbf{v},$$

que escribiremos

$$\mathbf{y}'\mathbf{y}=\mathbf{v}'\mathbf{v}+(\mathbf{y}-\mathbf{v})'(\mathbf{y}-\mathbf{v})$$

que es la primera igualdad. Para demostrar la segunda, partamos de la identidad

$$\mathbf{y}-\mathbf{z}=\mathbf{y}-\mathbf{v}+\mathbf{v}-\mathbf{z}$$

y multiplicando por el vector transpuesto y utilizando que $\mathbf{y}-\mathbf{v}$ debe ser ortogonal a $\mathbf{v}-\mathbf{z}$, por serlo a todos los vectores de E_p, el teorema queda demostrado. ■

Una consecuencia de este teorema es que podemos definir la proyección ortogonal de un vector $\mathbf{y}$ sobre un espacio E_p como aquel vector $\mathbf{v}$ de E_p tal que $\|\mathbf{y}-\mathbf{v}\|$ es mínimo. En este sentido el vector proyección es, el "más próximo" al original. En efecto, como, para cualquier vector $\mathbf{z}$ del plano:

$$\|\mathbf{y}-\mathbf{z}\|\geq\|\mathbf{y}-\mathbf{v}\|$$

el vector $\mathbf{v}$, proyección ortogonal, minimiza las distancias entre el espacio E_p y el vector $\mathbf{y}$.

Teorema 2.4 *Si $\mathbf{y}\in\Re^n$, el cuadrado de la norma de su proyección sobre un espacio E_p definido por las columnas de la matriz $\mathbf{X}$ vendrá dado por $\mathbf{y}'\mathbf{A}\mathbf{y}$, donde $\mathbf{A}$ es idempotente.*

Demostración El vector proyectado será $\mathbf{A}\mathbf{y}$, donde $\mathbf{A}$ es idempotente, y su norma será:

$$(\mathbf{A}\mathbf{y})'(\mathbf{A}\mathbf{y})=\mathbf{y}'\mathbf{A}\mathbf{y}.$$

■

Teorema 2.5 *Si $\mathbf{y}\in\Re^n$ y proyectamos este vector sobre espacios ortogonales, $E_1,...,E_h$, definidos por matrices de proyección , $A_1,...A_h$, donde:*

$$n=\sum_{i=1}^{h}rg(\mathbf{A}_i)$$

se verifica:

$$\mathbf{y}'\mathbf{y}=\mathbf{y}'\mathbf{A}_1\mathbf{y}+\mathbf{y}'\mathbf{A}_2\mathbf{y}+...+\mathbf{y}'\mathbf{A}_h\mathbf{y}.$$

Ejercicios 2.5

2.5.1 Calcule la proyección ortogonal del vector (1,1 3) sobre el espacio generado por las dos variables (1 ,1,1) y (0, 1,2).

2.5.2 Exprese al vector anterior como combinación lineal de las dos variables.

2.5.3 Obtener el vector ortogonal al vector proyección.

2.5.4 Demuestre que el resultado anterior es equivalente a realizar la regresión simple entre la variable (1,1 3) y la variable (0, 1,2).

2.5.5 Demostrar, utilizando el Teorema 2.1, que para calcular los coeficientes de regresión múltiple entre una variable y un conjunto de variables incorreladas basta con calcular los coeficientes de las regresiones simples.

2.6 (*)DERIVADAS MATRICIALES

Definición 2.1 *Dada un función f que depende de n variables, $x_1, ..., x_n$, que pueden considerarse componentes de un vector $\mathbf{x}$, la derivada de f respecto a $\mathbf{x}$ es un vector cuyos componentes son la derivada de f respecto a cada componente de $\mathbf{x}$.*

Ejemplo 2.1 *Si* $f = 5x_1 + 2x_2 + 3x_3$

$$\frac{\partial f}{\partial x} = \begin{bmatrix} 5 \\ 2 \\ 3 \end{bmatrix}$$

Los siguientes resultados son consecuencia de la definición

Corolario 2.1 *Si* $f = \mathbf{a}'\mathbf{x}$ *tendremos que:*

$$\frac{\partial(\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}$$

Corolario 2.2 *Si* $f = \mathbf{x}'\mathbf{A}\mathbf{x}$, *donde* $\mathbf{A}$ *es cuadrada y simétrica:*

$$\frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

Demostración Resulta de aplicar la definición anterior, como:

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{j>i} a_{ij}x_i x_j$$

tendremos que:

$$\frac{\partial(\mathbf{x}\mathbf{A}\mathbf{x})}{\partial x_1} = 2a_{11}x_1 + 2a_{12}x_2 + ... + 2a_{1n}x_n = 2\mathbf{a}_1'\mathbf{x}$$

donde $\mathbf{a}_1'$ es la primera fila de la matriz. Por tanto:

$$\frac{\partial(\mathbf{x}\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} 2\mathbf{a}_1'\mathbf{x} \\ 2\mathbf{a}_2'\mathbf{x} \\ \vdots \\ 2\mathbf{a}_n'\mathbf{x} \end{bmatrix} = 2\mathbf{A}\mathbf{x}$$

■

Definición 2.2 *Dada un función f que depende de np variables, $x_{11}, ..., x_{np}$, que son los componentes de una matriz rectangular $n \times p$, $\mathbf{X}$, la derivada de f respecto a $\mathbf{X}$ se define como la matriz cuyos componentes son la derivada de f respecto a cada componente de $\mathbf{X}'$. La derivada es pues una matriz $p \times n$ con las dimensiones de $\mathbf{X}'$.*

Los siguientes resultados se comprueban aplicando la definición

Corolario 2.3 ***Corolario 2.4*** *Si* $f = \mathbf{a}'\mathbf{X}\mathbf{b}$

$$\frac{\partial(\mathbf{a}'\mathbf{X}\mathbf{b})}{\partial\mathbf{X}} = \mathbf{b}\mathbf{a}'$$

Definición 2.3 ***Ejemplo 2.2 Corolario 2.5*** *Si* $f = \mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{b}$

$$\frac{\partial(\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{b})}{\partial\mathbf{X}} = (\mathbf{a}\mathbf{b}' + \mathbf{b}\mathbf{a}')\mathbf{X}'$$

Definición 2.4 *Dado un vector* $\mathbf{y}$ *cuyos componentes son funciones* f_i *de un vector de variables* $\mathbf{x}' = (x_1, ..., x_n)$*, definimos la derivada de* $\mathbf{y}$ *respecto a* $\mathbf{x}$ *como la matriz cuyas columnas son las derivadas de los componentes* f_i *respecto a* $\mathbf{x}$*. Es decir, si:*

$$\mathbf{y}' = (f_1(\mathbf{x}), ..., f_n(\mathbf{x}))$$

entonces:

$$\frac{\partial\mathbf{y}}{\partial\mathbf{x}} = \left[\frac{\partial f_1}{\partial\mathbf{x}}, \dots, \frac{\partial f_n}{\partial\mathbf{x}}\right] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

Corolario 2.6 *Si* $\mathbf{y} = \mathbf{A}\mathbf{x}$*, donde* $\mathbf{A}$ *es una matriz cualquiera.*

$$\frac{\partial(\mathbf{A}\mathbf{x})}{\partial\mathbf{x}} = \mathbf{A}'$$

Demostración Para deducir este resultado de la definición anterior, escribamos la matriz $\mathbf{A}$ como:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1' \\ \vdots \\ \mathbf{a}_n' \end{bmatrix}$$

donde cada $\mathbf{a}_1'$ es una fila de la matriz. Entonces:

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \begin{bmatrix} \mathbf{a}_1'\mathbf{x} \\ \vdots \\ \mathbf{a}_n'\mathbf{x} \end{bmatrix}$$

con lo que:

$$\frac{\partial f_i}{\partial\mathbf{x}} = \frac{\partial(\mathbf{a}_i'\mathbf{x})}{\partial\mathbf{x}} = \mathbf{a}_i$$

por tanto, según lo anterior:

$$\frac{\partial\mathbf{y}}{\partial\mathbf{x}} = [\mathbf{a}_1, \dots, \mathbf{a}_n] = \mathbf{A}'$$

■

Otras propiedades

Puede deducirse, extendiendo las definiciones anteriores, que, si los elementos de la matriz cuadrada y no singular $\mathbf{X}$ son distintos:

$$\begin{array}{ll} \text{a)} \ \dfrac{\partial ln\left|\mathbf{X}\right|}{\partial \mathbf{X}} = (\mathbf{X}')^{-1} & \text{d)} \ \dfrac{\partial tr\left(\mathbf{XB}\right)}{\partial \mathbf{X}} = \mathbf{B}' \\ \text{b)} \ \dfrac{\partial \left|\mathbf{X}\right|}{\partial \mathbf{X}} = \left|\mathbf{X}\right|(\mathbf{X}')^{-1} & \text{e)} \ \dfrac{\partial tr\left(\mathbf{X}'\mathbf{AXB}\right)}{\partial \mathbf{X}} = \mathbf{BX}'\mathbf{A} + \mathbf{B}'\mathbf{X}'\mathbf{A}' \\ \text{c)} \ \frac{\partial tr(\mathbf{BXC})}{\partial \mathbf{X}} = \mathbf{B}'\mathbf{C}' & \text{f)} \ \frac{\partial tr\left(\mathbf{BX}^{-1}\right)}{\partial \mathbf{X}} = -\left(\mathbf{X}^{-1}\mathbf{BX}^{-1}\right) \end{array}$$

además, si $\mathbf{X}$ es simétrica:

$$\begin{aligned} \frac{\partial tr(\mathbf{XB})}{\partial \mathbf{X}} &= \mathbf{B} + \mathbf{B}' - diag(\mathbf{B}) \\ \frac{\partial \left|\mathbf{X}\right|}{\partial \mathbf{X}} &= \left|\mathbf{X}\right|\left(2\mathbf{X}^{-1} - diag\left(\mathbf{X}^{-1}\right)\right) \end{aligned}$$

El lector interesado puede encontrar las demostraciones de estos resultados en Bibby y Toutenterg (1977), Graybill (1983) y Pollock (1979).

Ejercicios 2.6

2.6.1 Calcular la derivada con respecto al vector $\mathbf{x} = (x_1, x_2)'$ de las funciones siguientes

a) $f_1(\mathbf{x}) = 2x_1 + 3x_2$,

b) $f_2(\mathbf{x}) = 4x_1^2 - 3x_1x_2$,

c) $f_3(\mathbf{x}) = 3x_1^4x_2^3 + 2x_1^2x_2^2 - 7x_1x_2^3 + 6$

2.6.2 Calcular la derivada con respecto al vector $\mathbf{x} = (x_1, x_2)'$ de las funciones vectoriales siguientes, construidas con la notación de 2.6.1

a)$\mathbf{f}_1(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x}))'$,

b)$\mathbf{f}_2(\mathbf{x}) = (2f_1(\mathbf{x}) + 5f_2(\mathbf{x}), -6f_3(\mathbf{x}))'$,

2.6.3 Si $\mathbf{x} = (x_1, x_2, x_3, x_4)'$ y $\mathbf{X} = \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix}$, comprobar que $\dfrac{\partial ln\left|\mathbf{X}\right|}{\partial x_1} = \frac{x_4}{x_1x_4 - x_2x_3}$ y utilizar este resultado para confirmar la expresión de $\dfrac{\partial ln\left|\mathbf{X}\right|}{\partial \mathbf{X}}$.

2.6.4 En el ejercicio anterior comprobar que $\dfrac{\partial \left|\mathbf{X}\right|}{\partial \mathbf{X}} = \begin{bmatrix} x_4 & -x_3 \\ -x_2 & x_1 \end{bmatrix}$. Utilizar esta expresión para verificar la ecuación dada de la derivada del determinante de una matriz cuadrada.

2.6.5 Si $\mathbf{x} = (x_1, x_2, x_3)'$ y $\mathbf{X} = \begin{bmatrix} x_1 & x_2 \\ x_2 & x_3 \end{bmatrix}$, comprobar que $\dfrac{\partial \left|\mathbf{X}\right|}{\partial \mathbf{X}} = \begin{bmatrix} x_3 & -2x_2 \\ -2x_2 & x_1 \end{bmatrix}$, Utilizar este resultado para confirmar la expresión general de la derivada del determinante de una matriz cuadrada.

Capítulo 3

DESCRIPCIÓN DE DATOS MULTIVARIANTES

3.1 INTRODUCCIÓN

En este capítulo y en el siguiente vamos a estudiar como describir un conjunto de datos multivariantes. Supondremos que hemos observado un conjunto de variables en un conjunto de elementos de una población y en este capítulo presentaremos métodos para resumir los valores de las variables y describir su estructura de dependencia. En el capítulo siguiente completaremos el análisis descriptivo analizando como representar los datos gráficamente y decidir respecto a posibles transformaciones de las variables originales que conduzcan a una descripción más simple. También comentaremos el problema de limpiar los datos de valores atípicos, que son observaciones debidas a errores de medida o otras causas de heterogeneidad.

El análisis descriptivo que presentamos en este capítulo debe siempre aplicarse como primer paso para comprender la estructura de los datos y extraer la información que contienen, antes de pasar a los métodos más complejos de los capítulos siguientes. Las herramientas simples que describimos en estos dos capítulos pueden, en ocasiones, resolver el problema que ha motivado la recogida de los datos. En particular, cuando el interés se centra en la relación entre las variables o en la comparación de dos conjuntos de datos, los métodos descriptivos pueden ser de gran ayuda antes de emprender estudios más complejos.

3.2 DATOS MULTIVARIANTES

3.2.1 Tipos de variables

La información de partida para los métodos estudiados en este libro puede ser de varios tipos. La más habitual es una tabla donde aparecen los valores de p variables observadas sobre n elementos. Las variables pueden ser cuantitativas, cuando su valor se exprese numéricamente, como la edad de una persona, su estatura o su renta, o cualitativas, cuando su valor sea un atributo o categoría, como el género, el color de los ojos o el municipio de nacimiento. Las variables cuantitativas pueden a su vez clasificarse en continuas o de intervalo, cuando pueden tomar cualquier valor real en un intervalo, como la estatura, o discretas, cuando sólo toman

CO	x_1	x_2	x_3
A	1	0	0
V	0	1	0
C	0	0	1
N	0	0	0

Tabla 3.1: Codificación de variables categóricas

valores enteros, como el número de hermanos. Las variables cualitativas pueden clasificarse en binarias, cuando toman únicamente dos valores posibles, como el género (mujer, hombre) o generales, cuando toman muchos valores posibles, como el municipio de residencia.

Supondremos en adelante que las variables binarias se han codificado como numéricas (Por ejemplo, la variable género se convierte en numérica asignando el cero al varón y el uno a mujer). Las variables cualitativas pueden también codificarse numéricamente, pero requieren un tratamiento distinto. Si los valores de las categorías no tienen relación entre sí, la forma más útil de codificarlas es convirtiéndolas en variables binarias. Por ejemplo, supongamos la variable color de los ojos, CO, y para simplificar supongamos que las categorías posibles son azules (A), verdes (V), castaños (C) y negros (N). Tenemos $p = 4$ categorías que podemos representar con $p - 1 = 3$ variables binarias definidas como:

a) $x_1 = 1$ si CO=A, $x_1 = 0$ en otro caso.

b) $x_2 = 1$ si CO=V, $x_2 = 0$ en otro caso.

c) $x_3 = 1$ si CO=C, $x_3 = 0$ en otro caso.

La tabla 3.1 presenta la codificación de la variable atributo CO en las tres variables binarias cuantitativas, x_1, x_2, x_3

Si el número de clases posibles de una variable cualitativa es muy grande este procedimiento siempre puede aplicarse pero puede lógicamente dar lugar a muchas variables. Conviene entonces ver si podemos agrupar las clases o categorías para evitar tener variables que casi siempre toman el mismo valor (cero si la categoría es poco frecuente o uno si lo es mucho).

Naturalmente la variable CO podría también haberse codificado dando valores numéricos arbitrarios a las categorías, por ejemplo, A=1, V=2, C=3, N=4, pero esta codificación tiene el inconveniente de sugerir una graduación de valores que puede no existir. Sin embargo, cuando los atributos pueden interpretarse en función de los valores de una variable continúa tiene más sentido codificarla con números que indiquen el orden de las categorías. Por ejemplo, si tenemos empresas pequeñas, medianas y grandes, en función del número de trabajadores, tienen sentido codificarlas con los números 1, 2, y 3, aunque conviene siempre recordar que estos números sólo tienen un sentido de orden.

3.2.2 La matriz de datos

Supondremos en adelante que hemos observado p variables numéricas en un conjunto de n elementos. Cada una de estas p variables se denomina una variable **escalar o univariante** y el conjunto de las p variables forman una variable **vectorial o multivariante**. Los valores de las p variables escalares en cada uno de los n elementos pueden representarse en una matriz, $\mathbf{X}$, de dimensiones $(n \times p)$, que llamaremos **matriz de datos.** Denotaremos por x_{ij}

al elemento genérico de esta matriz, que representa el valor de la variable escalar j sobre el individuo i. Es decir:

$$\begin{aligned} \text{datos } x_{ij} \text{ donde } i &= 1, ..., n \text{ representa el individuo;} \\ j &= 1, ..., p \text{ representa la variable} \end{aligned}$$

Algunos ejemplos de datos que se utilizan en el análisis multivariante son:

1. En 100 estudiantes de una universidad medimos la edad, el género (1 mujer, 0 hombre), la calificación media, el municipio de residencia (que se codifica en 4 categorías en función del tamaño) y el curso más alto en que se encuentra matriculado. Los datos iniciales se representan en una tabla de 100 filas, cada una de ellas correspondiente a los datos de un estudiante. La tabla tendrá 5 columnas, cada una de ellas conteniendo los valores de una de las 5 variables definidas. De estas 5 variables 3 son cuantitativas, una binaria (el género) y otra cualitativa general (municipio de residencia, que tomará los valores 1, 2, 3, y 4). Alternativamente podríamos codificar el municipio de residencia con tres variables binarias, y entonces, la matriz de datos tendrá n =100 filas y p = 7 columnas correspondientes a las tres cuantitativas, el género, y las tres variables binarias adicionales para describir el tamaño del municipio de residencia.

2. En cada una de las 138 empresas de una zona medimos el número de trabajadores, la facturación, el sector industrial y la cantidad recibida en ayudas oficiales. Si clasificamos el sector en ocho clases con siete variables binarias la matriz de datos será de dimensiones 138×10 con tres variables cuantitativas y siete binarias (que describen el sector industrial).

3. En 400 puntos de una ciudad instalamos controles que proporcionan cada hora las medidas de 30 variables ambientales y de contaminación atmosférica en dicho punto. Cada hora tendremos una matriz de datos con 400 filas, los puntos de observación, y 30 columnas, las 30 variables observadas.

La matriz de datos, $\mathbf{X}$, puede representarse de dos formas distintas. Por filas, como:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \dots & \dots & x_{2p} \\ \vdots & & & \\ x_{n1} & \dots & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

donde cada variable $\mathbf{x}_i'$ es un vector fila, $p \times 1$, que representa los valores de las p variables sobre el individuo i. Alternativamente, podemos representar la matriz $\mathbf{X}$ por columnas:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{(1)} \dots \mathbf{x}_{(p)} \end{bmatrix}$$

donde ahora cada variable $\mathbf{x}_{(j)}$ es un vector columna, $n \times 1$, que representa la variable escalar x_j medida en los n elementos de la población. Llamaremos $\mathbf{x} = (x_1, ..., x_p)'$ a la variable multivariante formada por las p variables escalares que toma los valores particulares $\mathbf{x}_1, ..., \mathbf{x}_n$, en los n elementos observados.

3.2.3 Análisis univariante

Describir datos multivariantes supone estudiar cada variable aisladamente y además las relaciones entre ellas. Supondremos que el lector está familiarizado con el análisis descriptivo de una variable, y aquí expondremos únicamente las fórmulas que utilizaremos en otras partes del libro. El estudio univariante de la variable escalar x_j implica calcular su *media*:

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

que para una variable binaria es la frecuencia relativa de aparición del atributo y para una numérica es el centro de gravedad o geométrico de los datos. Se calcula una medida de variabilidad con relación a la media, promediando las desviaciones entre los datos y su media. Si definimos las *desviaciones* mediante $d_{ij} = (x_{ij} - \overline{x}_j)^2$, donde el cuadrado se toma para prescindir del signo, se define la *desviación típica* por:

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} d_{ij}}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}{n}} \tag{3.1}$$

y su cuadrado es la *varianza*, $s_j^2 = \sum_{i=1}^{n} d_{ij}/n$. Para comparar la variabilidad de distintas variables conviene construir medidas de variabilidad relativa que no dependan de las unidades de medida. Una de estas medidas es el *coeficiente de variación*

$$CV_j = \sqrt{\frac{s_j^2}{\overline{x}_j^2}}$$

donde de nuevo se toman los cuadrados para prescindir del signo y suponemos que $\overline{x}_j$ es distinto de cero. En tercer lugar, conviene calcular los *coeficientes de asimetría*, que miden la simetría de los datos respecto a su centro, y que se calculan como:

$$A_j = \frac{1}{n}\frac{\sum (x_{ij} - \overline{x}_j)^3}{s_j^3}.$$

Este coeficiente es cero para una variable simétrica. Cuando el valor absoluto del coeficiente es aproximadamente mayor que uno podemos concluir que los datos tienen una distribución claramente asimétrica.

Una característica importante de un conjunto de datos es su homogeneidad. Si las desviaciones d_{ij} son muy distintas, esto sugiere que hay datos que se separan mucho de la media y que tenemos por tanto alta heterogeneidad. Una posible medida de homogeneidad es la varianza de las d_{ij}, dada por:

$$\frac{1}{n}\sum_{i=1}^{n}(d_{ij} - s_j^2)^2$$

ya que, según (3.1), la media de las desviaciones $\overline{d_j} = s^2$. Se calcula una medida adimensional análoga al coeficiente de variación dividiendo la varianza de las desviaciones por el cuadrado

de la media, s^4, con lo que tenemos el *coeficiente de homogeneidad,* que puede escribirse

$$H_j = \frac{\frac{1}{n}\sum_{i=1}^{n}(d_{ij} - s_j^2)^2}{s_j^4}.$$

Este coeficiente es siempre mayor o igual a cero. Desarrollando el cuadrado del numerador como $\sum_{i=1}^{n}(d_{ij}-s_j^2)^2 = \sum_{i=1}^{n} d_{ij}^2 + ns_j^4 - 2s_j^2 \sum_{i=1}^{n} d_{ij}$ este coeficiente puede escribirse también como:

$$H_j = \frac{1}{n}\frac{\sum (x_{ij} - \overline{x}_j)^4}{s_j^4} - 1 = K_j - 1.$$

El primer miembro de esta expresión, K_j, es una forma alternativa de medir la homogeneidad y se conoce como *coeficiente de kurtosis.* Como $H_j \geq 0$, el coeficiente de kurtosis será igual o mayor que uno. Ambos coeficientes miden la relación entre la variabilidad de las desviaciones y la desviación media. Es fácil comprobar que :

1. Si hay unos pocos datos atípicos muy alejados del resto, la variabilidad de las desviaciones será grande, debido a estos valores y los coeficientes de kurtosis o de homogeneidad serán altos.

2. Si los datos se separan en dos mitades correspondientes a dos distribuciones muy alejadas entre sí, es decir, tenemos dos conjuntos separados de datos distintos, la media de los datos estará equidistante de los dos grupos de datos y las desviaciones de todos los datos serán similares, con lo que el coeficiente H_j será muy pequeño (cero en el caso extremo en que la mitad de los datos son iguales a cualquier número, $-a$, y la otra mitad igual a a).

Un objetivo central de la descripción de datos es decidir si los datos son una muestra homogénea de una población o corresponden a una mezcla de poblaciones distintas que deben estudiarse separadamente. Como veremos en el capítulo siguiente, un caso especialmente importante de heterogeneidad es la presencia de una pequeña proporción de observaciones atípicas (outliers), que corresponden a datos heterogéneos con el resto. La detección de estas observaciones es fundamental para una correcta descripción de la mayoría de los datos, ya que. como veremos, estos valores extremos distorsionan los valores descriptivos del conjunto. El coeficiente de kurtosis puede ayudar en este objetivo, ya que tomará un valor alto, mayor que 7 u 8. Por ejemplo, si contaminamos datos que provienen de una distribución normal con un 1% de atípicos generados por otra distribución normal con la misma media, pero una varianza 20 veces mayor, el coeficiente de kurtosis será alrededor de 10. Siempre que observemos un valor alto de la kurtosis para una variable esto implica heterogeneidad por uno pocos atípicos muy alejados del resto.

Aparece un tipo distinto de heterogeneidad cuando tenemos una mezcla de dos poblaciones, de manera que una proporción importante de los datos, entre el 25% y el 50%, son heterogeneos con el resto. En este caso, el coeficiente de kurtosis es pequeño, menor de dos, y es fácil comprobar que si mezclamos a partes iguales dos distribuciones muy distintas, la kurtosis de la distribución resultante tiende a uno, el valor mínimo del coeficiente, cuando aumenta la separación entre las poblaciones

La presencia posible de datos atípicos recomienda calcular junto a los estadísticos tradicionales medidas robustas de centralización y dispersión de los datos. Para centralización

conviene calcular la mediana, que es el valor que se encuentra en la posición central al ordenar los datos. Para la dispersión, la MEDA, que es la mediana de las desviaciones absolutas respecto a la mediana. Finalmente siempre conviene representar gráficamente las variables continuas mediante un histograma o un diagrama de caja (véase por ejemplo Peña, 2001). En el análisis inicial de los datos conviene siempre calcular la media y la mediana de cada variable. Si ambas son similares, la media es un buen indicador del centro de los datos. Sin embargo, si difieren mucho, la media puede no ser una buena medida del centro de los datos debido a: (1) una distribución asimétrica, (2) la presencia de valores atípicos (que afectaran mucho a la media y poco a la mediana) (3) heterogeneidad en los datos.

A continuación pasaremos al análisis multivariante de las observaciones. En este capítulo presentaremos como obtener medidas conjuntas de centralización y dispersión para el conjunto de variables y medidas de dependencia lineal entre pares de variables y entre todas ellas.

3.3 MEDIDAS DE CENTRALIZACIÓN: EL VECTOR DE MEDIAS

La medida de centralización más utilizada para describir datos multivariantes es el vector de medias, que es un vector de dimensión p cuyos componentes son las medias de cada una de las p variables. Puede calcularse, como el caso escalar, promediando las medidas de cada elemento, que ahora son vectores:

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{bmatrix} \tag{3.2}$$

Su expresión a partir de la matriz de datos es :

$$\overline{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}, \tag{3.3}$$

donde $\mathbf{1}$ representará siempre un vector de unos de la dimensión adecuada. En efecto, escribiendo la matriz $\mathbf{X}$ en términos de sus vectores fila, que son vectores de dimensión $1 \times p$ que contienen los valores de las p variables en cada elemento de la muestra, estos vectores son las columnas de $\mathbf{X}'$, y tendremos que:

$$\overline{\mathbf{x}} = \frac{1}{n}\left[\mathbf{x}_1 \ldots \mathbf{x}_n\right]\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \tag{3.4}$$

que conduce a (3.2). El vector de medias se encuentra en el centro de los datos, en el sentido de hacer cero la suma de desviaciones:

$$\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}}) = 0.$$

ya que esta suma es $\sum_{i=1}^{n} \mathbf{x}_i - n\overline{\mathbf{x}}$, y aplicando la definición (3.2) es inmediato que esta suma es cero.

Las medidas de centralización escalares basadas en el orden de las observaciones no pueden generalizarse fácilmente al caso multivariante. Por ejemplo, podemos calcular el vector de medianas, pero este punto no tiene necesariamente una situación como centro de los datos. Esta dificultad proviene de la falta de un orden natural de los datos multivariantes.

Ejemplo 3.1 *La tabla A.5 del Apéndice de Datos, MEDIFIS, presenta ocho variables físicas tomadas en un grupo de 27 estudiantes. Las variables son sexo (sex con 0 para mujer, 1 para varón), estatura (est, en cm.), peso (pes, en kgr.), longitud de pie (lpie, en cm), longitud de brazo (lbra, en cm), anchura de la espalda (aes, en cm), diámetro de cráneo (dcr, en cm) y longitud entre la rodilla y el tobillo (lrt, en cm).*

La tabla 3.2 presenta las medias y desviaciones típicas de las variables, así como otras medidas de la distribución univariante de cada variable.

	sex	est	pes	lpie	lbr	aes	dcr	lrt
Medias	.44	168.8	63.9	39.0	73.5	45.9	57.2	43.1
D. Típicas	.53	10.0	12.6	2.8	4.9	3.9	1.8	3.1
Coef. asimetría	.22	.15	.17	.27	.37	-.22	.16	.56
Coef. kurtosis	1.06	1.8	2.1	1.9	2.1	2.4	2.0	3.4
Coef. variación	1.2	.06	.20	.07	.07	.09	.03	.07

Tabla 3.2: Análisis descriptivo de las medidas físicas

En la variable binaria sexo la media es la proporción de unos (hombres) en los datos, la desviación típica es $\sqrt{p(1-p)}$, *donde* p *es la media. El lector puede comprobar que para variables binarias el coeficiente de kurtosis es*

$$\frac{p^3 + (1-p)^3}{p(1-p)}$$

y en este caso, como $p = .44$ *el coeficiente de kurtosis es 1.06. Para las variables continuas las medias describen los valores centrales. Si miramos los coeficientes de variación se observa que en las medidas de longitudes, como la estatura, la longitud del pie y las extremidades, que vienen determinadas más por la herencia genética que por nuestros hábitos, la variabilidad relativa es del orden del 7% . El diámetro del cráneo es mucho más constante, con una variabilidad relativa de menos de la mitad, el 3%. La variabilidad relativa de las variables que dependen más de nuestros hábitos, como el peso, es mucho mayor, del 20%. Las distribuciones son aproximadamente simétricas, a juzgar por los bajos valores de los coeficientes de asimetría. Los coeficientes de kurtosis son bajos, menores o iguales a dos para tres de las variables, lo que puede indicar la presencia de dos poblaciones mezcladas, como veremos en la sección 3.6. Ninguna variable tiene alta kurtosis, por lo que podemos descartar la presencia de unos pocos valores atípicos grandes.*

La tabla 3.3 presenta dos medidas robustas, las mediana (valor central de los datos) y la MEDA o mediana de las desviación absolutas para cada variable. Estas medidas confirman los comentarios anteriores.

	est	pes	lpie	lbr	aes	dcr	lrt
medianas	168	65	39	73	46	57	43
medas	8.51	10.50	2.38	3.96	3.26	1.52	2.39
meda/mediana	.05	.16	.05	.05	.07	.03	.06

Tabla 3.3: Análisis descriptivo robusto de las medidas físicas

Se observa que las medianas son muy similares a las medias y las medas a las desviaciones típicas, lo que sugiere falta de valores extremos. Los coeficientes de variación robustos, calculados como ratio entre la meda y la mediana son también básicamente similares a los anteriores. Hay que señalar que, en general, la meda es más pequeña que la desviación típica, y que, por tanto, estos coeficientes de variación serán más pequeños que los originales. Lo importante es que la estructura es similar entre las variables. La figura 3.1 muestra el histograma de la variable estatura donde se aprecia que los datos parecen ser la mezcla de dos distribuciones. Esto es esperable, ya que tenemos juntos hombres y mujeres.

Figura 3.1: Histograma de las estaturas donde se observa una distribución mezclada

3.4 LA MATRIZ DE VARIANZAS Y COVARIANZAS

Como hemos comentado, para variables escalares la variabilidad respecto a la media se mide habitualmente por la varianza, o su raíz cuadrada, la desviación típica. La relación lineal entre dos variables se mide por la covarianza. La *covarianza* entre dos variables (x_j, x_k) se

calcula con:

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)$$

y mide su dependencia lineal. Esta información para una variable multivariante puede presentarse de forma compacta en la *matriz de varianzas y covarianzas*. Definimos esta matriz como:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' \tag{3.5}$$

que es una matriz cuadrada y simétrica que contiene en la diagonal las varianzas y fuera de la diagonal las covarianzas entre las variables. En efecto, al multiplicar los vectores :

$$\begin{bmatrix} x_{i1} - \overline{x}_1 \\ \vdots \\ x_{ip} - \overline{x}_p \end{bmatrix} [x_{i1} - \overline{x}_1, \ldots, \overline{x}_{ip} - \overline{x}_p] = \begin{bmatrix} (x_{i1} - \overline{x}_1)^2 \ldots (x_{i1} - \overline{x}_1)(x_{ip} - \overline{x}_p) \\ \vdots \qquad\qquad\qquad \vdots \\ (x_{ip} - \overline{x}_p)(x_{i1} - \overline{x}_1) \ldots (x_{ip} - \overline{x}_p)^2 \end{bmatrix}$$

se obtiene la matriz de cuadrados y productos cruzados de las p variables en el elemento i. Al sumar para todos los elementos y dividir por n se obtienen las varianzas y covarianzas entre las variables. La matriz de varianzas y covarianzas, que llamaremos para simplificar matriz de covarianzas, es la matriz simétrica de orden p con forma:

$$\mathbf{S} = \begin{bmatrix} s_1^2 & \ldots & s_{1p} \\ \vdots & & \vdots \\ s_{p1} & \ldots & s_p^2 \end{bmatrix}.$$

3.4.1 Cálculo a partir de la matriz de datos centrados

La matriz $\mathbf{S}$ puede obtenerse directamente a partir de la *matriz de datos centrados* $\widetilde{\mathbf{X}}$, que se define como la matriz resultado de restar a cada dato su media:

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}'.$$

Sustituyendo el vector de medias por su expresión (3.3):

$$\widetilde{\mathbf{X}} = \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X} = \mathbf{P}\mathbf{X}, \tag{3.6}$$

donde la matriz cuadrada $\mathbf{P}$ está definida por

$$\mathbf{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$$

y es simétrica e idempotente (compruebe el lector que $\mathbf{PP} = \mathbf{P}$). La matriz $\mathbf{P}$ tiene rango $n-1$ (es ortogonal al espacio definido por el vector $\mathbf{1}$, ya que $\mathbf{P1} = 0$) y proyecta los datos ortogonalmente al espacio definido por el vector constante (con todas las coordenadas iguales). Entonces la matriz $\mathbf{S}$ puede escribirse:

$$\mathbf{S} = \frac{1}{n}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \frac{1}{n}\mathbf{X}'\mathbf{P}\mathbf{X}. \tag{3.7}$$

Algunos autores definen la matriz de covarianzas dividiendo por $n-1$ en lugar de n para tener un estimador insesgado de la matriz de la población. Este divisor aparece, como en el caso univariante, porque para calcular la variabilidad no tenemos n desviaciones independientes sino solamente $n-1$. En efecto, los n vectores de desviaciones $(\mathbf{x}_i - \overline{\mathbf{x}})$ están ligados por la ecuación

$$\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}}) = 0$$

y sólo podemos calcular $n-1$ desviaciones independientes. Si dividimos la suma por ese número se obtiene una estimación insesgada de la varianza. En este libro llamaremos *matriz de varianzas corregida*, $\widehat{\mathbf{S}}$ al estimador insesgado dado por

$$\widehat{\mathbf{S}} = \frac{1}{n-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$$

Ejemplo 3.2 *La tabla A.7 del apéndice de datos ACCIONES presenta tres medidas de rentabilidad de 34 acciones en bolsa durante un período de tiempo. La primera, x_1 es la rentabilidad efectiva por dividendos (dividendos repartidos por acción divididos por precio de la acción), x_2 es la proporción de beneficios que va a dividendos (beneficios repartidos en dividendos sobre beneficios totales) y x_3 es el cociente entre precio por acción y beneficios. La tabla 3.4 presenta las medidas descriptivas de las tres variables.*

	x_1 (rentab.)	x_2 (benef.)	x_3 (precio)
Medias	9.421	69.53	9.097
D. Típicas	5.394	24.00	4.750
Coef. asimetría	0.37	0.05	2.71
Coef. kurtosis	1.38	1.40	12.44

Tabla 3.4: Análisis descriptivo de la rentabilidad de las acciones

Las medidas de asimetría y kurtosis indican un alejamiento de la distribución normal para las tres variables: las dos primeras tienen valores muy bajos de la kurtosis, lo que indica alta heterogeneidad, posiblemente por la presencia de dos grupos de datos distintos, y la tercera tiene alta kurtosis, lo que sugiere la presencia de valores atípicos.

Estas características son muy claras en los histogramas de las variables. La primera variable, rentabilidad efectiva por dividendos, x_1, muestra dos grupos de acciones con comportamiento distinto. El histograma de la segunda variable, x_2, muestra también dos grupos

de acciones. Finalmente, la distribución de la tercera variable es muy asimétrica, con un valor atípico muy destacado. La evidencia disponible indica que las acciones pueden probablemente dividirse en dos grupos más homogéneos. Sin embargo, vamos a ilustrar el análisis de todos los datos.

Histograma de la rentabilidad por dividendos.

Figura 3.2: Histograma de la proporción de beneficios que va a dividendos

Figura 3.3: Histograma del precio por acción con relación a los beneficios (per)

La matriz de varianzas y covarianzas de estas tres variables se presenta en la tabla 3.5

X_1	X_2	X_3
29.1	100.4	-15.7
100.4	576	-18.5
-15.7	-18.5	22.6

Tabla 3.5: Matriz de covarianzas de las acciones

Los elementos diagonales de esta matriz son los cuadrados de las desviaciones típicas de la tabla 3.4. Como las dimensiones de las variables son distintas, no tiene sentido calcular medidas promedio.

Los histogramas de las tres variables han mostrado una clara falta de normalidad. Una posibilidad, que estudiaremos con más detalle en el capítulo siguiente, es transformar las variables para facilitar su interpretación. Tomando logaritmos, la matriz de covarianzas de las variables transformadas, se indica en la tabla 3.6

$\log x_1$	$\log x_2$	$\log x_3$
.35	.15	-.19
.15	.13	-.03
-.19	-.03	.16

Tabla 3.6: Matriz de covarianzas de las acciones

Se observa que los logaritmos modifican mucho los resultados. Los datos ahora son más homogéneos y la variable de mayor varianza pasa a ser la primera, el logaritmo de la rentabilidad efectiva, mientras que la menor es la segunda, el logaritmo de la proporción de beneficios que va a dividendos. La relación entre el logaritmo del ratio precio/beneficios (X_3) *y*

la rentabilidad efectiva es negativa. Las otras relaciones son débiles. Una ventaja adicional de los logaritmos es que hace las variables independientes de la escala de medida: Si multiplicamos las variables por una constante al tomar logaritmos esto es equivalente a sumar una cantidad y sumar una constante a los datos no altera su variabilidad. Por tanto, al tomar logaritmos en las variables las varianzas pueden compararse aunque los datos tengan distintas dimensiones. La varianza media de las tres variables es

$$Var = \frac{.35 + .13 + .16}{3} = .213$$

y parece describir razonablemente la variabilidad de las variables.

3.4.2 Propiedades

Así como la varianza es siempre un número no negativo, la matriz de varianzas y covarianzas tiene una propiedad similar: es semidefinida positiva. Esta propiedad asegura que si $\mathbf{y}$ es cualquier vector, $\mathbf{y}'\mathbf{S}\mathbf{y} \geq 0$. También la traza, el determinante y los valores propios de esta matriz son no negativos.

Demostración

Sea $\mathbf{w}$ cualquier vector de dimensión p, definamos la variable escalar:

$$v_i = \mathbf{w}'(\mathbf{x}_i - \overline{\mathbf{x}}). \tag{3.8}$$

La media de esta variable será:

$$\overline{v} = \frac{1}{n}\sum_{i=1}^{n} v_i = \frac{1}{n}\mathbf{w}'\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}}) = 0,$$

y la varianza debe ser forzosamente no negativa, con lo que:

$$\begin{aligned} Var(v) &= \frac{1}{n}\sum_{i=1}^{n} v_i^2 = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{w}'(\mathbf{x}_i - \overline{\mathbf{x}})][(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{w}] \geq 0 \\ &= \mathbf{w}'\mathbf{S}\mathbf{w} \geq \mathbf{0}. \end{aligned}$$

Como la ecuación anterior es válida para cualquier vector $\mathbf{w}$, concluimos que $\mathbf{S}$ es semidefinida positiva. Esta condición también implica que si $\mathbf{S}\mathbf{w}_i = \lambda_i \mathbf{w}_{i,}$, entonces $\lambda_i \geq 0$. Finalmente, todos los menores principales son no negativos (en particular $|\mathbf{S}| \geq 0$).

3.4.3 Variables redundantes: El caso con Matriz S singular

Vamos a analizar las consecuencias de que la matriz $\mathbf{S}$ sea singular. Observemos que si existe algún vector $\mathbf{w}$ tal que $\mathbf{w}'\mathbf{S}\mathbf{w} = 0$, entonces la variable (3.8) tiene varianza nula y al tener media cero esta variable siempre toma el valor cero. Por tanto para cualquier i:

$$\sum_{j=1}^{p} w_j(x_{ij} - \overline{x}_j) = 0 \qquad \forall\ i.$$

Esta ecuación implica que las p variables no son independientes, ya que podemos despejar una cualquiera en función de las demás:

$$x_{i1} = \overline{x}_1 - \frac{w_2}{w_1}(x_{i2} - \overline{x}_2) - ... - \frac{w_p}{w_1}(x_{ip} - \overline{x}_p).$$

Por tanto, si existe algún vector $\mathbf{w}$ que haga $\mathbf{w}'\mathbf{S}\mathbf{w} = 0$, existe una relación lineal exacta entre las variables. Lo contrario es también cierto. Si existe una relación lineal entre las variables podemos escribir $\mathbf{w}'(\mathbf{x}_i - \overline{\mathbf{x}}) = 0$, para todo elemento, es decir

$$\widetilde{\mathbf{X}}\mathbf{w} = \mathbf{0},$$

multiplicando esta expresión por la derecha por la matriz $\widetilde{\mathbf{X}}'$ y dividiendo por n:

$$\frac{1}{n}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{w} = \mathbf{S}\mathbf{w} = 0. \tag{3.9}$$

Esta condición implica la matriz $\mathbf{S}$ tiene una raíz característica o autovalor igual a cero y $\mathbf{w}$ es el vector característico asociado a la raíz característica cero. Multiplicado en (3.9) por $\mathbf{w}'$ se obtiene $(\widetilde{\mathbf{X}}\mathbf{w})'(\widetilde{\mathbf{X}}\mathbf{w}) = 0$, que implica $\widetilde{\mathbf{X}}\mathbf{w} = 0$, y concluimos que una variable es una combinación lineal exacta de las otras. En consecuencia, es posible reducir la dimensionalidad del sistema eliminando esta variable. Observemos, además, que las coordenadas del vector $\mathbf{w}$ indican la combinación lineal redundante.

Ejemplo 3.3 *La matriz de covarianzas siguiente corresponde a cuatro variables simuladas de manera que tres de ellas son linealmente independientes, pero la cuarta es el promedio de las dos primeras.*

$$\mathbf{S} = \begin{bmatrix} .0947 & .0242 & .0054 & .0594 \\ .0242 & .0740 & .0285 & .0491 \\ .0054 & .0285 & .0838 & .0170 \\ .0594 & .0491 & .0170 & .0543 \end{bmatrix}$$

Los autovalores de esta matriz calculados con Matlab son $(0,172\,97; 0,08\,762,\ 0,04\,617$ *y* $0,\ 00005)$. *El menor valor propio es prácticamente cero comparado con los otros tres, por lo que la matriz tiene, muy aproximadamente, rango 3. El vector propio asociado a este valor propio nulo es (.408 .408 .000 -.816). Dividiendo por el término mayor este vector propio puede escribirse como (.5 .5 0 -1), que revela que la falta de rango completo de la matriz de covarianzas es debido a que la cuarta variable es el promedio de las dos primeras.*

Ejemplo 3.4 *En la tabla A.2 del anexo de datos figura una base de datos denominada EUROSEC compuesta por 26 países en los que se ha medido el porcentaje de población que se dedica a cada uno de 9 sectores económicos. El cálculo de los autovalores de la matriz de varianzas covarianzas correspondiente a estos datos se presenta a continuación y se observa que hay un autovalor muy próximo a 0 (.0019). Si este autovalor no es exactamente cero es debido a que en la tabla de datos la suma de las filas no es exactamente 100% en todos los casos por errores de redondeo (varía entre 99,8% y 102%)*

0.0019	*0.0649*	*0.4208*	*1.0460*	*2.4434*	*5.6394*	*15.207*	*43.7017*	*303.458*

Sin embargo, este autovalor tan pequeño define una variable escalar que no tiene prácticamente variabilidad. Para determinarla, obtenemos el vector propio ligado a este valor propio, que es el vector (.335, .324, .337, .339, .325, .337, .334, .332, .334). Este vector es aproximadamente el vector constante, lo que indica que la suma de todas las variables da lugar a una variable escalar aproximadamente constante.

El segundo valor propio es también bastante pequeño. El vector propio correspondiente es (-0.07, -0.29, -0.07, 0.91, 0.00, -0.10, -0.12, -0.05, -0.22). Este vector propio está determinado por la variable cuarta, que tiene mucho más peso que las demás. Esto sugiere que la variable cuarta, sector de la energía, debe tener un peso muy similar en todos los países. La matriz de varianzas y covarianzas de las variables es:

$$\begin{bmatrix} 241.6 & & & & & & & & \\ 0.53 & 0.94 & & & & & & & \\ -73.11 & 3.02 & 49.10 & & & & & & \\ -2.33 & 0.14 & 1.01 & 0.14 & & & & & \\ -13.77 & -0.04 & 5.70 & 0.03 & 2.70 & & & & \\ -52.42 & 1.76 & 6.53 & 0.34 & 2.68 & 20.93 & & & \\ 9.59 & -1.20 & -3.06 & 0.11 & 0.07 & 4.69 & 7.87 & & \\ -79.29 & -1.86 & 7.37 & 0.34 & 1.77 & 17.87 & 2.06 & 46.64 & \\ -12.22 & 0.21 & 3.41 & 0.19 & 0.88 & 1.19 & -0.96 & 5.39 & 1.93 \end{bmatrix}$$

Se observa como la cuarta variable tiene mucha menor variabilidad que las demás.

Generalización

Este procedimiento puede extenderse para cualquier número de valores propios nulos: si $\mathbf{S}$ tiene rango $h < p$, existen $p - h$ variables redundantes que pueden eliminarse. Los vectores asociados a autovalores nulos indican la composición de estas variables redundantes. En efecto, si $\mathbf{S}$ tiene rango h ese será el número de valores propios no nulos, y existirán $r = p - h$ vectores que verifican :

$$\begin{aligned} \mathbf{S}\mathbf{w}_1 &= \mathbf{0} \\ \vdots &\quad \vdots \\ \mathbf{S}\mathbf{w}_r &= \mathbf{0} \end{aligned}$$

o lo que es equivalente, existen r relaciones del tipo

$$(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{w}_j = 0, \quad j = 1, ..., r$$

que implica la existencia de r combinaciones lineales exactas entre las variables. Podemos pues representar las observaciones con $h = p - r$ variables. Existen muchas posibles representaciones, ya que cualquier vector del subespacio definido por $(\mathbf{w}_1, ..., \mathbf{w}_r)$ puede expresarse como una combinación lineal de estos vectores y verifica:

$$\mathbf{S}(a_1\mathbf{w}_1 + ... + a_r\mathbf{w}_r) = \mathbf{0}$$

Los r vectores propios de $\mathbf{S}$ asociados a valores propios nulos constituyen una base ortonormal (vectores perpendiculares y de módulo unitario) en dicho espacio. Observemos que cuando hay más de una raíz nula, las relaciones lineales entre las variables no están definidas unívocamente, ya que dadas dos relaciones lineales nulas cualquier nueva relación que resulte combinando estas dos tendrá la misma propiedad.

Una forma alternativa de analizar el problema es la siguiente. Como

$$\mathbf{S} = \frac{1}{n}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}},$$

el rango de $\mathbf{S}$ coincide con la matriz $\widetilde{\mathbf{X}}$, ya que para cualquier matriz $\mathbf{A}$, si llamamos $rg(\mathbf{A})$ al rango de $\mathbf{A}$, se verifica siempre que:

$$rg(\mathbf{A}) = rg(\mathbf{A}') = rg(\mathbf{A}'\mathbf{A}) = rg(\mathbf{A}\mathbf{A}').$$

Por tanto si la matriz $\widetilde{\mathbf{X}}$ tiene rango p, éste será también el rango de $\mathbf{S}$. Sin embargo, si existen h combinaciones lineales entre las variables $\mathbf{X}$, el rango de la matriz $\widetilde{\mathbf{X}}$ será $p-h$, y éste será también el rango de la matriz $\mathbf{S}$.

Ejemplo 3.5 *Calculemos los vectores propios de la matriz de varianzas y covarianzas para los datos de ACCIONES de la tabla A.7, que fueron analizados en el ejemplo 3.2. Los valores propios de la matriz de las variables originales son (594.86, 29.82, 3.22) y vemos que existe un valor propio muy grande y dos pequeños, en particular el valor más pequeño esta ligado al vector propio (0.82, -0.13, 0.55). Para las variables en logaritmos los valores propios son* $(0,5208; 0,1127$ *y* $0.0065)$. *Ahora existe un valor propio mucho más pequeño que los otros dos, y su vector propio es (57,-.55,.60).*

Para interpretar la variable definida por este vector propio escribamos su expresión en función de las variables originales. Recordando la definición de las variables y llamando d a los dividendos, p al precio, B al beneficio y N al número de acciones, suponiendo que la gran mayoría de los beneficios que se reparten van a dividendos (lo que es sólo una aproximación) podemos escribir,

$y = .57\log(d/p) - .55\log(dN/B) + .60\log(p/B/N)$

y, redondeando, esta variable será, aproximadamente,

$y = .6\log(d/p)(B/dN)(pN/B) = .6\log 1 = 0$

Es decir, llamando X_i a las variables en logaritmos, la variable definida por la combinación $X_1 - X_2 + X_3$ debe tomar valores pequeños. Si construimos esta variable a partir de los datos, su media es .01 y su varianza .03, que es mucho menor que la de las variables originales. Comprobamos que esta variable tiene poca variabilidad pero, al no ser constante,

no hay una relación determinista entre las tres variables en logaritmos. Los beneficios repartidos aparte de los dividendos aunque pequeños en promedio, no son despreciables para algunas acciones. Observemos que esta información, que es revelada por el análisis de los vectores propios de la matriz de covarianzas, puede pasar fácilmente desapercibida al lector no experto que trabaja directamente con estas medidas de rentabilidad.

3.5 MEDIDAS GLOBALES DE VARIABILIDAD

Cuando las variables se miden en las mismas unidades (euros, km) o son adimensionales (porcentajes, proporciones, etc) interesa encontrar medidas de la variabilidad promedio que permitan comparar distintos conjuntos de variables. Vamos a obtener primero estas medidas globales como resumen de la matriz de varianzas y covarianzas y, en segundo lugar, interpretaremos estas medidas mediante el concepto de distancias entre puntos.

3.5.1 La variabilidad total y la varianza promedio

Una forma de resumir la variabilidad de un conjunto de variables es mediante la traza de su matriz de varianzas y covarianzas y se define la *variabilidad total* de los datos por:

$$T = tr(\mathbf{S}) = \sum_{i=1}^{p} s_i^2$$

y la varianza promedio por

$$\overline{s}^2 = \frac{1}{p} \sum_{i=1}^{p} s_i^2. \tag{3.10}$$

El inconveniente de esta medida es que no tienen en cuenta la estructura de dependencia entre las variables. Para ilustrar el problema supongamos $p = 2$ y el caso extremo en que ambas variables son la misma, pero en unidades distintas. Entonces, la variabilidad conjunta de las dos variables en el espacio es nula, porque los puntos están siempre forzados a estar sobre la recta que define la relación lineal entre las dos variables, y, sin embargo, $\overline{s}^2$ puede ser alta. En general, si la dependencia entre las variables es muy alta, intuitivamente la variabilidad conjunta es pequeña, ya que conocida una variable podemos determinar aproximadamente los valores de las demás. Este aspecto no queda recogido en esta medida, que prescinde de las relaciones de dependencia existentes.

3.5.2 La Varianza Generalizada

Una medida mejor de la variabilidad global es la *varianza generalizada*, que es el determinante de la matriz de varianzas y covarianzas, es decir

$$VG = |\mathbf{S}|$$

Su raíz cuadrada se denomina *desviación típica generalizada*, y tiene las propiedades siguientes:

a) Está bien definida, ya que el determinante de la matriz de varianzas y covarianzas es siempre no negativo.

b) Es una medida del área (para $p = 2$), volumen (para $p = 3$) o hipervolumen (para $p > 3$) ocupado por el conjunto de datos.

Para aclarar estas ideas, supongamos el caso $p = 2$. Entonces, $\mathbf{S}$ puede escribirse:

$$\mathbf{S} = \begin{bmatrix} s_x^2 & r s_x s_y \\ r s_x s_y & s_y^2 \end{bmatrix}$$

y la desviación típica generalizada es:

$$|\mathbf{S}|^{1/2} = s_x s_y \sqrt{1 - r^2} \tag{3.11}$$

Si las variables son independientes, la mayoría de sus valores estarán dentro de un rectángulo de lados $6s_x$, $6s_y$ ya que, por el teorema de Tchebychev, entre la media y 3 desviaciones típicas deben estar, aproximadamente, al menos el 90% de los datos. En consecuencia, el área ocupada por ambas variables es directamente proporcional al producto de las desviaciones típicas.

Si las variables están relacionadas linealmente y el coeficiente de correlación es distinto de cero, la mayoría de los puntos tenderán a situarse en una franja alrededor de la recta de regresión y habrá una reducción del área tanto mayor cuanto mayor sea r^2. En el límite, si $r^2 = 1$, todos los puntos están en una línea recta, hay una relación lineal exacta entre las variables y el área ocupada es cero. La fórmula (3.11) describe esta contracción del área ocupada por los puntos al aumentar el coeficiente de correlación.

Un inconveniente de la varianza generalizada es que no sirve para comparar conjuntos de datos con distinto número de variables, ya que tiene las dimensiones del producto de las variables incluidas. Si añadimos a un conjunto de p variables que tiene una varianza generalizada $|\mathbf{S}_p|$ una variable adicional, incorrelada con el resto y varianza s_{p+1}^2, es fácil comprobar, con los resultados del cálculo del determinante de una matriz particionada presentados en 2.3.5, que

$$|\mathbf{S}_{p+1}| = |\mathbf{S}_p| \, s_{p+1}^2$$

y eligiendo las unidades de medida de la variable $p + 1$ podemos hacer que la varianza generalizada aumente o disminuya a voluntad. Supongamos el caso más simple donde la matriz $\mathbf{S}$ es diagonal y las variables van expresadas en las mismas unidades, por ejemplo euros. Entonces

$$|\mathbf{S}_p| = s_1^2 .... s_p^2$$

Supongamos que todas las varianzas en euros son mayores que la unidad. Entonces, si añadimos una variable $p + 1$, la nueva varianza generalizada será

$$|\mathbf{S}_{p+1}| = s_1^2 .... s_p^2 s_{p+1}^2 = |\mathbf{S}_p| \, s_{p+1}^2 > |\mathbf{S}_p|$$

ya que $s^2_{p+1} > 1$. En este caso la varianza generalizada aumenta monotonamente al considerar nuevas variables, es decir, llamando $|\mathbf{S}_j|$ a la varianza generalizada de las primeras j variables, tenemos que

$$|\mathbf{S}_p| > |\mathbf{S}_{p-1}| .... > |\mathbf{S}_2| > s_1^2$$

Supongamos ahora que expresamos las variables en miles de euros y con este cambio todas las varianzas son ahora menores que la unidad. Entonces la varianza generalizada disminuye monotonamente al incluir variables.

3.5.3 La variabilidad promedio

Para evitar estos inconvenientes, Peña y Rodríguez (2000) han propuesto como medida global de variabilidad *la variabilidad promedio,* dada por

$$VP = |\mathbf{S}|^{1/p} \tag{3.12}$$

que tiene la ventaja de que cuando todas las variables van en las mismas dimensiones esta medida tiene las unidades de la varianza. Para matrices diagonales esta medida es simplemente la media geométrica de las varianzas. Observemos que, como el determinante es el producto de los valores propios, la variabilidad promedio es la media geométrica de los valores propios de la matriz $\mathbf{S}$, que por ser semidefinida positiva serán siempre no negativos.

Como la media geométrica de un conjunto de números es siempre menor que su media aritmética esta medida será siempre menor que la varianza media. La variabilidad promedio tiene en cuenta la dependencia conjunta, ya que si una variable fuese combinación lineal de las demás al existir un valor propio nulo, la medida (3.12) es nula, mientras que la varianza media, dada por (3.10) no lo será. Veremos en los capítulos siguientes que la variabilidad promedio y la varianza media tienen una gran importancia en los procedimientos multivariantes.

Análogamente podemos definir la *desviación promedio* mediante

$$DP = |\mathbf{S}|^{1/2p} .$$

Ejemplo 3.6 *Partiendo de la matriz de covarianzas* $\mathbf{S}$ *para los logaritmos de las acciones, datos A.7, ACCIONES, del ejemplo 3.5, obtenemos que*

$$|S| = 0.000382$$

La variabilidad promedio es

$$VP = |S|^{1/3} = .0726$$

que podemos comparar con la media aritmética de las tres varianzas que calculamos en el ejemplo 3.2:

$$tr(\mathbf{S})/3 = .2133$$

Como vemos, la fuerte dependencia entre las variables hace que la variabilidad real promedio, cuando se tienen en cuenta las covarianzas, sea mucho menor que cuando se prescinde de ellas y se calcula el promedio de las varianzas.

Para las desviaciones típicas

$$DP = |S|^{1/6} = .269$$

que podemos tomar como medida global de variabilidad en los datos originales.

Ejemplo 3.7 *La matriz de varianzas y covarianzas para los datos de las medidas físicas es*

$$\mathbf{S} = \begin{array}{ccccccc} 100.24 & 104.49 & 26.12 & 44.22 & 33.20 & 10.64 & 26.19 \\ 104.49 & 158.02 & 30.04 & 50.19 & 41.67 & 14.08 & 27.99 \\ 26.12 & 30.04 & 7.91 & 11.66 & 8.86 & 2.79 & 7.42 \\ 44.22 & 50.19 & 11.66 & 23.69 & 15.4 & 4.18 & 11.55 \\ 33.20 & 41.67 & 8.86 & 15.4 & 15.59 & 4.48 & 7.72 \\ 10.64 & 14.08 & 2.79 & 4.18 & 4.48 & 3.27 & 3.11 \\ 26.19 & 27.99 & 7.42 & 11.55 & 7.72 & 3.11 & 9.61 \end{array}$$

y la medida promedio $VP = |S|^{1/7}$ =5.7783 y $VP^{1/2}$ =2.4038. Como existe bastante dependencia estas medidas son mucho menores de los promedios de las varianzas. Por ejemplo tr(S)/7=45.48. Observemos que esta medida no tiene, en este ejemplo, clara interpretación, al estar las variables en distintas unidades.

3.6 VARIABILIDAD Y DISTANCIAS

Un procedimiento alternativo para estudiar la variabilidad de las observaciones es utilizar el concepto de distancias entre puntos. En el caso escalar, la distancia entre el valor de una variable x en un punto, x_i, y la media de la variable, $\overline{x}$, se mide de manera natural mediante $\sqrt{(x_i - \overline{x})^2}$, o, lo que es equivalente, por el valor absoluto de la diferencia, $|x_i - \overline{x}|$. La desviación típica es un promedio de estas distancias entre los puntos y su media. Cuando disponemos de una variable vectorial, cada dato es un punto en $\Re^p$, y podemos pensar en construir medidas de variabilidad promediando las distancias entre cada punto y el vector de medias. Esto requiere generalizar el concepto de distancia a espacios de cualquier dimensión. El concepto de distancia entre puntos será importante en los capítulos siguientes.

3.6.1 El concepto de distancia

Dados dos puntos $\mathbf{x}_i$, $\mathbf{x}_j$ pertenecientes a $\Re^p$, diremos que hemos establecido una distancia, o una métrica, entre ellos si hemos definido una función d con las propiedades siguientes:

1. $d : \Re^p \times \Re^p \rightarrow \Re^+$, es decir, dados dos puntos en el espacio de dimensión p su distancia con esta función es un número no negativo, $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$;
2. $d(\mathbf{x}_i, \mathbf{x}_i) = 0 \quad \forall i$, la distancia entre un elemento y sí mismo es cero.
3. $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$, la distancia es una función simétrica en sus argumentos.

4. $d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_p) + d(\mathbf{x}_p, \mathbf{x}_j)$, la distancia debe verificar que si tenemos tres puntos, la suma de las longitudes de dos lados cualesquiera del triángulo formado por los tres puntos debe siempre ser mayor que el tercer lado. Esta propiedad se conoce como la propiedad triangular.

Estas propiedades generalizan la noción intuitiva de distancia entre dos puntos sobre una recta. Una familia de medidas de distancia muy habituales en $\Re^p$ es la familia de métricas o distancias de Minkowski, que se define en función de un parámetro r por

$$d_{ij}^{(r)} = \left(\sum_{s=1}^{p} (x_{is} - x_{js})^r \right)^{1/r} \tag{3.13}$$

y las potencias más utilizadas son $r = 2$, que conduce a la distancia euclídea, o en L_2,

$$d_{ij} = \left(\sum_{s=1}^{p} (x_{is} - x_{js})^2 \right)^{1/2} = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)^{1/2},$$

y $r = 1$, que se denomina distancia en L_1:

$$d_{ij} = |\mathbf{x}_i - \mathbf{x}_j|'\mathbf{1},$$

donde $\mathbf{1}' = (1, \ldots, 1)$.

La distancia más utilizada es la euclídea pero tiene el inconveniente de depender de las unidades de medida de las variables. Por ejemplo, sea x la estatura de una persona en metros e y su peso en kilogramos. Compararemos la distancia entre tres personas: $A(1.80, 80)$, $B(1.70, 72)$ y $C(1.65, 81)$. El cuadrado de la distancia euclídea del individuo A al B será:

$$d^2(A, B) = (1.80 - 1.70)^2 + (80 - 72)^2 = .1^2 + 8^2 = 64.01$$

y, análogamente $d^2(A, C) = .15^2 + 1 = 1.225$. Por tanto, con la distancia euclídea el individuo A estará mucho más cerca del individuo C que del B. Supongamos que, para hacer los números más similares, decidimos medir la estatura en centímetros, en lugar de metros. Las nuevas coordenadas de los individuos son ahora $A(180, 80)$, $B(170, 72)$ y $C(165, 81)$, y las distancias euclídeas entre los individuos se transforman en $d^2(A, B) = 10^2 + 8^2 = 164$ y $d^2(A, C) = 15^2 + 1 = 226$. Con el cambio de unidades, el individuo A está con la distancia euclídea más cerca del B que del C. La distancia euclídea depende mucho de las unidades de medida, y cuando no existe una unidad fija natural, como en este ejemplo, no está justificado utilizarla.

Una manera de evitar el problema de las unidades es dividir cada variable por un término que elimine el efecto de la escala. Esto conduce a la familia de métricas euclídeas ponderadas, que se definen por

$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{M}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right]^{1/2} \tag{3.14}$$

donde $\mathbf{M}$ es una matriz diagonal que se utiliza para estandarizar las variables y hacer la medida invariante ante cambios de escala. Por ejemplo, si colocamos en la diagonal de $\mathbf{M}$ las desviaciones típicas de las variables, la expresión (3.14) se convierte en

$$d_{ij} = \left(\sum_{s=1}^{p} (\frac{x_{is} - x_{js}}{s_s})^2 \right)^{1/2} = \left(\sum_{s=1}^{p} s_s^{-2} (x_{is} - x_{js})^2 \right)^{1/2}$$

que puede verse como una distancia euclídea donde cada coordenada se pondera inversamente proporcional a la varianza. Por ejemplo, si suponemos que las desviaciones típicas de las variables altura y peso son 10 cm y 10 kgr, las distancias estandarizadas al cuadrado entre los individuos anteriores son

$$d^2(A, B) = (1 + 0, 8^2) = 1, 64$$

y

$$d^2(A, C) = (1, 5^2 + 0, 1^2) = 2, 26.$$

Con esta métrica, que es más razonable, A está más próximo a B que a C.

En general la matriz $\mathbf{M}$ puede no ser diagonal, pero siempre debe ser una matriz no singular y definida positiva para que $d_{ij} \geq 0$. En el caso particular en que tomemos $\mathbf{M} = \mathbf{I}$ se obtiene de nuevo la *distancia euclídea*. Si tomamos $\mathbf{M} = \mathbf{S}$ se obtiene la distancia de Mahalanobis que estudiamos a continuación.

3.6.2 La Distancia de Mahalanobis

Se define la distancia de Mahalanobis entre un punto y su vector de medias por

$$d_i = \left[(\mathbf{x}_i - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}) \right]^{1/2}$$

Es frecuente referirse al valor d_i^2 también como distancia de Mahalanobis, en lugar de como cuadrado de la distancia, y en este libro, para simplificar, utilizaremos a veces esta licencia, aunque estrictamente la distancia es d_i. Vamos a interpretar esta distancia y comprobar que es una medida muy razonable de distancia entre variables correladas. Consideremos el caso $p = 2$. Entonces, escribiendo $s_{12} = r s_1 s_2$, tenemos que

$$\mathbf{S}^{-1} = \frac{1}{(1 - r^2)} \begin{bmatrix} s_1^{-2} & -r s_1^{-1} s_2^{-1} \\ -r s_1^{-1} s_2^{-1} & s_2^{-2} \end{bmatrix}$$

y la distancia de Mahalanobis (al cuadrado) entre dos puntos (x_1, y_1), (x_2, y_2) puede escribirse :

$$d_M^2 = \frac{1}{(1 - r^2)} \left[\frac{(x_1 - x_2)^2}{s_1^2} + \frac{(y_1 - y_2)^2}{s_2^2} - 2r \frac{(x_1 - x_2)(y_1 - y_2)}{s_1 s_2} \right]$$

Si $r = 0$, esta distancia se reduce a la distancia euclídea estandarizando las variables por sus desviaciones típicas. Cuando $r \neq 0$ la distancia de Mahalanobis añade un término adicional

que es positivo (y por lo tanto "separa" los puntos) cuando las diferencias entre las variables tienen el mismo signo, cuando $r > 0$, o distinto cuando $r < 0$. Por ejemplo, entre el peso y la altura hay correlación positiva: al aumentar la estatura de una persona en promedio también lo hace su peso. Si consideramos las tres personas anteriores $A(180, 80)$, $B(170, 72)$ y $C(165, 81)$ con desviaciones típicas 10 cm y 10 kgr y el coeficiente de correlación 0,7, los cuadrados de las distancias de Mahalanobis serán

$$d^2_M(A, B) = \frac{1}{0,51}\left[1 + 0,8^2 - 1,4 \times 0,8\right] = 1.02$$

y

$$d^2_M(A, C) = \frac{1}{0,51}\left[1,5^2 + 0,1^2 + 1,4 \times 1,5 \times 0,1\right] = 4.84,$$

concluimos que el individuo A está más cerca del B que del C con esta distancia. La distancia de Mahalanobis tiene en cuenta que, aunque el individuo B es más bajo que el A, como hay correlación entre el peso y la altura si su peso también disminuye proporcionalmente, el aspecto físico de ambos es similar porque aunque cambia el tamaño global no cambia la forma del cuerpo. Sin embargo, el individuo C es todavía más bajo que el A y además pesa más, lo que implica que su aspecto físico es muy distinto del de A. Como consecuencia, la distancia de A a C es mayor que a B. La capacidad de esta distancia para tener en cuenta la forma de un elemento a partir de su estructura de correlación explica su introducción por P. C. Mahalanobis, un eminente estadístico indio, en los años 30 para comparar medidas físicas de razas en la India.

3.6.3 La distancia promedio

Podríamos plantearnos construir una medida global de la variabilidad respecto a la media de una variable vectorial escogiendo promediando las distancias entre los puntos y la media. Por ejemplo, si todas las variables van en las mismas unidades, podemos tomar la distancia euclídea al cuadrado y promediar por el número de términos en la suma:

$$V_m = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})'(\mathbf{x}_i - \overline{\mathbf{x}}). \tag{3.15}$$

Como un escalar es igual a su traza, podemos escribir

$$V_m = \sum_{i=1}^{n} tr\left[\frac{1}{n}(\mathbf{x}_i - \overline{\mathbf{x}})'(\mathbf{x}_i - \overline{\mathbf{x}})\right] = \sum_{i=1}^{n} tr\left[\frac{1}{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'\right] = tr(\mathbf{S})$$

y el promedio de distancias es la variabilidad total. Si promediamos las distancias también por la dimensión del vector, tenemos que:

$$V_{m,p} = \frac{1}{np}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})'(\mathbf{x}_i - \overline{\mathbf{x}}) = \overline{s}^2 \tag{3.16}$$

y el promedio estandarizado de las distancias euclídeas entre los puntos y la media es el promedio de las varianzas de las variables.

No tiene sentido definir una medida de distancia promediando las distancias de Mahalanobis, ya que es fácil comprobar (véase ejercicio 3.12) que el promedio de las distancias de Mahalanobis es siempre p, y el promedio estandarizado por la dimensión del vector es uno .

Ejemplo 3.8 *La tabla adjunta presenta para los datos de las medidas físicas, MEDIFIS, las distancias euclídeas al cuadrado de cada dato a su media, d_e^2, las distancias de Mahalanobis de cada dato a su media, D_M^2, la máxima distancia euclídea entre cada punto y otro de la muestra, d_{em}^2, el orden del dato más alejado con esta distancia, I_e , la máxima distancia de Mahalanobis entre cada punto y otro de la muestra, D_{Mm}^2, y el orden del dato más alejado con esta distancia, I_M.*

orden	d_e^2	D_M^2	d_{em}^2	I_e	D_{Mm}^2	I_M
1.0000	*3.8048*	*0.0226*	*29.0200*	*24.0000*	*29.0200*	*24.0000*
2.0000	*0.3588*	*0.0494*	*15.4800*	*24.0000*	*15.4800*	*24.0000*
3.0000	*0.2096*	*0.0447*	*10.0600*	*20.0000*	*10.0600*	*20.0000*
4.0000	*1.6899*	*0.0783*	*20.5925*	*24.0000*	*20.5925*	*24.0000*
5.0000	*2.2580*	*0.0759*	*23.8825*	*24.0000*	*23.8825*	*24.0000*
6.0000	*0.8336*	*0.0419*	*15.6000*	*24.0000*	*15.6000*	*24.0000*
7.0000	*2.8505*	*0.0830*	*23.5550*	*24.0000*	*23.5550*	*24.0000*
8.0000	*3.0814*	*0.0858*	*20.3300*	*20.0000*	*20.3300*	*20.0000*
9.0000	*3.6233*	*0.0739*	*21.7750*	*20.0000*	*21.7750*	*20.0000*
10.0000	*3.5045*	*0.0348*	*28.1125*	*24.0000*	*28.1125*	*24.0000*
11.0000	*2.0822*	*0.0956*	*20.2900*	*24.0000*	*20.2900*	*24.0000*
12.0000	*0.6997*	*0.1037*	*11.5425*	*20.0000*	*11.5425*	*20.0000*
13.0000	*6.2114*	*0.0504*	*34.7900*	*24.0000*	*34.7900*	*24.0000*
14.0000	*2.2270*	*0.0349*	*18.2700*	*20.0000*	*18.2700*	*20.0000*
15.0000	*4.2974*	*0.1304*	*23.2200*	*20.0000*	*23.2200*	*20.0000*
16.0000	*10.5907*	*0.1454*	*35.6400*	*20.0000*	*35.6400*	*20.0000*
17.0000	*1.7370*	*0.0264*	*16.9000*	*20.0000*	*16.9000*	*20.0000*
18.0000	*0.7270*	*0.0853*	*14.1100*	*24.0000*	*14.1100*	*24.0000*
19.0000	*4.5825*	*0.1183*	*30.5500*	*24.0000*	*30.5500*	*24.0000*
20.0000	*7.8399*	*0.0332*	*39.1100*	*24.0000*	*39.1100*	*24.0000*
21.0000	*4.4996*	*0.0764*	*23.9600*	*20.0000*	*23.9600*	*20.0000*
22.0000	*0.5529*	*0.0398*	*12.3100*	*20.0000*	*12.3100*	*20.0000*
23.0000	*3.9466*	*0.0387*	*29.3900*	*24.0000*	*29.3900*	*24.0000*
24.0000	*11.9674*	*0.0998*	*39.1100*	*20.0000*	*39.1100*	*20.0000*
25.0000	*0.4229*	*0.0745*	*10.6500*	*20.0000*	*10.6500*	*20.0000*
26.0000	*0.2770*	*0.0358*	*10.5850*	*20.0000*	*10.5850*	*20.0000*
27.0000	*0.9561*	*0.1114*	*17.6050*	*24.0000*	*17.6050*	*24.0000*

Se observa que con la distancia euclídea los puntos más alejados de la media son el 24 y el 16, seguidos del 20. El punto más extremo para cada uno es el 24 o el 20, lo que define a estos puntos como extremos en el espacio con esta medida. Con las distancias de Mahalanobis los más alejados de la media son los 15 y 16 pero, sin embargo, los puntos que

aparecen como extremos de la muestra son de nuevo los 20 y 24. Observando estos datos, el 24 correponde a un hombre muy alto, el mayor de la muestra, y el 20 a una mujer de baja estatura y delgada, que constituye el extremo opuesto de los datos.

3.7 MEDIDAS DE DEPENDENCIA LINEAL

Un objetivo fundamental de la descripción de los datos multivariantes es comprender la estructura de dependencias entre las variables. Estas dependencias pueden estudiarse: (1) entre pares de variables; (2) entre una variable y todas las demás; (3) entre pares de variables pero eliminando el efecto de las demás variables; (4) entre el conjunto de todas las variables. Vamos a analizar estos cuatro aspectos.

3.7.1 Dependencia por pares: La matriz de correlación

La dependencia lineal entre dos variables se estudia mediante el coeficiente de correlación lineal o simple. Este coeficiente para las variables x_j, x_k es:

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

y tiene las propiedades siguientes: (1) $0 \leq |r_{jk}| \leq 1$; (2) si existe una relación lineal exacta entre las variables, $x_{ij} = a + bx_{ik}$, entonces $|r_{jk}| = 1$; (3) r_{jk} es invariante ante transformaciones lineales de las variables.

La dependencia por pares entre las variables se mide por la matriz de correlación. Llamaremos matriz de correlación, $\mathbf{R}$, a la matriz cuadrada y simétrica que tiene unos en la diagonal principal y fuera de ella los coeficientes de correlación lineal entre pares de variables, escribiremos:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ \vdots & \vdots & \dots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

Esta matriz es también semidefinida positiva. Para demostrarlo, llamemos $\mathbf{D} = \mathbf{D}(\mathbf{S})$ a la matriz diagonal de orden p formada por los elementos de la diagonal principal de $\mathbf{S}$, que son las varianzas de las variables. La matriz $\mathbf{D}^{1/2}$ contendrá las desviaciones típicas y la matriz $\mathbf{R}$ esta relacionada con la matriz de covarianzas, $\mathbf{S}$, mediante:

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}, \tag{3.17}$$

que implica

$$\mathbf{S} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}. \tag{3.18}$$

La condición $\mathbf{w}'\mathbf{S}\mathbf{w} \geq 0$ equivale a:

$$\mathbf{w}'\mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}\mathbf{w} = \mathbf{z}'\mathbf{R}\mathbf{z} \geq \mathbf{0}$$

llamando $\mathbf{z} = \mathbf{D}^{1/2}\mathbf{w}$ al nuevo vector transformado por $\mathbf{D}^{1/2}$. Por tanto, la matriz $\mathbf{R}$ es, como la matriz $\mathbf{S}$, semidefinida positiva.

3.7.2 Dependencia de cada variable y el resto: Regresión Múltiple

Además de estudiar la relación entre pares de variables podemos estudiar la relación entre una variable y todas las demás. Hemos visto que si una variable es combinación lineal de las demás, y por lo tanto puede predecirse sin error con el resto, debemos eliminarla de consideración. Es posible que, sin llegar a esta situación extrema, haya variables que sean muy dependientes de las demás y conviene medir su grado de dependencia. Supongamos que x_j es la variable de interés y para simplificar la notacion la llamaremos variable explicativa o respuesta y la denotaremos por y. A continuación, consideremos su mejor predictor lineal a partir de las restantes variables, que llamaremos variables explicativas o regresores. Este predictor lineal tendrá la forma:

$$\widehat{y_i} = \overline{y} + \widehat{\beta}_1(x_{i1} - \overline{x}_1) + ... + \widehat{\beta}_p(x_{ip} - \overline{x}_p), \quad i = 1, ..., n \tag{3.19}$$

y se comprueba que cuando las variables explicativas toman un valor igual a su media la variable respuesta es también igual a su media. Los $p-1$ coeficientes $\widehat{\beta}_k$, para $k = 1, ..., p$ con $k \neq j$, se determinan de manera que la ecuación proporcione, en promedio, la mejor predicción posible de los valores de y_i. Llamando residuos a los errores de predicción, $e_i = y_i - \widehat{y}_i$, es inmediato, sumando para los n datos en (3.19), que la suma de los residuos para todos los puntos muestrales es cero. Esto indica que cualquiera que sean los coeficientes $\widehat{\beta}_j$ la ecuación (3.19) va a compensar los errores de predicción positivos con los negativos. Como queremos minimizar los errores con independencia del signo, los elevamos al cuadrado y calculamos los $\widehat{\beta}_j$ minimizando:

$$M = \sum_{i=1}^{n} e_i^2,$$

Derivando esta expresión respecto a los parámetros $\widehat{\beta}_j$, se obtiene el sistema de $p-1$ ecuaciones, para $k = 1, ..., p$ con $k \neq j$,:

$$2\sum_{i=1}^{n} \left[y_i - \overline{y} + \widehat{\beta}_1(x_{i1} - \overline{x}_1) + ... + \widehat{\beta}_p(x_{i,p} - \overline{x}_p) \right] (x_{ik} - \overline{x}_k)$$

que puede escribirse:

$$\sum e_i x_{ik} = 0 \qquad k = 1, ..., p; \quad k \neq j,$$

que tiene una clara interpretación intuitiva. Indica que los residuos, o errores de predicción, deben de estar incorrelados con las variables explicativas, de manera que la covarianza entre ambas variables sea cero. En efecto, si existiese relación entre ambas variables podría utilizarse para prever los errores de predicción y reducirlos, con lo que la ecuación de predicción no podría ser óptima. Geometricamente este sistema establece que el vector de residuos debe ser ortogonal al espacio generado por las variables explicativas. Definiendo una matriz $\mathbf{X}_R$ de datos para la regresión de dimensiones $(n \times p - 1)$ que se obtiene de la matriz de datos centrada, $\widetilde{\mathbf{X}}$, eliminando la columna de esta matriz que corresponde a la variable que

queremos prever, que llamaremos $\mathbf{y}$, el sistema de ecuaciones para obtener los parámetros es:

$$\mathbf{X}'_R\mathbf{y} = \mathbf{X}'_R\mathbf{X}_R\widehat{\boldsymbol{\beta}}$$

que conduce a :

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'_R\mathbf{X}_R\right)^{-1}\mathbf{X}'_R\mathbf{y} = \mathbf{S}^{-1}_{\mathbf{p-1}}\mathbf{S}_{\mathbf{xy}}.$$

donde $\mathbf{S_{p-1}}$ es la matriz de covarianzas de las $p-1$ variables explicativas y $\mathbf{S_{xy}}$ la columna de la matriz de covarianzas correspondiente a las covarianzas de la variable seleccionada como $\mathbf{y}$ con el resto. La ecuación obtenida con estos coeficientes se conoce como la *ecuación de regresión múltiple* entre la variable $y = x_j$ y las variables, x_k, con $k = 1, ..., p$;y $k \neq j$.

El promedio de los residuos al cuadrado con la ecuación de regresión múltiple para explicar x_j es:

$$s_r^2(j) = \frac{\sum e_i^2}{n} \tag{3.20}$$

y es una medida de la precisión de la regresión para prever la variable $y = x_j$. Una medida adimensional de la dependencia se construye partiendo de la identidad

$$y_i - \overline{y} = \widehat{y}_i - \overline{y} + e_i$$

y elevando al cuadrado y sumando para todos los puntos se obtiene la descomposición básica del análisis de la varianza, que podemos escribir como:

$$VT = VE + VNE$$

donde la variabilidad total o inicial de los datos, $VT = \sum \left(y_i - \overline{y}\right)^2$, se expresa como suma de la variabilidad explicada por la regresión, $VE = \sum \left(\widehat{y}_i - \overline{y}\right)^2$, y la residual o no explicada por la regresión, $VNE = \sum e_i^2$. Una medida descriptiva de la capacidad predictiva del modelo es el cociente entre la variabilidad explicada por la regresión y la variabilidad total. Esta medida se llama *coeficiente de determinación*, o *coeficiente de correlación múltiple* al cuadrado, y se define por:

$$R^2_{j.1,...,p} = \frac{VE}{VT} = 1 - \frac{VNE}{VT} \tag{3.21}$$

donde el subíndice indica la variable que estamos explicando y los regresores. Utilizando (3.20) podemos escribir

$$R^2_{j.1,...,p} = 1 - \frac{s_r^2(j)}{s_j^2} \tag{3.22}$$

Es immediato comprobar que en el caso de una única variable explicativa R^2 es el cuadrado del coeficiente de correlación simple entre las dos variables. También se comprueba que es el cuadrado del coeficiente de correlación simple entre las variables y y $\widehat{y}$. El coeficiente de

correlación múltiple al cuadrado puede ser mayor, menor o igual que la suma de los cuadrados de las correlaciones simples entre la variable y y cada una de las variables explicativas (véase Cuadras, 1993).

Según la ecuación (3.22) podemos calcular el coeficiente de correlación múltiple entre cualquier variable x_i y las restantes si conocemos su varianza y la varianza residual de una regresión de esta variable sobre las demás. Se demuestra en el apéndice 3.1 que los términos diagonales de la inversa de la matriz de covarianzas, $\mathbf{S}^{-1}$, son precisamente las inversas de las varianzas residuales de la regresión de cada variable con el resto. Por tanto podemos calcular fácilmente el coeficiente de correlación múltiple al cuadrado entre la variable x_j y las restantes como sigue:

(1) Tomar el elemento diagonal j de la matriz $\mathbf{S}$, s_{jj} que es la varianza s_j^2 de la variable.

(2) Invertir la matriz $\mathbf{S}$ y tomar el elemento diagonal j de la matriz $\mathbf{S}^{-1}$ que llamaremos s^{jj}. Este término es $1/s_r^2(j)$, la varianza residual de una regresión entre la variable j y el resto.

(3) Calcular R_j^2, la correlación múltiple como

$$R_j^2 = 1 - \frac{1}{s^{jj}s_{jj}}$$

Esta expresión permite obtener inmediatamente todos los coeficientes de correlación múltiple a partir de las matrices $\mathbf{S}$ y $\mathbf{S}^{-1}$.

Ejemplo 3.9 *La matriz de correlación para las 7 variables físicas, tabla A.5, MEDIFIS, del ejemplo 1.1. se presenta en la tabla 1.5. Las variables aparecen en el orden del ejemplo 1.1*

$$\mathbf{R} = \begin{bmatrix} 1 & 0.83 & 0.93 & 0.91 & 0.84 & 0.59 & 0.84 \\ 0.83 & 1 & 0.85 & 0.82 & 0.84 & 0.62 & 0.72 \\ 0.93 & 0.85 & 1 & 0.85 & 0.80 & 0.55 & 0.85 \\ 0.91 & 0.82 & 0.85 & 1 & 0.80 & 0.48 & 0.76 \\ 0.84 & 0.84 & 0.80 & 0.80 & 1 & 0.63 & 0.63 \\ 0.59 & 0.62 & 0.55 & 0.48 & 0.63 & 1 & 0.56 \\ 0.84 & 0.72 & 0.85 & 0.76 & 0.63 & 0.56 & 1 \end{bmatrix}$$

Se observa que la máxima correlación aparece entre la primera y la tercera variable (estatura y longitud del pie) y es 0,93. La mínima es entre la longitud del brazo y el diámetro del cráneo (0,48). En general las correlaciones más bajas aparecen entre el diámetro del cráneo y el resto de las variables. La matriz S^{-1} es:

$$\begin{array}{ccccccc} 0.14 & 0.01 & -0.21 & -0.11 & -0.07 & -0.05 & -0.07 \\ 0.01 & 0.04 & -0.08 & -0.03 & -0.04 & -0.04 & -0.00 \\ -0.21 & -0.08 & 1.26 & 0.06 & -0.05 & 0.18 & -0.29 \\ -0.11 & -0.03 & 0.06 & 0.29 & -0.04 & 0.13 & -0.04 \\ -0.07 & -0.04 & -0.05 & -0.04 & 0.34 & -0.13 & 0.15 \\ -0.05 & -0.04 & 0.18 & 0.13 & -0.13 & 0.64 & -0.15 \\ -0.07 & -0.00 & -0.29 & -0.04 & 0.15 & -0.15 & 0.50 \end{array}$$

y utilizando los elementos diagonales de esta matriz y de la matriz S podemos calcular las correlaciones múltiples al cuadrado de cada variable con el resto como sigue: (1) multiplicamos los elementos diagonales de las matrices S y S^{-1}. El resultado de esta operacion es el

vector (14.3672, 5.5415, 9.9898, 6.8536, 5.3549, 2.0784, 4.7560). (2) A continuación, calculamos las inversas de estos elementos, para obtener (0.0696 0.1805 0.1001 0.1459 0.1867 0.4811 0.2103). Finalmente, restamos a uno estos coeficientes para obtener (0.9304, 0.8195, 0.8999, 0.8541, 0.8133, 0.5189, 0.7897) y estos son los coeficientes de correlación múltiple entre cada variable y el resto. Vemos que la variable más previsible por las restantes es la estatura,(R^2 = 0.9304) , despues el pié (R^2 = 0.8999) y luego la longitud del brazo (R^2 = 0.8541). La menos predecible es dcr, que tiene un coeficiente de correlación múltiple con el resto de 0.5189, o en otros términos, el resto de las variables explica el 52% de la variabilidad de esta variable.

La ecuación para prever la estatura en función del resto de las variables se obtiene facilmente con cualquier programa de regresión. El resultado es

est = 0.9 - 0.094 peso+ 1.43pie + 0.733 lbr + 0.494 aes + 0.347 dcr + 0.506 lrt

que es la ecuación que permite prever con menor error la estatura de una persona dadas el resto de las medidas. El R^2 de esta regresión es = 0,93, resultado que habíamos obtenido anteriormente. La ecuación para prever la longitud del pié es:

pie = 8.14 + 0.162 est + 0.0617 pes - 0.051 lbr + 0.037 aes - 0.144 dcr + 0.229lrt

que indica que para prever el pie las variables más relevantes parecen ser la estatura y l alongitud rodilla tobillo. Podemos hacer regresiones tomando como variable explicativa el sexo, entonces:

sexo = - 3.54 - 0.0191 est - 0.0013 pes + 0.141 pie + 0.0291 lbr + 0.0268 aes

- 0.0439 dcr + 0.0219 lrt

La variable más importante para prever el sexo de una persona parece ser el pie que es la que tiene un coeficiente más alto.

3.7.3 Dependencia directa entre pares: Correlaciones parciales

La dependencia directa entre dos variables controlando el efecto de las restantes se mide por el *coeficiente de correlación parcial*. Se define el coeficiente de correlación parcial entre dos variables, (x_1, x_2), dadas las variables $(x_3, ..., x_p)$, y se denota por $r_{12.3..p}$, como el coeficiente de correlación entre las partes de x_1 y x_2 que están libres de los efectos de las variables $(x_3, ..., x_p)$. Este coeficiente se obtiene en dos etapas. Primero, hay que obtener la parte de cada variable que no es explicada por (o está libre de los efectos de) el grupo de variables que se controlan. Esta parte es el residuo de la regresión sobre el conjunto de variables $(x_3, ..., x_p)$, ya que, por construcción, el residuo es la parte de la respuesta que no puede preverse o es independiente de los regresores. Segundo, se calcula el coeficiente de correlación simple entre estos dos residuos. Se demuestra en el apéndice 3.3 que los coeficientes de correlación parcial entre cada par de variables se obtienen estandarizando los elementos de la matriz $\mathbf{S}^{-1}$. En concreto, si llamamos s^{ij} los elementos de $\mathbf{S}^{-1}$, el coeficiente de correlacion parcial entre las variables $x_j x_k$ se obtiene como

$$r_{jk.12,...,p} = -\frac{s^{ij}}{\sqrt{s^{ii} s^{jj}}} \tag{3.23}$$

Los coeficientes de correlación parcial pueden calcularse a partir de los coeficientes de correlación múltiple mediante la relación, que se demuestra en el apéndice 3.3:

$$1 - r^2_{12.3..p} = \frac{1 - R^2_{1.2,...,p}}{1 - R^2_{1.3,...,p}},$$

donde $r^2_{12.3..p}$ es el cuadrado del coeficiente de correlación parcial entre las variables (x_1, x_2) cuando se controlan las variables $(x_3, ..., x_p)$, $R^2_{1.2,...,p}$ es el coeficiente de determinación o coeficiente de correlación múltiple al cuadrado en la regresión de x_1 con respecto a $(x_2, x_3, ..., x_p)$ y $R^2_{1.3,...,p}$ es el coeficiente de determinación o coeficiente de correlación múltiple al cuadrado en la regresión de x_1 con respecto a $(x_3, ..., x_p)$. (El resultado es equivalente si intercambiamos x_1 por x_2). Esta expresión indica una relación simple entre términos del tipo $1 - r^2$, que, según la expresión (3.21), representan la proporción relativa de variabilidad no explicada.

Se define la matriz de correlaciones parciales, $\mathbf{P}$, como aquella que contiene los coeficientes de correlación parcial entre pares de variables eliminando el efecto de las restantes. Por ejemplo, para cuatro variables, la matriz de correlaciones parciales, :

$$\mathbf{P}_4 = \begin{bmatrix} 1 & r_{12.34} & r_{13.24} & r_{14.23} \\ r_{21.34} & 1 & r_{23.14} & r_{24.13} \\ r_{31.24} & r_{32.14} & 1 & r_{34.12} \\ r_{41.23} & r_{42.13} & r_{43.12} & 1 \end{bmatrix}$$

donde, por ejemplo, $r_{12.34}$ es la correlación entre las variables 1 y 2 cuando eliminamos el efecto de la 3 y la 4, o cuando las variables 3 y 4 permanecen constantes. De acuerdo con (3.23) está matriz se obtiene como

$$\mathbf{P} = (-1)^{diag}\mathbf{D}(\mathbf{S}^{-1})^{-1/2}\mathbf{S}^{-1}\mathbf{D}(\mathbf{S}^{-1})^{-1/2}$$

donde $\mathbf{D}(\mathbf{S}^{-1})$ es la matriz diagonal obtenida seleccionando los elementos diagonales de la matriz $\mathbf{S}^{-1}$ y el término $(-1)^{diag}$ indica que cambiamos el signo de todos los elementos de la matriz menos de los elementos diagonales que serán la unidad. La expresión (3.23) es similar a la (3.17), pero utilizando la matriz $\mathbf{S}^{-1}$ en lugar de $\mathbf{S}$. Observemos que $\mathbf{D}(\mathbf{S}^{-1})^{-1/2}$ no es la inversa de $\mathbf{D}(\mathbf{S})^{-1/2} = \mathbf{D}^{-1/2}$, y que, en consecuencia, $\mathbf{P}$ no es la matriz inversa de $\mathbf{R}$.

3.7.4 El coeficiente de Dependencia

Para obtener una medida conjunta de la dependencia entre las variables podemos utilizar el determinante de la matriz de correlación, que mide el alejamiento del conjunto de variables de la situación de perfecta dependencia lineal. Se demuestra en el apéndice 3.2 que $0 \leq |\mathbf{R}| \leq 1$ y:

(1) Si las variables están todas incorreladas $\mathbf{R}$ es una matriz diagonal con unos en la diagonal y $|\mathbf{R}| = 1$.

(2) Si una variable es combinación lineal del resto hemos visto que $\mathbf{S}$ y $\mathbf{R}$ son singulares y $|\mathbf{R}| = 0$

(3) En el caso general, se demuestra en el apéndice 3.3 que:

$$|\mathbf{R}_p| = \left(1 - R^2_{p.1...p-1}\right)\left(1 - R^2_{p-1.1\cdots p-2}\right)...\left(1 - R^2_{2.1}\right). \tag{3.24}$$

es decir, el determinante de la matriz de correlación es el producto de $p-1$ términos. El primero representa la proporción de variabilidad no explicada en una regresión múltiple entre la variable p y las restantes variables, $p-1, p-2, ..., 1$. El segundo la proporción de variabilidad no explicada en una regresión múltiple entre la variable $p-1$ y las variables restantes siguientes, $p-2, p-3, ..., 1$. El último representa la proporción de variabilidad no explicada en una regresión simple entre las variables dos y uno.

De acuerdo con la propiedad anterior $|\mathbf{R}_p|^{1/p-1}$ representa la media geométrica de la proporción de variabilidad explicada por todas las regresiones anteriores. Observemos que también es la media geométrica de los valores propios de la matriz $\mathbf{R}_p$, teniendo en cuenta que sólo tenemos $p-1$ valores propios independientes ya que están ligados por $\sum \lambda_i = p$.

A partir de estas propiedades Peña y Rodríguez (2000) han propuesto como medida de dependencia lineal global la *Dependencia,* definida por :

$$D(\mathbf{R}_p) = 1 - |\mathbf{R}_p|^{1/(p-1)} \tag{3.25}$$

Por ejemplo, para $p = 2$ como $|\mathbf{R}_2| = 1 - r_{12}^2$, esta medida coincide con el cuadrado del coeficiente de correlación lineal entre las dos variables. Para $p > 2$ podemos escribir de (3.24) y (3.25):

$$1 - D(\mathbf{R}_p) = \left[\left(1 - R^2_{p.1...p-1}\right)\left(1 - R^2_{p-1.1\cdots p-2}\right) ... \left(1 - R^2_{2.1}\right)\right]^{1/(p-1)}$$

y vemos que la dependencia es el coeficiente de correlación necesario para que la variabilidad no explicada en el problema sea igual a la media geométrica de todas las posibles variabilidades no explicadas. El coeficiente de correlación promedio estará dado por

$$\overline{\rho}(\mathbf{R}_p) = D(\mathbf{R}_p)^{1/2} = \sqrt{1 - |\mathbf{R}_p|^{1/(p-1)}}.$$

En el caso particular en que $p = 2$, el coeficiente de correlación promedio coincide con el valor absoluto del coeficiente de correlación simple.

Ejemplo 3.10 *Vamos a construir la matriz de correlaciones parciales para las 7 variables físicas, tabla A.5, MEDIFIS. Podemos construir la matriz de correlaciones parciales a partir de* S^{-1} *estandarizandola por los elementos diagonales para obtener:*

$$\mathbf{P} = \begin{array}{ccccccc} 1.00 & -0.19 & 0.48 & 0.52 & 0.32 & 0.17 & 0.27 \\ -0.19 & 1.00 & 0.37 & 0.30 & 0.34 & 0.26 & 0.00 \\ 0.48 & 0.37 & 1.00 & -0.11 & 0.07 & 0.20 & 0.37 \\ 0.52 & 0.30 & -0.11 & 1.00 & 0.13 & -0.31 & 0.10 \\ 0.32 & 0.34 & 0.07 & 0.13 & 1.00 & 0.29 & -0.37 \\ 0.17 & 0.26 & 0.20 & -0.31 & 0.29 & 1.00 & 0.27 \\ 0.27 & 0.00 & 0.37 & 0.10 & -0.37 & 0.27 & 1.00 \end{array}$$

Esta matriz muestra que las relaciones parciales más fuertes se dan entre la estatura y las longitudes del pie (0,48) y del brazo (0,52). Por ejemplo este coeficiente se interpreta que si consideramos personas con el mismo peso, pie, anchura de espalda, diámetro del cráneo y longitud rodilla tobillo, hay una correlación positiva entre la estatura y la longitud del brazo de 0,52. La tabla muestra que para personas de la misma estatura, peso y demás medidas físicas, la correlación entre la anchura de la espalda y la longitud rodilla tobillo es negativa.

Para obtener una medida de dependencia global, como el determinante de $\mathbf{R}$ *es* 1.42×10^{-4} *y el coeficiente global de dependencia es*

$$D = 1 - |\mathbf{R}|^{1/6} = 1 - \sqrt[6]{1.42 \times 10^{-4}} = 0.771$$

Podemos concluir que, globalmente, la dependencia lineal explica 77% de la variabilidad de este conjunto de datos.

Ejemplo 3.11 *Calcularemos el coeficiente global de dependencia para los datos del Anexo de Datos en las unidades originales en que se presentan,*

	EUROALI	*EUROSEC*	*EPF*	*INVEST*	*MUNDODES*	*ACCION*
D	.51	.80	.62	.998	.82	.61

Se observa que en INVEST la dependencia conjunta es muy fuerte. Esto sugiere que puede reducirse el número de variables necesarias para describir la información que contienen.

3.8 La matriz de precisión

Se denomina matriz de precisión a la inversa de la matriz de varianzas y covarianzas. Esta matriz juega un papel importante en muchos procedimientos estadísticos, como veremos en capítulos sucesivos. Un resultado importante es que la matriz de precisión contiene la información sobre la relación multivariante entre cada una de las variable y el resto. Este resultado es a primera vista sorprendente, ya que la matriz de varianzas y covarianzas sólo contiene la información sobre las relaciones por pares de las variables, pero se explica por las propiedades de la matriz inversa (véase 2.3.4). Puede demostrarse, (véase el apéndice 3.1) que la inversa de la matriz de covarianzas contiene :

(1) Por filas, y fuera de la diagonal términos proporcionales a los coeficientes de regresión múltiple de la variable correspondiente a esa fila explicada por todas las demás. Los términos de la matriz son estos coeficientes cambiados de signo y multiplicados por la inversa de la varianza residual en esa regresión. Es decir, si llamamos s^{ij} a los elementos de la matriz de precisión:

$$s^{ij} = -\widehat{\beta}_{ij}/s_r^2(i)$$

donde $\widehat{\beta}_{ij}$ es el coeficiente de regresión de la variable j para explicar la variable i, y $s_r^2(i)$ la varianza residual de la regresión.

(2) En la diagonal las inversas de las varianzas residuales de cada variable en su regresión con el resto. Es decir:

$$s^{ii} = 1/s_r^2(i)$$

(3) Si estandarizamos los elementos de esta matriz para que tenga unos en la diagonal, los elementos fuera de la diagonal son los coeficientes de correlación parcial entre estas variables. Es decir

$$r_{ij.R} = -\frac{s^{ij}}{\sqrt{s^{ii}s^{jj}}}$$

donde R se refiere al resto de las variables, es decir el conjunto de $p-2$ variables x_k con $k=1,...,p$ y $k \neq i,j$.

Por ejemplo, con cuatro variables, la primera fila de la matriz inversa de varianzas y covarianzas es

$$s_R^{-2}(1), -s_R^{-2}(1)\widehat{\beta}_{12}, -s_R^{-2}(1)\widehat{\beta}_{13}, -s_R^{-2}(1)\widehat{\beta}_{14}$$

donde $s_R^2(1)$ es la varianza residual de una regresión entre la primera variable y las otras tres y $\widehat{\beta}_{12}$, $\widehat{\beta}_{13}$, $\widehat{\beta}_{14}$ son los coeficientes de regresión en la ecuación

$$\widehat{x}_1 = \widehat{\beta}_{12}x_2 + \widehat{\beta}_{13}x_3 + \widehat{\beta}_{14}x_4$$

donde hemos supuesto, sin pérdida de generalidad, que las variables tienen media cero. Por tanto, la matriz $\mathbf{S}^{-1}$ contiene toda la información de las regresiones de cada variable en las demás.

Ejemplo 3.12 *Calculemos e interpretemos la matriz de precisión de los datos de los logaritmos de las acciones, tabla A: ACCIONES, del ejemplo 3.5. Esta matriz es*

$$S^{-1} = \begin{bmatrix} 52.0942 & -47.9058 & 52.8796 \\ -47.9058 & 52.0942 & -47.1204 \\ 52.8796 & -47.1204 & 60.2094 \end{bmatrix}$$

Por ejemplo, la primera fila de esta matriz puede escribirse como $52.0942 \times (1.0000, -0.9196, 1.0151)$ *que indica que la varianza residual de una regresión entre la primera variables y las otras dos es* $1/52.0942 = .0192$, *y los coeficientes de regresión de las variables* X_2 *y* X_3 *en una regresión para explicar* X_1 *son* -0.9196 *y* 1.0151 *respectivamente. Observemos que, de nuevo, aparece que la relación* $z = X_1 - X_2 + X_3$ *tiene poca variabilidad. La varianza de la regresión , 0.019, es menor que la de la variable* z, *ya que representa una variabilidad condicionada cuando se conocen las variables* X_2 *y* X_3.

3.9 COEFICIENTES DE ASIMETRÍA Y KURTOSIS

La generalización de los coeficientes de asimetría y kurtosis al caso multivariante no es inmediata. Una de las propuestas más utilizadas es debida a Mardia (1970), que propone calcular las distancias de Mahalanobis para cada par de elementos muestrales (i,j) :

$$d_{ij} = (\mathbf{x}_i - \overline{\mathbf{x}})' \, \mathbf{S}^{-1} \, (\mathbf{x}_j - \overline{\mathbf{x}}) \, .$$

y define el coeficiente de asimetría multivariante en la distribución conjunta de las p variables como

$$A_p = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^3,$$

y el de kurtosis

$$K_p = \frac{1}{n}\sum_{i=1}^{n} d_{ii}^2.$$

Estos coeficientes tienen las propiedades siguientes:

1. Para variables escalares $A_p = A^2$. En efecto, entonces

$$A_p = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \overline{x})^3(x_j - \overline{x})^3/s^6 = \frac{(\sum_{i=1}^{n}(x_i - \overline{x})^3)^2}{n^2 s^6} = A^2$$

2. El coeficiente de asimetría es no negativo y será cero si los datos están distribuidos homogéneamente en una esfera.

3. Para variables escalares $K = K_p$. El resultado es inmediato porque entonces $d_{ii}^2 = (x_i - \overline{x})^4/s^4$.

4. Los coeficientes son invariantes ante transformaciones lineales de los datos. Si $\mathbf{y} = \mathbf{Ax} + \mathbf{b}$, los coeficientes de asimetría y kurtosis de $\mathbf{y}$ y de $\mathbf{x}$ son idénticos.

Ejemplo 3.13 *Calcularemos los coeficientes de asimetría y kurtosis multivariantes para los datos sobre de rentabilidad de las acciones . Se comprueba que si tomamos los datos en su métrica original, el coeficiente de asimetría multivariante es:* $A_p = 16.76$.. *Este valor será, en general, mayor que los coeficientes univariantes, que son, respectivamente, 0,37, 0,04, y 2,71. Si tomamos logaritmos a los datos* $A_p = 7.5629$ *, mientras que los univariantes son 0,08, -0,25 y 1,02. Podemos concluir que , efectivamente, la transformación logarítmica ha servido para simetrizar más estos datos. El coeficiente de kurtosis multivariante es* $K_p = 31.26$, *que debe compararse con los valores univariantes de 1,38, 1,40, y 12,44. Al tomar logaritmos el coeficiente multivariante es* $K_p = 21.35$, *mientras que los univariantes son 1,43, 1,75, y 4,11, con lo que vemos que también se reduce la kurtosis tomando logaritmos.*

EJERCICIOS

Ejercicio 3.1 *Calcular el vector de medias y el de medianas para las tres variables de las ACCIONES, tabla A.7. Comparar sus ventajas como medidas de centralización de estas variables.*

Ejercicio 3.2 *Se dispone de 3 indicadores económicos* X_1, X_2, X_3, *que se miden en cuatro países, con los resultados siguientes:*

X_1	X_2	X_3
2	*3*	*-1*
1	*5*	*-2*
2	*2*	*1*
2	*3*	*1*

Calcular el vector de medias, la matriz de varianzas y covarianzas, la varianza generalizada, la matriz de correlación y la raíz y vector característico mayor de dichas matrices.

Ejercicio 3.3 *A partir de los tres indicadores económicos X_1, X_2, X_3 del problema 1 se construyen dos nuevos indicadores*

$$y_1 = (1/3)x_1 + (1/3)x_2 + (1/3)x_3$$

$$y_2 = x_1 - 0,5x_2 - 0,5x_3$$

Calcular el vector de medias para $\mathbf{y}' = (y_1, y_2)$, su matriz de varianzas y covarianzas, la matriz de correlación y la varianza generalizada.

Ejercicio 3.4 *Demostrar que la matriz* $\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$ *tiene autovalores* $1+r$ *y* $1-r$ *y autovectores* $(1,1)$ *y* $(1,-1)$.

Ejercicio 3.5 *Demostrar que si una matriz es de la forma* $C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ *los autovectores son de la forma* $(u_1, 0)$ *y* $(0,\ u_2)$, *donde* u_1 *y* u_2 *son autovectores de* A *y* B, *respectivamente.*

Ejercicio 3.6 *Cuál es la relación entre los autovalores de C y los de A y B en el ejercicio 5?.*

Ejercicio 3.7 *Demostrar que si $Y = XA$ donde Y es $n \times m$ y X es $n \times p$ las matriz de covarianzas de Y está relacionada con la de X por $S_y = A' S_x A$.*

Ejercicio 3.8 *Calcular los coeficientes de correlación múltiple entre cada variable y todas las demás para los datos de INVES.*

Ejercicio 3.9 *Calcular la matriz de correlaciones parciales para los datos de INVES.*

Ejercicio 3.10 *Demostrar que la varianza residual de una regresión múltiple entre una variable y y un conjunto de x puede escribirse como $s_y^2(1 - R^2)$ donde s_y^2 es la varianza de la variable y y R^2 el coeficiente de correlación múltiple.*

Ejercicio 3.11 *Calcular los coeficientes de correlación parcial entre las variables del conjunto de acciones mediante regresiones y utilizando los elementos de la matriz S^{-1} y comprobar la equivalencia.*

Ejercicio 3.12 *Calcular el coeficiente de asimetría multivariante para un vector de dos variables incorreladas entre si. ¿Cuál es la relación entre el coeficiente de asimetría multivariante y los univariantes?*

Ejercicio 3.13 *Repetir el ejercicio anterior para los coeficientes de kurtosis.*

Ejercicio 3.14 *Demostrar que para un conjunto de datos* $\frac{1}{np}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) = 1$ (*sugerencia, tome trazas y utilice que* $tr[\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})] = tr\,[\mathbf{S}^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})']$).

Ejercicio 3.15 *Demostrar que podemos calcular la matriz de distancias euclídeas entre los puntos con la operación* $diag(\mathbf{XX\prime})\mathbf{1} + \mathbf{1}diag(\mathbf{XX\prime}) - \mathbf{2X'X}$, *donde* $\mathbf{X}$ *es la matriz de datos,* $diag(\mathbf{XX\prime})$ *el vector que tiene por componentes los elementos diagonales y* $\mathbf{1}$ *es un vector de unos.*

Ejercicio 3.16 *Demostrar que podemos calcular la matriz de distancias de Mahalanobis entre los puntos con la operación* $diag(\mathbf{XS^{-1}X\prime})\mathbf{1} + \mathbf{1}diag(\mathbf{XS^{-1}X\prime}) - \mathbf{2X'}S^{-1}\mathbf{X}$, *donde* $\mathbf{X}$ *es la matriz de datos,* $diag(\mathbf{XS^{-1}X\prime})$ *el vector que tiene por componentes los elementos diagonales de la matriz* $\mathbf{XS^{-1}X\prime}$, *y* $\mathbf{1}$ *es un vector de unos.*

APÉNDICE 3.1: LA ESTRUCTURA DE LA MATRIZ DE PRECISIÓN

Particionemos la matriz $\mathbf{S}$ separando las variables en dos bloques: la variable 1 que llamaremos y, y el resto, que llamaremos R. Entonces:

$$\mathbf{S} = \begin{bmatrix} s_1^2 & \mathbf{c}'_{1R} \\ \mathbf{c}_{1R} & \mathbf{S}_R \end{bmatrix}$$

donde s_1^2 es la varianza de la primera variable, $\mathbf{c}_{1R}$ el vector de covarianzas entre la primera y el resto y $\mathbf{S}_R$ la matriz de varianzas y covarianzas del resto. Su inversa, utilizando los resultados del capítulo anterior sobre la inversa de una matriz particionada, será:

$$\mathbf{S}^{-1} = \begin{bmatrix} \left(s_1^2 - \mathbf{c}'_{1R}\mathbf{S}_R^{-1}\mathbf{c}_{1R}\right)^{-1} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}.$$

Supongamos para simplificar que la media de todas las variables es cero. Entonces la regresión de la primera variable sobre el resto tiene de coeficientes:

$$\widehat{\boldsymbol{\beta}}_{1R} = \mathbf{S}_R^{-1}\mathbf{c}_{1R},$$

Para encontrar la relación que buscamos, utilizaremos la identidad básica del análisis de la varianza (ADEVA):

$$\frac{1}{n}VT = \frac{1}{n}VE + \frac{1}{n}VNE$$

Apliquemos esta descomposición a la primera variable. El primer término es s_1^2, la varianza de la primera variable, y el segundo, como $\widehat{\mathbf{y}} = \mathbf{X}_R\widehat{\boldsymbol{\beta}}_{1R}$, puede escribirse:

$$\begin{aligned} \frac{1}{n}VE &= \frac{1}{n}\left(\widehat{\mathbf{y}}'\widehat{\mathbf{y}}\right) = \widehat{\boldsymbol{\beta}}'_{1R}\mathbf{S}_R\widehat{\boldsymbol{\beta}}_{1R} \\ &= \mathbf{c}'_{1R}\mathbf{S}_R^{-1}\mathbf{S}_R\mathbf{S}_R^{-1}\mathbf{c}_{1R} = \mathbf{c}'_{1R}\mathbf{S}_R^{-1}\mathbf{c}_{1R}, \end{aligned}$$

y el tercero, $VNE/n = \sum e_{1R}^2/n = s_r^2(1)$, donde hemos llamado e_{1R} a los residuos de la regresión de la primera variable respecto a las demás, y $s_r^2(1)$ a la varianza residual, sin

corregir por grados de libertad, de esta regresión. Sustituyendo estos términos en la identidad básica de ADEVA, obtenemos que la varianza residual puede calcularse como:

$$s_r^2(1) = s_1^2 - \mathbf{c}'_{1R}\mathbf{S}_R^{-1}\mathbf{c}_{1R}.$$

Si comparamos esta expresión con el primer término de la matriz $\mathbf{S}^{-1}$concluimos que el término diagonal primero de $\mathbf{S}^{-1}$ es la inversa de la varianza de los residuos (dividida por n y sin corrección por grados de libertad) en una regresión entre la primera variable y el resto. Como este análisis puede hacerse para cualquiera de las variables, concluimos que los términos diagonales de $\mathbf{S}^{-1}$ son las inversas de las varianzas residuales en las regresiones entre cada variable y el resto.

Para obtener la expresión de los términos de fuera de la diagonal en $\mathbf{S}^{-1}$ aplicaremos la fórmula para la inversa de una matriz particionada:

$$\mathbf{A}_{12} = -\left[s_r^2(1)\right]^{-1}\mathbf{c}'_{1R}\mathbf{S}_R^{-1} = -\left[s_r^2(1)\right]^{-1}\widehat{\boldsymbol{\beta}}'_{1R},$$

y, por tanto, las filas de la matriz $\mathbf{S}^{-1}$ contienen los coeficientes de regresión (cambiados de signo) de cada variable con relación a las restantes divididos por la varianza residual de la regresión (sin corregir por grados de libertad).

En resumen, $\mathbf{S}^{-1}$ puede escribirse:

$$\mathbf{S}^{-1} = \begin{bmatrix} s_r^{-2}(1) & -s_r^{-2}(1)\widehat{\boldsymbol{\beta}}'_{1R} \\ \cdots & \cdots \\ \cdots & \cdots \\ -s_r^{-2}(p)\widehat{\boldsymbol{\beta}}'_{pR} & s_r^{-2}(p) \end{bmatrix},$$

donde $\widehat{\boldsymbol{\beta}}_{jR}$ representa el vector de coeficientes de regresión al explicar la variable j por las restantes. Observemos que en esta matriz el subíndice R se refiere al conjunto de $p-1$ variables que queda al tomar como variable respuesta la que ocupa el lugar de la fial correspondiente ne la matriz. Por ejemplo, $\widehat{\boldsymbol{\beta}}'_{pR}$ es el vector de coeficientes de regresión entre la p y las $(1, ...p-1)$.

APÉNDICE 3.2 LOS DETERMINANTES DE S Y R.

Vamos a obtener expresiones para los determinantes de la matriz de varianzas y covarianzas y de correlación, utilizando los resultados para matrices particionadas del capítulo 2. Escribamos la matriz de varianzas y covarianzas como:

$$\mathbf{S}_p = \begin{bmatrix} s_1^2 & \mathbf{c}'_{1R} \\ \mathbf{c}_{1R} & \mathbf{S}_{p-1} \end{bmatrix}$$

donde s_1^2 es la varianza de la primera variable, $\mathbf{c}'_{1R}$ contiene las covarianzas entre la primera y utilizamos ahora la notación $\mathbf{S}_p$ para referirnos a la matriz de varianzas y covarianzas de las correspondientes p variables. Aplicando la fórmula para el determinante de una matriz particionada, podemos escribir

$$|\mathbf{S}_p| = |\mathbf{S}_{p-1}|\, s_1^2\left(1 - R^2_{1.2...p}\right)$$

donde $R^2_{1.2...p}$ es el coeficiente de correlación múltiple entre la primera variable y el resto que viene dado, utilizando los resultados del apéndice 3.1, por

$$R^2_{1.2...p} = \frac{1}{s_1^2}\mathbf{c}'_{1R}\mathbf{S}_{p-1}\mathbf{c}_{1R}$$

Análogamente si escribimos la matriz de correlación particionada como

$$\mathbf{R}_p = \begin{bmatrix} 1 & \mathbf{r}'_{1R} \\ \mathbf{r}_{1R} & \mathbf{R}_{p-1} \end{bmatrix}$$

donde $\mathbf{r}_{1R}$ y $\mathbf{R}_{p-1}$ son, respectivamente, el vector de correlaciones de la primera variable con el resto y la matriz de correlación entre el resto de las variables. Entonces,

$$|\mathbf{R}_p| = |\mathbf{R}_{p-1}| \left(1 - R^2_{1.2...p}\right), \tag{3.26}$$

ya que, también

$$R^2_{1.2...p} = \mathbf{r}'_{1R}\mathbf{R}_{p-1}\mathbf{r}_{1R}.$$

Para demostrar esta igualdad, observemos que la relación entre los vectores de correlaciones y covarianzas es $\mathbf{r}_{1R} = \mathbf{D}_{p-1}^{-1/2}\mathbf{c}_{1R}/s_1$, donde $\mathbf{D}_{p-1}^{-1/2}$ contiene las inversas de las desviaciones típicas de las $p-1$ variables. Como $\mathbf{R}_{p-1} = \mathbf{D}_{p-1}^{-1/2}\mathbf{S}_{p-1}\mathbf{D}_{p-1}^{-1/2}$, tenemos que

$$\mathbf{r}'_{1R}\mathbf{R}_{p-1}\mathbf{r}_{1R} = (\mathbf{c}'_{1R}/s_1)\mathbf{D}_{p-1}^{-1/2}\mathbf{D}_{p-1}^{1/2}\mathbf{S}_{p-1}^{-1}\mathbf{D}_{p-1}^{1/2}\mathbf{D}_{p-1}^{-1/2}(\mathbf{c}_{1R}/s_1) = \frac{1}{s_1^2}\mathbf{c}'_{1R}\mathbf{S}_{p-1}^{-1}\mathbf{c}_{1R} = R^2_{1.2...p}$$

Aplicando sucesivamente la ecuación (3.26), se obtiene que

$$|R_p| = \left(1 - R^2_{1.2...p}\right)\left(1 - R^2_{2.3...p}\right)...\left(1 - r^2_{p-1.p}\right).$$

APÉNDICE 3.3 CORRELACIONES PARCIALES

El coeficiente de correlación parcial es el coeficiente de correlación simple en una regresión entre residuos. Su cuadrado puede interpretarse de la forma habitual como la proporción de variación explicada respecto al total, siendo en este caso la variación total la no explicada por otra regresión previa. Vamos a utilizar esta interpretación para obtener la relación entre los coeficientes de correlación parcial y múltiple.

Supongamos p variables y vamos a obtener el coeficiente de correlación parcial entre las variables x_1, y x_2, cuando se controlan $x_3, ..., x_p$. Para ello haremos una regresión simple entre dos variables: la primera es $e_{1.3..p}$, los residuos de una regresión entre x_1 y $x_3, ..., x_p$, y la segunda $e_{2.3..p}$, los residuos de una regresión entre x_2 y $x_3, ..., x_p$. El coeficiente de correlación simple de esta regresión entre residuos, $r_{12.3...p}$, es el coeficiente de correlación parcial. Por construcción este coeficiente es simétrico entre el par de variables, pero suponiendo que tomamos la primera variable como dependiente en la regresión, la ecuación estimada entre los residuos es

$$e_{1.3,...,p} = b_{12.3,...,p}e_{2.3,...,p}$$

y el coeficiente de correlación de esta regresión, que es el de correlación parcial, será

$$r_{12.3...p} = b_{12.3,...,p}\frac{s(e_{1.3,...,p})}{s(e_{2.3,...,p})}$$

Vamos a comprobar que estos términos los podemos obtener de la matriz $\mathbf{S}^{-1}$. En esta matriz $s^{12} = -s_r^{-2}(1)\widehat{\beta}_{12.3,...,p} = s^{21} = -s_r^{-2}(2)\widehat{\beta}_{12.3,...,p}$ ya que la matriz es simétrica. dividiendo por la raiz de los elementos s^{11} y s^{22}. Se obtiene

$$-\frac{s^{12}}{s^{22}s^{11}} = \frac{s_r^{-2}(1)\widehat{\beta}_{12.3,...,p}}{s_r^{-1}(1)s_r^{-1}(2)} = \widehat{\beta}_{12.3,...,p}\frac{s_r(2)}{s_r(1)}$$

y puede comprobarse que esta expresión es $r_{12.3...p}$, el coeficiente de correlación parcial.

En la regresión entre los residuos el cociente entre la variabilidad no explicada y la total es uno menos el cuadrado del coeficiente de correlación. La variabilidad no explicada en esta regresión es la variabilidad no explicada de la primera variable respecto a todas, que llamaremos $VNE_{1.23...p}$ ($e_{1.3..p}$ contenía la parte no explicada por las variables $3, .., p$ y ahora hemos añadido la x_2). La variabilidad total de la regresión es la de los residuos, $e_{1.3..p}$, es decir la no explicada en la regresión de x_1 respecto a $x_2, ..., x_p$. Por tanto, podemos escribir:

$$1 - r^2_{12.3...p} = \frac{VNE_{1.2,3,...,p}}{VNE_{1.3...p}}$$

Vamos a expresar estas VNE en función de los coeficientes de correlación múltiple de las correspondientes regresiones. Llamando $R^2_{1.3...p}$ al coeficiente de determinación en la regresión múltiple de x_1 respecto a $x_3, ..., x_p$:

$$1 - R^2_{1.3...p} = \frac{VNE_{1.3,...,p}}{VT_1}$$

donde VT_1 es la variabilidad de la primera variable. Analogamente, en la regresión múltiple entre la primera variable y todas las demás, x_2, $x_3, ..., x_p$ tenemos que

$$1 - R^2_{1.23...p} = \frac{VNE_{1.2,3,...,p}}{VT_1}.$$

De estas tres ecuaciones deducimos que

$$1 - r^2_{12.3...p} = \frac{1 - R^2_{1.23...p}}{1 - R^2_{1.3...p}} \tag{3.27}$$

que permite calcular los coeficientes de correlación parcial en función de los coeficientes de correlación múltiple. Aplicando reiteradamente esta expresión podemos también escribir

$$\left(1 - R^2_{1.23...p}\right) = \left(1 - r^2_{12.3...p}\right)\left(1 - r^2_{13.4...p}\right)...\left(1 - r^2_{1p-1.p}\right)\left(1 - r^2_{1p}\right)$$

También puede demostrarse (véase Peña, 2002) que el coeficiente de correlación parcial entre las variables (x_1, x_2) cuando se controlan las variables $(x_3, ..., x_p)$ puede expresarse en

función del coeficiente de regresión de la variable x_2 en la regresión de x_1 con respecto a $(x_2, x_3, ..., x_p)$, y su varianza. La expresión es:

$$r_{12.3...p} = \widehat{\beta}_{12.3...p} \left(\sqrt{\widehat{\beta}^2_{12.3...p} + (n-p-1)s^2 \left[\widehat{\beta}_{12.3...p}\right]} \right)$$

donde $\widehat{\beta}_{12.3...p}$ y su varianza, $s^2 \left[\widehat{\beta}_{12.3...p}\right]$ se obtienen en la regresión entre variables de media cero:

$$\widehat{x}_1 = \widehat{\beta}_{12.3...p} x_2 + \widehat{\beta}_{13.2...p} x_3 + ... + \widehat{\beta}_{1p.2...p-1} x_p$$

Capítulo 4

ANALISIS GRAFICO Y DATOS ATIPICOS

4.1 INTRODUCCIÓN

En este capítulo vamos a continuar la descripción de datos multivariantes, estudiando su representación gráfica y posibles transformaciones de las variables que conduzcan a una descripción más simple de los datos. También introduciremos un análisis inicial de la homogeneidad de la muestra mediante el estudio de los posibles valores atípicos, debidos a errores de medida, o otras causas de heterogeneidad.

Obtener buenas representaciones gráficas de datos multivariantes es un problema difícil, y en este capítulo introduciremos los métodos más simples que se complementarán con los análisis gráficos presentados en capítulos posteriores. Recordemos que las correlaciones miden las relaciones lineales entre las variables, y pueden ser má interpretadas cuando las relaciones son no lineales. Por esa razón se intenta transformar las variables para que las variables transformadas tengan relaciones aproximadamente lineales, y veremos como generalizar las transformaciones univariantes para conseguir este objetivo. Por último, los datos multivariantes contienen con freuencia observaciones que son heterogeneas con el resto y, que si no son detectadas, pueden alterar completamente el análisis descriptivo de las variables originales. En este capítulo presentaremos métodos para detectar los datos atípicos.

4.2 REPRESENTACIONES GRÁFICAS

4.2.1 Histogramas y diagramas de dispersión

El primer paso de cualquier análisis multivariante es representar gráficamente las variables individualmente, mediante un histograma o un diagrama de caja. Estas representaciones son muy útiles para detectar asímetrías, heterogeneidad, datos atípicos etc. En segundo lugar conviene construir los diagramas de dispersión de las variables por pares, y esta posibilidad se incluye ya en muchos programas de ordenador. Con p variables existen $p(p-1)/2$ gráficos posibles que pueden disponerse en forma de matriz y son muy útiles para entender el tipo de relación existente entre pares de variables, e identificar puntos atípicos en la relación

bivariante. En particular, estos gráficos son importantes para apreciar si existen relaciones no lineales, en cuyo caso la matriz de covarianzas puede no ser un buen resumen de la dependencia entre las variables.

Podemos simular gráficos de tres variables presentando en la pantalla de un ordenador proyecciones adecuadas de esta relación girando el punto de vista del observador para dar idea del espacio tridimensional. Estas representaciones gráficas se conocen con el nombre de *Gran Tour* de los datos y pueden ser muy útiles, utilizados interactivamente con un ordenador, pero no pueden construirse para dimensiones superiores a tres. También para variables discretas podemos construir diagramas de barras tridimensionales y para variables continuas podemos construir los equivalentes multidimensionales de los histogramas. La figura 4.3 presenta un ejemplo de estas representaciones.

Ejemplo 4.1 *La figura muestra los gráficos de dispersión de los datos de medidas de desarrollo del mundo, MUNDODES, del Anexo I.*

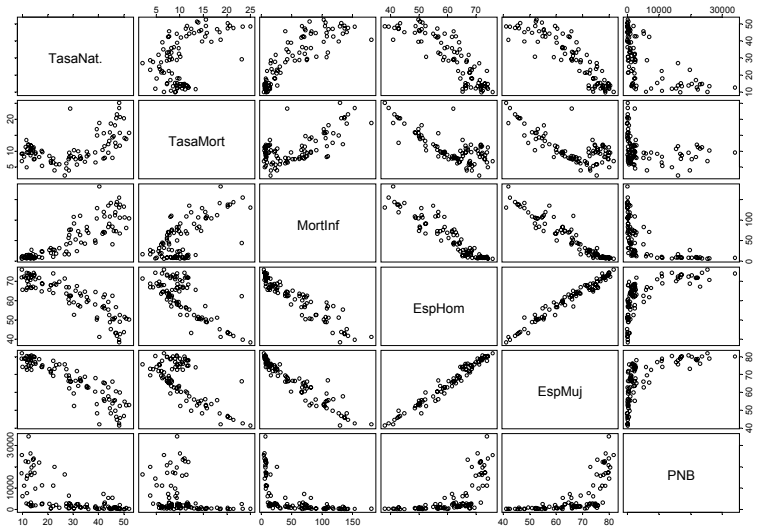


Figura 4.1: Matriz de dispersión para los datos MUNDOES.

La figura 4.1 ilustra claramente que existen relaciones de dependencia fuerte entre las variables, muchas de carácter no lineal. Por ejemplo, la relación entre las variables primera y segunda, tasa de natalidad y de mortalidad, es claramente no lineal y se observa un valor atípico muy destacado en la relación. En toda la primera fila (o columna) que indica las relaciones de la primera variable (tasa de natalidad) con las restantes las relaciones parecen no lineales y , en algunos casos, heterocedástica. Comentarios similares se aplican a la segunda variable. En otros casos parece que la relación entre dos variables es diferente para distintos grupos de países. Por ejemplo, en prácticamente todas las relaciones en que aparece la sexta variable, riqueza del país medida por el PNB, parecen existir dos tipos de países. En unos parece no existir relación entre la variable demográfica y el PNB, mientras que en los otros parece existir una clara relación positiva (como con la tasa de mortalidad) o negativa (como con la mortalidad infantil) entre las variables demográficas y el PNB.

Esta figura muestra además que algunas de las relaciones son heterocedásticas, es decir, que la variabilidad de la relación aumenta al aumentar los niveles de las variables. Por ejemplo, en la relación entre tasa de natalidad y mortalidad infantil, donde además se aprecia claramente un valor atípico. Este punto aparece muy claramente en la relación entre las dos primeras variables (posiciones 1,2 y 2,1 de la matriz) y en los gráficos de la segunda fila y columna, indicando que el punto es atípico en las dos primeras variables.

Finalmente algunas relaciones son muy fuertes y lineales como entre las esperanzas de vida y la mortalidad infantil.

Ejemplo 4.2 *El gráfico siguiente presenta los diagramas de dispersión para los datos de las ACCIONES.*

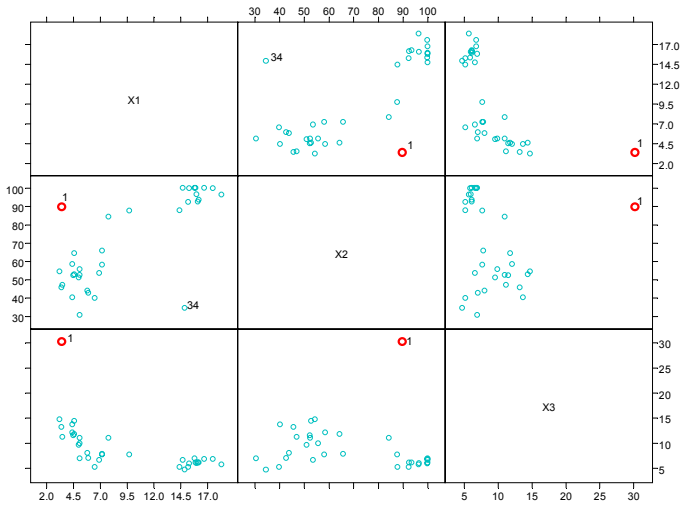


Figura 4.2: Matriz de dispersión de los datos de ACCIONES

En la figura 4.2 se observa como la primera observación aparece como un valor atípico en todos los diagramas de dispersión. En los gráficos con la tercera variable este punto es un valor muy extremo en la relación, mientras que en los gráfico de las otras dos variables aparece como atípico, pero no parece muy influyente en la relación lineal entre ambas variables. La acción 34 aparece, al igual que la 1, como una observación aislada heterogénea con el resto.

En este caso podemos hacer una representación en tres dimensiones de las tres variables. En la figura 4.3 se observa que la relación entre las variables x_1 y x_2 depende del nivel de la x_3. El gráfico ilustra también con claridad el carácter claramente atípico de las observaciones 1 y 34 que aparecen muy separadas del resto. Por otro lado, se observa en el gráfico tridimendional que las observaciones se agrupan en dos conjuntos distintos. Esta característica, que se apunta en los gráficos bidimensionales, aparece claramente de manifiesto en la representación tridimendional, ilustrando las ventajas de construir estas representaciones cuando sea posible.

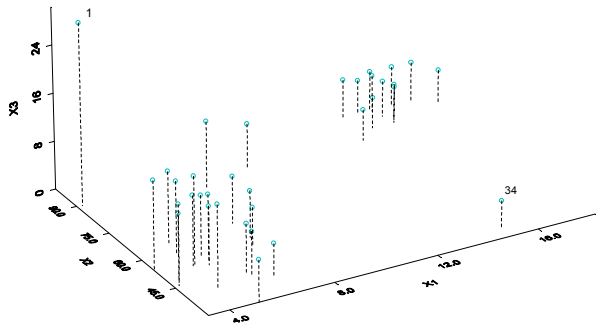


Figura 4.3: Representacion Tridimensional de los datos de ACCIONES

En la figura 4.4 se presenta la matriz de datos de dispersión para los datos de las acciones ahora en los logaritmos de las variables. Se observa que la transformación en logaritmos aporta mayor linealidad a las relaciones entre variables dos a dos y reduce algo el efecto de la primera observación que es atípica.

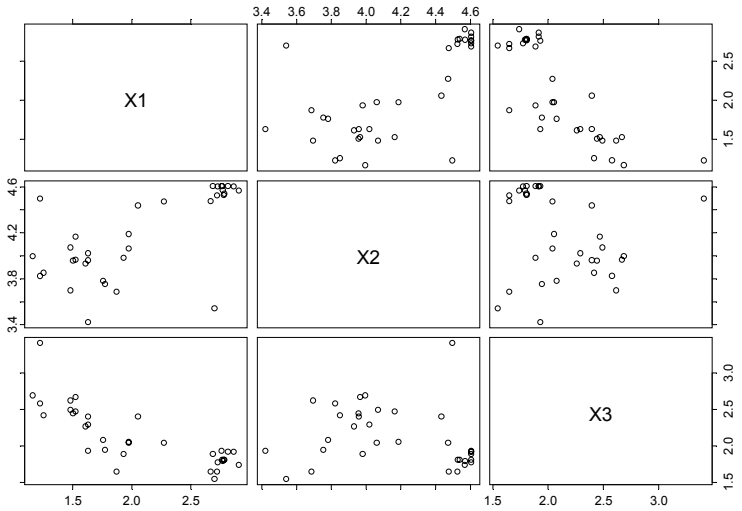


Figura 4.4: Matriz de dispersión para los logartimos de los datos de ACCIONES.

Ejemplo 4.3 *La figura 4.5 representa los datos A.4, INVEST, para las publicaciones científicas. Se observa que existe una fuerte relación entre todas las variables. Las relaciones son aproximadamente lineales, si bien en algunos casos se observa cierta curvatura que podría*

resolverse tomando logaritmos. No hay valores atípicos muy destacados.

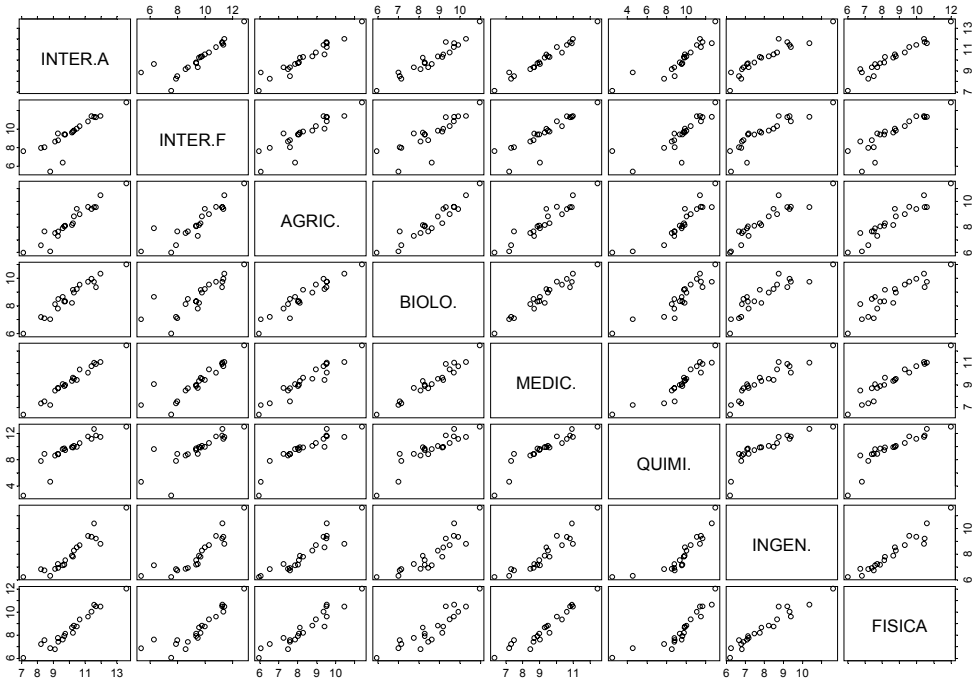


Figura 4.5: Representación como matriz de dispersión de los datos de INVES

Ejemplo 4.4 *La figura 4.6 presenta los gráficos de dispersión para los datos de medidas fisicas del banco de datos MEDIFIS. Las relaciones son aproximadamente lineales y no se detecta la presencia de datos atípicos destacados.*

4.2.2 Representación mediante figuras

Para más de tres variables se utilizan principalmente dos tipos de métodos gráficos. El primero, es mostrar los datos mediante figuras planas, asociando cada variable a una característica del gráfico. El segundo, es buscar conjuntos de proyecciones en una y dos dimensiones que revelen aspectos característicos de los datos. Vamos a presentar en esta sección el primer enfoque, en la sección siguiente hablaremos del segundo.

Existen muchas alternativas posibles para representar los datos mediante figuras. Chernoff ha propuesto la utilización de caras, que tienen la ventaja de nuestra facilidad para reconocer patrones en este formato y el inconveniente de que la representación es muy dependiente de las variables escogidas para representar cada rasgo. Por ejemplo, la boca y la forma de la cabeza son rasgos más llamativos que las orejas o la longitud de la nariz, y el mismo conjunto de datos puede sugerir distintos patrones de similitud entre las observaciones según la asociación elegida entre rasgos y variables. La figura 4.7 presenta un ejemplo.

Si asociamos cada variable a un rasgo de una figura plana, podemos representar cada elemento en la muestra por una figura geométrica. En estas representaciones las similitudes

entre figuras indican las similitudes entre los elementos, y los valores atípicos aparecerán como figuras discordantes con el resto. Una figura muy utilizada es la estrella. Por ejemplo, para representar cinco variables, podemos escoger una estrella de cinco radios y asociar cada variable a cada uno de estos radios o ejes. Cada observación dará lugar a una estrella. Normalmente las variables se estandarizan de manera que tengan media cero y desviación típica unitaria. Entonces, se marca el cero sobre cada eje y se representa el valor de la variable en unidades de desviaciones típicas. La figura 4.8 presenta un ejemplo de su uso.

Ejemplo 4.5 *La figura 4.7 presenta una representación gráfica de los datos de investigación, INVEST, con las* Caras de Chernoff. *Cada observación es convertida en una cara, y cada variable es asignada a una característica de la misma. Para el siguiente ejemplo se ha utilizado el programa Splus, que asigna las variables a las siguientes características: (1) área de la cara; (2) forma de la cara; (3) longitud de la nariz; (4) localización de la boca; (5) curva de la sonrisa; (6) grosor de la boca; (7a 11) localización, separación, ángulo, forma y grosor de los ojos, etc. Se puede representar más de 15 características, y, originalmente, Chernoff logró representar 18 variables en una cara. En la figura 4.7 se han representado los países contenidos en la base de datos INVEST pero eliminado EEUU, ya que este país distorsionaría la representación gráfica por tomar un valor muy extremo en todas las variables. En este tipo de gráficos podemos o bien ver el comportamiento por separado de cada variable o bien la similitud global de cada dato multivariado. Por ejemplo, la variable MEDIC se ha asignado a la curva de la sonrisa y vemos que los primeros cuatro países son claramente diferentes en cuanto a esta característica. Sin embargo, juzgando globalmente, notamos que el comportamiento más parecido lo presentan los cinco primeros países.*

La representación de las caras de Chernoff nos permite observar las diferencias entre los países en cuanto al volumen de publicaciones. Para observar las diferencias entre los patrones de publicación de los distintos países deberíamos aplicar logaritmos a los datos para reducir la asimetría en las distribuciones univariantes, observada en la figura 4.10, y para linealizar más las relaciones.

Ejemplo 4.6 *Para representar los países con las variables en logaritmos se ha optado por un gráfico de estrellas. Como se explicó anteriormente, cada radio de la estrella está asociado a una variable, en el ejemplo que trataremos fue utilizado Splus, este programa comienza a asignar variables desde la derecha en el sentido opuesto a las agujas del reloj. En la figura 4.8 se presenta cómo es esta asignación para las variables de la base INVEST.En la figura 4.9 se siguen observando diferencias de tamaño entre los primeros cinco países y el resto, pero se aprecian ciertos patrones en los que se distinguen países con tendencia a la investigación en algunas áreas frente a otras.*

4.2.3 (*)Representación de Proyecciones

En lugar de intentar representar las variables originales por pares podríamos intentar representar parejas de variables que resuman en algún sentido el conjunto de variables. Esperamos así obtener una mayor intuición visual de las propiedades de los datos. Una forma simple de resumir un vector de variables es construir una variable escalar como combinación lineal de

sus valores. Por ejemplo, si $\mathbf{x}' = (x_1, ..., x_p)$ representa el precio de un conjunto de productos en un mercado, una medida resumen de estos precios es:

$$y = \mathbf{a}'\mathbf{x} = \sum_{j=1}^{p} a_j x_j. \tag{4.1}$$

Si $a_j = 1/p$, la combinación lineal resultante es la media de los precios. Si $a_j \neq 1/p$, pero $a_j \geq 0$ y $\sum a_j = 1$ la nueva variable es una media ponderada de las variables originales con pesos a_j. En general (4.1) define una nueva variable que informa globalmente del conjunto de variables $\mathbf{X}$.

La variable escalar obtenida mediante una combinación lineal puede siempre interpretarse geométricamente como una proyección. El producto escalar del vector $\mathbf{x}$, en $\Re^p$, por otro vector $\mathbf{a}$ de $\Re^p$viene dado por:

$$\mathbf{a}'\mathbf{x} = |\mathbf{a}||\mathbf{x}| \ \cos\alpha \tag{4.2}$$

y si el vector de ponderación, $\mathbf{a}$, se toma de manera que su norma sea uno, $|\mathbf{a}| = 1$, el producto escalar es directamente la proyección del vector $\mathbf{x}$ sobre la dirección del vector $\mathbf{a}$. En consecuencia, si elegimos una dirección con $|\mathbf{a}| = 1$, la nueva variable escalar

$$y_i = \mathbf{a}'\mathbf{x}_i \tag{4.3}$$

que tomará valores $(y_1, ..., y_n)$, puede interpretarse como la proyección de los datos $\mathbf{X}$ sobre la dirección indicada por el vector $\mathbf{a}$. El conjunto de los n valores de la nueva variable $\mathbf{y}$ pueden englobarse en un vector $\mathbf{y}$ $(n \times 1)$ que vendrá dado por

$$\mathbf{y} = \mathbf{X}\mathbf{a}, \tag{4.4}$$

donde $\mathbf{X}$ es la matriz de datos $n \times p$.

Como construir un indicador a partir de variables multivariantes puede interpretarse como proyectar los datos sobre cierta dirección, es natural preguntarse por direcciones de proyección que sean informativas para revelarnos la disposición de los puntos en el espacio. Para ello tenemos que definir un criterio de proyección y encontrar la dirección donde ese criterio se maximiza. Las técnicas diseñadas con este objetivo se conocen como *búsqueda de proyecciones* (projection pursuit), y se aplican como sigue:

1. Escoger la dimensión del espacio sobre el que vamos a proyectar (normalmente 2), y el criterio que se desea maximizar.

2. Encontrar la dirección que maximiza el criterio analiticamente. Si no es posible encontrar la dirección de forma analítica hacerlo de manera aproximada, por ejemplo seleccionando un número grande de direcciones $(\mathbf{a}_1, ..., \mathbf{a}_N)$, evaluando el criterio en cada una y seleccionando la dirección de este conjunto donde el criterio toma el valor máximo.

3. Encontrar una dirección ortogonal a la primera que maximice el criterio. Esto puede hacerse por ejemplo proyectando los datos sobre el espacio ortogonal a la primera dirección, $\mathbf{a}$, lo que supone transformales con $\mathbf{Y} = (\mathbf{I} - \mathbf{a}\mathbf{a}\prime)\mathbf{X}$ y aplicar el algortimo del punto 2 a los nuevos datos $\mathbf{Y}$.

4. Representar los datos sobre el plano definido por las dos direcciones de proyección.

Se suelen considerar interesantes las proyecciones que muestren relaciones no lineales entre las variables, o distribuciones multimodales que pueden indicar la presencia de clusters o grupos de observaciones. Inicialmente las funciones objetivo utilizadas se basaban en la teoría de la información. Por ejemplo, una medida de diversidad o heterogeneidad es la entropía de Shannon

$$I(x) = \int \log f(x) f(x) dx$$

que, entre las distribuciones continuas, se minimiza con la distribución normal. Si maximizamos esta función esperamos obtener proyecciones donde la distribución resultante se aparte más de la normal, en cierto sentido, lo que puede resultar en combinaciones interesantes y estructuras inesperadas entre las variables. Naturalmente otros muchos criterios son posibles, y en la sección 4.5 utilizaremos otro criterio para buscar direcciones que muestren la presencia de atípicos. En el capítulo siguiente utilizaremos estas ideas para obtener proyeciones que mantengan lo más posible las distancias entre los puntos en el espacio. Los capítulos 5, 6 y 7 presentan más ejemplos de estas técnicas gráficas.

4.3 TRANSFORMACIONES LINEALES

4.3.1 Consecuencias

Muchas propiedades importantes de los datos son independientes de las unidades de medida de las variables y no cambiarán si pasamos de euros a dólares o de centímetros a metros. Vamos a estudiar como afectan cambios en las unidades de medida a los estadísticos estudiados en el capítulo 3. Por ejemplo, supongamos que en lugar de medir una variable bidimensional $\mathbf{x} = (x_1, x_2)'$ en euros y en unidades lo hacemos en dólares y en miles de unidades, $\mathbf{y} = (y_1, y_2)'$. La relación entre ambas variables será:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{4.5}$$

donde $\mathbf{A}$ es una matriz diagonal que tiene como términos diagonales los factores de conversión de euros a dólares y de unidades a miles de unidades (1/1000). Para el conjunto de las n observaciones la relación será:

$$\mathbf{Y} = \mathbf{X}\ \mathbf{A} \tag{4.6}$$

donde $\mathbf{X}$ e $\mathbf{Y}$ son $n \times p$, y $\mathbf{A}$ es una matriz diagonal $p \times p$. Aplicando la definición de vector de medias

$$\overline{\mathbf{y}} = \frac{1}{n}\mathbf{Y}'\mathbf{1} = \mathbf{A}'\frac{\mathbf{1}}{\mathbf{n}}\mathbf{X}'\mathbf{1} = \mathbf{A}'\overline{\mathbf{x}} \tag{4.7}$$

y como $\mathbf{A} = \mathbf{A}'$, el vector de medias se transforma de la misma forma que los hacen las variables.

Las matrices de varianzas y covarianzas estarán relacionadas por:

$$\mathbf{S}_y = \frac{1}{n}\mathbf{Y}'\mathbf{P}\mathbf{Y} = \mathbf{A}'(\frac{1}{n}\mathbf{X}'\mathbf{P}\mathbf{X})\mathbf{A} = \mathbf{A}'\mathbf{S}_x\mathbf{A}. \tag{4.8}$$

El cambio de unidades es un caso particular de una transformación lineal de las variables para simplificar su interpretación. Una transformación lineal importante es la estandarización de las variables, que puede hacerse de dos formas distintas, como veremos a continuación.

4.3.2 Estandarización univariante

Llamando $\mathbf{x}$ al vector $p \times 1$ de la variable vectorial, la transformación lineal

$$\mathbf{y} = \mathbf{D}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}})$$

donde la matriz $\mathbf{D}^{-1/2}$ es cuadrada y diagonal con términos:

$$\mathbf{D}^{-1/2} = \begin{bmatrix} s_1^{-1} & 0 & \dots & 0 \\ 0 & s_2^{-1} & \dots & 0 \\ 0 & 0 & \dots & s_p^{-1} \end{bmatrix},$$

convierte las variables originales, $\mathbf{x}$, en otras nuevas variables, $\mathbf{y}$, de media cero y varianza unidad. Cada componente del vector $\mathbf{x}$, x_j para $j = 1, ..., p$, se transforma con $y_j = (x_j - \overline{x_j})/s_j$. La matriz de varianzas y covarianzas de las nuevas variables será la matriz de correlación de las variables primitivas. Esta transformación es la estandarización univariante de las variables.

4.3.3 (*)Estandarización multivariante

Dada una matriz definida positiva, $\mathbf{S}_x$, puede definirse su raíz cuadrada $\mathbf{S}_x^{1/2}$, por la condición

$$\mathbf{S}_x = \mathbf{S}_x^{1/2}(\mathbf{S}_x^{1/2})' \tag{4.9}$$

La matriz $\mathbf{S}_x^{1/2}$ no es única (véase 2.4.2). En efecto si $\mathbf{S}_x^{1/2}$ verifica la condición (4.9) también la verifica $\mathbf{S}_x^{1/2}\mathbf{M}$, donde $\mathbf{M}$ es cualquier matriz ortogonal. La matriz $\mathbf{S}_x^{1/2}$ puede construirse a partir de la descomposición espectral

$$\mathbf{S}_x = \mathbf{A}\mathbf{D}\mathbf{A}'$$

donde $\mathbf{D}$ es diagonal y contiene los valores propios de $\mathbf{S}_x$ y $\mathbf{A}$ es ortogonal y contiene los vectores propios. Sea $\mathbf{D}^{1/2}$ la matriz diagonal cuyos términos son las raíces cuadradas de los términos de $\mathbf{D}$, que son positivos. Definiendo la raiz cuadrada por la matriz simétrica:

$$\mathbf{S}_x^{1/2} = \mathbf{A}\mathbf{D}^{1/2}\mathbf{A}' \tag{4.10}$$

la variable

$$\mathbf{y} = \mathbf{S}_x^{-1/2}(\mathbf{x}-\overline{\mathbf{x}})$$

tiene media cero y matriz de varianzas y covarianzas identidad, ya que

$$\mathbf{S}_y = \mathbf{S}_x^{-1/2}\mathbf{S}_x\mathbf{S}_x^{-1/2} = \mathbf{I}$$

Con esta transformación pasamos de variables correladas, con matriz de covarianza $\mathbf{S}_x$, a variable incorreladas, con matriz de varianzas identidad. El nuevo conjunto de variables viene dado por

$$\mathbf{Y} = \widetilde{\mathbf{X}}\mathbf{S}_x^{-1/2} = \widetilde{\mathbf{X}}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{A}'$$

Esta estandarización se denomina multivariante, ya que utiliza todas las covarianzas para estandarizar cada variable. Observemos que la estandarización univariante utiliza sólo los términos diagonales de $\mathbf{S}_x$ para construir $\mathbf{D}^{-1/2}$, y no tiene en cuenta las covarianzas, mientras que la multivariante utiliza toda la matriz.

Ejemplo 4.7 *La tabla A.4 de los daos de INVEST presenta el número de publicaciones recogidas en un trienio en 8 bases de datos de producción científica para los países de la OCDE. (La descripción de las fuentes se encuentra en el apéndice de datos). En la figura 4.10 se presenta un diagrama de cajas múltiple (Boxplot) que permite, además de la exploración de cada una de las variables, comparar los rangos de todas ellas de forma conjunta.*

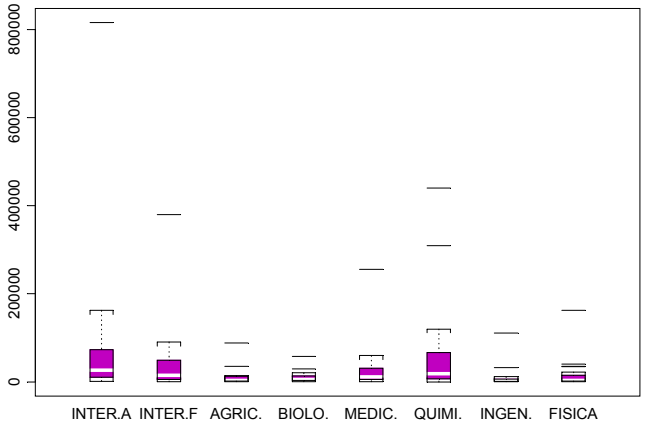


Figura 4.10: Diagrama de cajas de las variables de INVEST.

En el gráfico se observa la existencia de un atípico en todas las variables (EEUU) y una asimetría en la distribución de todas las variables que puede estar producida por éste dato.

	INTER.A	INTER.F	AGRIC.	BIOLO.	MEDIC.	QUIMI.	INGEN.	FISICA
EE.UU	4.2223	4.0650	3.9773	3.5825	4.1091	3.3889	4.1696	4.0846
UK	0.4845	0.5640	1.2398	1.4429	0.5513	0.2697	-0.1532	0.4831
JP	0.1627	0.4247	0.1562	0.4567	0.4788	2.2109	0.9060	0.6573
F	0.2375	0.3930	0.1480	0.0406	0.3755	0.5152	-0.0054	0.5237
G	0.0782	0.5000	0.0417	0.7273	0.2177	0.0305	0.0491	0.1381
C	-0.0269	0.0698	0.1594	0.4540	-0.1104	0.3521	0.0793	-0.0716
I	-0.1975	-0.1687	-0.1545	0.2062	0.0336	-0.2367	-0.1770	-0.1643
A	-0.2363	-0.2594	0.0765	-0.0645	-0.3089	-0.3865	-0.2156	-0.3065
H	-0.2719	-0.3102	-0.2232	-0.2395	-0.2811	-0.3561	-0.2611	-0.2931
S	-0.2796	-0.3325	-0.3551	-0.0918	-0.2606	-0.3982	-0.3194	-0.3839
CH	-0.2914	-0.3527	-0.3861	-0.5353	-0.3287	-0.3895	-0.3124	-0.3210
E	-0.3490	-0.3854	-0.4009	-0.5092	-0.3994	-0.4237	-0.3660	-0.4081
B	-0.3440	-0.3857	-0.3932	-0.5069	-0.3831	-0.4554	-0.3448	-0.3877
D	-0.3590	-0.5216	-0.4241	-0.3817	-0.3782	-0.4348	-0.3686	-0.4276
AU	-0.3803	-0.3692	-0.4856	-0.6308	-0.4224	-0.5026	-0.3636	-0.4197
FI	-0.3800	-0.4502	-0.4552	-0.4506	-0.4260	-0.5032	-0.3767	-0.4369
N	-0.3911	-0.4626	-0.4667	-0.5608	-0.4428	-0.5150	-0.3803	-0.4598
Y	-0.4162	-0.4925	-0.4550	-0.7199	-0.4971	-0.4996	-0.3849	-0.4315
GR	-0.4217	-0.4950	-0.5235	-0.7124	-0.5024	-0.5412	-0.3810	-0.4454
IR	-0.4042	-0.5257	-0.5368	-0.7256	-0.5053	-0.5620	-0.3964	-0.4574
P	-0.4360	-0.5050	-0.5391	-0.7810	-0.5197	-0.5627	-0.3976	-0.4722

Tabla 4.1: Estandarización univariante de INVEST

Vamos a estudiar para estos datos las dos estandarizaciones propuestas:

Se observa que la estandarización univariante resalta el valor atípico de EEUU, pero mantiene sin cambios importantes las variables, que sufren solamente un cambio de escala.

La estandarización multivariante transforma totalmente las variables originales. En la primera variable EEUU sigue siendo atípico, pero en las siguientes esta característica desaparece. En el capítulo siguiente, componentes principales, interpretaremos las propiedades de estas nuevas variables transformadas.

4.4 TRANSFORMACIONES NO LINEALES

4.4.1 Simplicidad en las distribuciones

El análisis de un conjunto de datos multivariante es más simple cuando su distribución es simétrica y las relaciones entre las variables son lineales, y la mayoría de los métodos multivariantes hacen estas hipótesis. En estas condiciones, la matriz de varianzas y covarianzas es un buen resumen de las relaciones de dependencia existentes.

Al elegir las variables conviene tener en cuenta que la misma variable puede medirse de muchas formas, en principio igualmente válidas. Por ejemplo, el consumo de gasolina de un automóvil se expresa en Europa en litros cada 100 kilómetros (x) mientras que en EE.UU

	y_1	y_2	y_3·	y_4	y_5	y_6	y_7	y_8
EE.UU	4.15	-0.36	1.53	-0.22	-0.46	-0.12	0.12	0.05
UK	0.64	-2.14	-2.70	2.18	0.00	-0.25	0.18	0.44
JP	0.70	3.81	-1.77	0.18	-0.43	-0.34	-0.60	0.25
F	0.29	0.58	0.02	1.78	3.29	0.11	0.32	-0.81
G	0.23	-0.70	-1.11	-2.88	1.90	1.81	-0.65	-0.57
C	0.11	0.18	-1.40	-0.64	-1.47	1.72	1.96	-0.36
I	-0.11	-0.48	-0.61	-1.33	0.34	-2.79	-1.84	0.19
A	-0.22	-0.66	-0.28	0.52	-1.76	0.92	-2.36	-0.56
H	-0.29	-0.23	0.05	-0.03	-0.42	-0.29	-0.66	-1.06
S	-0.32	-0.35	-0.28	-1.14	-0.06	-1.25	1.71	0.94
CH	-0.38	0.08	0.62	0.32	0.24	-0.06	0.22	0.92
E	-0.42	0.01	0.40	0.18	-0.01	0.34	0.36	2.03
B	-0.42	-0.05	0.48	0.12	-0.04	0.14	-0.18	0.91
D	-0.43	-0.07	0.12	-0.19	-0.48	-1.67	1.41	-1.15
AU	-0.47	0.05	0.73	-0.02	0.34	0.80	-0.50	2.10
FI	-0.46	-0.12	0.31	-0.32	-0.17	-0.33	0.65	-0.96
N	-0.48	-0.03	0.51	-0.04	-0.20	-0.07	0.26	0.90
Y	-0.51	0.14	0.77	0.68	-0.27	0.47	-0.59	-0.62
GR	-0.53	0.12	0.81	0.23	-0.11	0.33	-0.17	-1.38
IR	-0.54	0.09	0.86	0.26	-0.14	-0.00	0.70	-1.31
P	-0.55	0.15	0.93	0.36	-0.09	0.52	-0.35	0.06

Tabla 4.2: Estandarización multivariante de INVEST

se expresa en km recorridos con 1 litro (o galón) de gasolina (y). La relación entre ambas medidas es no lineal, ya que $y = 100/x$. Como segundo ejemplo, para medir el crecimiento de una variable C_t en el tiempo podemos calcular las diferencias $C_t - C_{t-1}$, pero en general resulta más relevante considerar las diferencias relativas $(C_t - C_{t-1})/C_{t-1}$ o $(C_t - C_{t-1})/C_t$. Si expresamos la variable en logaritmos, sus diferencias en dicha escala son una buena medida del crecimiento relativo, ya que:

$$\ln C_t - \ln C_{t-1} = \ln \frac{C_t}{C_{t-1}} = \ln \left(1 + \frac{C_t - C_{t-1}}{C_{t-1}}\right) \simeq \frac{C_t - C_{t-1}}{C_{t-1}}$$

utilizando que $\ln(1+x)$ es aproximadamente x, si x es pequeño. Además, es fácil demostrar que, supuesto $C_t \geq C_{t-1}$:

$$\frac{C_t - C_{t-1}}{C_t} \leq \ln \frac{C_t}{C_{t-1}} \leq \frac{C_t - C_{t-1}}{C_{t-1}}$$

y las diferencias de las variables en logaritmos son una medida promedio de las dos formas posibles de medir el crecimiento relativo. El logaritmo es una de las transformaciones más utilizadas para datos positivos ya que:

(1) Las distribuciones que describen el tamaño de las cosas (renta de países o familias habitantes en las principales ciudades del mundo, tamaño de empresas, consumo de energía

en hogares, etc), son generalmente muy asimétricas, pero se convierten en aproximadamente simétricas al expresar la variable en logaritmos.

(2) Cuando las diferencias relativas entre los valores de la variable sean importantes, conviene expresar las variables en logaritmos, ya que las diferencias entre logaritmos equivalen a diferencias relativas en la escala original.

(3) La variabilidad de las variable transformada es independiente de las unidades de medida.

Para comprobar esta última propiedad, supongamos una variable escalar x que transformamos con $y = \log x$ y la variable transformada tiene media $\overline{y}$ y varianza s_y^2. Si cambiamos las unidades de medida de x multiplicando por una constante, $z = kx$, entonces la variable $\log z$ tiene media $\overline{y} + \log k$ y la misma varianza que la variable $\log x$.

4.4.2 Simplicidad en las relaciones

Es frecuente con datos económicos observar fuertes relaciones no lineales entre las variables. En estos casos, el análisis de los datos se simplifica mucho si transformamos las variables de manera que las nuevas variables tengan relaciones lineales. Por ejemplo, una relación frecuente es del tipo proporcional

$$y = kx^b \tag{4.11}$$

que implica que si la variable x aumenta en una unidad la variable y aumenta (supuesto $b > 0$) una cantidad que depende del valor de x, pero el incremento proporcional de y cuando x aumenta un 1% es constante e igual al b%. Esta relación suele ir unida a heterocedasticidad en la relación, manifestada en una mayor variabilidad en el gráfico de dispersión cuando las variables toman valores altos que en la zona de valores bajos. La relación puede convertirse en lineal y homocedástica (varianza constante) transformando las variables en logaritmos. En efecto, tomando logaritmos en (4.11) y llamando $y^* = \log y$, $x^* = \log x$ tenemos una relación lineal entre las nuevas variables $(x^* y^*)$. A la hora de decidir si transformar o no las variables es importante tener en cuenta la interpretación de las nuevas variables.

Las transformaciones habituales de las variables individuales pueden escribirse mediante la familia potencial de Box-Cox:

$$\begin{aligned} y^{(\lambda)} &= \frac{x^\lambda - 1}{\lambda}, \quad \text{para } \lambda \neq 0 \\ y^{(\lambda)} &= \log x, \quad \text{para } \lambda = 0. \end{aligned}$$

Un estudio más detallado de esta transformación incluyendo la estimación del parámetro λ se realizará en el capítulo 10. La transformación puede extenderse para tratar de transformar conjuntamente el vector de variables para que todas las distribuciones conjuntas de grupos de variables sean simétricas. (vease Gnanadesikan, 1997).

Ejemplo 4.8 *La figura 4.11 presenta los diagramas de dispersión de las variables de INVES en logaritmos con los histogramas de las variables en la diagonal principal. Este gráfico se ha hecho con Matlab. Se observa que la transformación logarítmica hace las relaciones más lineales y los histogramas de todas las variables más simétricos.*

Ejemplo 4.9 *La figura 4.12 muestra la representación de los datos de EPF en logaritmos. Se observa que las relaciones son aproximadamente lineales y los histogramas simétricos.*

4.5 DATOS ATÍPICOS

4.5.1 Definición

Llamaremos datos atípicos a aquellas observaciones que parecen haberse generado de forma distinta al resto de los datos. Pueden ser causadas por errores de medición o transcripción, cambios en el instrumento de medición o a heterogeneidad intrínseca de los elementos observados. Por ejemplo, supongamos que estamos estudiando las características de las viviendas en una zona urbana donde la gran mayoría son pisos, pero se ha incluido en la muestra una gran vivienda unifamiliar con jardín. Esta observación será atípica y corresponde a una heterogeneidad real de los datos. Es importante detectarla ya que obtendremos una mejor descripción de los datos separando ambos tipos de viviendas.

Los análisis efectuados sobre datos recogidos en condiciones de estrecho control, revelan que es frecuente que aparezcan entre un 1% y un 3% de observaciones atípicas respecto al resto de la muestra. Cuando los datos se han recogido sin un cuidado especial, la proporción de datos atípicos puede llegar al 5% y ser incluso mayor.

La caracterización de un sólo valor atípico es simple ya que, por definición, debe estar alejado del resto, con lo que la distancia entre el atípico y el resto de las observaciones será grande. Alternativamente, podemos definir una atípico como aquel punto que se encuentra lejos del centro de los datos. Llamando $\overline{\mathbf{x}}$ al vector de medias y utilizando como medida de distancia la distancia euclídea, una observación $\mathbf{x}_i$ será atípica en esta métrica si

$$d_E(\mathbf{x}_i,\overline{\mathbf{x}}) = \left[(\mathbf{x_i}-\overline{\mathbf{x}})'(\mathbf{x_i}-\overline{\mathbf{x}})\right]^{1/2}$$

es grande. Para identificar las observaciones atípicas podríamos hacer un histograma de estas distancias y ver si existen puntos mucho más alejados que los demás. Sin embargo, como hemos visto, esta medida de distancia no es razonable cuando exista dependencia entre las observaciones. La figura 4.13 ilustra una situación donde el punto + es claramente atípico y, sin embargo, ni está a una distancia euclídea grande del centro de los datos, ni aparecerá como atípico al analizar cada variable aisladamente. El problema es que, como vimos, la distancia euclídea no tiene en cuenta la estructura de correlación de los datos, y una posibilidad mejor es estandarizar previamente los datos de forma multivariante. De esta manera los datos transformados tienen media cero y matriz de covarianzas identidad, y podemos buscar atípicos con la distancia euclídea, eliminando el problema de la correlación entre las variables. Definiendo, como antes, las variables estandarizadas multivariantemente por:

$$\mathbf{y} = \mathbf{S}_x^{-1/2}(\mathbf{x}-\overline{\mathbf{x}})$$

La distancia euclídea al cuadrado entre una observación, $\mathbf{y}_i$, y su media, cero, será

$$d_E^2(\mathbf{y}_i,\mathbf{0}) = \mathbf{y_i'y_i} = (\mathbf{x_i}-\overline{\mathbf{x}})'\mathbf{S_x^{-1}}(\mathbf{x_i}-\overline{\mathbf{x}}) = d_M^2(\mathbf{x}_i,\overline{\mathbf{x}})$$

y la distancia euclídea entre las variables incorreladas equivale a la distancia de Mahalanobis entre las variables originales. Podríamos entonces identificar datos atípicos calculando las distancias de Mahalanobis para todos ellos y viendo si existe algún punto con una distancia mucho mayor que el resto.

4.5.2 Los efectos de los atípicos

Las consecuencias de una sola observación atípica pueden ser graves: distorsionar las medias y desviaciones típicas de las variables y destruir las relaciones existentes entre ellas. Para ilustrar este problema, supongamos que en una muestra multivariante de tamaño n se introduce un valor atípico, $\mathbf{x}_a$, donde $\mathbf{x}_a$ es un vector de falsas observaciones. Llamando $\overline{\mathbf{x}}$ y $\mathbf{S}$ al vector de medias y matriz de covarianzas sin la observación $\mathbf{x}_a$, y $\overline{\mathbf{x}}_c$ y $\mathbf{S}_c$ a los de la muestra contaminada con este dato atípico, es fácil comprobar (veáse ejercicio 4.4) que

$$\overline{\mathbf{x}}_c = \overline{\mathbf{x}} + \frac{(\mathbf{x_a} - \overline{\mathbf{x}})}{n+1} \tag{4.12}$$

y

$$\mathbf{S}_c = \frac{n}{n+1}\mathbf{S} + \frac{(\mathbf{x_a} - \overline{\mathbf{x}})(\mathbf{x_a} - \overline{\mathbf{x}})'}{n+1}\left(\frac{n}{(n+1)}\right). \tag{4.13}$$

Estas fórmulas indican que un solo dato atípico puede afectar mucho al vector de medias y a todas las varianzas y covarianzas entre las variables. El efecto del atípico depende de su tamaño, medido por su distancia euclídea al centro del resto de las observaciones, pero también de su posición, ya que los términos más afectados de la matriz $\mathbf{S}$ dependen de la posición del atípico en el espacio. En general, si el tamaño del atípico es grande, lo que supone $|\mathbf{x_a} - \overline{\mathbf{x}}|$ grande, la media, varianzas y covarianzas de las variables pueden estar muy distorsionadas.

Para analizar con más detalle la distorsión de los coeficientes de correlación, consideremos el caso más simple de $p = 2$ y supongamos que $\overline{\mathbf{x}} = 0$, $\mathbf{S} = \mathbf{I}$, y n no muy pequeño de manera que, para simplificar la presentación, tomaremos $n \backsimeq n+1$. Sea $\mathbf{x}_a = (a_1, a_2)'$ y supongamos para simplificar que $\overline{\mathbf{x}}_c \backsimeq \overline{\mathbf{x}} = 0$. Llamando s_{ij}^c a los elementos de $\mathbf{S}_c$ y tomando $n \backsimeq n+1$, en (4.13) tendremos que

$$s_{ii}^c \backsimeq 1 + \frac{a_i^2}{n}, \quad i = 1, 2 \tag{4.14}$$

y

$$s_{ij}^c \backsimeq \frac{a_i a_j}{n}, \quad i \neq j$$

con lo que el coeficiente de correlación entre las dos variables será:

$$r_c \backsimeq \frac{a_1 a_2}{(n + a_1^2)^{1/2}(n + a_2^2)^{1/2}}.$$

Esta expresión muestra que si a_1 y a_2 tienen el mismo signo y son grandes con relación a $\sqrt{n}$ el coeficiente tiende a uno, mientras que si tienen signos opuestos, el coeficiente tiende hacia

menos uno. Vemos que la distorsión que produce el atípico depende no sólo de su tamaño sino también de su posición en el espacio.

La conclusión de este ejercicio es que una sola observación puede distorsionar arbitrariamente los coeficientes de correlación entre las variables. En la figura 4.14 hemos añadido a dos variables incorreladas una observación atípica, marcada por a, con $\mathbf{a} = (9, 9)'$. Como indica la teoría que hemos visto, esta única observación introduce una alta correlación entre las variables, creando una relación inexistente.

La figura 4.15 ilustra cómo una única observación puede ocultar una relación existente: la observación atípica **a** destruye la fuerte correlación existente entre las variables.

Cuando existe más de un atípico en los datos, puede producirse el efecto conocido como *enmascaramiento,* que consiste en que observaciones atípicas similares se ocultan entre sí. Por ejemplo, supongamos que en la figura 4.13 en la posición del atípico hay tres puntos idénticos. Aunque eliminemos el primero, los otros dos continuarán distorsionando el cálculo de las medias y varianzas, haciendo muy difícil su identificación, ya que cada punto enmascara a los otros.

4.5.3 (*)Identificación de grupos de atípicos

Hay dos filosofías para tratar con la heterogeneidad. La primera es utilizar estimadores robustos, que son estimadores diseñados para verse poco afectados por cierta contaminación de atípicos. Comentaremos estos estimadores en el capítulo 11. La segunda es detectar los atípicos, y aplicar el cálculo de los estimadores a las muestras limpias de atípicos. Ambos enfoques son complementarios, y en esta sección introduciremos el segundo.

El procedimiento para detectar grupos de atípicos es eliminar de la muestra todos los puntos sospechosos, de manera que evitemos el enmascaramiento y podamos calcular el vector de medias y la matriz de covarianzas sin distorsiones. A continuación identificaremos con estos estimadores la distancia de cada punto sospechoso respecto al centro de los datos, y consideraremos atípicos a los muy alejados. El primer paso para identificar las observaciones sospechosas es detectar aquellas que lo sean claramente respecto a una variable. Para ello podemos utilizar el histograma o los diagramas de caja, como hemos visto en los ejemplos anteriores. Una regla simple y automática es considerar sospechosas aquellas observaciones tales que

$$\frac{|x_i - med(x)|}{Meda(x)} > 4,5,$$

donde $med(x)$ es la mediana de las observaciones, que es un estimador robusto del centro de los datos, y $Meda(x)$ es la mediana de las desviaciones absolutas $|x_i - med(x)|$, que es una medida robusta de la dispersión. Este método puede verse como una estandarización robusta de los datos.

Esta detección univariante no identificará muchos atípicos multivariantes. Por ejemplo, el punto (-1,1) marcado con + en el gráfico 4.13 es claramente atípico, pero no aparecerá como tal en los análisis univariantes. Con frecuencia los atípicos multivariantes corresponden a situaciones con efectos pequeños sobre todas las variables, como un error sistemático de observación en todas ellas, en lugar de un efecto importante sobre una variable. Si el número

de variables no es muy grande, los diagramas de dispersión pueden ayudar visualmente a determinar datos atípicos en dos dimensiones. Para dimensiones mayores no es recomendable utilizar la distancia de Mahalanobis, ya que si existen grupos de atípicos, pueden distorsionar la estimación del centro y la dispersión de los datos enmascarando los atípicos y quizás señalando como atípicos a puntos que no lo son.

Para evitar este problema podemos buscar proyecciones de los datos que muestren las observaciones atípicas. Observemos que cualquier observación atípica multivariante debe aparecer como atípica al menos en una dirección de proyección: la definida por la recta que une el centro de los datos con el dato atípico. En base a esta idea, Stahel (1981) y Donoho (1982) propusieron generar muchas direcciones al azar, proyectar los puntos sobre estas direcciones y marcar como datos atípicos a aquellas observaciones que aparecen como extremas en estas proyecciones. Para generar direcciones al azar pueden tomarse muestras al azar de p puntos, calcular el plano que las contiene y tomar como dirección el vector ortogonal al plano.

Este método funciona bien con pocas variables, pero al aumentar la dimensión del problema el número de direcciones que necesitamos generar para cubrir razonablemente el espacio y tener garantías de éxito aumenta exponencialmente. Una solución propuesta por Peña y Prieto (2001), es proyectar los datos sobre ciertas direcciones específicas, escogidas de manera que tengan alta probabilidad de mostrar los atipicos cuando existan. Hemos comentado que en muestras univariantes una pequeña proporción de atípicos hace aumentar el coeficiente de kurtosis, lo que sugiere investigar las direcciones donde los puntos proyectados tengan máxima kurtosis univariante. Por otro lado, un grupo grande de atípicos puede producir bimodalidad y baja kurtosis, por lo que conviene también explorar las direcciones donde los puntos proyectados tengan mínima kurtosis. La idea del procedimiento es buscar p direcciones ortogonales de máxima kurtosis y p direcciones ortogonales de mínima kurtosis, eliminar provisionalmente los datos extremos en estas direcciones, calcular la media y la matriz de covarianzas con los datos no sospechosos y después identificar los datos atípicos como aquellos que son extremos con la distancia de Mahalanobis calculada con las estimaciones no contaminadas. Dada la muestra multivariante $(\mathbf{x}_1, ..., \mathbf{x}_n)$, el proceso se realiza como sigue:

1. Sean $\overline{\mathbf{x}}$ y $\mathbf{S}$ el vector de medias y la matriz de covarianzas de los datos. Estandarizar los datos de forma multivariante y sean $\mathbf{z}_i = \mathbf{S}_x^{-1/2}(\mathbf{x}_i - \overline{\mathbf{x}})$ los datos estandarizados con media cero y matriz de covarianzas identidad. Tomar $j = 1$ y $\mathbf{z}_i^{(1)} = \mathbf{z}_i$.

2. Calcular la dirección $\mathbf{d}_j$ con norma unidad que maximiza el coeficiente de kurtosis univariante de los datos proyectados. Llamando $y_i^{(j)} = \mathbf{d}_j' \mathbf{z}_i^{(j)}$, a las datos proyectado sobre la dirección $\mathbf{d}_j$, esta dirección se obtiene como solución de:

$$\max \sum (y_i^{(j)} - \overline{y}^{(j)})^4 + \lambda(\mathbf{d}'\mathbf{d} - 1)$$

   que puede resolverse como se indica en el apéndice 4.1.

3. Proyectar los datos sobre un espacio de dimensión $p - j$ definido como el espacio ortogonal a la dirección $\mathbf{d}_j$. Para ello tomar $\mathbf{z}^{(j+1)} = (\mathbf{I} - \mathbf{d}_j \mathbf{d}_j')\mathbf{z}^{(j)}$. Hacer $j = j + 1$.

4. Repetir (2) y (3) hasta obtener las p direcciones, $\mathbf{d}_1, ..., \mathbf{d}_p$.

5. Repetir (2) y (3) pero ahora minimizando la kurtosis en lugar de maximizarla para obtener otras p direcciones, $\mathbf{d}_{p+1}, ..., \mathbf{d}_{2p}$

6. Considerar como sospechosos aquellos puntos que en alguna de estas $2p$ direcciones están claramente alejados del resto, es decir, verifican

$$\frac{\left|y_i^{(j)} - med(y^{(j)})\right|}{Meda(y^{(j)})} > 5$$

A continuación se eliminan todos los valores sospechosos detectados y se vuelve a 2 para analizar los datos restantes. La estandarización multivariante ahora se realizará con la nueva media y matriz de covarianzas de los datos restantes. Los pasos 2 a 6 se repiten hasta que no se detecten más datos atípicos o se haya eliminado una proporción de datos prefijada, por ejemplo un máximo del 40% de los datos.

Una vez que la muestra no contenga más valores sospechosos con el criterio anterior se calcula el vector de medias, $\overline{\mathbf{x}}_R$, y la matriz de covarianzas, $\mathbf{S_R}$, de los datos no sospechosos, y las distancias de Mahalanobis para los sospechosos como:

$$d_R^2(\mathbf{x}_i,\overline{\mathbf{x}}_R) = (\mathbf{x_i}-\overline{\mathbf{x}}_R)\mathbf{S_R^{-1}}(\mathbf{x_i}-\overline{\mathbf{x}}_R)'$$

Por razones que veremos más adelante al estudiar contrastes de valores atípicos en el capítulo 10, aquellos valores mayores que $p + 3\sqrt{2p}$ se consideran atípicos (recordemos que el valor promedio de la distancia de Mahalanobis es p). Algunos puntos del conjunto de sospechosos serán atípicos y otros no. Los atípicos son desechados, y los buenos incoporados al conjunto de puntos. Finalmente, se calculará un vector de medias, $\overline{\mathbf{x}}_f$, y una matriz de covarianzas, $\mathbf{S_f}$, con los puntos no atípicos, que serán las estimaciones finales obtenidas de los datos.

En el capítulo 11 presentaremos métodos formales para contrastar si unos datos son atípicos respecto a un modelo. En el apéndice 4.1 se detalla el cálculo de las direcciones que maximizan la kurtosis. El procedimiento converge rápidamente en general. Un programa de ordenador en Matlab para ejecutar este algoritmo puede bajarse de la dirección http:/******

Los datos detectados como potencialmente atípicos deben ser estudiadas con detalle para determinar las causas de la heterogeneidad. Si estos datos no tienen un error detectable, conviene, cuando sea posible, investigar las causas de su tamaño anómalo ya que puede llevar a importantes descubrimientos. Si no hay un error en el dato y, sin embargo, es muy distinto de los demás, hay que sospechar que sobre esa observación ha actuado alguna causa que no ha estado activa en el resto de las observaciones. Por ejemplo, una variable no incluída en el estudio ha tomado un valor distinto en esa observación y es responsable del cambio observado. El descubrimiento de esta variable insospechada puede ser el resultado más importante del estudio estadístico. Muchos descubrimientos científicos importantes, (por ejemplo la penicilina) y muchas patentes industriales, han surgido de la investigación para determinar las razones de un dato anómalo.

Ejemplo 4.10 *Buscaremos datos atípicos en los datos de la EPF. En primer lugar calculamos las distancias de Mahalanobis de cada dato al centro de todos ellos. Estas distancias se presentan en el histograma de la figura 4.16. Las provincias más alejadas del centro de*

los datos son, por este orden, Madrid (D=4.29), Gerona (D=3.98) y Navarra (3.97). Si aplicamos ahora el procedimiento de buscar direcciones extremas en la kurtosis, obtenemos los gráficos de las figuras 4.17, 4.18, 4.19 y 4.20. Si eliminamos estos nueve posibles datos extremos y calculamos las distancias de Mahalanobis con las medias y covarianzas calculadas sin estos datos obtenemos el histograma de la figura 4.21 . Las dos observaciones claramente extremas corresponden a Madrid y Barcelona. Observemos que en el análisis inicial Barcelona quedaba oculta (enmascarada) por la presencia de Madrid, pero aparece claramente como atípica cuando se elimina el efecto de Madrid.

Ejemplo 4.11 *Vamos analizar los datos de los sectores industriales en Europa. EUROSEC. Como las variables suman 100 aproximadamente eliminaremos la última y trabajaremos por tanto con ocho variables. La figura 4.22 presenta el histograma de estas distancias. Hay tres países muy alejados del resto que son Yugoeslavia ($D = 4.17$), Luxemburgo ($D = 4.16$) y Turquía ($D = 4.02$). Estos tres países están separados del resto y son atípicos en su estructura de empleo.*

Para entender la razón dividiremos los valores de cada uno de estos tres países por la media. La tabla siguiente presenta los valores medios, el país más próximo en la distancia de Mahalanobis a esta estructura media (Francia, $D = 1.4$) y los cocientes entre los valores del país más extremo y los valores medios.

Media	*19.1*	*1.25*	*27.00*	*0.91*	*8.16*	*12.95*	*4.00*	*20.02*
Francia /Med	*0.56*	*0.64*	*1.01*	*.99*	*1.09*	*1.30*	*1.50*	*1.13*
Yugoes/Med	*2.54*	*1.19*	*0.62*	*1.21*	*0.60*	*0.49*	*2.82*	*0.26*

En esta tabla aparece claramente el carácter atípico de Yugoslavia: tiene más del doble de población empleada en Agricultura y finanzas que el país medio y la mitad de empleo en los servicios.

Vamos a comparar este resultado con el que se obtendría buscando posibles grupos de atípicos. Las figuras 4.23, 4.24, y **??** *presentan las proyecciones sobre la dirección que maximiza la kurtosis de los datos y dos direcciones ortogonales a ella.*

En la primera dirección aparece como extremo el punto 7 (Luxemburgo), en la segunda el 26 (Yugoeslavia) y en la tercera el 15 (España) y el 18 (Turquía). Es interesante que si eliminamos estos cuatro puntos y calculamos las distancias de Mahalanobis del resto a estos cuatro países, España aparece más alejada que Luxemburgo.

4.6 Lecturas complementarias

El libro de Gnanadesikan (1997) amplía el material de este capítulo incluyendo otros métodos gráficos para los datos, como las curvas de Andrews, donde cada observación se representa for una función $f(t)$. Este libro también considera con detalle la transformación Box- Cox multivariante, que ha sido estudiada, entre otros por Velilla (1993,1995) y Atkinson (19). La detección de atípicos multivariantes ha sido objeto de numeros trabajos. Algunas referencias recientes son Rousseeuw y van Zomeren (1990), Atkinson (1994), Maronna y Yohai (1995), , Rocke y Woodruff (1996) y Juan y Prieto (2001). Volveremos sobre este tema al presentar los estimadores robustos en el capítulo 11.

EJERCICIOS

4.1 Construir los diagramas de dispersión con un programa de ordenador como Matlab, Minitab o Spss para los datos de EUROSEC.

4.2 Demostrar que un cambio de medida de las variables que equivale a una transformación lineal no modifica su matriz de correlación.

4.3 Demostrar que la estandarización univariante no modifica la matriz de correlación de las variables.

4.3 Demostrar que la estandarización multivariante hace cero los coeficientes de correlación parcial entre las nuevas variables.

4.4 Demostrar que si introducimos un dato atípico en una muestra con vector de medias

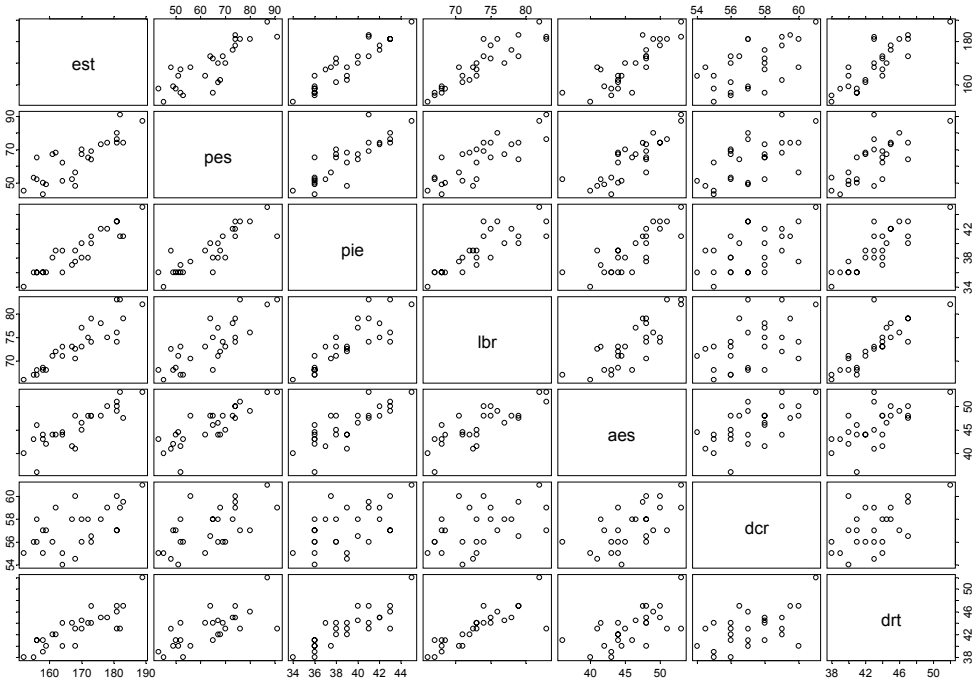


Figura 4.6: Matriz de dispersión para los datos de la medidas físicas (MEDIFIS)

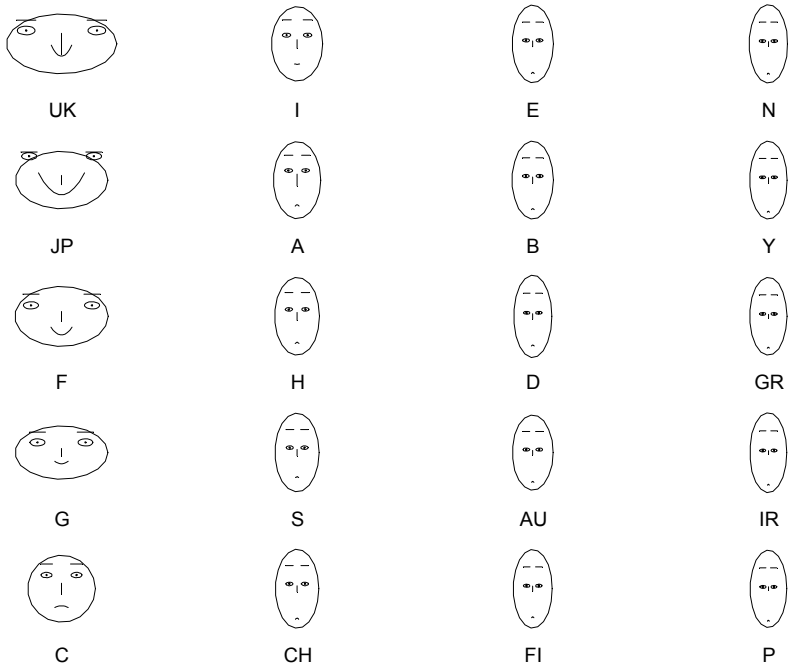


Figura 4.7: Representación de las contribuciones científicas de los países de INVEST en caras de Chernoff

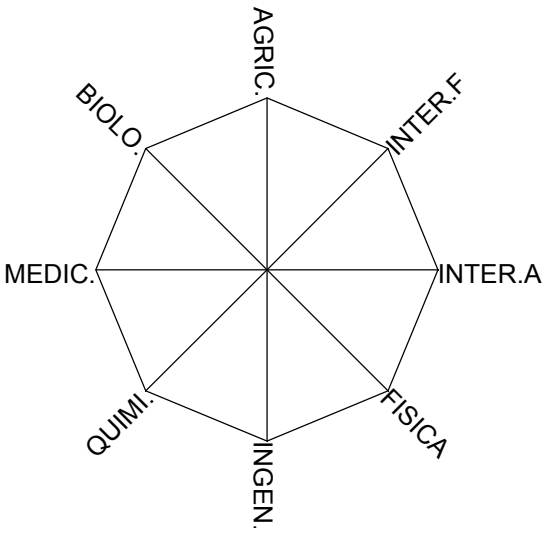


Figura 4.8: Esquema de asignación de los radios de la estrella a la variables para los datos de la investigación de los países de la OCEDE

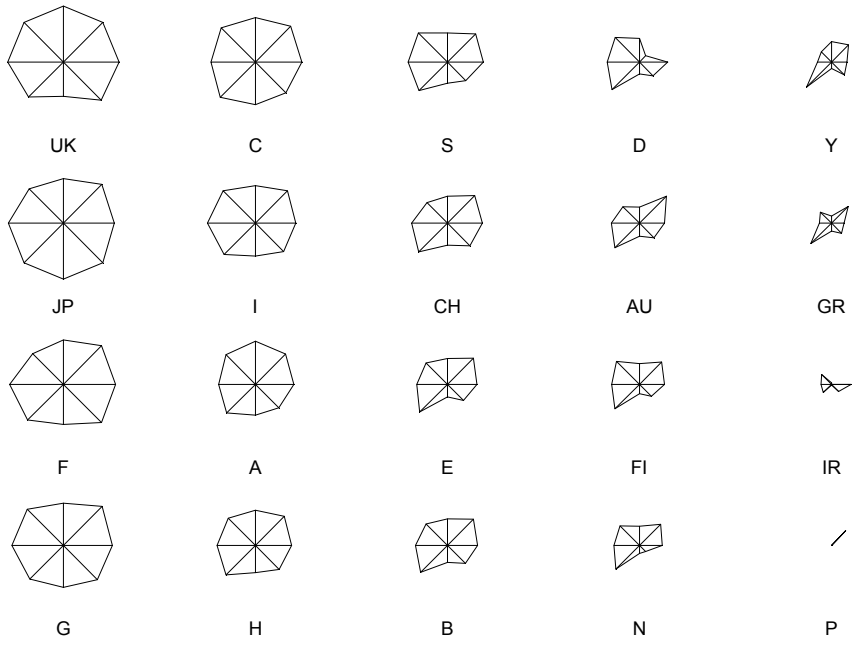


Figura 4.9: Representación mediante estrellas de los países de INVEST en logaritmos una vez eliminado EEUU.

Figura 4.11: Distribuciones conjuntas y marginales de las variables de INVES en logaritmos.

Figura 4.12: Los datos de EPF en logaritmos. Representaciones bivariantes e histogramas.

Figura 4.13: Una observación atípica multivariante que no aparece como tal en los análisis univariantes.

Figura 4.14: En esta figura las variables estaban originalmente casi incorreladas ($r = -.11$), pero la presencia del valor atípico ha creado una fuerte correlación positiva ($r = .71$).

Figura 4.15: En esta figura el coeficiente de correlación sin el dato atipico es de 0,91 y disminuye hasta 0,41 por la presencia del atípico marcado con a.

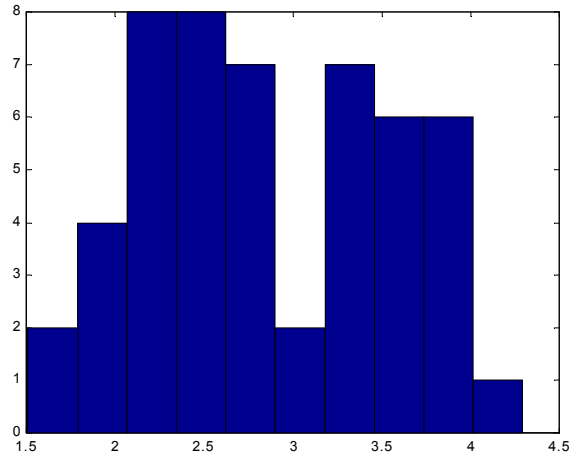


Figura 4.16: Distribución de las distancias de Mahalanobis entre cada dato y el centro para los datos de la EPF

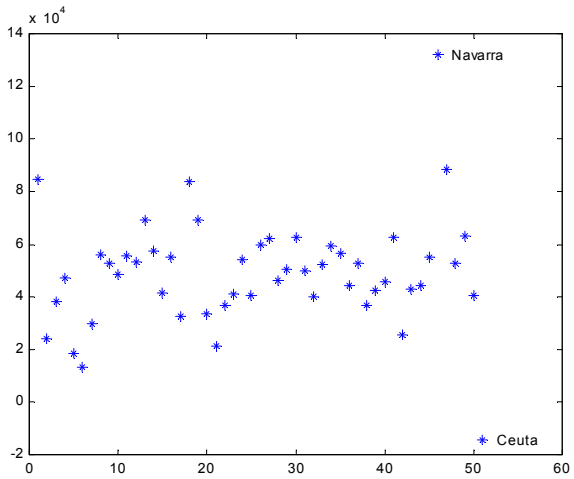


Figura 4.17: Primera proyección en la dirección de máxima kurtosis para los datos de la EPF

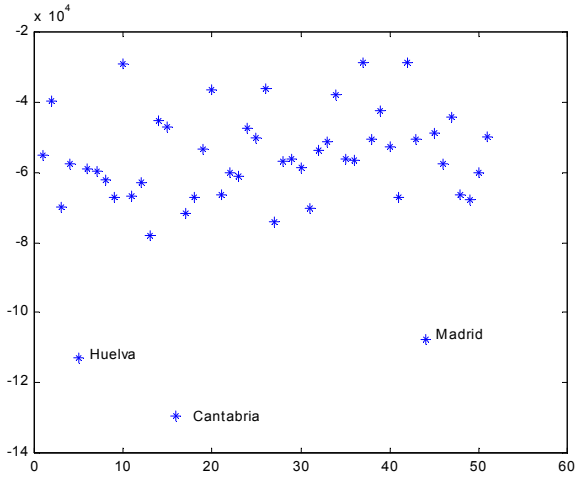


Figura 4.18: Segunda proyección en la dirección de máxima curtosis para los datos de la EPF

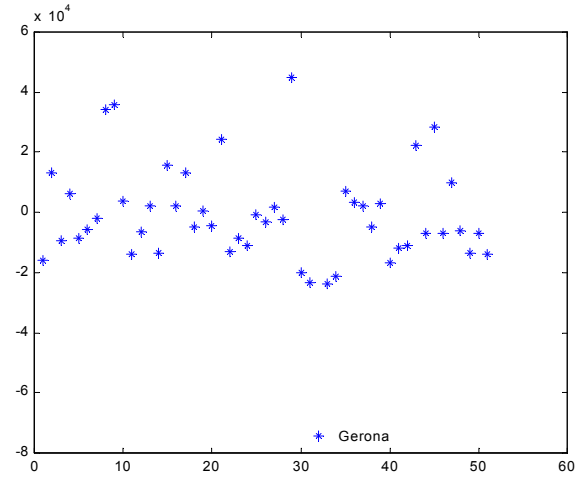


Figura 4.19: Tercera proyección en la dirección de máxima curtosis para los datos de la EPF

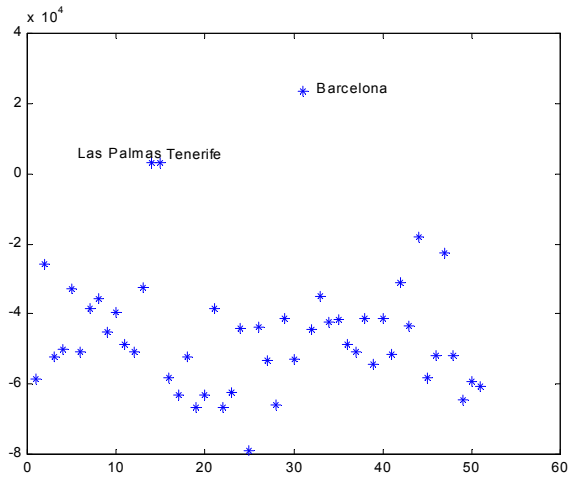


Figura 4.20: Cuarta proyección sobre la dirección de máxima curtosis para los datos de la EPF

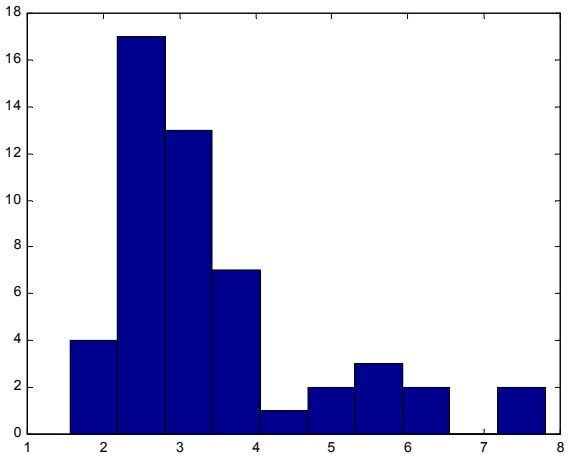


Figura 4.21: Distancias de Mahalanobis para los datos de la EPF calculadas de manera robusta, eliminando los datos extremos.

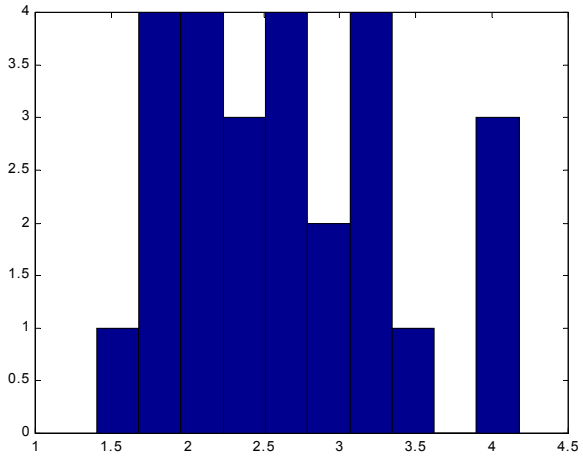


Figura 4.22: Distancias de Mahalanobis de cada dato al centro de la muestra para los datos de EUROSEC.

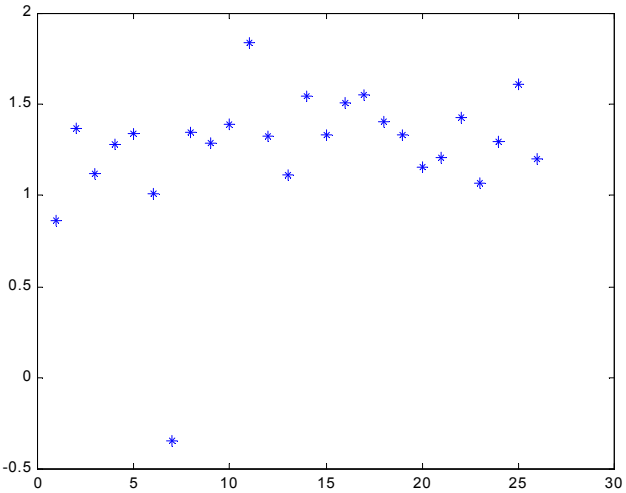


Figura 4.23: Proyección sobre la dirección de máxima curtosis

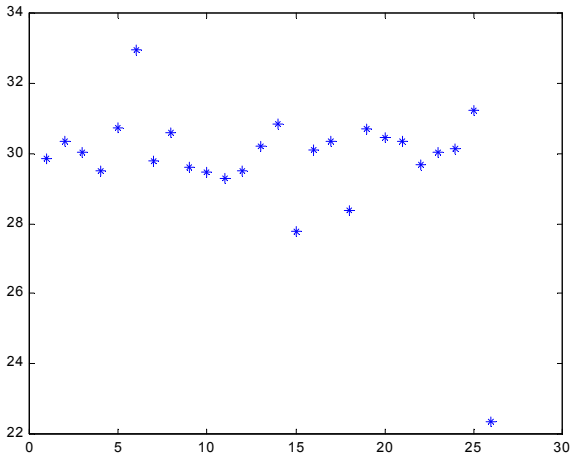


Figura 4.24: Proyección sobre la segunda dirección de máxima curtosis ortogonal a la primera

Capítulo 5

COMPONENTES PRINCIPALES

5.1 INTRODUCCIÓN

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad: si es posible describir con precisión los valores de p variables por un pequeño subconjunto $r < p$ de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información.

El análisis de componentes principales tiene este objetivo: dadas n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales. Por ejemplo, con variables con alta dependencia es frecuente que un pequeño número de nuevas variables (menos del 20% de las originales) expliquen la mayor parte (más del 80%) de la variabilidad original.

La técnica de componentes principales es debida a Hotelling (1933), aunque sus orígenes se encuentran en los ajustes ortogonales por mínimos cuadrados introducidos por K. Pearson (1901). Su utilidad es doble:

1. Permite representar óptimamente en un espacio de dimensión pequeña, observaciones de un espacio general p-dimensional. En este sentido componentes principales es el primer paso para identificar posibles variables "latentes" o no observadas, que están generando la variabilidad de los datos.

2. Permite transformar las variables originales, en general correladas, en nuevas variables incorreladas, facilitando la interpretación de los datos.

En este capítulo presentamos únicamente esta técnica como una herramienta exploratoria para facilitar la descripción e interpretación de los datos. El problema de inferir si las propiedades de reducción de la dimensión encontradas en los datos puede extenderse a una población se estudiara en el capítulo de análisis factorial.

5.2 PLANTEAMIENTO DEL PROBLEMA

Supongamos que se dispone de los valores de p-variables en n elementos de una población dispuestos en una matriz $\mathbf{X}$ de dimensiones $n \times p$, donde las columnas contienen las variables y las filas los elementos. Supondremos en este capítulo que previamente hemos restado a cada variable su media, de manera que las variables de la matriz $\mathbf{X}$ tienen media cero y su matriz de covarianzas vendrá dada por $1/n\ \mathbf{X}'\mathbf{X}$.

El problema que se desea resolver es cómo encontrar un espacio de dimensión más reducida que represente adecuadamente los datos. El problema puede abordarse desde tres perspectivas equivalentes.

a) Enfoque descriptivo

Se desea encontrar un subespacio de dimensión menor que p tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible. Veamos cómo convertir esta noción intuitiva en un criterio matemático operativo. Consideremos primero un subespacio de dimensión uno, una recta. Se desea que las proyecciones de los puntos sobre esta recta mantengan, lo más posible, sus posiciones relativas. Para concretar, consideremos el caso de dos dimensiones ($p = 2$). La figura 5.1 indica el diagrama de dispersión y una recta que, intuitivamente, proporciona un buen resumen de los datos, ya que las proyecciones de los puntos sobre ella indican aproximadamente la situación de los puntos en el plano. La representación es buena porque la recta pasa cerca de todos los puntos y estos se deforman poco al proyectarlos. Esta propiedad puede concretarse exigiendo que las distancias entre los puntos originales y sus proyecciones sobre la recta sean lo más pequeñas posibles. En consecuencia, si consideramos un punto $\mathbf{x}_i$ y una dirección $\mathbf{a}_1 = (a_{11}, ..., a_{1p})'$, definida por un vector $\mathbf{a}_1$ de norma unidad, la proyección del punto $\mathbf{x}_i$ sobre esta dirección es el escalar:

$$z_i = a_{11}x_{i1} + \ldots + a_{1p}x_{ip} = \mathbf{a}_1'\mathbf{x}_i \tag{5.1}$$

y el vector que representa esta proyección será $z_i\mathbf{a}_1$. Llamando r_i a la distancia entre el punto $\mathbf{x}_i$, y su proyección sobre la dirección $\mathbf{a}_1$, este criterio implica:

$$minimizar\ \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} |\mathbf{x}_i - z_i\mathbf{a}_1|^2\,, \tag{5.2}$$

donde $|\mathbf{u}|$ es la norma euclídea o módulo del vector $\mathbf{u}$.

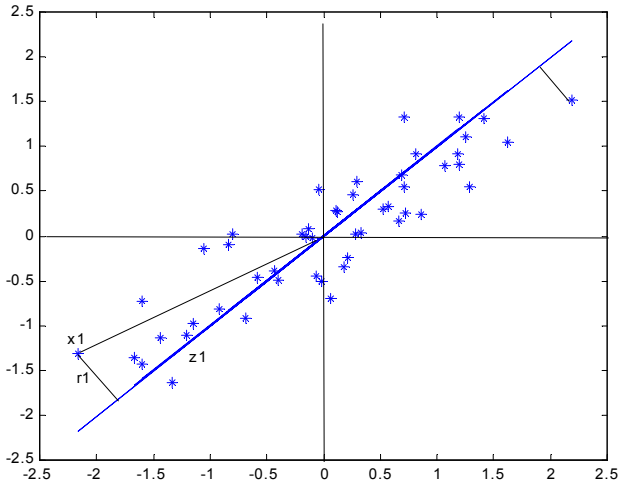


Figura 5.1: Ejemplo de la recta que minimiza las distancias ortogonales de los puntos a ella.

La figura (5.1) muestra que al proyectar cada punto sobre la recta se forma un triángulo rectángulo donde la hipotenusa es la distancia al origen del punto al origen, $(\mathbf{x}_i'\mathbf{x}_i)^{1/2}$, y los catetos la proyeccion del punto sobre la recta (z_i) y la distancia entre el punto y su proyección (r_i). Por el teorema de Pitágoras, podemos escribir:

$$\mathbf{x}_i'\mathbf{x}_i = {z_i}^2 + r_i^2, \tag{5.3}$$

y sumando esta expresión para todos los puntos, se obtiene:

$$\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2. \tag{5.4}$$

Como el primer miembro es constante, minimizar $\sum_{i=1}^{n} r_i^2$, la suma de las distancias a la recta de todos los puntos, es equivalente a maximizar $\sum_{i=1}^{n} z_i^2$, la suma al cuadrado de los valores de las proyecciones. Como las proyecciones z_i son, por (9.21) variables de media cero, maximizar la suma de sus cuadrados equivale a mazimizar su varianza. Este resultado es intuitivo: la recta de la figura 5.1 parece adecuada porque conserva lo más posible la variabilidad original de los puntos. El lector puede convencerse considerando una dirección de proyección perpendicular a la de la recta en esta figura: los puntos tendrían muy poca variabilidad y perderiamos la información sobre sus distancias en el espacio.

El objetivo de proyectar los puntos con mínima deformación puede abordarse desde otro punto de vista que conduce al mismo resultado final. En el espacio de p-dimensiones, lo característico de la nube de puntos son sus distancias relativas. Tratemos de encontrar un subespacio de dimensión 1, es decir, un recta tal que los puntos proyectados conserven lo más posible sus distancias relativas. Si llamamos $d_{ij}^2 = \mathbf{x}_i'\mathbf{x}_j$ a los cuadrados de las distancias originales entre los puntos y $\widehat{d_{ij}^2} = (z_i - z_j)^2$ a las distancias entre los puntos proyectados

sobre una recta, deseamos que

$$D = \sum_i \sum_j (d_{ij}^2 - \widehat{d}_{ij}^2)$$

sea mínima. Como la suma de las distancias originales es fija, minimizar D require maximizar $\sum_i \sum_j \widehat{d}_{ij}^2$, las distancias entre los puntos proyectados. Se demuestra en el apéndice 5.1 que la dirección es la misma que proporciona una variable escalar de varianza máxima.

b) Enfoque estadístico:

Representar puntos p dimensionales con la mínima pérdida de información en un espacio de dimensión uno es equivalente a sustituir las p variables originales por una nueva variable, z_1, que resuma óptimamente la información. Esto supone que la nueva variable debe tener globalmente máxima correlación con las originales o, en otros términos, debe permitir prever las variables originales con la máxima precisión. Esto no será posible si la nueva variable toma un valor semejante en todos los elementos, y, se demuestra en el apéndice 5.2, que la condición para que podamos prever con la mínima pérdida de información los datos observados, es utilizar la variable de máxima variabilidad.

Volviendo a la figura 5.1 se observa que la variable escalar obtenida al proyectar los puntos sobre la recta sirve para prever bien el conjunto de los datos. La recta indicada en la figura no es la línea de regresión de ninguna de las variables con respecto a la otra, que se obtienen minimizando las distancias verticales u horizontales, sino que al minimizar las distancias ortogonales o de proyección se encuentra entre ambas rectas de regresión.

Este enfoque puede extenderse para obtener el mejor subespacio resumen de los datos de dimensión 2. Para ello calcularemos el plano que mejor aproxima a los puntos. El problema se reduce a encontrar una nueva dirección definida por un vector unitario, $\mathbf{a_2}$, que, sin pérdida de generalidad, puede tomarse ortogonal a $\mathbf{a_1}$, y que verifique la condición de que la proyección de un punto sobre este eje maximice las distancias entre los puntos proyectados. Estadísticamente esto equivale a encontrar una segunda variable z_2, incorrelada con la anterior, y que tenga varianza máxima. En general, la componente $z_r (r < p)$ tendrá varianza máxima entre todas las combinaciones lineales de las p variables X originales, con la condición de estar incorrelada con las $z_1, ..., z_{r-1}$ previamente obtenidas.

c) enfoque geométrico

El problema puede abordarse desde un punto de vista geométrico con el mismo resultado final. Si consideramos la nube de puntos de la figura 5.1 vemos que los puntos se sitúan siguiendo una elipse y podemos describir su orientación dando la dirección del eje mayor de la elipse y la posición de los punto por su proyección sobre esta dirección. Puede demostrarse que este eje es la recta que minimiza las distancias ortogonales y volvemos al problema que ya hemos resuelto. En varias dimensiones tendremos elipsoides y la mejor aproximación a los datos es la proporcionada por el eje mayor del elipsoide. Considerar los ejes del elipsoide como nuevas variables originales supone pasar de variables correladas a variables ortogonales, como veremos a continuación.

5.3 CALCULO DE LOS COMPONENTES

5.3.1 Cálculo del primer componente

El primer componente principal será la combinación lineal de las variables originales que tenga varianza máxima. Los valores de este primer componente en los n individuos se representarán por un vector $\mathbf{z}_1$, dado por

$$\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1.$$

Como las variables originales tienen media cero también $\mathbf{z}_1$ tendrá media nula. Su varianza será:

$$Var(\mathbf{z}_1) = \frac{1}{n}\mathbf{z}_1'\mathbf{z}_1 = \frac{1}{n}\mathbf{a}_1'\mathbf{X}'\mathbf{X}\mathbf{a}_1 = \mathbf{a_1'}\mathbf{S}\mathbf{a_1} \tag{5.5}$$

donde $\mathbf{S}$ es la matriz de varianzas y covarianzas de las observaciones. Es obvio que podemos maximizar la varianza sin limite aumentando el módulo del vector $\mathbf{a}_1$. Para que la maximización de (5.5) tenga solución debemos imponer una restricción al módulo del vector $\mathbf{a}_1$, y, sin pérdida de generalidad, impondremos que $\mathbf{a}_1'\mathbf{a}_1 = 1$. Introduciremos esta restricción mediante el multiplicador de Lagrange:

$$M = \mathbf{a_1'}\mathbf{S}\mathbf{a}_1 - \lambda(\mathbf{a_1'}\mathbf{a_1} - 1)$$

y maximizaremos esta expresión de la forma habitual derivando respecto a los componentes de $\mathbf{a}_1$ e igualando a cero. Entonces

$$\frac{\partial M}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = 0$$

cuya solución es:

$$\mathbf{S}\mathbf{a_1} = \lambda\mathbf{a}_1, \tag{5.6}$$

que implica que $\mathbf{a_1}$ es un vector propio de la matriz $\mathbf{S}$, y λ su correspondiente valor propio. Para determinar qué valor propio de $\mathbf{S}$ es la solución de la ecuación (5.6) tendremos en cuenta que, multiplicando por la izquierda por $\mathbf{a}_1'$ esta ecuación,

$$\mathbf{a_1'}\mathbf{S}\mathbf{a_1} = \lambda\mathbf{a_1'}\mathbf{a_1} = \lambda$$

y concluimos, por (5.5), que λ es la varianza de $\mathbf{z_1}$. Como esta es la cantidad que queremos maximizar, λ será el mayor valor propio de la matriz $\mathbf{S}$. Su vector asociado, $\mathbf{a_1}$, define los coeficientes de cada variable en el primer componente principal.

Ejemplo 5.1 *Ilustraremos con detalle el cálculo de la primera componente principal con los datos de los logaritmos de las ACCIONES, tabla A.7. Los paquetes estadísticos habituales (Minitab, SPSS, Statgraphics, etc) proporcionan directamente los componentes principales, pero vamos a indicar con detalle como se realizan los cálculos para el lector interesado.*

La matriz de varianzas y covarianzas de estos datos en logaritmos, que ya utilizamos en el ejemplo 3.5, es,

$$S = \begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix}$$

Para el cálculo de los autovalores tenemos que calcular las raíces de la ecuación:

$$\begin{aligned} 0 &= |\mathbf{S} - \boldsymbol{\lambda}\mathbf{I}| = \\ &= \left| \begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} - \begin{bmatrix} \boldsymbol{\lambda} & 0 & 0 \\ 0 & \boldsymbol{\lambda} & 0 \\ 0 & 0 & \boldsymbol{\lambda} \end{bmatrix} \right| = \\ &= 0,000382 - 0,0628\lambda + 0,64\lambda^2 - \lambda^3 \end{aligned}$$

Las raíces del polinomio, obtenidas con MATLAB son $\lambda_1 = 0.521$, $\lambda_2 = 0.113$, $\lambda_3 = 6.51 \times 10^{-3}$. *El autovector asociado a* λ_1 *nos da los pesos de la primera componente principal. Para calcular el primer autovector resolvemos el sistema*

$$\mathbf{S}\mathbf{a_1} = \lambda_1 \mathbf{a}_1$$

que conduce a:

$$\begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \end{bmatrix} = 0.521 \times \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \end{bmatrix}$$

$$\begin{bmatrix} -0.171a_{11} + 0.15a_{12} - 0.19a_{13} \\ 0.15a_{11} - 0.391a_{12} - 0.03a_{13} \\ -0.19a_{11} - 0.03a_{12} - 0.361a_{13} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

el sistema es compatible indeterminado. Para encontrar una de las infinitas soluciones tomemos la primera variable como parámetro, x, y resolvamos el sistema en función de x. La solución es,

$$\{a_{11} = x,\ a_{12} = 0.427x,\ a_{13} = -0.562x\}$$

El valor de x se obtiene ahora imponiendo que el vector tenga norma unidad, con lo que resulta:

$$\mathbf{a}_1 = \begin{bmatrix} -0.817 \\ -0.349 \\ 0.459 \end{bmatrix}$$

y el primer componente es

$$Z_1 = -0.817X_1 - 0.349X_2 + 0.459X_3$$

donde X_1, X_2 y X_3 son las variables en logaritmos. Por ejemplo, el valor de esta nueva variable, la primera componente principal, para la primera observación (la primera acción) es

$$z_1 = -0.817 \times \log(3.4) - 0.349 \times \log(89.7) + 0.459 \times \log(30.2) = -1.0049$$

El primer componente principal puede aproximadamente escribirse

$$Z_1 \cong -0.82X_1 + 0.35(X_3 - X_2) + 0.11X_3$$

y utilizando la definición de las variables originales este componente puede escribirse

$$Z_1 \cong -0.82 \log(d/p) + 0.35 \log(p/d) + 0.11 \log(pN/b)$$

es decir,

$$Z_1 \cong -1.17 \log(d/p) + 0.11 \log(pN/b)$$

que indica que este primer componente depende basicamente de la variable X_1, la rentabilidad por dividendos. Llamando $z_1 = \log Z_1$ este primer componente puede escribirse también como

$$z_1 = \frac{p^{1.27}}{d^{1.16}}(\frac{N}{B})^{.09}$$

que es, aproximadamente, de nuevo la variable x_1, el cociente entre el precio de la acción y los dividendos recibidos. Esta variable es la que explica mejor la variabilidad conjunta de las acciones.

Ejemplo 5.2 *La encuesta de presupuestos familiares en España (Tabla A.3) presenta los gastos medios de las familias españolas en nueve epígrafes: X_1 = alimentación, X_2 = vestido y calzado, X_3 = vivienda, X_4 = mobiliario doméstico, X_5 = gastos sanitarios, X_6 = transportes, X_7 = enseñanza y cultura, X_8 = turismo y ocio, X_9 = otros gastos, para las 51 provincias españolas (Ceuta y Melilla aparecen unidas como una provincia). La matriz de covarianzas resume la variabilidad de estas 9 variables en los 51 elementos observados. Como las distribuciones de los gastos son muy asimétricas, las variables se han expresado en logaritmos. El vector propio asociado al mayor valor propio, 0,348, define la siguiente variable:*

$$\begin{aligned} z_1 \;=\; & 0,12x_1 + 0,18x_2 + 0,30x_3 + 0,31x_4 + 0,46x_5 + 0,34x_6 \\ & +0,50x_7 + 0,31x_8 + 0,31x_9 \end{aligned}$$

Se observa que z_1 es una suma ponderada de todos los gastos con mayor peso en los gastos en enseñanza y cultura (x_7) y gastos sanitarios (x_5). El menor peso lo tiene el gasto en alimentación (x_1).

Si calculamos las coordenadas z_1 para las provincias españolas y las ordenamos por esta nueva variable las provincias quedan prácticamente ordenadas por su renta. La primera componente principal tiene pues en este caso una explicación inmediata: redescubre la renta de cada provincia.

5.3.2 Cálculo del segundo componente

Vamos a obtener el mejor plano de proyección de las variables $\mathbf{X}$. Lo calcularemos estableciendo como función objetivo que la suma de las varianzas de $\mathbf{z_1}= \mathbf{Xa_1}$ y $\mathbf{z_2}= \mathbf{Xa_2}$ sea máxima, donde $\mathbf{a_1}$ y $\mathbf{a_2}$ son los vectores que definen el plano. La función objetivo será:

$$\phi = \mathbf{a}_1'\mathbf{S}\mathbf{a_1} + \mathbf{a_2'}\mathbf{S}\mathbf{a_2} - \lambda_1(\mathbf{a}_1'\mathbf{a}_1 - 1) - \lambda_2(\mathbf{a_2'}\mathbf{a}_2 - 1) \tag{5.7}$$

que incorpora las restricciones de que las direcciones deben de tener módulo unitario $(\mathbf{a}_i'\mathbf{a}_i) = 1,\ i = 1, 2$. Derivando e igualando a cero:

$$\frac{\partial \phi}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1 = 0$$

$$\frac{\partial \phi}{\partial \mathbf{a}_2} = 2\mathbf{S}\mathbf{a}_2 - 2\lambda_2\mathbf{a}_2 = 0$$

La solución de este sistema es:

$$\mathbf{S}\mathbf{a}_1 = \lambda_1\mathbf{a}_1, \tag{5.8}$$

$$\mathbf{S}\mathbf{a}_2 = \lambda_2\mathbf{a}_2 \tag{5.9}$$

que indica que $\mathbf{a_1}$ y $\mathbf{a_2}$ deben ser vectores propios de $\mathbf{S}$. Tomando los vectores propios de norma uno y sustituyendo en (5.7), se obtiene que, en el máximo, la función objetivo es

$$\phi = \lambda_1 + \lambda_2 \tag{5.10}$$

es claro que λ_1 y λ_2 deben ser los dos autovalores mayores de la matriz $\mathbf{S}$ y $\mathbf{a}_1$ y $\mathbf{a}_2$ sus correspondientes autovectores. Observemos que la covarianza entre $\mathbf{z}_1$ y $\mathbf{z}_2$, dada por $\mathbf{a}_1'\ \mathbf{S}\ \mathbf{a}_2$ es cero ya que $\mathbf{a}_1'\mathbf{a}_2 = 0$, y las variables $\mathbf{z}_1$ y $\mathbf{z}_2$ estarán incorreladas. Puede demostrarse (véase el ejercicio 5.7) que si en lugar de maximizar la suma de varianzas, que es la traza de la matriz de covarianzas de la proyección, se maximiza la varianza generalizada (el determinante de la matriz de covarianzas) se obtiene el mismo resultado.

Ejemplo 5.3 *El segundo componente principal para las variables de gastos de la EPF definidas en el ejemplo 5.1 es el asociado al segundo valor propio mayor que es 0,032. El vector propio asociado a este valor propio define la nueva variable:*

$$\begin{aligned} z_2 \ = \ & 0,05x_1 + 0,16x_2 - 0,17x_3 + 0,07x_4 - 0,21x_5 + 0,29x_6 - \\ & 0,40x_7 - 0,17x_8 + 0,78x_9 = \\ & (0,05x_1 + 0,16x_2 + 0,07x_4 + 0,29x_6 + 0,78x_9) - \\ & (0,17x_3 + 0,21x_5 + 0,40x_7 + 0,17x_8) \end{aligned}$$

Esta variable puede verse como la diferencia entre dos medias ponderadas de los gastos. La primera da sobre todo peso a otros gastos (x_9), y transporte (x_6). En la variable otros gastos

están incluidas las transferencias fuera de la provincia a miembros de la familia mayores de 14 años que no residan en ella, podemos conjeturar esta variable separa las provincias que reciben transferencias de las que las envian. Es también significativo que estas provincias tienen altos gastos en transporte. La primera media ponderada puede considerarse un indicador de como esta provincia envía recursos a otras. La segunda media da mayor peso a las variables enseñanza y cultura (x_7) y gastos sanitarios (x_5).

Este segundo componente va a separar a provincias que envian recursos a otras (alto valor de x_9) y que tienen también altos gastos de transporte, respecto a las que transfieren relativamente poco y tienen altos gastos de educación y sanidad. Las provincias con valores más altos de este componente son Zamora, León, Lugo, Toledo, Huesca, Lérida, Segovia, Soria y Palencia. Estas provincias no han tenido tradicionalmente universidad, por lo que tienen que enviar los estudiantes fuera y tienen bajos costes de educación. Por el contrario, las provincias con valores bajos de este componente z_2 incluyen a Madrid y Barcelona, centros receptores netos de estudiantes de otras provincias, así como a Salamanca, Zaragoza y Tenerife. La Tabla 5.1 presenta la ordenación de las provincias según el primer y segundo componente. La figura **??** *representa cada provincia en el plano de las dos primeras componentes principales. Cada punto aparece representado por sus coordenadas respecto a los ejes definidos por las componentes principales y puede interpretarse como la proyección de los puntos, que están en un espacio de dimensión 9, tantos como variables, sobre el plano que mejor mantiene sus distancias relativas, que es el definido por las dos primeras componentes.*

Proyección de los datos de la EPF sobre el plano definido por las dos primeras componentes principales

5.3.3 Generalización

Puede demostrarse análogamente que el espacio de dimensión r que mejor representa a los puntos viene definido por los vectores propios asociados a los r mayores autovalores de $\mathbf{S}$. Estas direcciones se denominan direcciones principales de los datos y a las nuevas variables por ellas definidas componentes principales. En general, la matriz $\mathbf{X}$ (y por tanto la $\mathbf{S}$) tiene

Comp. 1	Comp. 2
Navarra	Zamora
Madrid	León
Barcelona	Lugo
Lérida	Toledo
Vizcaya	Huesca
Gerona	Murcia
Baleares	Navarra
Tarragona	Lérida
Guipuzcoa	Segovia
Las Palmas	Soria
⋮	⋮
Ciudad Real	Málaga
Cuenca	Salamanca
Ávila	Cádiz
Teruel	Madrid
Castellón	Badajoz
Orense	Jaén
Zamora	Ceuta y Melilla
Badajoz	Zaragoza
Ceuta y Melilla	Huelva
Salamanca	Tenerife
Jaén	Barcelona

Tabla 5.1: Ordenación de las provincias de la EPF, según los dos primeros componentes

rango p, existiendo entonces tantas componentes principales como variables que se obtendrán calculando los valores propios o raíces características, $\lambda_1, \ldots, \lambda_p$, de la matriz de varianzas y covarianzas de las variables, $\mathbf{S}$, mediante:

$$|\mathbf{S} - \lambda \mathbf{I}| = 0 \tag{5.11}$$

y sus vectores asociados son:

$$(\mathbf{S} - \lambda_i \mathbf{I})\mathbf{a}_i = 0. \tag{5.12}$$

Los términos λ_i son reales, al ser la matriz $\mathbf{S}$ simétrica, y positivos, ya que $\mathbf{S}$ es definida positiva. Por ser $\mathbf{S}$ simétrica si λ_j y λ_h son dos raíces distintas sus vectores asociados son ortogonales. En efecto:

$$\begin{aligned} \mathbf{a}_h'\mathbf{S}\mathbf{a}_j &= (\mathbf{a}_h'\mathbf{S}\mathbf{a}_j)' = \mathbf{a}_j'\mathbf{S}\mathbf{a}_h \\ \mathbf{a}_h'\mathbf{S}\mathbf{a}_j &= \mathbf{a}_j'\lambda_h\mathbf{a}_h \end{aligned}$$

y si $\lambda_j \neq \lambda_h, \mathbf{a}_h'\mathbf{a}_j = \mathbf{a}_j'\mathbf{a}_h = 0$ y son ortogonales.

Si $\mathbf{S}$ fuese semidefinida positiva de rango $p < p$, lo que ocurriría si $p - p$ variables fuesen combinación lineal de las demás, habría solamente p raíces características positivas y el resto serían ceros.

Llamando $\mathbf{Z}$ a la matriz cuyas columnas son los valores de los p componentes en los n individuos, estas nuevas variables están relacionadas con las originales mediante:

$$\mathbf{Z} = \mathbf{XA}$$

donde $\mathbf{A}'\mathbf{A} = \mathbf{I}$. Calcular los componentes principales equivale a aplicar una transformación ortogonal $\mathbf{A}$ a las variables $\mathbf{X}$ (ejes originales) para obtener unas nuevas variables $\mathbf{Z}$ incorreladas entre sí. Esta operación puede interpretarse como elegir unos nuevos ejes coordenados, que coincidan con los "ejes naturales" de los datos.

Ejemplo 5.4 *Los restantes valores propios de la matriz de covarianzas de los datos de la EPF son 0.027, 0.0175, 0.0126, 0.0107, 0.010, 0.0059, y 0.00526. A partir del tercero son muy pequeños y de valor similar. El tercer componente principal es*

$$\begin{aligned} z_3 &= 0,12x_1 + 0,05x_2 + 0,34x_3 + 0,11x_4 - 0,85x_5 + 0,04x_6 - \\ &\quad 0,30x_7 + 0,20x_8 + 0,003x_9 = \\ &\quad (0,12x_1 + 0,05x_2 + 0,34x_3 + 0,11x_4 + 0,04x_6 + 0,20x_8) - \\ &\quad (0,85x_5 + 0,30x_7) \end{aligned}$$

y puede de nuevo interpretarse como la diferencia entre dos medias ponderadas. La primera da sobre todo peso a las variables 3, vivienda, 8, turismo y ocio, 1, alimentación y 4 , mobiliario doméstico. La segunda a la 5, gastos sanitarios, y a la 7, enseñanza y cultura. Separá provincias con bajos costes en sanidad y altos en vivienda y ocio de las que tengan la estructura opuesta. La figura **??** *representa las observaciones proyectadas sobre el plano de las componentes primera y tercera. Se observa que la tercera dimensión es independiente de la primera (riqueza o renta) y separa provincias con altos gastos en sanidad, como Salamanca y Palencia, de otras de aquellas con gastos relativamente bajos en esta magnitud y más en vivienda y ocio.*

Representación de los datos de la EPF em el plano definido por loso componentes primero y tercero.

Ejemplo 5.5 *La tabla 5.2 presenta la matriz de varianzas y covarianzas entre nueve indicadores económicos medidos en distintas empresas.*

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$
177	179	95	96	53	32	-7	-4	-3
	419	245	131	181	127	-2	1	4
		302	60	109	142	4	.4	11
			158	102	42	4	3	2
				137	96	4	5	6
					128	2	2	8
						34	31	33
							39	39
								48

Tabla 5.2: Matriz de varianzas covarianzas de los nueve indicadores

Las raíces características de esta matriz se presentan en la tabla 5.3.

Componente	1	2	3	4	5	6	7	8	9
λ_i	878,5	196,1	128,6	103,4	81,2	37,8	7,0	5,7	3,5

Tabla 5.3: Autovalores de la matriz tabla 5.2

La suma de los valores propios de la matriz es $1441,8$, *prácticamente igual, salvo por errores de redondeo, a la suma de las varianzas de las variables, que es 1442. Ya veremos que esta concordancia ocurre siempre. Los vectores propios de los tres primeros componentes se indican en la tabla 5.4. Se observa que el primer componente principal es una media ponderada de las primeras seis variables. El segundo contrapone la primera, la segunda y la cuarta a la tercera y la sexta. El tercer componente contrapone las tres primeras al resto de las variables.*

Estos resultados son consistentes con la matriz de la tabla 5.2. El rasgo más característico de esta tabla es la distinta magnitud de las seis primeras variables respecto al resto. Esto lo recoge el primer componente principal. El segundo rasgo es la presencia de covarianzas negativas en las filas de las dos primeras variables y esto se recoge en el segundo componente. El tercero incorpora por un lado las tres últimas variables y, por otro, contrapone las tres primeras variables frente al resto.

Componente	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	0.30	0.66	0.48	0.26	0.32	0.27	0.00	0.00	0.01
2	-0.48	-0.15	0.58	-0.49	-0.04	0.37	0.06	0.04	0.08
3	-0.41	-0.18	-0.23	0.45	0.49	0.27	0.26	0.28	0.29

Tabla 5.4: Vectores propios de la matriz tabla 5.2

5.4 PROPIEDADES DE LOS COMPONENTES

Los componentes principales como nuevas variables tienen las propiedades siguientes:

1. Conservan la variabilidad inicial: la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.
   Comprobemos el primer punto. Como $Var\ (z_h) = \lambda_h$ y la suma de las raíces características es la traza de la matriz:

   $$tr(\mathbf{S}) = Var\ (x_1) + \ldots + Var\ (x_p) = \lambda_1 + \ldots + \lambda_p$$

   por tanto $\sum_{i=1}^{p} Var(x_i) = \sum \lambda_i = \sum_{i=1}^{p} Var(z_i)$. Las nuevas variables z_i tienen conjuntamente la misma variabilidad que las variables originales, la suma de varianzas es la misma, pero su distribución es muy distinta en los dos conjuntos.
   Para comprobar que los componentes principales también conservan la *Varianza generalizada*, valor del determinante de varianzas y covarianzas de las variables, como el determinante es el producto de las raíces características, tenemos que, llamando $\mathbf{S}_z$ a la matriz de covarianzas de los componentes, que es diagonal con términos λ_i :

   $$|\mathbf{S}_x| = \lambda_1 \ldots \lambda_p = \Pi_{i=1}^{p} Var(z_i) = |\mathbf{S}_z|\,.$$

2. La proporción de variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado al vector propio que lo define, y la suma de los valores propios de la matriz.

   En efecto, como la varianza del componente h es λ_h, el valor propio que define el componente, y la suma de todas las varianzas de las variables originales es $\sum_{i=1}^{p} \lambda_i$, igual como acabamos de ver a la suma de las varianzas de los componentes, la proporción de variabilidad total explicada por el componente h es $\lambda_h / \sum \lambda_i$.

3. Las covarianzas entre cada componente principal y las variables X vienen dadas por el producto de las coordenadas del vector propio que define el componente por el valor propio:

   $$Cov(z_i; x_1, \ldots x_p) = \lambda_i \mathbf{a}_i = (\lambda_i a_{i1}, \ldots\ , \lambda_i a_{ip})$$

   donde $\mathbf{a}_i$ es el vector de coeficientes de la componente z_i.
   Para justificar este resultado, vamos a calcular la matriz $p \times p$ de covarianzas entre los componentes y las variables originales. Esta matriz es:

   $$Cov(z, x) = 1/n \mathbf{Z}'\mathbf{X}$$

   y su primera fila proporciona las covarianzas entre la primera componente y las p variables originales. Como $\mathbf{Z} = \mathbf{X}\mathbf{A}$, sustituyendo

   $$Cov(z, x) = 1/n \mathbf{A}'\mathbf{X}'\mathbf{X} = \mathbf{A}'\mathbf{S} = \mathbf{D}\mathbf{A}',$$

donde $\mathbf{A}$ contiene en columnas los vectores propios de $\mathbf{S}$ y $\mathbf{D}$ es la matriz diagonal de los valores propios. En consecuencia, la covarianza entre, por ejemplo, el primer componente principal y las p variables vendrá dada por la primera fila de $\mathbf{A'S}$, es decir $\mathbf{a}_1'\mathbf{S}$ o también $\lambda_1\mathbf{a}_1'$, donde $\mathbf{a}_1'$ es el vector de coeficientes de la primera componente principal.

4. Las correlación entre un componente principal y una variable X es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.

   Para comprobarlo:

$$Corr(z_i; x_j) = \frac{Cov(z_i x_j)}{\sqrt{Var(z_i)Var(x_j)}} = \frac{\lambda_i a_{ij}}{\sqrt{\lambda_i s_j^2}} = a_{ij}\frac{\sqrt{\lambda_i}}{s_j}$$

5. Las r componentes principales (r < p) proporcionan la predicción lineal óptima con r variables del conjunto de variables X.

   Esta afirmación puede expresarse de dos formas. La primera demostrando que la mejor predicción lineal con r variables de las variables originales se obtiene utilizando las r primeras componentes principales. La segunda demostrando que la mejor aproximación de la matriz de datos que puede construirse con una matriz de rango r se obtiene construyendo esta matriz con los valores de los r primeros componentes principales. La demostración de estas propiedades puede verse en el apéndice 5.1.

6. Si estandarizamos los componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales.

Estandarizando los componentes $\mathbf{Z}$ por sus desviaciónes típicas, se obtienen las nuevas variables

$$\mathbf{Y}_c = \mathbf{ZD}^{-1/2} = \mathbf{XAD}^{-1/2}$$

donde $\mathbf{D}^{-1/2}$ es la matriz que contienen las inversas de las desviaciónes típicas de las componentes. Hemos visto en el capítulo anterior que la estandarización multivariante de una matriz de variables $\mathbf{X}$ de media cero viene dada por se define como:

$$\mathbf{Y}_s = \mathbf{XAD}^{-1/2}\mathbf{A'}$$

y ambas variables están incorreladas y tienen matriz de covarianzas identidad. Se diferencian en que unas pueden ser una rotación de las otras, lo que es indiferente al tener todas las mismas varianzas. Por tanto,la estandarización multivariante puede interpretarse como :

(1) obtener los componentes principales;

(2) estandarizarlos para que tengan todos la misma varianza.

Esta relación se presenta gráficamente en la figura 5.2. La transformación mediante componentes principales conduce a variables incorreladas pero con distinta varianza, puede

interpretarse como rotar los ejes de la elipse que definen los puntos para que coincidan con sus ejes naturales. La estandarización multivariane produce variables incorreladas con varianza unidad, lo que supone buscar los ejes naturales y luego estandarizarlos. En consecuencia, si estandarizamos los componentes se obtiene las variables estandarizadas de forma multivariante.

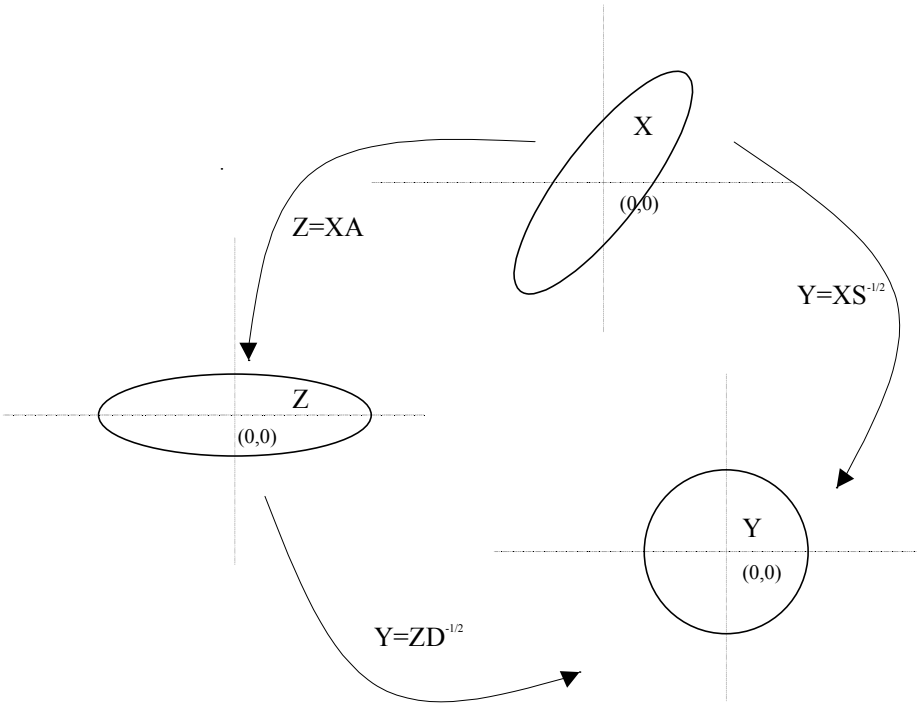


Figura 5.2: Representación gráfica de la relación entre componentes principales y estandarización multivariante.

5.5 ANÁLISIS NORMADO O CON CORRELACIONES

Los componentes principales se obtienen maximizando la varianza de la proyección. En términos de las variables originales esto supone maximizar:

$$M = \sum_{i=1}^{p} a_i^2 s_i^2 + 2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} a_i a_j s_{ij} \tag{5.13}$$

con la restricción $\mathbf{a}'\mathbf{a} = 1$. Si alguna de las variables, por ejemplo la primera, tiene una varianza s_1^2, mayor que las demás, la manera de aumentar M es hacer tan grande como podamos la coordenada a_1 asociada a esta variable. En el límite si una variable tiene una varianza mucho mayor que las demás el primer componente principal coincidirá muy aproximadamente con esta variable.

Cuando las variables tienen unidades distintas esta propiedad no es conveniente: si disminuimos la escala de medida de una variable cualquiera, de manera que aumenten en

magnitud sus valores numéricos (pasamos por ejemplo de medir en km. a medir en metros), el peso de esa variable en el análisis aumentará, ya que en (5.13):

(1) su varianza será mayor y aumentará su coeficiente en el componente, a_i^2, ya que contribuye más a aumentar M;

(2) sus covarianzas con todas las variables aumentarán, con el consiguiente efecto de incrementar a_i.

En resumen, cuando las escalas de medida de las variables son muy distintas, la maximización de (5.13) dependerá decisivamente de estas escalas de medida y las variables con valores más grandes tendrán más peso en el análisis. Si queremos evitar este problema, conviene estandarizar las variables antes de calcular los componentes, de manera que las magnitudes de los valores numéricos de las variables X sean similares.

La estandarización resuelve otro posible problema. Si las variabilidades de las X son muy distintas, las variables con mayor varianza van a influir más en la determinación de la primera componente. Este problema se evita al estandarizar las variables, ya que entonces las varianzas son la unidad, y las covarianzas son los coeficientes de correlación. La ecuación a maximizar se transforma en:

$$M' = 1 + 2\sum_{i=1}^{p}\sum_{j=i+1}^{p} a_i a_j r_{ij} \tag{5.14}$$

siendo r_{ij} el coeficiente de correlación lineal entre las variables ij. En consecuencia la solución depende de la correlaciones y no de las varianzas.

Los componentes principales normados se obtiene calculando los vectores y valores propios de la matriz $\mathbf{R}$, de coeficientes de correlación. Llamando λ_p^R a las raíces características de esa matriz, que suponemos no singular, se verifica que:

$$\sum_{i=1}^{p} \lambda_i^R = traza(R) = p \tag{5.15}$$

Las propiedades de los componentes extraídos de $\mathbf{R}$ son:

1. La proporción de variación explicada por λ_p^R será:

$$\frac{\lambda_p^R}{p} \tag{5.16}$$

2. Las correlaciones entre cada componente z_j y las variables X originales vienen dados directamente por $a'_j\sqrt{\lambda_j}$ siendo $\mathbf{z}_j = \mathbf{X}\mathbf{a}_j$.

Estas propiedades son consecuencia inmediata de los resultados de la sección 5.4.

Cuando las variables X originales están en distintas unidades conviene aplicar el análisis de la matriz de correlaciones o análisis normado. Cuando las variables tienen las mismas

unidades, ambas alternativas son posibles. Si las diferencias entre las varianzas de las variables son informativas y queremos tenerlas en cuenta en el análisis no debemos estandarizar las variables: por ejemplo, supongamos dos índices con la misma base pero uno fluctua mucho y el otro es casi constante. Este hecho es informativo, y para tenerlo en cuenta en el análisis, no se deben estandarizar las variables, de manera que el índice de mayor variabilidad tenga más peso. Por el contrario, si las diferencias de variabilidad no son relevantes podemos eliminarlas con el análisis normado. En caso de duda, conviene realizar ambos análisis, y seleccionar aquel que conduzca a conclusiones más informativas.

Ejemplo 5.6 *La matriz de correlación de los nueve indicadores económicos del ejemplo5.4 es*

$$\mathbf{R} = \begin{bmatrix} 1 & .66 & .41 & .57 & .34 & .21 & -.09 & -.05 & -.03 \\ & 1 & .69 & .51 & .76 & .55 & -.01 & .01 & .03 \\ & & 1 & .28 & .54 & .72 & .04 & .00 & .09 \\ & & & 1 & .69 & .30 & .05 & .03 & .02 \\ & & & & 1 & .73 & .06 & .07 & .07 \\ & & & & & 1 & .03 & .03 & .10 \\ & & & & & & 1 & .85 & .82 \\ & & & & & & & 1 & .90 \\ & & & & & & & & 1 \end{bmatrix}$$

Los valores propios son:

λ_i	*3.70*	*2.72*	*1.06*	*.70*	*.30*	*.23*	*.16*	*.09*	*.03*

y los vectores propios asociados a los tres primeros valores propios son:

λ	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
3.7	.34	.46	.41	.36	.46	.40	.06	.06	.08
2.72	-.11	-.07	-.03	-.04	-.02	-.01	.56	.58	.57
1.06	-.54	-.05	.38	-.52	.07	.53	-.04	-.07	.00

Tabla 5.5: Vectores propios de la matriz de correlaciones

Si comparamos estos resultados con los del ejemplo 5.4 vemos que el primer vector propio cambia apreciablemente. Con la matriz de varianzas las variables con más peso en el componente eran las que tenían una mayor varianza: la 2, luego la 3 y finalmente las 1,4,5 y 6 con un peso parecido. Estos pesos siguen estrechamente la relación relativa entre las varianzas de las variables. Sin embargo, al utilizar la matriz de correlaciones este efecto desaparece, y el peso de las variables está más relacionado con las correlaciones. La proporción de variabilidad explicada por el primer componente cambia mucho: de $878,5/1441,8 = 60,9\%$ *a* $3.7/9 = 41\%$

El segundo componente cambia completamente: ahora está prácticamente asociado a las tres últimas variables. La proporción de variabilidad que explica ha aumentado considerablemente, del $196/1441,8 = 13,6\%$ *a* $2.72/9 = 30\%$ *. El tercer vector propio es también distinto en ambas matrices.*

Ejemplo 5.7 *Consideremos los datos de INVEST publicaciones científicas en los países de la OCDE. Los datos tienen magnitudes muy distintas (unos bancos de datos tienen muchos más trabajos que otros). Si deseamos conservar esta propiedad, que esta asociada a que en algunos campos científicos se publica mucho más que en otros, haremos el análisis sobre la matriz de covarianzas. Si no queremos dar más peso a unos campos que a otros, es conveniente realizar el análisis normado o sobre la matriz de correlación. Los resultados en este último caso se indican en la tabla 5.6*

Comp.	λ_h	P_h	$\sum_{i=1}^{h} P_h$
1	7.630	0.954	0.954
2	0.207	0.026	0.980
3	0.121	0.015	0.995
4	0.019	0.002	0.997
5	0.017	0.002	0.999
6	0.004	0.001	1.000
7	0.001	0.000	1.000
8	0.000	0.000	1.000

Tabla 5.6: Variabilidad explicada por los componentes principales

Se observa que el primer componente principal explica una proporción muy alta de la variabilidad, el 95,4%. Con los tres primeros componentes se explica el 99,5% de la variabilidad. Además, después del tercer vector propio la variabilidad explicada disminuye claramente, (véase la tabla 5.6 y la figura 5.3)lo que indica que sólo debemos preocuparnos de los tres primeros componentes ya que los siguientes tienen poca capacidad explicativa. En la tabla 5.7 se indican los valores de los componentes para estos tres vectores propios.

	Comp. 1	Comp. 2	Comp. 3
INTER.A	0.358	-0.173	0.36
INTER.F	0.360	-0.098	0.08
AGRIC.	0.355	-0.366	-0.10
BIOLO.	0.346	-0.359	-0.69
MEDIC.	0.361	-0.070	0.15
QUIMI.	0.334	0.786	-0.41
INGEN.	0.354	0.268	0.40
FISICA	0.361	0.054	0.17

Tabla 5.7: Vectores propios de los tres primeros componentes

Ejemplo 5.8 *Para interpretar los componentes consideramos sus coordenadas en las variables. Estas se indican en la tabla 5.7 y en la figura 5.4. Se observa que el primer componente es un factor de tamaño, ya que es una media ponderada de todas las variables con mayor peso de los bancos interdisciplinarios y del banco médico. El segundo componente es un factor de forma y contrapone la investigación en Química e Ingeniería frente a la realizada en*

Figura 5.3: Gráfico para la selección del número de componentes.

Agricultura y Biología. El tercero contrapone ingeniería, física y el banco interA con respecto a Biología y Química.

5.6 INTERPRETACIÓN DE LOS COMPONENTES

Componentes de tamaño y forma

Cuando existe una alta correlación positiva entre todas las variables, el primer componente principal tiene todas sus coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas las variables (véase el ejercicio 5.2). Se interpreta entonces como un factor global de "tamaño". Los restantes componentes se interpretan como factores "de forma" y típicamente tienen coordenadas positivas y negativas, que implica que contraponen unos grupos de variables frente a otros. Estos factores de forma pueden frecuentemente escribirse como medias ponderadas de dos grupos de variables con distinto signo y contraponen las variables de un signo a las del otro. Por ejemplo el segundo componente principal de los datos de la EPF del ejercicio 5.3 puede escribirse aproximadamente, despreciando los coeficiente pequeños (menores que 0,1):

$$z_2 = (0,05x_1 + 0,16x_2 + 0,07x_4 + 0,29x_6 + 0,78x_9) - (0,17x_3 + 0,21x_5 + 0,40x_7 + 0,17x_8) \simeq I_0 - I_S$$

Figura 5.4: Representación de los pesos de las dos componentes.

donde

$$I_0 = 0,16x_2 + 0,29x_6 + 0,78x_9$$

es un indicador de los gastos de transporte y transferencias a otras provincias y

$$I_S = 0,17x_3 + 0,21x_5 + 0,40x_7 + 0,17x_8$$

es un indicador de gastos en servicios (educación y sanidad). Además, cuando las variables van en logaritmos, los componentes suelen poder escribirse como ratios de promedios geométricos de las variables. Por ejemplo, supongamos que un componente tiene la expresión

$$z_1 = -0.5 \log x_1 + 0.3 \log x_2 + 0.2 \log x_3$$

este componente puede escribirse también como

$$z_1 = 0.3 \log \frac{x_2}{x_1} + 0.2 \log \frac{x_3}{x_1}$$

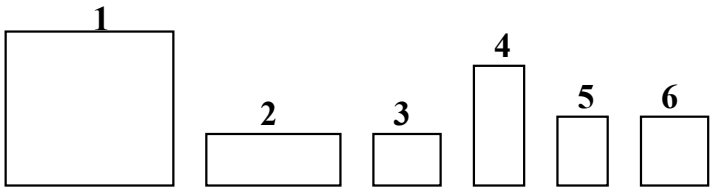
que indica que es un promedio de estos dos ratios (véase el ejemplo 5.1).

La interpretación de los componentes se simplifica suponiendo que los coeficientes pequeños son cero y redondeando los coeficientes grandes para expresar el componente como cocientes, diferencias o sumas entre variables. Estas aproximaciones son razonables si modifican poco la estructura del componente y mejoran su interpretación. Una medida del cambio introducido al modificar un vector propio de $\mathbf{a}_i$ a $\mathbf{a}_{iM}$ es el cambio en la proporción de variabilidad explicada por el componente. Si el valor propio asociado a $\mathbf{a}_i$ es λ_i, el componente explica el

$\lambda_i / \sum \lambda_j$ de la variabilidad. Si ahora modificamos el vector a $\mathbf{a}_{iM}$, la varianza de la proyección de los datos sobre este componente es $\lambda_{iM} = \mathbf{a}'_{iM}\mathbf{S}\mathbf{a}_{iM} = (\widetilde{\mathbf{X}}\mathbf{a}_{iM})'(\widetilde{\mathbf{X}}\mathbf{a}_{iM})/n$, la varianza del componente, y la proporción de variabilidad explicada será $\lambda_{iM}/\sum \lambda_j$. El cambio relativo será $(\lambda_i - \lambda_{iM})/\lambda_i$, ya que siempre $\lambda_i \geq \lambda_{iM}$, y si este cambio es pequeño, esta justificada la modificación si favorece la interpretación.

Ejemplo 5.9 *Vamos a calcular el cambio relativo que experimenta el segundo componente principal de los datos de la EPF si despreciamos los coeficientes más pequeños, la varianza del segundo componente modificado es 0,0319. La varianza del componente original es 0,0320, por lo que el cambio de explicación por tomar el coeficiente simplificado es sólo de (0,0320-0,0319)/0,0320=1/320=0,0031.*

Ejemplo 5.10 *Supongamos 6 observaciones $x_1, \ldots, x_6$ en dos dimensiones, cada observación corresponde a un rectángulo y las variables son longitud de la base y altura del rectángulo. Gráficamente las observaciones son,*



que corresponden a la matriz de datos,

$$X = \begin{bmatrix} 2 & 2 \\ 1.5 & 0.5 \\ 0.7 & 0.5 \\ 0.5 & 1.5 \\ 0.5 & 0.7 \\ 0.7 & 0.7 \end{bmatrix}$$

aplicamos logaritmos a estos datos para facilitar la interpretación de las componentes,

$$\log(X) = \begin{bmatrix} 0.301 & 0.301 \\ 0.176 & -0.301 \\ -0.155 & -0.301 \\ -0.301 & 0.176 \\ -0.301 & -0.155 \\ -0.155 & -0.155 \end{bmatrix}$$

cuya matriz de varianzas covarianzas es,

$$S = \begin{bmatrix} 6.39 & 1.41 \\ 1.41 & 6.39 \end{bmatrix} .10^{-2}$$

Los autovalores y autovectores de la descomposición espectral de esta matriz son,

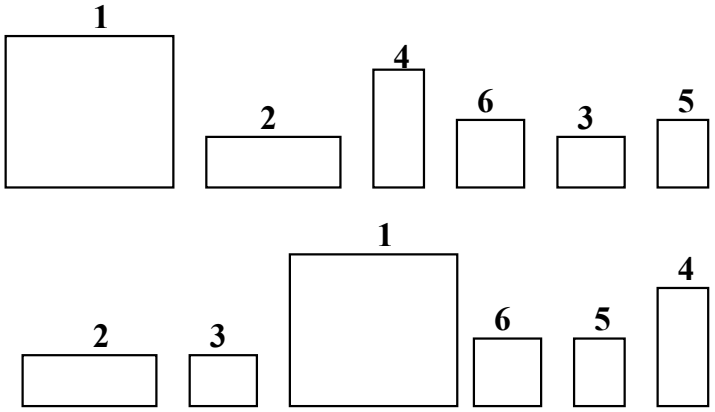
$$\begin{aligned} \lambda_1 &= 0.78 \qquad \lambda_2 = 0,0498 \\ a_1 &= \begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix} \qquad a_2 = \begin{bmatrix} 0.707 \\ -0.707 \end{bmatrix} \end{aligned}$$

las dos primeras componentes son

$$Z_1 = Xa_1 = 0.707\log(X_1) + 0.707\log(X_2) = 0.707\log(X_1X_2) = \begin{bmatrix} 0.426 \\ -0.088 \\ -0.322 \\ -0.088 \\ -0.322 \\ -0.219 \end{bmatrix}$$

$$Z_2 = Xa_2 = 0.707\log(X_1) - 0.707\log(X_2) = 0.707\log(\frac{X_1}{X_2}) = \begin{bmatrix} 0 \\ 0.337 \\ 0.103 \\ -0.337 \\ -0.103 \\ 0 \end{bmatrix}$$

Si ordenamos los rectángulos según el valor de la primera y segunda componente obtenemos,



La primera ordenación coincide con la inducida por el volumen de los rectángulos, es una transformación creciente del producto de la base por la altura, y el primer componente describe el tamaño. El segundo componente relaciona la base con la altura y ordena las observaciones en función de su forma.

5.6.1 Selección del número de componentes

Se han sugerido distintas reglas para seleccionar el número de componentes a mantener:

(1) Realizar un gráfico de λ_i frente a i. Comenzar seleccionando componentes hasta que los restantes tengan aproximadamente el mismo valor de λ_i. La idea es buscar un "codo" en el gráfico, es decir, un punto a partir del cual los valores propios son aproximadamente iguales. El criterio es quedarse con un número de componentes que excluya los asociados a valores pequeños y aproximadamente del mismo tamaño.

(2) Seleccionar componentes hasta cubrir una proporción determinada de varianza, como el 80% o el 90%. Esta regla es arbitraria y debe aplicarse con cierto cuidado. Por ejemplo,

es posible que un único componente de "tamaño" recoja el 90% de la variabilidad y sin embargo pueden existir otros componentes que sean muy adecuados para explicar la "forma" de las variables.

(3) Desechar aquellos componentes asociados a valores propios inferiores a una cota, que suele fijarse como la varianza media, $\sum \lambda_i/p$. En particular, cuando se trabaja con la matriz de correlación, el valor medio de los componentes es 1, y esta regla lleva a seleccionar los valores propios mayores que la unidad. De nuevo esta regla es arbitraria: una variable que sea independiente del resto suele llevarse un componente principal (véase ejercicio 5.8) y puede tener un valor propio mayor que la unidad. Sin embargo, si esta incorrelada con el resto puede ser una variable poco relevante para el análisis, y no aportar mucho a la comprensión del fenómeno global.

5.6.2 Representación gráfica

La interpretación de los componentes principales se favore representando las proyecciones de las observaciones sobre un espacio de dimensión dos, definido por parejas de los componentes principales más importantes. Este punto se ha ilustrado en los ejemplos anteriores, donde se ha indicado que la proyección de cualquier observación sobre un componente es directamente el valor del componente para esa observación. La representación habitual es tomar dos ejes ortogonales que representen los dos componentes considerados, y situar cada punto sobre ese plano por sus coordendas con relación a estos ejes, que son los valores de los dos componentes para esa observación. Por ejemplo, en el plano de los dos primeros componentes, las coordenadas del punto $\mathbf{x}_i$ son $z_{1i} = \mathbf{a}_1'\mathbf{x}_i$ y $z_{2i} = \mathbf{a}_2'\mathbf{x}_i$.

La interpretación se favorece representando en el mismo plano además de las observaciones las variables originales. Esto puede hacerse utilizando como coordenadas su coeficiente de correlación con cada uno de los ejes. El vector de correlaciones entre el primer componente y las variables originales viene dado por $\lambda_1^{1/2}\mathbf{a}_1'\mathbf{D}$, donde $\mathbf{D}$ es una matriz diagonal cuyos términos son las inversas de las desviaciones típicas de cada variable. La matriz de correlaciones $\mathbf{R}_{cv}$ entre los p componentes y las p variables tendrá como filas los términos $\lambda_j^{1/2}\mathbf{a}_j'\mathbf{D}$ y puede escribirse

$$\mathbf{R}_{cv} = \Lambda^{1/2}\mathbf{A}\mathbf{D}$$

donde $\mathbf{A}$ es la matriz de vectores propios, $\Lambda^{1/2}$ es la matriz diagonal con términos $\sqrt{\lambda_i}$ y En el análisis normado como las variables se estandarizan a varianza unidad las correlaciones será simplemente $\Lambda^{1/2}\mathbf{A}$.

Una representación equivalente es el biplot que presentamos en la sección siguiente. Tiene la ventaja de representar al mismo tiempo las variables y las observaciones en un mismo gráfico.

Conviene investigar si transformando las variables se obtiene una interpretación más simple. Como regla general, cuando al tomar logaritmos las variables $\mathbf{X}$ tienen una distribución aproximadamente simétrica, conviene realizar el análisis de componentes principales sobre los logaritmos de las variables.

Es importante recordar que las covarianzas (o correlaciones) miden únicamente las relaciones lineales entre las variables. Cuando entre ellas existan relaciones fuertes no lineales el análisis de componentes principales puede dar una información muy parcial de las variables.

Ejemplo 5.11 *La figura 5.5 presenta la proyección de los datos de INVEST, los países de la OCDE, sobre el plano formado por los dos primeros componentes principales extraídos de la matriz de correlación, que se estudiaron en el ejemplo 5.6. Se observa que el primer eje ordena a los países por su cantidad de investigación, mientras que el segundo tiene en cuenta sus características: separa a Japón, con gran énfasis en investigación tecnológica, del Reino Unido, que tiene más énfasis en la investigación biomédica*

Figura 5.5: Proyección de las observaciones en las dos primeras componentes principales.

Como indicamos en el Capítulo la observación de EEUU es atípica y existe una marcada asimetría en las distribuciones de las variables. Vamos a presentar los datos excluyendo a EEUU y con una transformación logarítmica de las variables para reducir la asimetría. La figura 5.6 muestra el nuevo diagrama de cajas múltiple.Como la varianza de las nuevas variables transformadas es similar, el análisis de componentes principales se realizará directamente sobre la matriz de varianzas covarianzas. Los resultados obtenidos figuran en las tablas 5.8 y5.9

Los tres primeros componentes explican el 97% de la variabilidad y tienen la siguiente interpretación. El primero es una media ponderada de todos los bancos con mayo peso del

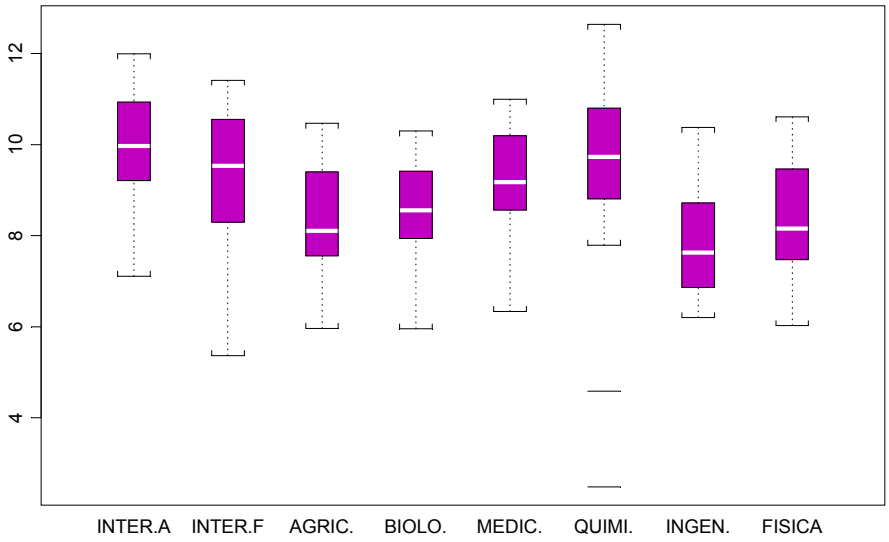


Figura 5.6: Diagrama de cajas de los logaritmos de las variables de INVEST una vez eliminado EEUU.

	λ_h	P_h	$\sum_{i=1}^{h} P_h$
Comp. 1	14.98	0.90	0.90
Comp. 2	0.83	0.05	0.94
Comp. 3	0.50	0.03	0.97
Comp. 4	0.21	0.01	0.99
Comp. 5	0.10	0.01	0.99
Comp. 6	0.08	0.00	1.00
Comp. 7	0.02	0.00	1.00
Comp. 8	0.02	0.00	1.00

Tabla 5.8: Variabilidad explicada por los componentes principales

banco químico. El segundo, contrapone la investigación en Química frente a la general del banco INTER.F y a la de ingeniería y física. El tercero contrapone el banco INTER.F y Química al resto.

	Comp. 1	Comp. 2	Comp. 3
INTER.A	0,31	0,05	-0,40
INTER.F	0,37	0,63	0,63
AGRIC.	0,30	0,07	-0,14
BIOLO.	0,27	-0,06	-0,30
MEDIC.	0,32	0,01	-0,25
QUIMI.	0,56	-0,70	0,41
INGEN.	0,28	0,25	-0,18
FÍSICA	0,32	0,21	-0,26

Tabla 5.9: Pesos de las tres primeras componentes principales

Los países proyectados en estos tres componentes se presentan en la figura 5.7. Se ha añadido también la proyección sobre el cuarto componente, que separa completamente a UK de Japón.

5.6.3 Datos atípicos

Antes de obtener los componentes principales conviene asegurarse de que no existen datos atípicos, ya que, como hemos visto en el capítulo anterior, los atípicos pueden distorsionar totalmente la matriz de covarianzas.

Para ilustrar su efecto sobre los componentes, supongamos el caso más simple en que un error de medida en una variable introduce un valor atípico grande en la primera variable. Su efecto será aumentar mucho la varianza de esta variable y disminuir las covarianzas con las restantes, con lo que, si hacemos el atípico muy grande, la matriz $\mathbf{S}$ será, aproximadamente:

$$\begin{bmatrix} \sigma_1^2 & \dots & \mathbf{0}' \\ \mathbf{0} & & \mathbf{S}_{22} \end{bmatrix}$$

donde $\mathbf{0}'=(0,0,...,0)$. Esta matriz tiene un vector propio $(1, 0, \dots, 0)$ unido al valor propio σ_1^2 y si σ_1^2 es muy grande este será el primer componente principal. Por tanto, un valor atípico suficientemente grande distorsiona todos los componentes que podemos obtener de la matriz afectada (véase el ejemplo 5.9).

El resultado anterior sugiere que las componentes principales podrían utilizarse para detectar datos atípicos multivariantes, ya que un valor muy extremo se llevara un componente principal y aparecerá como extremo sobre esta componente. Desgraciadamente, aunque los componentes pueden identificar atípicos aislados, no hay garantía de que funcionen cuando existen grupos de atípicos, debido al problema de enmascaramiento. Por esta razón conviene utilizar para detectarlos el método presentado en el capítulo anterior, basado en proyecciones sobre las direcciones extremas de kurtosis, que al ser capaz de identificar todos los posibles atípicos permite calcular una la matriz de covarianzas libre de distorsiones graves.

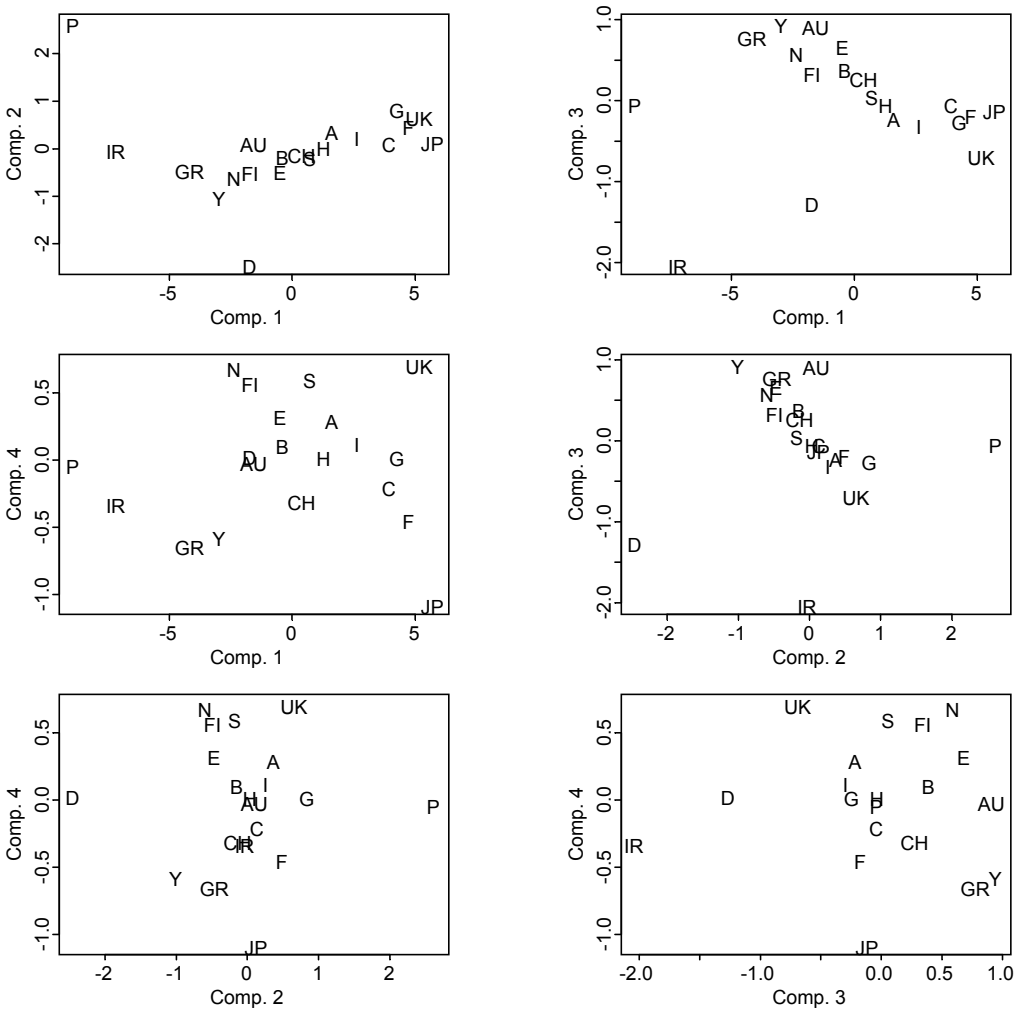


Figura 5.7: Representación de las observaciones de INVEST en losplanos definidos por las cuatro primeras componentes.

5.6.4 Distribución de los componentes

Los componentes principales pueden verse como un conjunto nuevo de variables y estudiar su distribución individual y conjunta. Por construcción estarán incorrelados, pero pueden exitir fuertes relaciones no lineales entre ellos.

Ejemplo 5.12 *Vamos a calcular los componentes principales de la matriz de correlación de las 27 medidas físicas, MEDIFIS. Aunque todas las variables van en centímetros, los tamaños de las variables son muy distintos, lo que aconseja utilizar la matriz de correlación. La proporción de varianza que explica cada vector propio se indica en la tabla 5.10*

Para decidir cuántos componentes tomar utilizaremos la figura 5.8 que indica que a partir del tercer componente hay una caída en la capacidad predictiva. Los tres primeros componentes explican conjuntamente el 93.5% de la variabilidad.

λ_h	5.56	0.62	0.39	0.17	0.14	0.10	0.05
$P_h\%$	78.96	8.87	5.65	2.48	1.98	1.37	0.68

Tabla 5.10: Variabilidad explicada por las componentes

Figura 5.8: Gràfico para seleccionar el número de componentes.

Los tres primeros vectores propios son:

	est	*pes*	*pie*	*lbr*	*aes*	*dcr*	*drt*
Comp. 1	*.41*	*.39*	*.40*	*.39*	*.38*	*.29*	*.37*
Comp. 2	*-.16*	*.04*	*-.20*	*-.30*	*.11*	*.89*	*-.15*
Comp. 3	*.04*	*-.29*	*.13*	*-.15*	*-.57*	*.20*	*.71*

El primer componente es una media de todas las medidas físicas, y por tanto una medida del tamaño del cuerpo, siendo la variable con menor peso el diámetro del cráneo. La segunda variable es de forma, y esta dominada por el diámetro del cráneo. Observemos que esta variable está poco correlada con el resto y, por lo tanto, arrastra ella sola un componente principal, ya que no puede explicarse como combinación de otras. El tercer componente principal diferencia longitud frente a anchura: da mayor peso a la longitud de la pierna (drt) y lo contrapone al peso y a la anchura de la espalda.

La figura 5.9 presenta un gráfico de las observaciones sobre el plano de los dos primeros componentes principales. Las coordenadas son las puntuaciones estandarizadas $z_i^* = X^* a_i$, $i = 1, 2$, *donde* X^* *es la matriz de variables estandarizadas (de media cero y varianza uno). En este gráfico cada punto se indica con un 1, cuando la observación corresponde a un varón y un 0 cuando es mujer. Puede verse que la primera componente de "tamaño" separa casi perfectamente los hombres de las mujeres. El segundo componente no parece reflejar ningún efecto del sexo. Observemos que la primera componente es capaz, por si misma, de explicar casi el 80% de variabilidad. Dado que el diámetro del cráneo está poco correlado con el resto*

de las variables, siendo casi en exclusiva responsable de una dimensión, vamos a repetir el análisis eliminando esta variable.

Figura 5.9: Proyección de las observaciones en las dos primeras componentes principales.

Los resultados de eliminar la variable diámetro del cráneo del análisis se presentan en la tabla siguiente. Se incluyen los dos primeros valores y vectores propios que explican por sí mismos el 92% de la variabilidad.

λ_h	$P_h\%$	*est*	*pes*	*pie*	*lbr*	*aes*	*drt*
5.1	*85*	*.43*	*.41*	*.42*	*.41*	*.39*	*.38*
.4	*7*	*.08*	*-.32*	*.17*	*-.04*	*-.60*	*.71*
$Corr(z_1x_i)$		*.97*	*.93*	*.95*	*.93*	*.88*	*.86*
$Corr(z_2x_i)$		*.05*	*-.20*	*.11*	*-.030*	*-.38*	*.45*

El primer componente es de nuevo una media ponderada que indica el tamaño de las personas, dando el mayor peso a la estatura de la persona. El segundo es de forma, ya que contrapone la longitud de la pierna a la anchura de la espalda y tiene peso positivo en las longitudes (del pie y estatura), y negativo en el peso. La proyección de los datos sobre el plano definido por los dos componentes se presenta en la figura 5.10. Se observa que el primer componente de "tamaño" separa como antes los hombres de las mujeres, y que el segundo componente al ser

ortogonal al tamaño no parece depender del sexo. Este componente separa para ambos sexos pesonas con constitución delgada de gruesa.

La figura 5.11 presenta de forma gráfica las correlaciones entre el primer y segundo componente y cada variable, calculadas como $\sqrt{\lambda_h a_{hj}}$. *Se observa que el primer componente está correlado con la altura y las restantes longitudes, mientras que el segundo está especialmente relacionado con la longitud de la pierna y la anchura de la espalda.*

Figura 5.10: Proyección de las observaciones en las dos primeras componentes principales.

Ejemplo 5.13 *Vamos a analizar la base de datos de MUNDODES (tabla A.6 del Anéxo). Esta matriz de datos está constituida por 91 países en los que se han observado 9 variables:* X_1*: ratio de natalidad,* X_2*: ratio de mortalidad,* X_3*: mortalidad infantil,* X_4*: esperanza de vida en hombres* X_5*: esperanza de vida de mujeres y* X_6*: PNB per capita.*

La representación gráfica de las variables dos a dos, presentada en el capítulo anterior, muestra relaciones claramente no lineales. Aplicando transformaciones logarítmicas a las variables mejoramos la linealidad en estas relaciones dos a dos.

Como las variables están medidas en distintas unidades se debe realizar un análisis de componentes principales normado (basado en la matriz de correlaciones), los resultados se presentan en la figura 5.12.

La figura 5.13 presenta el gráfico en forma de codo para seleccionar el número de componentes. El primer valor propio es 4.7278, y explica el 78,8% de la variabilidad. El segundo es 0.7261, y explica el 12%. Hay un valor propio de 0,002 que corresponde a una variable que es practicamente constante. Los vectores propios se presentan a continuación.

Figura 5.11: Correlación de las variables con las componentes principales.

variable	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*
X_1	*-0.454*	*0.034*	*-0.130*	*0.159*	*0.378*	*0.780*
X_2	*0.416*	*0.196*	*0.513*	*0.683*	*0.233*	*0.067*
X_3	*0.341*	*-0.680*	*-0.524*	*0.307*	*0.225*	*-0.031*
X_4	*0.440*	*-0.052*	*0.222*	*-0.632*	*0.578*	*0.145*
X_5	*-0.452*	*0.085*	*-0.029*	*0.114*	*0.639*	*-0.605*
X_6	*-0.326*	*-0.699*	*0.628*	*-0.039*	*-0.100*	*0.002*

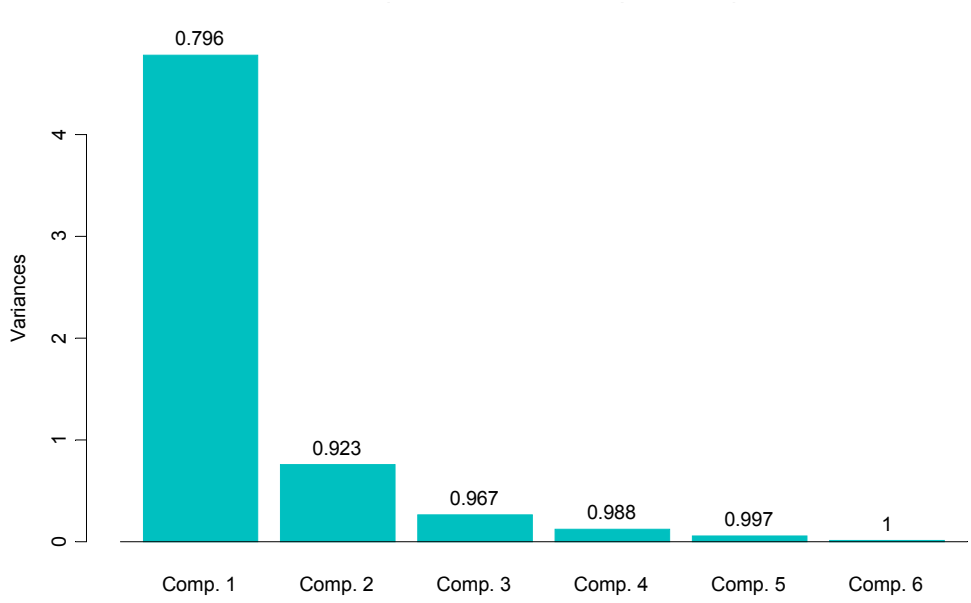


Figura 5.12: Proporción de variabilidad explicada por cada componente para los datos de MUNDODES.

Figura 5.13:

El primer componente explica el 79% de la variabilidad, el segundo corresponde a un valor propio inferior a 1, pero lo incluiremos para interpretarlo. La primera componente se puede interpretar como una medida de desarrollo de un país, dado que las variables con peso

positivo son las esperanzas de vida de hombres y mujeres y la renta, mientras que las de peso negativo son la mortalidad infantil y las tasas de natalidad y mortalidad, que son bajas en los países más desarrollados. El segundo componente esta asociado a la mortalidad infantil y a la renta, con lo que resulta de dificil interpretación ya que mide una dimensión que está incorrelada con el primer término de desarrollo. Para interpretarla, la figura 5.14 muestra los paises en el plano de los dos componentes. Se observa que existe una fuerte relación no lineal entre ambos y aunque los componentes están incorrelados no son claramente independientes. El primer componente podemos suponer que ordena a los países por desarrollo y el segundo tiene en cuanta la mortalidad infantil y tiene una relación no lineal con la renta.

Figura 5.14: Representacion de los dos primeros componentes para los datos de Mundodes

En los diagramas de dispersión vimos que relaciones entre las variables eran no lineales, por lo que vamos a repetir el análisis para las variables en logaritmos. Los valores propios de la matriz de correlaciones de las variables en logaritmos no cambian mucho, pero los vectores propios sí lo hacen. Son ahora:

PC1 PC2 PC3 PC4 PC5 PC6
0.403 0.435 -0.376 -0.436 -0.562 0.033
0.307 -0.831 0.011 -0.457 -0.077 -0.020
0.433 0.267 -0.023 -0.331 0.793 0.051
-0.441 0.147 0.224 -0.531 0.019 -0.672
-0.446 0.071 0.213 -0.454 -0.012 0.738
-0.403 -0.149 -0.873 -0.057 0.223 -0.008

El primero sigue siendo una medida de desarrollo pero ahora el segundo esta sobre todo ligado a la tasa de mortalidad. Separa paises con alta tasa de mortalidad de los de baja. Vemos que el último vector propio también tiene una interesante interpretación. Nos dice que la diferencia en logaritmos entre las esperanzas de vida de hombres y mujeres es prácticamente constante en todos los países, ya que el valor propio que corresponde a este vector

propio es muy pequeño (0,015). Los pesos asociados a cada una de las variables se presentan en la figura 5.15

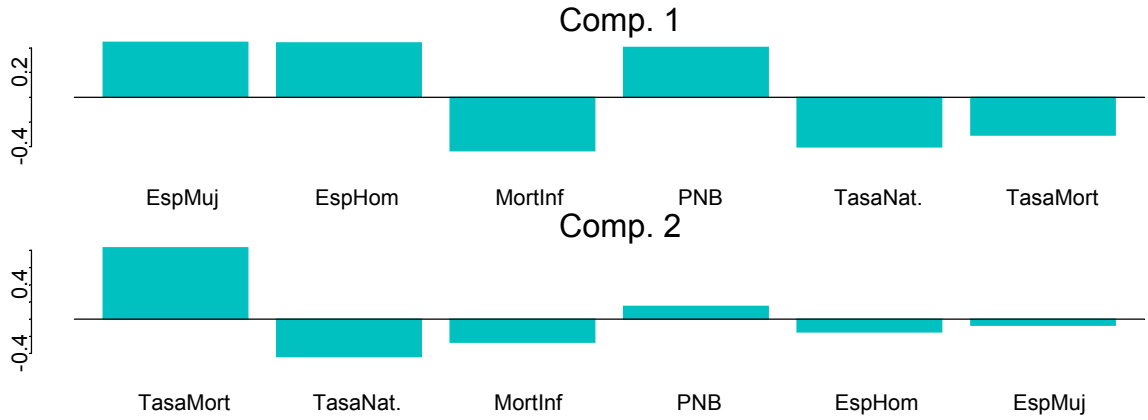


Figura 5.15: Pesos de las variables en los dos primeros componentes para los datos de MUNDODES

La figura 5.16 presenta la representación de los paises en los dos primeros componentes. El primero es una medida del desarrollo y el segundo depende principalmente de la tasa de mortalidad, y separa paises que tienen alto (o bajo) valor aparente de desarrollo de otros que tienen una mortalidad mucho mayor de la que correspondería de acuerdo a su nivel de desarrollo. Ambas dimensiones estan incorreladas pero no son independientes, como se observa en la figura. Sin embargo, el grado de dependencia entre las variables es menor que con las variables sin transformar.

Figura 5.16: Grafico de los datos de Mundodes sobre los dos primeros componentes principales de los datos en logaritmos.

5.7 Generalizaciones

La idea de componentes principales puede extenderse para buscar representaciones no lineales de los datos que expliquen su estructura. Este enfoque es especialmente interesante si sospechamos que los datos pueden disponerse siguiendo una determinada superficie en el espacio. Como hemos visto los vectores propios ligados a valores propios próximos a cero sonmuy importantes porque revelan relaciones de poca variabilidad de los datos. Por ejemplo, supongamos para simplificar una variable bidimensional donde, aproximadamente, $f(x_1) + f(x_2) = c$. Entonces, si hacemos componentes principales de las cuatro variables $(x_1, x_2, f(x_1), f(x_2),)$ encontraremos un valor propio muy próximo a cero con un vector propio de la forma (0,0, 1,1).

Generalizando esta idea, si existe una relación cualquiera no lineal entre las variables, como esta relación podemos aproximarla por una relación polinómica

$$f(x_1, ..., x_p) = \sum a_i x_i + \sum b_{ij} x_i x_j + \sum c_{ijk} x_i x_j x_k + ...$$

si incluimos nuevas variables adicionales como $x_1^2, ..., x_p^2$ o productos de variables $x_1 x_2$ etc y extraemos los componentes principales de la matriz de correlaciones entre todas estas variables, si los puntos tienen una relación no lineal esta se detectará ligada a un valor propio próximo a cero. Este enfoque se conoce a veces como componentes principales generalizados, y el lector interesado puede encontrar ejemplos de su aplicación en Gnandesikan (1977). El inconveniente de introducir nuevas variables, transformaciones de las iniciales, es que inmediatamente aumenta mucho la dimensión del problema con lo que si la muestra no es muy grande podemos tener una matriz de correlaciones singular. Por otro lado la interpretación de los resultados de este análisis, salvo en casos muy especiales, no suele ser fácil, con lo que esta herramienta no suele ayuda mucho en para la exploración de datos multivariantes.

5.8 Lecturas complementarias

Todos los textos generales de análisis multivariante que se indican en las referencias estudian componentes principales. Johnson y Wichern (1998) y Rechner (1998) son buenas presentaciones con similar filosofía a la utilizada en el libro mientras que Flury (1997) presenta un enfoque distinto al aquí expuesto. Componentes principales es un caso particular de los métodos de proyección introducidos en la sección 4.2.3 que se conocen como *Projection Pursuit* (Búsqueda de la Proyección). Véase Krzanowski y Marriot (1994) para más detalles. Un excelente tratado sobre componentes principales y sus extensiones es el libro de Jackson (1991), que contiene numerosas referencias. La idea de componentes principales puede extenderse al caso no lineal, y Gnanadesikan (1997) es una buena referencia. Los componentes principales puede aplicarse para investigar si varios grupos de datos tienen componentes comunes. Este aspecto ha sido investigado por Krzanowski (1979) y Flury (1984, 1986). Cuadras, C.M. (1991) y Aluja, T. y Morineau, A. (1999) son buenas referencias en español.

EJERCICIOS

Ejercicio 5.1 *Dada la matriz de covarianzas*

$$\mathbf{S} = \begin{bmatrix} 1+d & 1 & 1 \\ 1 & 1+d & 1 \\ 1 & 1 & 1+d \end{bmatrix}$$

encontrar los componentes principales. Calcular la proporción de variabilidad explicada por cada uno y las correlaciones entre los componentes y las variables. Interpretar los componentes en función del tamaño de d.

Ejercicio 5.2 *Dada la matriz de correlación:*

$$S = \begin{bmatrix} 1 & d & d & d \\ d & 1 & d & d \\ d & d & 1 & d \\ d & d & d & 1 \end{bmatrix}$$

encontrar la primera componente principal. (Nota, utilizar que $\sum = [d.1.1' + (1-d)I]$ *para encontrar los componentes y discutir su interpretación).*

Ejercicio 5.3 *Supongamos que* $Z, X_1, ..., X_p$ *tienen una distribución normal* $(p+1)$ *dimensional. Sean* $Y_1, ..., Y_p$ *los componentes principales de* $X_1, ..., X_p$*. Demostrar que el coeficiente de correlación múltiple de las regresiones:*

$$\begin{aligned} Z &= \sum a_i X_i \\ Z &= \sum b_i Y_i \end{aligned}$$

es idéntico.

Ejercicio 5.4 *Demostrar que si* $\mathbf{S} = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$*, donde A y B son no singulares de rango* r_A *y* r_B *los vectores propios de* $\mathbf{S}$ *son de la forma* $(u_1, 0)$ *y* $(0, u_2)$*, donde* u_1 *es un vector propio de A y* u_2 *un vector propio de B.*

Ejercicio 5.5 *Indicar las implicaciones del resultado del ejercicio 5.4 para calcular componentes principales.*

Ejercicio 5.6 *Demostrar que si* $S = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ *los valores propios de S son los de A más los de B.*

Ejercicio 5.7 *Demostrar que el espacio que maximiza la varianza generalizada de la proyección es el definido por* $z_1 = Xa_1$ *y* $z_2 = Xa_2$ *donde* z_1 *y* z_2 *son los dos primeros componentes principales.*

Ejercicio 5.8 *Demostrar que si una variable x_1 está incorrelada con el resto de manera que la matriz* $\mathbf{S}$ *tiene la forma* $\mathbf{S} = \begin{bmatrix} s_1^2 & \mathbf{0}' \\ 0 & S_2 \end{bmatrix}$ *donde* 0 *y* $\mathbf{0}'$ *son vectores de ceros, la matriz* $\mathbf{S}$ *tiene un componente principal asociado únicamente a la primera variable, es decir, el vector (1, 0...0) es un vector propio de* $\mathbf{S}$.

Ejercicio 5.9 *Demostrar que la dirección donde la variabilidad de la proyección es mínima es la dada por el vector propio ligado al menor valor propio de la matriz de covarianzas.*

Ejercicio 5.10 *Demostrar la siguiente acotación para formas cuadráticas :* $\lambda_{\min} w'w \leq w'Bw \leq \lambda_{\max} w'w$, *donde* $\lambda_{\min}$ *y* $\lambda_{\max}$ *son el menor y el mayor valor propio de la matriz* B. *(Sugerencia, maximizar la forma cuadrática como se hizo para obtener el primer componente principal)*

APÉNDICE 5.1. DISTANCIAS ENTRE PUNTOS Y PROYECCIONES

Vamos a demostrar que maximizar las distancias al cuadrado entre los puntos proyectados equivale a maximizar la varianza de la variable definida por las proyecciones de los puntos. Sea $z_i = \mathbf{a}_1'\mathbf{x}_i$ la proyección de una observación sobre la dirección $\mathbf{a}_1$, donde suponemos $\mathbf{a}_1'\mathbf{a}_1 = 1$. La variable z_i tendrá media cero ya que si las x tienen media cero $\sum_{i=1}^n z_i = \sum_{i=1}^n \mathbf{a}_1'\mathbf{x}_i = \mathbf{a}_1' \sum_{i=1}^n \mathbf{x}_i = 0$. La suma de las distancias al cuadrado entre los puntos proyectados es

$$D_p = \sum_{i=1}^{n} \sum_{h=i+1}^{n} (z_i - z_h)^2.$$

Para interpretar este sumatorio observemos que cada término z_i aparece al cuadrado $n-1$, veces ya que cada punto se compara con los otros $n-1$, y que habrá tantos dobles productos como parejas de puntos, es decir $\binom{n}{2} = n(n-1)/2$. Por tanto:

$$D_p = (n-1)\sum_{i=1}^{n} z_i^2 - 2\sum_{i=1}^{n}\sum_{h=i+1}^{n} z_i z_h = n\sum z_i^2 - B$$

siendo B :

$$B = \sum_{i=1}^{n} z_i^2 + 2\sum_{i=1}^{n}\sum_{h=i+1}^{n} z_i z_h$$

que puede escribirse,

$$\begin{aligned} B &= z_1(z_1 + z_2 + \ldots z_n) + z_2(z_1 + \ldots + z_n) + \ldots z_n(z_1 + \ldots + z_n) \\ &= \sum_{i=1}^{n} z_i \sum_{i=1}^{n} z_i = 0. \end{aligned}$$

Por tanto, maximizar las distancias entre los puntos equivale a maximizar:

$$A = n \sum z_i^2$$

que es el criterio de maximizar la varianza de la nueva variable, obtenida anteriormente.

Algunos autores han propuesta minimizar

$$\sum \sum w_{ij}(d_{ij} - \widehat{d}_{ij})^2$$

donde w_{ij} es una función de ponderación. El problema así planteado no tiene una solución simple y debe resolverse mediante un algoritmo iterativo no lineal. Veáse por ejemplo Krzanowski (1990, cap2).

APÉNDICE 5.2. LOS COMPONENTES COMO PREDICTORES ÓPTIMOS

Demostraremos que los componentes principales son predictores óptimos de las $\mathbf{X}$. Comencemos demostrando que si queremos aproximar la matriz $\mathbf{X}$, de rango p, por otra matriz $\widehat{\mathbf{X}}_r$ de rango $r < p$, la aproximación óptima es $\mathbf{X}\mathbf{A}_r\mathbf{A}_r' = \mathbf{Z}_r\mathbf{A}_r'$, donde la matriz $\mathbf{A}_r$ es $p \times r$ y sus columnas son los vectores propios asociados a los r mayores valores propios de la matriz $\mathbf{S}$.

El problema de aproximar la matriz $\mathbf{X}$ puede establecerse así: Consideremos un espacio de dimensión r definido por una base $\mathbf{U}_r$ ortonormal, donde $\mathbf{U}_r$ es $p \times r$ y $\mathbf{U}_r'\mathbf{U}_r = \mathbf{I}$. Se desea encontrar una aproximación de la matriz $\mathbf{X}$ utilizando una base de ese espacio, es decir, queremos prever cada una de las filas $(\mathbf{x}_1, ..., \mathbf{x}_n)$ de la matriz, donde $\mathbf{x}_i$ es el vector $p \times 1$ de observaciones en el elemento i de la muestra, mediante los vectores $\mathbf{U}_r$. La predicción de la variable $\mathbf{x}_i$ será la proyección ortogonal sobre el espacio generado por estos vectores que es

$$\widehat{x}_i = \mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i$$

y queremos determinar los vectores $\mathbf{U}_r$ tal que el error cuadrático de aproximación total para todas las filas de la matriz, dado por

$$E = \sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \widehat{x}_{ij})^2 = \sum_{i=1}^{n}(\mathbf{x}_i - \widehat{\mathbf{x}}_i)'(\mathbf{x}_i - \widehat{\mathbf{x}}_i) \tag{5.17}$$

sea mínimo. El error puede escribirse

$$E = \sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i - \sum_{i=1}^{n}\mathbf{x}_i'\mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i \tag{5.18}$$

y minimizar el error equivale a maximizar el segundo término. Utilizando que un escalar es igual a su traza, $\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i = tr(\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i) = \sum_{i=1}^{n}tr(\mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i\mathbf{x}_i') = tr(\mathbf{U}_r\mathbf{U}_r'\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i')$. Introduciendo que $\mathbf{S} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'/n$ y sustituyendo en $tr(\mathbf{U}_r\mathbf{U}_r'\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i')$, tenemos que esta expresión es $ntr(\mathbf{U}_r\mathbf{U}_r'\mathbf{S}) = ntr(\mathbf{U}_r'\mathbf{S}\mathbf{U}_r)$. Por tanto:

$$\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{U}_r\mathbf{U}_r'\mathbf{x}_i = ntr(\mathbf{U}_r'\mathbf{S}\mathbf{U}_r) \tag{5.19}$$

Según esta expresión, minimizar el error (5.18) implica encontrar un conjunto de vectores $\mathbf{U}_r = [\mathbf{u}_1, ..., \mathbf{u}_r]$ que maximicen la suma de los elementos diagonales de $\mathbf{U}_r'\mathbf{S}\mathbf{U}_r$, es decir, $\sum_{j=1}^{r} \mathbf{u}_j'\mathbf{S}\mathbf{u}_j$. Si $r = 1$, este es el problema que se ha resuelto para encontrar el primer componente. Si $r = 2$, como el nuevo vector debe ser ortogonal al primero, obtenemos el segundo componente, y así sucesivamente. Por tanto, $\mathbf{U}_r = \mathbf{A}_r$, y la aproximación óptima a la matriz $\mathbf{X}$ vendrá dada por $\widehat{\mathbf{X}}_r = \mathbf{X}\mathbf{A}_r\mathbf{A}_r'$. Además, como en (5.18) el primer término es

$$\begin{aligned} \sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x} &= \mathbf{tr}(\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}) = \sum_{i=1}^{n} tr(\mathbf{x}_i'\mathbf{x}) = \\ tr \sum_{i=1}^{n} (\mathbf{x}\mathbf{x}_i') &= ntr(\mathbf{S}) = n \sum_{i=1}^{p} \lambda_i \end{aligned}$$

y el segundo es, según (5.19), igual a $n \sum_{i=1}^{r} \lambda_i$, tenemos que el error de la aproximación será $n \sum_{i=r+1}^{p} \lambda_i$.

Es interesante señalar que esta aproximación a una matriz es la que proporciona la descomposición en valores singulares, es decir la mejor aproximación a la matriz $\mathbf{X}$ por otra matriz $\widehat{\mathbf{X}}_r$ de rango $r < p$ es

$$\widehat{\mathbf{X}}_r = \mathbf{U}_r\mathbf{D}_r^{1/2}\mathbf{V}_r' = \sum_{i=1}^{r} \lambda_1^{1/2}\mathbf{u}_i\mathbf{v}_i'$$

donde $\mathbf{U}_r$ es la matriz de los r mayores vectores propios de $\mathbf{X}\mathbf{X}'$, $\mathbf{D}_r^{1/2}$ contiene los r mayores valores propios y $\mathbf{V}_r$ contiene los vectores propios de $\mathbf{X}'\mathbf{X}$·. En efecto, según hemos visto en la sección 5.7 $\widehat{\mathbf{X}}_r = \mathbf{Z}_r\mathbf{A}_r'$, que es el resultado anterior.

El problema puede enfocarse desde otro punto de vista. Busquemos unas variables $[\mathbf{z}_1, ..., \mathbf{z}_r]$ que sean combinaciones lineales de las originales y que tengan la propiedad de preverlas de manera óptima. Por ejemplo, si $r = 1$, buscamos un vector $\mathbf{a}_1$ de manera que la nueva variable:

$$\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1$$

permita prever con mínimo error los valores observados para el conjunto de variables que forman las columnas de la matriz $\mathbf{X}$. Por ejemplo, el valor previsto para la variable x_j en el individuo i, $\widehat{x}_{ij}$, conocido el valor de la variable z_1 para ese individuo, z_{1i} será:

$$\widehat{x}_{ij} = b_j z_{1i}$$

y el error de predicción será $\mathrm{e}_{ij} = x_{ij} - \widehat{x}_{ij}$. Vamos a demostrarlo para simplificar en el caso $r = 1$. Calcularemos el vector $\mathbf{a}_1$ para que minimice estos errores de predicción. Es conocido que el coeficiente de regresión b_j viene dado por:

$$b_j = \frac{\sum_{i=1}^{n} x_{ij} z_{1i}}{\sum z_{1i}^2} \tag{5.20}$$

como $1/n \sum z_{1i}^2 = 1/n\mathbf{a'X'Xa} = \mathbf{a'Sa}$, la varianza de z_1 puede crecer indefinidamente si no imponemos ninguna restricción. Exigiremos que sea unitaria, es decir que:

$$\mathbf{a'Sa} = 1 = (1/n) \sum z_{1i}^2 \tag{5.21}$$

Entonces:

$$b_j = 1/n \sum x_{ij} z_{1i} = 1/n \mathbf{X}'_j \mathbf{X} \mathbf{a}_1 = \mathbf{V}'_j \mathbf{a}_1 \tag{5.22}$$

donde $\mathbf{V}_j$ es el vector fila j de la matriz $\mathbf{S}$ de varianzas y covarianzas. Impongamos la condición mínimo cuadrática para obtener $\mathbf{a_1}$:

$$\frac{1}{n} \sum_{i=1}^{n} e_{ij}^2 = Mínimo = \frac{1}{n} \sum_{i=1}^{n} \left(x_{ij} - \mathbf{V}'_j \mathbf{a}_1 z_{1i} \right)^2$$

y el segundo miembro puede escribirse:

$$\frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 + \frac{1}{n} \mathbf{a}'_1 \mathbf{V}_j \mathbf{V}'_j \mathbf{a}_1 \sum_{i=1}^{n} z_{1i}^2 - 2\mathbf{V}'_j \mathbf{a} \frac{1}{n} \sum_{i=1}^{n} x_{ij} z_{1i}$$

utilizando ahora (5.21) y (5.22), se obtiene

$$\frac{1}{n} \sum_{i=1}^{n} e_{ij}^2 = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 - \mathbf{a}'_1 \mathbf{V}_j \mathbf{V}'_j \mathbf{a}_1.$$

Aplicando este mismo razonamiento a las otras variables X y sumando para todas ellas:

$$M = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} e_{ij}^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2 - \sum_{j=1}^{p} \mathbf{a}'_1 \mathbf{V}_j \mathbf{V}'_j \mathbf{a}_1$$

como el primer miembro es la traza de $\mathbf{S}$ que es fija, maximizar M equivale a minimizar:

$$\mathbf{a}'_1 \sum_{j=1}^{p} \mathbf{V}_j \mathbf{V}'_j \mathbf{a}_1 = \mathbf{a}'_1 \mathbf{S} \mathbf{S}' \mathbf{a}'_1 = \mathbf{a}'_1 \mathbf{S}^2 \mathbf{a}_1 \tag{5.23}$$

ya que $\mathbf{S}$ es simétrica. Por lo tanto, el problema es minimizar la expresión (9.14) con la restricción (5.21):

$$L = \mathbf{a}'_1 \mathbf{S}^2 \mathbf{a}_1 - \lambda(\mathbf{a}_1 \mathbf{S} \mathbf{a}_1 - 1)$$

$$\frac{\partial L}{\partial \mathbf{a}} = 2\mathbf{S}^2\mathbf{a} - \lambda 2\mathbf{S}\mathbf{a} = 0$$

$$\mathbf{S}^2\mathbf{a} = \lambda\mathbf{S}\mathbf{a}$$

de donde incluimos que a debe de ser un vector propio de $\mathbf{S}$ y λ un valor propio, ya que si:

$$\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$$

multiplicando por $\mathbf{S}$

$$\mathbf{S}^2\mathbf{a} = \lambda\mathbf{S}\mathbf{a}$$

Con lo que finaliza la demostración. Es interesante resaltar que este resultado es simplemente la implicación estadística de la propiedad que tienen los vectores y raíces característicos de "generar" la matriz de base.

Capítulo 6

ESCALADO MULTIDIMENSIONAL

6.1 INTRODUCCIÓN

Las técnicas de escalado multidimensional son una generalización de la idea de componentes principales cuando en lugar de disponer de una matriz de observaciones por variables, como en componentes principales, se dispone de una matriz, $\mathbf{D}$, cuadrada $n \times n$ de distancias o disimilaridades entre los n elementos de un conjunto. Por ejemplo, esta matriz puede representar las similitudes o distancias entre n productos fabricados por una empresa, las distancias percibidas entre n candidatos políticos, las diferencias entre n preguntas de un cuestionario o las distancias o similitudes entre n sectores industriales. Estas distancias pueden haberse obtenido a partir de ciertas variables, o pueden ser el resultado de una estimación directa, por ejemplo preguntando a un grupo de jueces por sus opiniones sobre las similaridades entre los elementos considerados.

El objetivo que se pretende es representar esta matriz mediante un conjunto de variables ortogonales $y_1, \ldots, y_p$, donde $p < n$, de manera que las distancias euclídeas entre las coordenadas de los elementos respecto a estas variables sean iguales (o lo más próximas posibles) a las distancias o disimilaridades de la matriz original. Es decir, a partir de la matriz $\mathbf{D}$ se pretende obtener una matriz $\mathbf{X}$, de dimensiones $n \times p$, que pueda interpretarse como la matriz de p variables en los n individuos, y donde la distancia euclídea entre los elementos reproduzca, aproximadamente, la matriz de distancias $\mathbf{D}$ inicial. Cuando $p > 2$, las variables pueden ordenarse en importancia y suelen hacerse representaciones gráficas en dos y tres dimensiones para entender la estructura existente.

Este planteamiento presenta dos interrogantes: ¿Es siempre posible encontrar estas variables? ¿Cómo construirlas? En general *no* es posible encontrar p variables que reproduzcan *exactamente* las distancias iniciales, sin embargo es frecuente encontrar variables que reproduzcan aproximadamente las distancia iniciales. Por otro lado, si la matriz de distancias se ha generado calculando las distancias euclídeas entre las observaciones definidas por ciertas variables, recupereraremos las componentes principales de estas variables.

El escalado multidimensional comparte con componentes principales el objetivo de describir e interpretar los datos. Si existen muchos elementos, la matriz de similaridades será muy grande y la representación por unas pocas variables de los elementos nos permitirá entender su estructura: qué elementos tienen propiedades similares, si aparecen grupos entre

los elementos, si hay elementos atípicos, etc. Además, si podemos interpretar las variables aumentará nuestro conocimiento del problema, al entender cómo se han generado los datos. Por ejemplo, supongamos que se realiza una encuesta para determinar que similitudes encuentran los consumidores entre n productos o servicios, y que la información se resume en una matriz cuadrada de similitudes entre los productos. Supongamos que descubrimos que estas similitudes pueden generarse por dos variables. Entonces, es razonable suponer que los consumidores han estimado la similitud entre los productos utilizando estas dos variables.

El escalado multidimensional representa un enfoque complementario a componentes principales en el sentido siguiente. Componentes principales considera la matriz $p \times p$ de correlaciones (o covarianzas) entre variables, e investiga su estructura. El escalado multidimensional considera la matriz $n \times n$ de correlaciones (o covarianzas) entre individuos, e investiga su estructura. Ambos enfoques están claramente relacionados, y existen técnicas gráficas, como el biplot que estudiaremos en este capítulo, que aprovechan esta dualidad para representar conjuntamente las variables y los individuos en un mismo gráfico.

El escalado multidimensional (Multidimensional Scaling) tiene sus orígenes en los estudios de psicología experimental, en los años 50, para descubrir la similaridad entre estímulos aplicados a distintos individuos. Su desarrollo actual es debido a las investigaciones de Torgerson, Shepard, Kruskal y Gower, entre otros, y se han aplicado, preferentemente, en las ciencias sociales. Los métodos existentes se dividen en métricos, cuando la matriz inicial es propiamente de distancias, y no métricos, cuando la matriz es de similaridades. Los métodos métricos, también llamados coordenadas principales, utilizan las diferencias entre similitudes mientras que los no métricos parten de que si A es más similar a B que a C, entonces A esta más cerca de B que de C, pero las diferencias entre las similitudes AB y AC no tienen interpretación.

6.2 ESCALADOS MÉTRICOS: COORDENADAS PRINCIPALES

6.2.1 Construcción de variables a partir de las distancias

Vimos en el Capítulo 3 que dada una matriz $\mathbf{X}$ de individuos por variables obtenemos variables con media cero mediante la operación:

$$\widetilde{\mathbf{X}} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X} = \mathbf{P}\mathbf{X}$$

A partir de esta matriz $\widetilde{\mathbf{X}}$, de variables con media cero y dimensiones $n \times p$, podemos construir dos tipos de matrices cuadradas y semidefinidas positivas: la matriz de covarianzas, $\mathbf{S}$, definida por $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}/n$ y la matriz de productos cruzados, $\mathbf{Q} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$, que vamos a ver que puede interpretarse como una matriz de similitud (covarianzas) entre los n elementos. En efecto, los términos de esta matriz, q_{ij}, contienen el producto escalar por pares de elementos:

$$q_{ij} = \sum_{s=1}^{p} x_{is}x_{js} = \mathbf{x}_i'\mathbf{x}_j, \tag{6.1}$$

donde hemos llamado $\mathbf{x}_i'$ a la fila i de la matriz $\widetilde{\mathbf{X}}$. Por la expresión del producto escalar, $q_{ij} = |\mathbf{x}_i|\,|\mathbf{x}_j|\cos\theta_{ij}$, si los dos elementos tienen coordenadas similares, $\cos\theta_{ij} \simeq 1$ y q_{ij} será grande. Por el contrario, si los dos elementos son muy distintos, $\cos\theta_{ij} \simeq 0$ y q_{ij} será pequeño. En este sentido podemos interpretar la matriz $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$ como la matriz de similitud entre elementos.

Las distancias entre las observaciones se deducen inmediatamente de esta matriz de similitud. La distancia euclídea al cuadrado entre dos elementos es:

$$d_{ij}^2 = \sum_{s=1}^{p}(x_{is} - x_{js})^2 = \sum_{s=1}^{p} x_{is}^2 + \sum_{s=1}^{p} x_{js}^2 - 2\sum_{s=1}^{p} x_{is}x_{js} \tag{6.2}$$

que puede calcularse en función de los términos de la matriz $\mathbf{Q}$, por la expresión

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}. \tag{6.3}$$

Por tanto, dada la matriz $\widetilde{\mathbf{X}}$ podemos construir la matriz de similitud $\mathbf{Q} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$ y, a partir de ella, la matriz $\mathbf{D}$ de distancias al cuadrado entre elementos con ayuda de (6.3). Llamando $diag(\mathbf{Q})$ al vector que contiene los términos diagonales de la matriz $\mathbf{Q}$, y $\mathbf{1}$ al vector de unos, la matriz $\mathbf{D}$ viene dada por

$$\mathbf{D} = diag(\mathbf{Q})\mathbf{1}' + \mathbf{1}diag(\mathbf{Q})' - 2\mathbf{Q}$$

El problema que vamos a abordar es el inverso: reconstruir la matriz $\widetilde{\mathbf{X}}$ a partir de una matriz de distancias al cuadrado, $\mathbf{D}$, con elementos d_{ij}^2. Para ello, obtendremos primero la matriz $\mathbf{Q}$, y a continuación la $\widetilde{\mathbf{X}}$.

Comencemos estudiando cómo obtener la matriz $\mathbf{Q}$ dada la matriz $\mathbf{D}$. En primer lugar, observemos que no hay pérdida de generalidad en suponer que las variables tienen media cero. Esto es consecuencia de que las distancias entre dos puntos, d_{ij}^2 no varían si expresamos las variables en desviaciones a la media, ya que

$$d_{ij}^2 = \sum_{s=1}^{p}(x_{is} - x_{js})^2 = \sum_{s=1}^{p}[(x_{is} - \overline{x}_s) - (x_{js} - \overline{x}_s)]^2. \tag{6.4}$$

Dado que estamos suponiendo que la única información existente son las distancias entre elementos, para resolver esta indeterminación vamos a buscar una matriz $\widetilde{\mathbf{X}}$ con variables de media cero. En consecuencia, como $\widetilde{\mathbf{X}}'\mathbf{1} = 0$ también $\mathbf{Q1} = 0$, es decir, la suma de todos los elementos de una fila de la matriz de similitudes, $\mathbf{Q}$, (y de una columna ya que la matriz es simétrica) debe de ser cero. Para imponer estas restricciones, sumemos en (6.3) por filas:

$$\sum_{i=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} q_{ii} + n q_{jj} = t + n q_{jj} \tag{6.5}$$

donde $t = \sum_{i=1}^{n} q_{ii}$ = traza ($\mathbf{Q}$), y hemos utilizado que la condición $\mathbf{Q1} = 0$ implica $\sum_{i=1}^{n} q_{ij} = 0$. Sumando (6.3) por columnas

$$\sum_{j=1}^{n} d_{ij}^2 = t + n q_{ii} \tag{6.6}$$

y sumando ahora (6.5) por filas de nuevo

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2nt. \tag{6.7}$$

Sustituyendo en (6.3) q_{jj} obtenida en (6.5) y q_{ii} en (6.6), tenemos que

$$d_{ij}^2 = \frac{1}{n} \sum_{i=1}^{n} d_{ij}^2 - \frac{t}{n} + \frac{1}{n} \sum_{j=1}^{n} d_{ij}^2 - \frac{t}{n} - 2q_{ij}, \tag{6.8}$$

y llamando $d_{i.}^2 = \frac{1}{n} \sum_{j=1}^{n} d_{ij}^2$ y $d_{.j}^2 = \frac{1}{n} \sum_{i=1}^{n} d_{ij}^2$, a las medias por filas y por columnas y utilizando (6.7), tenemos que

$$d_{ij}^2 = d_{i.}^2 + d_{.j}^2 - d_{..}^2 - 2q_{ij}. \tag{6.9}$$

donde $d_{..}^2$ es la media de todos los elementos de $\mathbf{D}$, dada por

$$d_{..}^2 = \frac{1}{n^2} \sum \sum d_{ij}^2,$$

Finalmente, de (6.9) resulta que

$$q_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2) \tag{6.10}$$

expresión que indica cómo construir la matriz de similitud $\mathbf{Q}$ a partir de la matriz $\mathbf{D}$ de distancias.

Pasemos ahora al problema de obtener la matriz $\mathbf{X}$ dada la matriz $\mathbf{Q}$. Suponiendo que la matriz de similitud es definida positiva de rango p, puede representarse por

$$\mathbf{Q} = \mathbf{V} \Lambda \mathbf{V}'$$

donde $\mathbf{V}$ es $n \times p$ y contiene los vectores propios correspondientes a valores propios no nulos de $\mathbf{Q}$, Λ es diagonal $p \times p$ y contiene los valores propios y $\mathbf{V}'$ es $p \times n$. Escribiendo:

$$\mathbf{Q} = (\mathbf{V}\Lambda^{1/2})(\Lambda^{1/2}\mathbf{V}') \tag{6.11}$$

y tomando

$$\mathbf{Y} = \mathbf{V}\Lambda^{1/2}$$

hemos obtenido una matriz $n \times p$ con p variables incorreladas que reproducen la métrica inicial. Observemos que si partimos de unas variables $\mathbf{X}$ y calculamos a partir de estas variables la matriz de distancias con (6.2) y luego aplicamos el método descrito a esta matriz de distancias no obtendremos las variables originales, $\mathbf{X}$, sino sus componentes principales. Esto es inevitable, ya que existe una indeterminación en el problema cuando la única información disponible son las distancias. En efecto, las distancias entre elementos no varían si:

(1) modificamos las medias de las variables

(2) rotamos los puntos, es decir multiplicamos por una matriz ortogonal.

Las distancias son función, por (6.3) de los términos de la matriz de similitud, $\mathbf{Q}$, y esta matriz es invariante ante rotaciones de las variables. En efecto:

$$\mathbf{Q} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}' = \widetilde{\mathbf{X}}\mathbf{A}\mathbf{A}'\widetilde{\mathbf{X}}'$$

para cualquier matriz $\mathbf{A}$ ortogonal. La matriz $\mathbf{Q}$ sólo contiene información sobre el espacio generado por las variables $\mathbf{X}$. Cualquier rotación preserva las distancias. En consecuencia, cualquier rotación de las variables originales podría ser solución.

6.3 Matrices compatibles con métricas euclídeas

Para poder calcular la raíz cuadrada de la matriz de similitud mediante (6.11) es necesario que los valores propios de la matriz $\mathbf{Q}$, que construimos a partir de la matriz $\mathbf{D}$ original, sean no negativos. Dada una matriz de distancias, $\mathbf{D}$, diremos que ésta matriz es compatible con una métrica euclídea si la matriz de similitud que se obtiene a partir de ella

$$\mathbf{Q} = -\frac{1}{2}\mathbf{PDP}$$

es semidefinida positiva, donde $\mathbf{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$.

Vamos a demostrar que esta condición es necesaria y suficiente, es decir, si $\mathbf{D}$ se ha construido a partir de una métrica euclídea $\mathbf{Q}$ es no negativa y si $\mathbf{Q}$ es no negativa es posible encontrar una métrica euclídea que reproduzca $\mathbf{D}$.

Demostración

Demostraremos primero que si $\mathbf{D}$ se ha construido a partir de una métrica euclídea $\mathbf{Q}$ es no negativa. Para ello comprobaremos en primer lugar que la matriz $-\frac{1}{2}\mathbf{PDP}$ tiene los términos (6.10). En efecto, los términos de la matriz $\mathbf{Q}$ serán

$$\mathbf{Q} = -\frac{1}{2}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{D}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}') = -\frac{1}{2}(\mathbf{D} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{D} - \frac{1}{n}\mathbf{D}\mathbf{1}\mathbf{1}' + \frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{D}\mathbf{1}\mathbf{1}') \tag{6.12}$$

y llamando q_{ij} a los elementos de $\mathbf{Q}$:

$$q_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2). \tag{6.13}$$

Vamos a comprobar ahora que $\mathbf{Q}$ puede expresarse como $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$ y por tanto es semidefinida positiva. Como ahora, por hipótesis, los términos d_{ij}^2 son los cuadrados de distancias euclídeas, por (6.2), podemos escribir

$$d_{.j}^2 = \frac{1}{n}\sum_i\sum_s x_{is}^2 + \sum_s x_{js}^2 - \frac{2}{n}\sum_i\sum_s x_{is}x_{js}$$

$$d_{i.}^2 = \sum_s x_{is}^2 + \frac{1}{n}\sum_j\sum_s x_{js}^2 - \frac{2}{n}\sum_j\sum_s x_{is}x_{js}$$

$$d_{..}^2 = -\frac{1}{n}\sum_i\sum_s x_{is}^2 + \frac{1}{n}\sum_j\sum_s x_{js}^2 - \frac{1}{n^2}\sum_i\sum_j\sum_s x_{is}x_{js}.$$

Como

$$\frac{1}{n}\sum_s(\sum_i x_{is})x_{js} = \sum_s \overline{x}_s x_{js} \tag{6.14}$$

$$\frac{1}{n^2}\sum_s(\sum_i x_{is})(\sum_j x_{js}) = \sum_s \overline{x}_s^2 \tag{6.15}$$

se verifica que

$$q_{ij} = \sum_p x_{pi}x_{pj} - \sum_p \overline{x}_p x_{pj} - \sum_p \overline{x}_p x_{pi} + \sum_p \overline{x}_p^2 = (\mathbf{x}_i - \overline{\mathbf{x}})'(\mathbf{x}_j - \overline{\mathbf{x}}) \tag{6.16}$$

y, por tanto, en general

$$Q = \begin{bmatrix} (\mathbf{x}_1 - \overline{\mathbf{x}})' \\ \vdots \\ (\mathbf{x}_n - \overline{\mathbf{x}})' \end{bmatrix} [(\mathbf{x}_1 - \overline{\mathbf{x}}) \dots (\overline{\mathbf{x}}_n - \overline{\mathbf{x}})] = \tilde{\mathbf{X}}\tilde{\mathbf{X}}', \tag{6.17}$$

que es siempre semidefinida positiva, de rango p.

Vamos a demostrar ahora que si $\mathbf{Q}$ es semidefinida positiva podemos encontrar ciertas variables, $y_1, \dots, y_p$, que reproduzcan las distancias observadas. Si $\mathbf{Q}$ es semidefinida positiva de rango p podemos expresarla como:

$$\mathbf{Q} = \sum^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i'$$

donde λ_i son sus valores propios y $\mathbf{v}_i$ los vectores propios. Llamando $\mathbf{y}_i = \sqrt{\lambda_i}\mathbf{v}_i$ a la estandarización de los vectores propios para que tengan varianza unidad, podemos escribir

$$\mathbf{Q} = \sum \mathbf{y}_i \mathbf{y}_i' \tag{6.18}$$

Las variables $\mathbf{y}_i$ representan la solución buscada: son un conjunto de p variables n-dimensionales incorreladas entre sí y tales que el cuadrado de la distancia euclídea que inducen entre dos puntos es:

$$\delta_{ij}^2 = (\mathbf{z}_i - \mathbf{z}_j)'(\mathbf{z}_i - \mathbf{z}_j) \tag{6.19}$$

donde $\mathbf{z}_i' = (y_{i1}, \dots, y_{ip})$ es igual a las distancias originales observadas d_{ij}^2. Para demostrarlo observemos que (6.18) implica que la matriz cuadrada de similitud $\mathbf{Q}$ puede también escribirse:

$$\mathbf{Q} = [\mathbf{y}_1, \dots, \mathbf{y}_p] \begin{bmatrix} \mathbf{y}_1' \\ \vdots \\ \mathbf{y}_p' \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1' \\ \vdots \\ \mathbf{z}_n' \end{bmatrix} [\mathbf{z}_1, \dots, \mathbf{z}_n]$$

donde estamos llamando $\mathbf{y}$ a las p variables n-dimensionales y $\mathbf{z}$ al vector de dimensión p formado por los valores de estas variables en un individuo de la población. Entonces:

$$q_{ij} = \mathbf{z}_i' \mathbf{z}_j \tag{6.20}$$

La distancia al cuadrado entre dos puntos es, por (6.19)

$$\delta_{ij}^2 = \mathbf{z}_i' \mathbf{z}_i + \mathbf{z}_j' \mathbf{z}_j - 2\mathbf{z}_i' \mathbf{z}_j$$

y, por (6.20)

$$\delta_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij},$$

y como esta expresión es idéntica a (6.3), concluimos que:

$$\delta_{ij}^2 = d_{ij}^2$$

y las nuevas variables reproducen exactamente las distancias euclídeas.

6.3.1 Construcción de las Coordenadas Principales

En general la matriz de distancias no será compatible con una métrica euclídea, pero es frecuente que la matriz de similitud obtenida a partir de ella tenga p valores propios positivos y más grandes que el resto. Si los restantes $n-p$ valores propios no nulos son mucho menores que los demás, podemos obtener una representación aproximada de los puntos utilizando los p vectores propios asociados a valores propios positivos de la matriz de similitud. En este caso, las representaciones gráficas conservarán sólo aproximadamente la distancia entre los puntos.

Supongamos que tenemos una matriz de distancias al cuadrado $\mathbf{D}$. El procedimiento para obtener las coordenadas principales es:

1. Construir la matriz $\mathbf{Q} = -\frac{1}{2}\mathbf{PDP}$, de productos cruzados.

2. Obtener los valores propios de $\mathbf{Q}$. Tomar los r mayores valores propios, donde r se escoge de manera que los restantes $n - r$ valores propios sean próximos a cero. Observemos que como $\mathbf{P1} = \mathbf{0}$, donde $\mathbf{1}$ es un vector de unos, la matriz $\mathbf{Q}$ tiene rango máximo $n - 1$ y siempre tendrá el vector propio $\mathbf{1}$ unido al valor propio cero.

3. Obtener las coordenadas de los puntos en las variables mediante $\mathbf{v}_i\sqrt{\lambda_i}$, donde λ_i es un valor propio de $\mathbf{Q}$ y $\mathbf{v}_i$ su vector propio asociado. Esto implica aproximar $\mathbf{Q}$ por

$$\mathbf{Q} \approx (\mathbf{V}_r\Lambda_r^{1/2})(\Lambda_r^{1/2}\mathbf{V}_r')$$

   y tomar como coordenadas de los puntos las variables

$$\mathbf{Y}_r = \mathbf{V}_r\Lambda_r^{1/2}.$$

El método puede también aplicarse si la información de partida es directamente la matriz de similitud entre elementos. Diremos que se ha definido una función de similitud entre elementos si existe una función, s_{ij}, con las propiedades siguientes:
(1) $s_{ii} = 1$,
(2) $0 \leq s_{ij} \leq 1$,
(3) $s_{ij} = s_{ji}$.

La similaridad es pues una función no negativa y simétrica. Si la matriz de partida, $\mathbf{Q}$, es una matriz de similitud, entonces $q_{ii} = 1, q_{ij} = q_{ji}$ y $0 \leq q_{ij} \leq 1$. La matriz de distancias asociadas será, por (6.3),

$$d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij} = 2(1 - q_{ij})$$

y puede comprobarse que $\sqrt{2(1 - q_{ij})}$ es una distancia y verifica la desigualdad triangular al corresponder a la distancia euclídea para cierta configuración de puntos.

Pueden obtenerse medidas de la precisión conseguida mediante la aproximación a partir de los p valores propios positivos de la matriz de similitud. Mardia ha propuesto el coeficiente:

$$m_{1,p} = 100 \times \frac{\sum^p \lambda_i}{\sum_1^p |\lambda_i|}$$

Ejemplo 6.1 *Las distancias en kilométros por carretera entre las ciudades españolas siguientes se encuentran en el cuadro adjunto, que llamaremos matriz* $\mathbf{M}$*, donde las ciudades se han representado por las letras siguientes: M es Madrid, B Barcelona, V Valencia, S Sevilla, SS San Sebastián y LC La Coruña.*

	M	*B*	*V*	*S*	*SS*	*LC*
M	*0*	*627*	*351*	*550*	*488*	*603*
B	*627*	*0*	*361*	*1043*	*565*	*1113*
V	*351*	*361*	*0*	*567*	*564*	*954*
S	*550*	*1043*	*567*	*0*	*971*	*950*
SS	*488*	*565*	*564*	*971*	*0*	*713*
LC	*603*	*1113*	*954*	*950*	*713*	*0*

Llamando $\mathbf{D}$ *a esta matriz de distancias, la matriz de similitud es* $\mathbf{Q} = -.5\mathbf{PDP}$ *y dividiendo cada término por 10,000 se obtiene la matriz:*

0.1176	*-0.3908*	*-0.1795*	*0.3856*	*-0.3180*	*0.3852*
-0.3908	*3.0321*	*1.2421*	*-2.0839*	*0.7338*	*-2.5333*
-0.1795	*1.2421*	*0.7553*	*0.6095*	*-0.3989*	*-2.0285*
0.3856	*-2.0839*	*0.6095*	*3.6786*	*-2.0610*	*-0.5288*
-0.3180	*0.7338*	*-0.3989*	*-2.0610*	*1.6277*	*0.4165*
0.3852	*-2.5333*	*-2.0285*	*-0.5288*	*0.4165*	*4.2889*

que tiene los siguientes vectores propios, por columnas:

-0.0960	*-0.0443*	*-0.2569*	*0.1496*	*0.8566*	*0.4082*
0.6270	*0.1400*	*-0.4155*	*-0.4717*	*-0.1593*	*0.4082*
0.2832	*-0.2584*	*-0.0094*	*0.7670*	*-0.3130*	*0.4082*
-0.2934	*-0.7216*	*0.2205*	*-0.4017*	*-0.1285*	*0.4082*
0.1241	*0.4417*	*0.7812*	*-0.0687*	*0.0885*	*0.4082*
-0.6449	*0.4426*	*-0.3198*	*0.0255*	*-0.3443*	*0.4082*

ligados a los siguientes valores propios:

7.3792	*5.9106*	*0.5947*	*-0.3945*	*0.0104*	*0.0000*

La matriz $\mathbf{Q}$ *tiene dos valores propios grandes y los otros tres son muy pequeños. Además tenemos el autovalor cero ligado al vector propio unidad. Esto sugiere que las distancias*

pueden explicarse aproximadamente mediante dos variables. Tomando los dos vectores propios asociados a los mayores valores propios, y estandarizándoles por la raíz de su valor propio, resultan las siguientes coordenadas para cada ciudad

Madrid	*-82.44*	*-34.05*
Barcelona	*538.61*	*107.67*
Valencia	*243.29*	*-198.62*
Sevilla	*-252.04*	*-554.79*
San Sebastián	*106.60*	*339.55*
La Coruña	*-554.02*	*340.25*

Si representamos estas coordenadas se obtiene la figura 6.1. Se observa que las coordenadas de las ciudades reproducen, con cierta aproximación, el mapa de España.

Figura 6.1: Representación de las coordenadas principales de seis ciudades españolas

El grado de bondad de esta representación puede medirse por el coeficiente

$$m = 100\frac{7.3792 + 5.9106}{7.3792 + 5.9106 + 0.5947 + 0.3945 + 0.0104} = 93\%$$

y vemos que la representacion en dos dimensiones es muy adecuada para estos datos.

Ejemplo 6.2 *La matriz adjunta indica las similitudes encontradas por un grupo de consumidores entre 7 productos de consumo.*

	A	B	C	D	E	F	G
A	0	7	5	9	5	7	9
B	7	0	4	6	4	6	7
C	5	4	0	3	4	5	6
D	9	6	3	0	3	2	2
E	5	4	4	3	0	5	4
F	7	6	5	2	5	0	4
G	9	7	6	2	4	4	0

Aplicando la transformacion a la matriz Q, los valores propios son 6.24, 3.37, 2.44, 2.04, 1.25, -.06, 0. La representacion de los productos correspondiente a los dos vectores principales se presenta en el figura 6.2. El grado de ajuste de esta representación es

$$m = 100\frac{9.61}{15.4} = 62.4\%$$

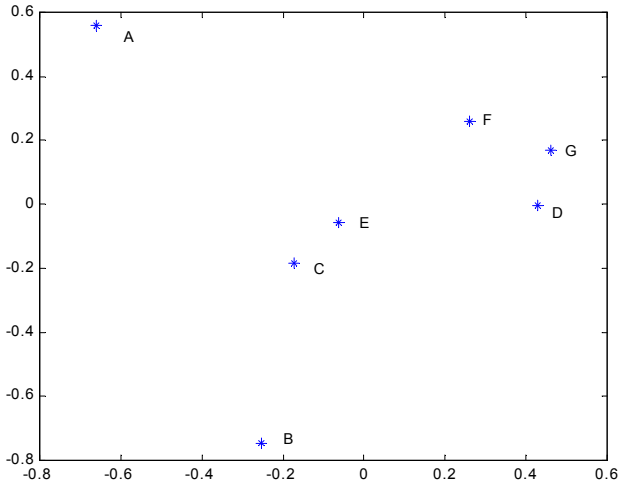


Figura 6.2: Representación de los productos en el plano de las dos primeras coordenadas principales.

Podemos concluir que los consumidores parecen utilizar dos dimensiones de valoración que explican el 62.4% de la variabilidad, aunque existen otras dimensiones que se tienen en cuenta con menor peso.

6.4 RELACIÓN ENTRE COORDENADAS Y COMPONENTES PRINCIPALES

Cuando los datos originales forman una matriz $\widetilde{\mathbf{X}}$ de individuos por variables y construimos la matriz $\mathbf{D}$ de distancias utilizando las distancias euclídeas entre los puntos a partir de dichas variables originales, las coordenadas principales obtenidas de la matriz $\mathbf{D}$ son equivalentes a los componentes principales de las variables.

En efecto, con variables de media cero los componentes principales son los autovectores de $\frac{1}{n}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$, mientras que, como hemos visto en la sección 6.3 (ecuación 6.18), las coordenadas principales son los vectores propios estandarizados por $\sqrt{\lambda_i}$ de los autovalores de $\mathbf{Q} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$. Vamos a comprobar que $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ y $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$ tienen el mismo rango y los mismos autovalores no nulos. Si $\mathbf{a}_i$ es un autovector de $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ con autovalor λ_i,

$$\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{a}_i = \lambda_i\mathbf{a}_i \tag{6.21}$$

y multiplicando por $\widetilde{\mathbf{X}}$ ambos miembros,

$$\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\mathbf{a}_i = \lambda_i\widetilde{\mathbf{X}}\mathbf{a}_i \tag{6.22}$$

es decir, $\widetilde{\mathbf{X}}\mathbf{a}_i$ es un autovector de $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$ con el mismo valor propio λ_i. Si $n > p$ y la matriz $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ tiene rango completo tendrá p autovalores no nulos que será los autovalores no nulos de $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$. Los vectores propios de $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$ son las proyecciones de la matriz $\widetilde{\mathbf{X}}$ sobre la dirección de los vectores propios de $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$.

Por otro lado, la matriz $n \times p$ que proporciona los valores de los p componentes principales en los n individuos es:

$$\mathbf{Z} = \widetilde{\mathbf{X}}\mathbf{A} \tag{6.23}$$

donde $\mathbf{Z}$ es $n \times p$ y tiene por columnas los componentes principales y $\mathbf{A}$ es $p \times p$ y contiene en columnas los vectores propios de $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$. La matriz $n \times p$ de coordenadas principales viene dada por:

$$\mathbf{Y} = [\mathbf{v}_1, \ldots, \mathbf{v}_p] \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_p} \end{bmatrix} = \mathbf{VL} \tag{6.24}$$

donde $\mathbf{v}_i$ es un vector propio de $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$, la matriz $\mathbf{V}$ es $n \times p$ y contiene los p autovectores no nulos de $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$, y $\mathbf{L}$ es $p \times p$ y diagonal. Como $\mathbf{V} = \widetilde{\mathbf{X}}$, es claro que, aparte de un factor de escala, ambos procedimientos conducen al mismo resultado.

El análisis en coordenadas principales o escalado multidimensional, está muy relacionado con componentes principales. En ambos casos tratamos de reducir la dimensionalidad de los datos. En componentes partimos de la matriz $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$, obtenemos sus valores propios, y luego proyectamos las variables sobre estas direcciones para obtener los valores de los componentes, que son idénticas a las coordenadas principales, que se obtienen directamente como vectores propios de la matriz $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'$. Si la matriz de similaridades proviene de una métrica euclídea ambos métodos conducirán al mismo resultado. Sin embargo, el concepto de coordenadas principales o escalado multidimensional puede aplicarse a una gama más amplia de problemas que componentes, ya que las coordenadas principales pueden obtenerse siempre, aunque las distancias de partida no hayan sido exactamente generadas a partir de variables, como veremos en el caso de escalado no métrico.

6.5 BIPLOTS

Se conocen como biplots a las representaciones gráficas conjuntas en un plano de las filas y de las columnas de una matriz. En el caso de una matriz de datos, el biplot es un gráfico conjunto de las observaciones y las variables. La representación se obtiene a partir de la

descomposición en valores singulares de una matriz (veáse la sección 2.4.2). Una matriz $\mathbf{X}$ de dimensiones $n \times p$ puede siempre descomponerse como

$$\mathbf{X} = \mathbf{V}\mathbf{D}^{1/2}\mathbf{A}'$$

o gráficamente

$$\begin{bmatrix} x_{11} & . & x_{1p} \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ x_{n1} & . & x_{np} \end{bmatrix} = \begin{bmatrix} v_{11} & . & v_{1p} \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ v_{np} & . & v_{np} \end{bmatrix} \begin{bmatrix} \lambda^{1/2} & . & 0 \\ 0 & . & 0 \\ 0 & . & \lambda^{1/2} \end{bmatrix} \begin{bmatrix} a_{11} & . & a_{1p} \\ . & . & . \\ a_{p1} & . & a_{pp} \end{bmatrix}$$

donde $\mathbf{V}$ es $n \times p$ y contiene en columnas los vectores propios asociados a valores propios no nulos de la matriz $\mathbf{X}\mathbf{X}'$, $\mathbf{D}$ es una matriz diagonal de orden p que contiene las raices cuadradas de los valores propios no nulos de $\mathbf{X}\mathbf{X}'$ o $\mathbf{X}'\mathbf{X}$ y $\mathbf{A}'$ es una matriz ortogonal de orden p y contiene por filas los vectores propios de $\mathbf{X}'\mathbf{X}$. Las matrices de vectores propios verifican $\mathbf{V}'\mathbf{V} = \mathbf{I}$, $\mathbf{A}'\mathbf{A} = \mathbf{I}$.

La descomposición en valores singulares tiene gran importancia práctica porque, como se demuestró en el apéndice 5.2, la mejor aproximación de rango $r < p$ a la matriz $\mathbf{X}$ se obtiene tomando los r mayores valores propios de $\mathbf{X}'\mathbf{X}$ y los correspondientes vectores propios de $\mathbf{X}\mathbf{X}'$ y $\mathbf{X}'\mathbf{X}$ y construyendo

$$\widehat{\mathbf{X}} = \mathbf{V}_r\mathbf{D}_r^{1/2}\mathbf{A}_r'$$

donde $\mathbf{V}_r$ es $n \times r$ y contiene las primeras r columnas de $\mathbf{V}$ correspondientes a los r mayores valores propios de $\mathbf{X}\mathbf{X}'$, $\mathbf{D}_r^{1/2}$ es diagonal de orden r y contiene estos r valores propios y $\mathbf{A}_r'$ es $r \times p$ y contiene las r primeras filas de $\mathbf{A}'$ que corresponden a los r vectores propios de $\mathbf{X}'\mathbf{X}$ ligados a los r mayores valores propios.

La representación biplot de una matriz $\mathbf{X}$ consiste en aproximarla mediante la descomposición en valores singulares de rango dos, tomando $r = 2$:

$$\mathbf{X} \approx \mathbf{V}_2\mathbf{D}_2^{1/2}\mathbf{A}_2' = (\mathbf{V}_2\mathbf{D}_2^{1/2-c/2})(\mathbf{D}_2^{c/2}\mathbf{A}_2') = \mathbf{F}\mathbf{C}$$

donde $\mathbf{V}_2$ es $n \times 2$, $\mathbf{D}_2^{1/2}$ es diagonal de orden 2 y $\mathbf{A}_2'$ es $2 \times p$. Tomando $0 \leq c \leq 1$ se obtienen distintas descomposiciones de la matriz $\mathbf{X}$ en dos matrices. La primera, $\mathbf{F}$ representa las n filas de la matriz $\mathbf{X}$ en un espacio de dos dimensiones y la segunda, $\mathbf{C}$, representa en el mismo espacio las columnas de la matriz. Según el valor de c se obtienen distintos biplots. Los más utilizados son para $c = 0, 0, 5$, y 1.

Vamos a interpretar el biplot cuando $c = 1$, que es el caso más interesante. Entonces representaremos las observaciones, filas de $\mathbf{X}$, por las filas de la matriz $\mathbf{V}_2$, y las variables, columnas de $\mathbf{X}$, por las columnas de la matriz $\mathbf{D}_2^{1/2}\mathbf{A}_2'$. Para distinguir ambas representaciones las observaciones se dibujan como puntos y las variables como vectores en el plano. Se verifica que:

(1) La representación de las observaciones como puntos en un plano mediante las filas de $\mathbf{V}_2$, equivale a proyectar las observaciones sobre el plano de las dos componentes principales estandarizadas para que tengan varianza unidad.

(2) Las distancias euclídeas entre los puntos en el plano equivale, aproximadamente, a las distancias de Mahalanobis entre las observaciones originales.

(3) La representación de las variables mediante vectores de dos coordenadas es tal que el ángulo entre los vectores equivale, aproximadamente, a la correlación entre las variables.

Para demostrar estas propiedades utilizaremos la relación entre los componentes y los vectores propios de $\mathbf{XX}'$. Las coordenadas de los componentes principales son $\mathbf{Z} = \mathbf{XA}$, y, como hemos visto en la sección anterior, los vectores que forman las columnas de $\mathbf{Z}$ son vectores propios sin normalizar de $\mathbf{XX}'$. En efecto, los vectores propios de $\mathbf{X}'\mathbf{X}$ verifican

$$\mathbf{X}'\mathbf{X}\mathbf{a}_i = \lambda_i \mathbf{a}_i$$

y multiplicando por $\mathbf{X}$ tenemos que

$$\mathbf{XX}'(\mathbf{X}\mathbf{a}_i) = \lambda_i(\mathbf{X}\mathbf{a}_i)$$

por tanto, $\mathbf{z}_i = \mathbf{X}\mathbf{a}_i$ es un vector propio de la matriz $\mathbf{XX}'$, pero no esta normalizado a norma unidad. El vector propio normalizado será

$$\mathbf{v}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{z}_i.$$

Generalizando, la matriz de vectores propios de $\mathbf{XX}'$ normalizados a norma unidad será

$$\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_p] = \left[\frac{1}{\sqrt{\lambda_i}}\mathbf{z}_1, ..., \frac{1}{\sqrt{\lambda_i}}\mathbf{z}_p\right] = \mathbf{ZD}^{-1/2}$$

y es inmediato que con esta normalización $\mathbf{V}'\mathbf{V} = \mathbf{D}^{-1/2}\mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\mathbf{D}\mathbf{D}^{-1/2} = \mathbf{I}$. Por tanto si representamos los puntos por $\mathbf{V}_2$ tenemos las proyecciones estandarizadas a varianza uno de las observaciones sobre los dos primeros componentes.

Vamos a comprobar la segunda propiedad. Una observación se representa por los componentes principales por $\mathbf{x}_i'\mathbf{A}$, y si estandarizamos los componentes a varianza uno $\mathbf{x}_i'\mathbf{A}\mathbf{D}^{-1/2}$. Las distancias euclídeas al cuadrado entre dos observaciones en términos de sus coordenadas en los componentes estandarizados serán:

$$\left\|\mathbf{x}_i'\mathbf{A}\mathbf{D}^{-1/2} - \mathbf{x}_j'\mathbf{A}\mathbf{D}^{-1/2}\right\|^2 = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{A}\mathbf{D}^{-1}\mathbf{A}'(\mathbf{x}_i - \mathbf{x}_j)'$$

y como $\mathbf{S} = \mathbf{ADA}'$ entonces $\mathbf{S}^{-1} = \mathbf{A}\mathbf{D}^{-1}\mathbf{A}'$ y obtenemos la distancia de Mahalanobis entre las observaciones originales. Si en lugar de tomar los p componentes tomamos sólo los dos más importantes esta relación será aproximada y no exacta.

Comprobaremos por último que si representamos las variables como vectores con coordenadas $\mathbf{D}_2^{1/2}\mathbf{A}_2' = \mathbf{C}$ los ángulos entre los vectores representan, aproximadamente, la correlación entre las variables. Para ello escribiremos

$$\mathbf{S} \simeq \mathbf{A}_2\mathbf{D}_2\mathbf{A}_2' = \mathbf{CC}' = \begin{bmatrix} \mathbf{c}_1' \\ ... \\ \mathbf{c}_p' \end{bmatrix} \begin{bmatrix} \mathbf{c}_1 & ... & \mathbf{c}_p \end{bmatrix}$$

donde $\mathbf{c}_1$ es un vector 2×1 correspondiente a la primera columna de la matriz $\mathbf{C}$. De esta expresión es inmediato que

$$\mathbf{c}_i'\mathbf{c}_i = s_i^2$$

y

$$\mathbf{c}_i'\mathbf{c}_j = s_{ij}$$

y finalmente

$$r_{ij} = \frac{\mathbf{c}_i'\mathbf{c}_j}{\|\mathbf{c}_i\| \, \|\mathbf{c}_j\|} = \cos(\mathbf{c}_i\mathbf{c}_j)$$

Por tanto, aproximadamente el ángulo entre estos vectores es el coeficiente de correlación entre las variables.

La precisión de la representación del biplot depende de la importancia de los dos primeros valores propios respecto al total. Si $(\lambda_1 + \lambda_2)/tr(\mathbf{S})$ es próximo a uno la representación será muy buena. Si este valor es pequeño el biplot no proporciona una representación fiable de los datos.

Ejemplo 6.3 *Vamos a utilizar la base de datos de MUNDODES (tabla A.6 del Anéxo), cuyos componentes principales se obtuvieron en el capítulo anterior (véae el ejemplo ***). Esta matriz de datos está constituida por 91 países en los que se han observado 9 variables: X_1: ratio de natalidad, X_2: ratio de mortalidad, X_3: mortalidad infantil, X_4: esperanza de vida en hombres X_5: esperanza de vida de mujeres y X_6: PNB per capita.*

La figura 6.3 es un biplot donde se han representado conjuntamente las observaciones por su proyección estandarizada en el plano de los dos componentes principales. El lector debe observar que ahora la variabilidad en ambos componentes es la misma como consecuencia de la estandarización, lo que no ocurría en los gráficos anteriores donde las escalas eran muy diferentes. Se han representado también las variables como vectores de manera que el ángulo entre las variables sea aproximadamente igual a sus correlaciones. En el biplot se observa una separación de los países en dos grupos y, por otro lado, una división de las variables en tres grupos: en el primero están las tasas de mortalidad infantil y natalidad que están muy correladas entre sí, por otro la tasa de mortalidad, que tiene baja correlación con el resto de variables, y por otro la renta y las esperanzas de vida de hombres y mujeres que están muy correladas con la renta.

En el gráfico 6.4 se muestra la misma representación conjunta que en la figuras 6.3 en el caso de realizar el análisis normado de componentes principales en las variables originales. Se aprecia una relación no lineal entre las dos primeras componentes.

6.6 ESCALADO NO MÉTRICO

En los problemas de escalado no métrico se parte de una matriz de diferencias o disimilitudes entre objetos que se ha obtenido generalmente por consultas a jueces, o a partir de

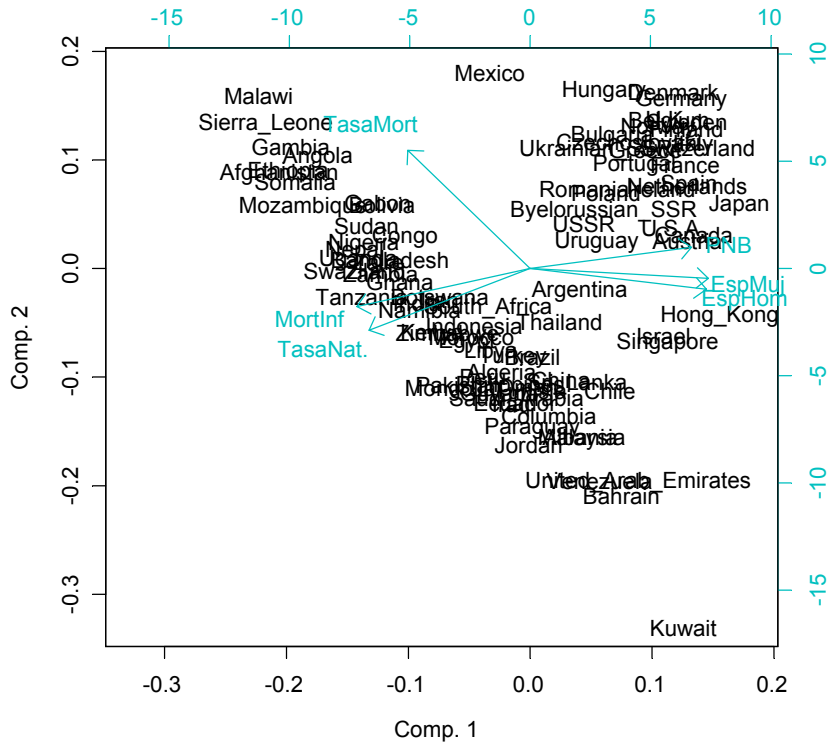


Figura 6.3: Representación de observaciones y variables en el plano de las dos primeras componentes, variables en logaritmos.

procedimientos de ordenación de los elementos. Por ejemplo, el escalado no métrico se ha aplicado para estudiar las semejanzas entre las actitudes, preferencias o percepciones de personas sobre asuntos políticos o sociales o para evaluar preferencias respecto a productos y servicios en marketing y en calidad. Los valores de una tabla de similaridades o distancias se obtienen habitualmente por alguno de los procedimientos siguientes:

1. Estimación directa. Un juez, o un conjunto de jueces, estiman directamente las distancias entre los elementos. Una escala muy utilizada es la escala 0-100, de manera que la distancia o disimilaridad entre un elemento y sí mismo sea cero y la distancia entre dos elementos distintos refleje la percepción de sus diferencias. Con n elementos esto requiere $n(n-1)/2$ evaluaciones.

2. Estimación de rangos. Se selecciona un elemento y se pide al juez, o grupo de jueces, que ordene los $n-1$ restantes por mayor o menor proximidad al seleccionado. A continuación se selecciona el siguiente y se ordenan los $n-2$ restantes, y así sucesivamente. Existen algoritmos de cálculo que transforman estas ordenaciones en una matriz de distancias (véase Green y Rao, 1972).

3. Rangos por pares. Se presentan al juez los $n(n-1)/2$ pares posibles y se le pide que los ordene de mayor a menor distancia. Por ejemplo, con cuatro objetos supongamos que se obtienen los resultados en orden de distancia: (3,4), (2,3), (2,4), (1,4), (1,2) y (1,3). Entonces, los más próximos son los objetos 3 y 4, y a esta pareja se le asigna el rango 1.

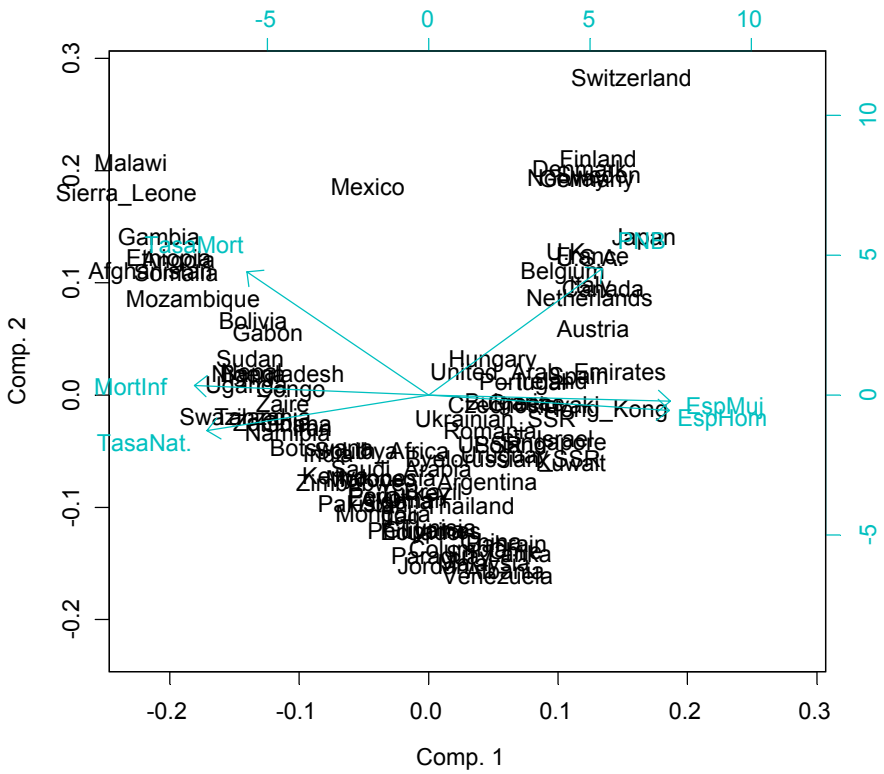


Figura 6.4: Representación de observaciones y variables en el plano de las dos primeras componentes, variables originales.

A la pareja siguiente, (2,3), se le asigna rango dos y así sucesivamente hasta la pareja de los elementos más alejados, el 1 y el 3, que reciben rango $n(n-1)/2$, que es 6 en este caso. A continuación se calcula un rango medio para cada objeto, promediando los rangos de los pares donde aparece. Por ejemplo, el objeto 1 aparece en pares que tienen rango 4, 5 y 6, con lo que el rango del objeto 1 es :

$$rango(1) = \frac{4+5+6}{3} = 5.$$

Igualmente obtenemos que rango(2)= $(2+3+5)/2 = 3,3$; rango(3)= $(1+2+6)/3 = 3$ y rango(4) $= (1+3+4)/3 = 2,7$. Las diferencias entre los rangos se toman ahora como medidas de distancia entre los objetos.

Se supone que la matriz de similaridades está relacionada con un matriz de distancias, pero de una manera compleja. Es decir, se acepta que los jueces utilizan en las valoraciones ciertas variables o dimensiones, pero que, además, los datos incluyen elementos de error y variabilidad personal. Por tanto, las variables que explican las similitudes entre los elementos comparados determinarán una distancias euclídeas entre ellos, d_{ij}, que están relacionadas con las similitudes dadas, δ_{ij} , mediante una función desconocida

$$\delta_{ij} = f(d_{ij})$$

donde la única condición que se impone es que f es una función monótona, es decir, si

$$\delta_{ij} > \delta_{ih} \Leftrightarrow d_{ij} > d_{ih} .$$

El objetivo que se pretende es encontrar unas coordenadas que sean capaces de reproducir estas distancias a partir únicamente de la condición de monotonía. Para ello hay que definir:

(1) Un criterio de bondad del ajuste que sea invariante ante transformaciones monótonas de los datos.

(2) Un algoritmo para obtener las coordenadas, optimizando el criterio establecido.

Estos problemas no tienen solución única y se han presentado muchos procedimientos alternativos. El más utilizado es minimizar las diferencias entre las distancias derivadas de las coordenadas principales, $\widehat{d}_{ij}$, y las similitudes de partida δ_{ij} , es decir minimizar $\sum\sum(\delta_{ij} - \widehat{d}_{ij})^2$ para todos los términos de la matriz. Esta cantidad se estandariza para favorecer las comparaciones, con lo que se obtiene el criterio de ajuste denominado STRESS, dado por:

$$S^2 = \frac{\sum_{i<j}(\delta_{ij} - \widehat{d}_{ij})^2}{\sum_{i<j}\delta_{ij}^2} \tag{6.25}$$

Un criterio alternativo es minimizar las distancias al cuadrado, con lo que se obtiene el criterio S-STRESS. Se han propuesto otros criterios que el lector puede consultar en Cox y Cox (1994). Las distancias $\widehat{d}_{ij}$ se determinarán encontrando p coordendas principales que se utilizan como variables implicitas $y_{ij}, i = 1, ..., n,\ j = 1, ...p$, que determinarán unas distancias euclídeas entre dos elementos:

$$\widehat{d}_{ij}^2 = \sum_{s=1}^{p}(y_{is} - y_{js})^2 \tag{6.26}$$

El método de cálculo es partir de la solución proporcionada por las coordenadas principales e iterar para mejorar esta solución minimizando el criterio (6.25). Normalmente se toma $p = 2$ para facilitar la representación gráfica de los datos, pero el número de dimensiones necesario para una buena representación de los datos puede estimarse probando distintos valores de p y estudiando la evolución del criterio de forma similar a como se determina el número de componentes principales. Fijado p el problema es minimizar (6.25) donde las distancias se calculan por (6.26). Derivando respecto a los valores de las coordenadas en los individuos (véase apéndice 6.1) se obtiene un sistema de ecuaciones no lineales en las variables y cuya solución requiere un algoritmo de optimización no lineal. Suele tomarse como solución inicial la obtenida con las coordenadas principales. Remitimos al lector interesado en los detalles de los algoritmos a Cox y Cox (1994).

Ejemplo 6.4 *Utilizaremos la matriz de similitudes entre productos. Con el programa SPSS se obtiene la solución indicada en la figura 6.5. Los productos A, B, C, etc se han representado en el gráfico como m1, m2, m3,... Como puede verse la solución es similar a la obtenida con coordenadas principales, pero no idéntica.*

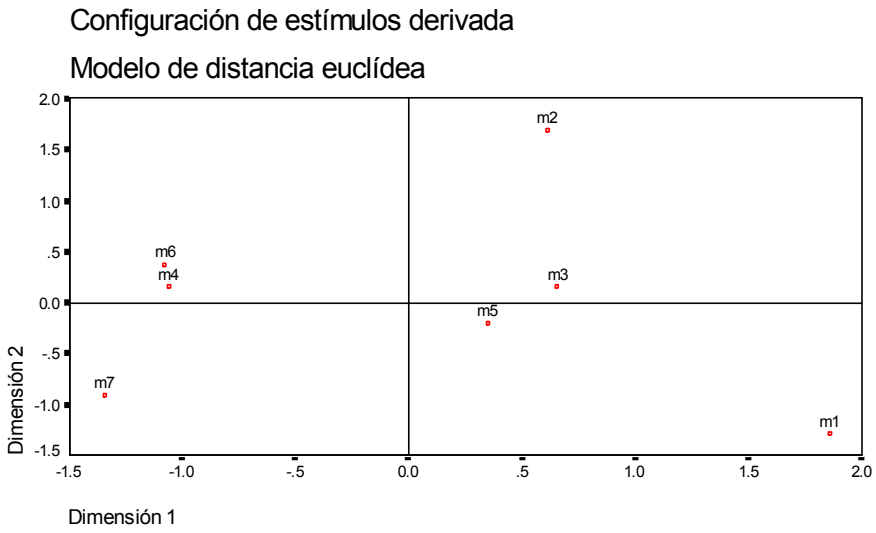


Figura 6.5: Representación de los productos con el escalado no métrico

El valor del coeficiente al finalizar la estimación no lineal es Stress = .14134 y la proporción de variabilidad explicada, RSQ = .87957.

La figura 6.6 presenta la relación entre las distancias obtenidas y las observaciones. Se aprecia que la relación es monótona, aunque no lineal.

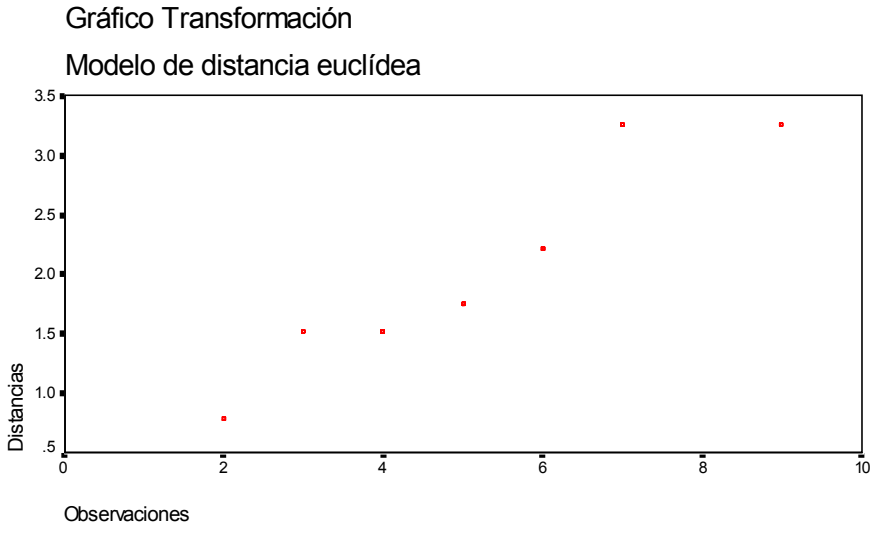


Figura 6.6: Relación entre las distancias originales y las calculadas por el escalado multidimensional

6.7 Lecturas complementarias

Los capítulos 11 y 12 de Jackson (1991) contienen ampliaciones sobre este tema y muchas referencias. El libro de Dillon y Goldstein (1984) presenta una introducción clara y simple del escalado multidimensional no métrico. Gnanadesikan (1997) presenta también una buena introducción al tema. Libros dedicados al escalado multidimensional son los de Schiffman et al (1981), Coxon (1982), Davidson (1983), Kruskal y Wish (1978), Green et al (1989) y Cox y Cox (1994). Young (1987) contiene muchos ejemplos de aplicación. Gower y Hand (1996) esta integramente dedicado a los biplots.

EJERCICIOS

Ejercicio 6.1 *Si la distancia euclídea entre dos elementos se define por* $d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$ *demostrar que puede escribirse como* $d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$ *donde los* q_{ij} *son elementos de la matriz* $\mathbf{XX}'$.

Ejercicio 6.2 *Demostrar que* $d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$ *puede escribirse como* $d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$ *donde ahora los* q_{ij} *son elementos de la matriz* $\mathbf{XPX}'$, *siendo* $\mathbf{P}$ *la matriz proyección que elimina las medias definida en 6.12*

Ejercicio 6.3 *Demostrar que si tenemos una solución* $\mathbf{Y}_r$ *de coordenadas principales también es solución* $\mathbf{Z}_r = \mathbf{Y}_r C + b$, *donde* C *es una matriz ortogonal y* b *cualquier vector.*

Ejercicio 6.4 *Demostrar que si la matriz* $\mathbf{Q}$ *es semidefinida positiva se verifica que* $q_{ii} + q_{jj} - 2q_{ij} \geq 0$. *(Ayuda: utilice que si* $\mathbf{Q}$ *es definida positiva* $\mathbf{u'Qu} \geq 0$ *para cualquier vector* u *y tome* $u = (0, ..., 1, -1, 0, ..., 0)'$*)*

Ejercicio 6.5 *Demostrar que si* Q *es semidefinida positiva las magnitudes* $d_{ij}^2 = q_{ii} + q_{jj} - 2q_{ij}$ *verifican las propiedades de una distancia. (Ayuda: para comprobar la propiedad triangular utilice que para tres puntos* $u'Qu \geq 0$ *con* $u = (1, -1, -1)'$ *implica* $q_{11} + q_{22} + q_{33} - 2q_{12} - 2q_{13} + 2q_{32} \geq 0$*)*

Ejercicio 6.6 *Demostrar que se verifica la relación* $\mathbf{Q} = \mathbf{PQP}$.

Ejercicio 6.7 *Demostrar que la descomposición biplot puede escribirse como* $\mathbf{Y}_r \mathbf{A}_r'$ *donde el primer término contiene las coordenadas principales y el segundo las componentes principales.*

Apéndice 6.1Maximización del STRESS

El procedimiento de optimización del criterio se obtiene derivando el STRESS respecto a cada término, y_{ip}, que nos indica como se modifica el criterio si modificamos el valor de la variable p en el elemeto i, lo que conduce a las ecuaciones

$$\frac{\partial S^2}{\partial y_{ip}} = 2\sum_{j=1}^{n}(\delta_{ij} - \widehat{d}_{ij})\frac{\partial \widehat{d}_{ij}}{\partial y_{ip}} = 0 \tag{6.27}$$

El cambio en las distancias del punto i a todos los demás cuando cambiamos la coordenada p de este punto es, por (6.26):

$$\frac{\partial \widehat{d}_{ij}}{\partial y_{ip}} = \frac{(y_{ip} - y_{jp})}{\widehat{d}_{ij}}$$

y sustituyendo en (6.27) tenemos que la ecuación a resolver es

$$y_{ip} \sum_{j=1}^{n} \frac{(\delta_{ij} - \widehat{d}_{ij})}{\widehat{d}_{ij}} - \sum_{j=1}^{n} \frac{(\delta_{ij} - \widehat{d}_{ij})}{\widehat{d}_{ij}} y_{jp} = 0.$$

Si derivamos para los np valores de las coordenadas principales, el sistema de ecuaciones resultante puede escribirse conjuntamente como

$$\mathbf{FX} = \mathbf{0}$$

donde $\mathbf{F}$ es una matriz cuadrada y simétrica de orden n con coeficientes

$$\begin{aligned} f_{ij} &= -\frac{(\delta_{ij} - \widehat{d}_{ij})}{\widehat{d}_{ij}}, \quad i \neq j \\ f_{ii} &= \sum_{j=1, j\neq i}^{n} f_{ij}, \quad i = j \end{aligned}$$

Capítulo 7

ANÁLISIS DE CORRESPONDENCIAS

7.1 INTRODUCCIÓN

El análisis de correspondencias es una técnica descriptiva para representar tablas de contingencia, es decir, tablas donde recogemos las frecuencias de aparición de dos o más variables cualitativas en un conjunto de elementos. Constituye el equivalente de componentes principales y coordenadas principales para variables cualitativas. La información de partida ahora es una matriz de dimensiones $I \times J$, que representa las frecuencias absolutas observadas de dos variables cualitativas en n elementos. La primera variable se representa por filas, y suponemos que toma I valores posibles, y la segunda se representa por columnas, y toma J valores posibles. Por ejemplo, la tabla 7.1 presenta la clasificación de $n = 5387$ escolares escoceses por el color de sus ojos, que tiene cuatro categorías posibles y $I = 4$, y el color de su cabello, que tiene cinco categorías posibles y $J = 5$. Esta tabla tiene interés histórico ya que fué utilizada por Fisher en 1940 para ilustrar un método de análisis de tablas de contingencia que está muy relacionado con el que aquí presentamos.

En general, una tabla de contingencia es un conjunto de números positivos dispuestos en una matriz, donde el número en cada casilla representa la frecuencia absoluta observada para esa combinación de las dos variables.

Una manera de llegar a una tabla de contingencia $I \times J$ es definir I variables binarias para

		Color	del	pelo		
C. ojos	rubio	pelirrojo	castaño	oscuro	negro	total
claros	688	116	584	188	4	1580
azules	326	38	241	110	3	718
castaños	343	84	909	412	26	1774
oscuros	98	48	403	618	85	1315
total	1455	286	2137	1391	118	5387

Tabla 7.1: Tabla de Contingencia del color de los ojos y el color del pelo de escolares escoceses. Recogida por Fisher en 1940

las categorías de las filas y J para las de las columnas y diponer estas variables en matrices $\mathbf{X}_a$ para las filas y $\mathbf{X}_b$ para las columnas. Por ejemplo, la matriz $\mathbf{X}_a$ para la variable color de los ojos contendrá 4 variables en columnas correspondientes a las 4 categorías consideradas para indicar el color de ojos, y en cada fila sólo una columna tomará el valor uno, la que corresponda al color de ojos de la persona. La matriz tendrá 5387 filas correpondientes a las personas incluidas en la muestra. Por tanto, la matriz $\mathbf{X}_a$ de dimensiones 5387×4 será de la forma:

$$\mathbf{X}_a = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ . & . & . & . \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

donde hemos tamado las categorías para el color de ojos en el mismo orden que aparecen en las filas de la tabla 7.1. Por ejemplo, el primer dato corresponde a una persona de ojos claros, ya que tiene un uno en la primera columna. El segundo dato tiene un uno en la cuarta categoría, que corresponde a ojos oscuros. Finalmente, el último elemento de la matriz corresponde a una persona de ojos azules. De la misma forma, la matriz $\mathbf{X}_b$ tendrá dimensiones 5387×5 y las columnas indicarán el color del cabello de cada persona. Observemos que estas matrices $\mathbf{X}$ de variables binarias tienen tantas columnas como categorías y sus variables son linealmente dependientes, ya que siempre la suma de los valores de una fila es uno, al ser las categorías excluyentes y exhaustivas. Al realizar el producto $\mathbf{X}_a\mathbf{X}_b$ sumaremos todas las personas que tienen cada par de características y se obtiene la tabla de contingencia.

El análisis de correspondencias es un procedimiento para resumir la información contenida en una tabla de contingencia. Puede interpretarse de dos formas equivalentes. La primera, como una manera de representar las variables en un espacio de dimensión menor, de forma análoga a componentes principales, pero definiendo la distancia entre los puntos de manera coherente con la interpretación de los datos y en lugar de utilizar la distancia euclídea utilizamos la distancia ji-cuadrado. Desde este enfoque, el análisis de correspondencias es el equivalente de componentes principales para datos cualitativos. La segunda interpretación está más próxima al escalado multidimensional: es un procedimiento objetivo de asignar valores numéricos a variables cualitativas. Vamos a analizar estos dos aspectos.

7.2 BÚSQUEDA DE LA MEJOR PROYECCIÓN

En adelante trabajaremos con la matriz $\mathbf{F}$ de frecuencias relativas obtenida dividiendo cada casilla por n, el total de elementos observados. Llamaremos f_{ij} a las frecuencias relativas que verifican

$$\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} = 1$$

La matriz $\mathbf{F}$ puede considerarse por filas o por columnas. Cualquier análisis lógico de esta matriz debe de ser equivalente al aplicado a su transpuesta, ya que la elección de la variable

	Sobre.	Not.	Aprob.	Sus.	Total
Zona A	0,03	0,06	0,15	0,06	0,3
Zona B	0,07	0,14	0,35	0,14	0,7
Total	0,1	0,2	0,5	0,2	1

Tabla 7.2: Clasificación de estudiantes por zona geográfica y calificación obtenida

que se coloca en filas, en lugar de en columnas, es arbitraria, y no debe influir en el análisis. Vamos a presentar primero el análisis por filas de esta matriz, que será simétrico al análisis por columnas, que estudiaremos a continuación.

7.2.1 Proyección de las Filas

Vamos a analizar la matriz de frecuencias relativas, $\mathbf{F}$, por filas. Entonces las I filas pueden tomarse como I puntos en el espacio $\Re^J$. Vamos a buscar una representación de estos I puntos en un espacio de dimensión menor que nos permita apreciar sus distancias relativas. El objetivo es el mismo que con componentes principales, pero ahora tendremos en cuenta las peculiaridades de este tipo de datos. Estas peculiaridades provienen de que la frecuencia relativa de cada fila es distinta, lo que implica que:

(1) Todos las filas (puntos en $\Re^J$) no tienen el mismo peso, ya que algunas continen más datos que otras. Al representar el conjunto de las filas (puntos) debemos dar más peso a aquellas filas que contienen más datos.

(2) La distancia euclídea entre puntos no es una buena medida de su proximidad y debemos modificar esta distancia, como veremos a continuación.

Comenzando con el primer punto, cada fila de la matriz $\mathbf{F}$ tiene una frecuencia relativa $f_{i.} = \sum_{j=1}^{J} f_{ij}$, y el conjunto de estas frecuencias relativas se calcula con:

$$\mathbf{f} = \mathbf{F}'\mathbf{1}$$

debemos dar a cada fila un peso proporcional a su frecuencia relativa y los términos del vector $\mathbf{f}$ pueden directamente considerarse como pesos, ya que son números positivos que suman uno.

Con relación a la medida de distancia a utilizar entre las filas, observemos que la distancia euclídea no es una buena medida de las diferencias reales entre las estructuras de las filas. Por ejemplo, supongamos la tabla 7.2 donde se presentan las frecuencias relativas de estudiantes clasificados por su procedencia geográfica, (A ó B) y sus calificaciones. Aunque las frecuencias relativas de las dos filas son muy distintas, las dos filas tienen exactamente la misma estructura relativa: simplemente, hay más del doble de estudiantes de la zona B que de la A, pero la distribución de calificaciones es idéntica en ambas zonas. Si calculamos la distancia euclídea entre las zonas obtendremos un valor alto, que no refleja una estructura distinta de las filas sino sólo que tienen distinta frecuencia relativa. Suponganos que dividimos cada casilla por la frecuencia relativa de la fila, $f_{i.}$. Con esto se obtiene la tabla 7.3 donde los números que aparecen en las filas representan la frecuencia relativa de la variable columna condicionada a la variable fila. Ahora las dos filas son idénticas, y esto es coherente con una distancia euclídea cero entre ambas.

	Sobre.	Not.	Aprob.	Sus.	Total
Zona A	0,1	0,2	0,5	0,2	1
Zona B	0,1	0,2	0,5	0,2	1

Tabla 7.3: Clasificación de estudiantes por zona geográfica y calificación obtenida

		Color	del	cabello		
C. ojos	rubio	pelirrojo	castaño	oscuro	negro	total
claros	0.435	0.073	0.369	0.119	0.003	1
azules	0.454	0.053	0.336	0.153	0.004	1
castaños	0.193	0.047	0.512	0.232	0.015	1
oscuros	0.075	0.037	0.307	0.518	0.065	1

Tabla 7.4: Tabla de frecuencias relativas del color del cabello condicionada al color de los ojos para los escolares escoceses

Para analizar que medida de distancia debemos utilizar, llamaremos $\mathbf{R}$ a la matriz de frecuencias relativas condicionadas al total de la fila, que se obtiene con:

$$\mathbf{R} = \mathbf{D}_f{}^{-1}\mathbf{F} \tag{7.1}$$

donde $\mathbf{D}_f$ es una matriz diagonal $I \times I$ con los términos del vector $\mathbf{f}$, $f_{i.}$, frecuencias relativas de las filas, en la diagonal principal. Esta operación transforma la matriz original de frecuencias relativas, $\mathbf{F}$, en otra matriz cuyas casillas por filas suman uno. Cada fila de esta matriz representa la distribución de la variable en columnas condicionada al atributo que representa la fila. Por ejemplo, la tabla 7.4 presenta las frecuencias relativas condicionadas para la tabla 7.1. En este caso $I = 4$, $J = 5$. Esta tabla permite apreciar mejor la asociación entre las características estudiadas.

Llamaremos $\mathbf{r}'_i$ a la fila i de la matriz $\mathbf{R}$ de frecuencias relativas condicionadas por filas, que puede considerarse un punto (o un vector) en el espacio $\Re^J$. Como la suma de los componentes de $\mathbf{r}'_i$ es uno, todos los puntos están en un espacio de dimensión $J-1$. Queremos proyectar estos puntos en un espacio de dimensión menor de manera que las filas que tengan la misma estructura estén próximas, y las que tengan una estructura muy diferente, alejadas. Para ello, debemos definir una medida de distancia entre dos filas $\mathbf{r}_a, \mathbf{r}_b$. Una posibilidad es utilizar la distancia euclídea, pero esta distancia tiene el inconveniente de tratar igual a todos los componentes de estos vectores. Por ejemplo, en la tabla 7.1 las personas de cabello rubio tienen una diferencia en frecuencia relativa entre los ojos azules y claros de 0,454-0,435= 0,019, y las personas de cabello negro tienen un diferencia en frecuencia relativa entre los ojos castaños y azules de 0,015 - 0,004=0,011. Hay una diferencia mayor en el primer caso que en el segundo y, sin embargo, intuitivamente vemos que la segunda diferencia es mayor que la primera. La razón es que en el primer caso el cambio relativo es pequeño, del orden del 4% (0,019/0,454), mientras que en el segundo caso el cambio relativo es muy grande: las personas de cabello negro tienen ojos castaños casi cuatro veces más frecuentemente (0,015/0,004=3,75 veces) que ojos azules. Como los componentes representan frecuencias relativas, no parece adecuado que una diferencia de 0,01 se considere igual en un atributo

de alta frecuencia (por ejemplo, pasar de 0,60 a 0,61) que en un atributo de baja frecuencia (por ejemplo, pasar de 0,0001 a 0,0101).

Para obtener comparaciones razonables entre estas frecuencias relativas tenemos que tener en cuenta la frecuencia relativa de aparición del atributo que estudiamos. En atributos raros, pequeñas diferencias absolutas pueden ser grandes diferencias relativas, mientras que en atributos con gran frecuencia, la misma diferencia será poco importante. Una manera intuitiva de construir las comparaciones es ponderar las diferencias en frecuencia relativa entre dos atributos inversamente proporcional a la frecuencia de este atributo. Es decir, en lugar de sumar los términos $(r_{aj} - r_{bj})^2 = (f_{aj}/f_{a.} - f_{bj}/f_{b.})^2$ que miden la diferencia que las filas a y b tienen en la columna j sumaremos los términos $(r_{aj} - r_{bj})^2/f_{.j}$ donde $f_{.j.} = \sum_{i=1}^{I} f_{ij}$ es la frecuencia relativa de la columna j . La expresión de la distancia entre dos filas, $\mathbf{r}_a$ y $\mathbf{r}_b$ de $\mathbf{R}$ vendrá dada en esta métrica por

$$D^2(\mathbf{r}_a, \mathbf{r}_b) = \sum_{j=1}^{J} (\frac{f_{aj}}{f_{a.}} - \frac{f_{bj}}{f_{b.}})^2 \frac{1}{f_{.j.}} = \sum_{j=1}^{J} \frac{(r_{aj} - r_{bj})^2}{f_{.j.}} \tag{7.2}$$

que puede escribirse matricialmente como

$$D^2(\mathbf{r}_a, \mathbf{r}_b) = (\mathbf{r}_a - \mathbf{r}_b)' \mathbf{D}_c^{-1} (\mathbf{r}_a - \mathbf{r}_b) \tag{7.3}$$

donde $\mathbf{D}_c$ es una matriz diagonal con términos $f_{.j.}$. A la distancia (7.2) ó (7.3) se la conoce como distancia χ^2, y se analizará con más detalle en la sección siguiente.

Observemos que está distancia equivale a la distancia euclídea entre los vectores transformados $\mathbf{y}_i = \mathbf{D}_c^{-1/2} \mathbf{r}_i$. Podemos pues simplificar el problema definiendo una matriz de datos transformada, sobre la que tiene sentido considerar la distancia euclídea entre filas. Llamando:

$$\mathbf{Y} = \mathbf{R}\ {\mathbf{D}_c}^{-1/2} = \mathbf{D}_f^{-1} \mathbf{F}\ \mathbf{D}_c^{-1/2} \tag{7.4}$$

obtenemos una matriz $\mathbf{Y}$ que contiene términos del tipo

$$y_{ij} = \left\{ \frac{f_{ij}}{f_{i.} {f_{.j}}^{1/2}} \right\} \tag{7.5}$$

que ya no suman uno ni por filas ni por columnas. Las casillas de esta matriz representan las frecuencias relativas condicionadas por filas, $f_{ij}/f_{i.}$, pero estandarizadas por su variabilidad, que depende de la raíz cuadrada de la frecuencia relativa de la columna. De esta manera las casillas son directamente comparables entre sí. La tabla 7.5 indica esta matriz resultado de estandarizar las frecuencias relativas de la tabla 7.1 dividiendo cada casilla por la raíz cuadrada de la frecuencia relativa de la columna correspondiente, que se obtiene de la tabla 7.1. Por ejemplo, el primer elemento de la tabla 7.5 se obtiene como $0.435/\sqrt{(1455/5387)} =$

.83 7	.31 6	.587	.235	.015
.873	.228	.536	.301	.02 9
.374	.205	.815	.455	.09 5
.14 7	.161	.484	1.022	.440

Tabla 7.5: Matriz estandarizada por fila y por variabilidad del color de los ojos y el color del pelo de escolares

0.0114. En esta tabla la estructura de las columnas es similar a la de la tabla 7.1 de frecuencias relativas, ya que hemos dividido todas las casillas de cada columna por la misma cantidad.

Podríamos tratar a esta matriz como una matriz de datos estándar, con observaciones en filas y variables en columnas, y preguntarnos como proyectarla de manera que se preserven las distancias relativas entre las filas, es decir, las filas con estructura similar aparezcan próximas en la proyección. Esto implica encontrar una dirección $\mathbf{a}$ de norma unidad,

$$\mathbf{a}'\mathbf{a} = 1 \tag{7.6}$$

tal que el vector de puntos proyectados sobre esta dirección,

$$\mathbf{y}_p(\mathbf{a}) = \mathbf{Y}\ \mathbf{a} \tag{7.7}$$

tenga variabilidad màxima. El vector $\mathbf{a}$ se encontrará maximizando $\mathbf{y}_p(\mathbf{a})'\mathbf{y}_p(\mathbf{a}) = \mathbf{a}'\mathbf{Y}'\mathbf{Y}\ \mathbf{a}$ con la condición (7.6), y este problema se ha resuelto en el capítulo 5 al estudiar componentes principales: el vector $\mathbf{a}$ es un vector propio de la matriz $\mathbf{Y}'\mathbf{Y}$. Sin embargo, este tratamiento de la matriz $\mathbf{Y}$ como una matriz de variables continuas no es del todo correcto porque las filas tienen una distinta frecuencia relativa, $f_{i.}$, y por tanto deben tener distinto peso. Aquellas filas con mayor frecuencia relativa deben de tener más peso en la representación que aquellas otras con frecuencia relativa muy baja, de manera que las filas con gran número de individuos estén bien representadas, aunque esto sea a costa de representar peor las filas con pocos elementos. En consecuencia, daremos a cada fila un peso proporcional al número de datos que contiene. Esto puede hacerse maximizando la suma de cuadrados ponderada.

$$m = \mathbf{a}'\mathbf{Y}'\mathbf{D}_f\ \mathbf{Y}\ \mathbf{a} \tag{7.8}$$

sujeto a (7.6), que equivale a

$$m = \mathbf{a}'\mathbf{D}_c^{-1/2}\mathbf{F}'\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}\ \mathbf{a}. \tag{7.9}$$

Alternativamente, podemos construir una matriz de datos $\mathbf{Z}$ definida por

$$\mathbf{Z} = \mathbf{D}_f^{-1/2}\mathbf{F}\mathbf{D}_c^{-1/2} \tag{7.10}$$

cuyos componentes son

$$z_{ij} = \left\{ \frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}} \right\}$$

y que estandariza las frecuencias relativas en cada casilla por el producto de las raíces cuadradas de las frecuancias relativas totales de la fila y la columna, y escribir el problema de encontrar el vector $\mathbf{a}$ como el problema de maximizar $m = \mathbf{a}'\ \mathbf{Z}'\mathbf{Z}\mathbf{a}$ sujeto a la restricción (7.6). Este es el problema resuelto en componentes principales, cuya solución es

$$\mathbf{D}_c^{-1/2}\mathbf{F}'\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}\mathbf{a} = \lambda\mathbf{a} \tag{7.11}$$

y $\mathbf{a}$ debe ser un vector propio de la matriz $\mathbf{Z}'\mathbf{Z}$ donde $\mathbf{Z}$ está dado por (7.9) y λ su valor propio.

Vamos a comprobar que la matriz $\mathbf{Z}'\mathbf{Z}$ tiene como mayor valor propio siempre el 1 y como vector propio $\mathbf{D}_c^{1/2}$. Multiplicando por la izquierda en (7.11) por $\mathbf{D}_c^{-1/2}$se obtiene:

$$\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_f^{-1}\mathbf{F}(\mathbf{D}_c^{-1/2}\mathbf{a}) = \lambda(\mathbf{D}_c^{-1/2}\mathbf{a})$$

Las matrices $\mathbf{D}_f^{-1}\mathbf{F}$ y $\mathbf{F}\mathbf{D}_c^{-1}$ representan matrices de frecuencias relativas por filas y por columnas y su suma por filas y columnas respectivamente es uno. Por tanto $\mathbf{D}_f^{-1}\mathbf{F}\mathbf{1} = \mathbf{1}$ y $\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{1} = \mathbf{1}$, que implica que la matriz $\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_f^{-1}\mathbf{F}$ tiene un valor propio 1 unido a un vector propio $\mathbf{1}$. En consecuencia, haciendo $(\mathbf{D}_c^{-1/2}\mathbf{a}) = \mathbf{1}$ concluimos que la matriz $\mathbf{Z}'\mathbf{Z}$ tiene un valor propio igual a uno con vector propio $\mathbf{D}_c^{1/2}$.

Olvidando esta solución trivial, que no da información sobre la estructura de las filas, tomaremos el valor propio mayor menor que la unidad y su vector propio asociado $\mathbf{a}$. Entonces, proyectando la matriz $\mathbf{Y}$ sobre la dirección $\mathbf{a}$ encontrada:

$$\mathbf{y}_f(\mathbf{a}) = \mathbf{Y}\mathbf{a} = \mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}\mathbf{a} \tag{7.12}$$

y el vector $\mathbf{y}_f(\mathbf{a})$ es la mejor representación de las filas de la tabla de contingencia en una dimensión. Análogamente, si extraemos el vector propio ligado al siguiente mayor valor propio obtenemos una segunda coordenada y podemos representar las filas en un espacio de dimensión dos. Las coordenadas de la representación de cada fila vendrán dadas por las filas de la matriz

$$\mathbf{C}_f = \mathbf{Y}\mathbf{A_2} = \mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}\mathbf{A}_2$$

donde $\mathbf{A}_2 = [\mathbf{a}_1\mathbf{a}_2]$ contiene en columnas los dos vectores propios $\mathbf{Z}'\mathbf{Z}$. La matriz $\mathbf{C}_f$ es $I \times 2$ y las dos coordenadas de cada fila proporcionan la mejor representación de las filas de la matriz $\mathbf{F}$ en un espacio de dos dimensiones. El procedimiento se extiende sin dificultad para representaciones en más dimensiones, calculando vectores propios adicionales de la matriz $\mathbf{Z}'\mathbf{Z}$.

En resumen el procedimiento que hemos presentado para buscar una buena representación de las filas de la tabla de contingencia es:

(1) Caracterizar las filas por sus frecuencias relativas condicionadas, y considerarlas como puntos en el espacio.

(2) Definir la distancia entre los puntos por la distancia χ^2, que tiene en cuenta que cada coordenada de las filas tiene distinta precisión.

(3) Proyectar los puntos sobre las direcciónes de máxima variabilidad, teniendo en cuenta que cada fila tiene un peso distinto e igual a su frecuencia relativa.

El procedimiento operativo para obtener la mejor representación bidimensional de las filas de la tabla de contingencia es:

(1) Calcular la matriz $\mathbf{Z}'\mathbf{Z}$ y obtener sus vectores y valores propios.

(2) Tomar los dos vectores propios, $\mathbf{a}_1, \mathbf{a}_2$, ligados a los mayores valores propios menores que la unidad de esta matriz.

(3) Calcular las proyecciones $\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}\mathbf{a}_i$, $i = 1, 2$, y representarlas gráficamente en un espacio bidimensional.

Ejemplo 7.1 *Aplicaremos este análisis a la matriz de la tabla 7.1. La matriz de frecuencias relativas estandarizada por filas,* $\mathbf{R}$*, se presenta en la tabla 7.4.*

La variable transformada, $\mathbf{Y}$, se calcula como

$$\mathbf{Y} = \mathbf{R}\ \mathbf{D_c}^{-\mathbf{1/2}} = \mathbf{R}(\frac{1}{5387}\begin{bmatrix} 1455 & & & & \\ & 286 & & & \\ & & 2137 & & \\ & & & 1391 & \\ & & & & 118 \end{bmatrix})^{-1/2}$$

dando lugar a

$$\mathbf{Y} = \begin{array}{ccccc} .83\,7 & .31\,6 & .587 & .235 & .015 \\ .873 & .228 & .536 & .301 & .02\,9 \\ .374 & .205 & .815 & .455 & .09\,5 \\ .14\,7 & .161 & .484 & 1.022 & .440 \end{array}$$

Esta matriz puede interpretarse como una matriz de datos donde por filas tenemos observaciones y por columnas variables. Para obtener la mejor representación de las filas en un espacio de dimensión dos, vamos a obtener los vectores propios de la matriz $\mathbf{Y}\mathbf{D}_f\mathbf{Y}$. Los tres primeros valores y vectores propios de esta matriz se presentan en la tabla siguiente por filas:

valor propio		vector		propio	
1	-0.5197	-0.2304	-0.6298	-0.5081	-0.1480
0.1992	-0.6334	-0.1204	-0.0593	0.6702	0.3629
0.0301	-0.5209	-0.0641	0.7564	-0.3045	-0.2444

Los otros dos valores propios de esta matriz son 0,0009 0,0000. La proyección de los puntos sobre el espacio definido por los valores propios .1992 y .0301 se presenta en la figura 7.1

El eje de abscisas contiene la primera dimensión que explica el .1992/(.1992+ .0301+.0009)=.8653. Vemos que se separan claramente los ojos claros y azules frente a castaños y oscuros. La primera dimensión es pues claro frente a oscuro. La segunda dimensión separa las características puras, ojos claros o azules y negros, frente a la mezclada, castaños.

Ejemplo 7.2 *En un estudio de mercado 4 evaluadores han indicado que características consideran importantes en un tipo de producto. El resultado es la matriz* $\mathbf{F}$ *donde en columnas se representan los evaluadores y en filas los productos.*

Figura 7.1: Proyección de las filas de la matriz de los colores de ojos y pelo sobre el mejor espacio de dimensión 2.

$F=$

	1	2	3	4
c_1	0	0	1	0
c_2	1	1	0	0
c_3	0	1	0	1
c_4	0	0	0	1
c_5	0	1	0	0
c_6	1	1	1	0

Esta matriz es una tabla de contingencia muy simple donde las frecuencias posibles son cero o uno. La matriz $\mathbf{Z}$ es

$$\mathbf{Z} = \begin{bmatrix} 0 & 0 & .707 & 0 \\ .5 & .35 & 0 & 0 \\ 0 & .35 & 0 & .50 \\ 0 & 0 & 0 & .707 \\ 0 & .5 & 0 & 0 \\ .408 & .289 & .408 & 0 \end{bmatrix}$$

y los valores propios de $\mathbf{Z}'\mathbf{Z}$ son (1, 0.75, 0.50, 0.17). El vector propio asociado al mayor valor propio menor que uno es $v = (0.27, 0, 0.53, -0.80)$. La proyección de las filas de $\mathbf{Y}$ sobre las dos direcciones principales conduce a la figura 7.2

Se observa que las características más próximas son la 2 y la 5. Las elecciones de los evaluadores parecen ser debidas a dos dimensiones. La primera explica el 0,75/(0,75+0,50+0,17)=52,83% de la variabilidad y la segunda el 35%. La primera dimensión tiene en cuenta las similitudes

Figura 7.2: Proyección de las características de los productos

aparentes por las elecciones de las personas: las características c3 y c4 son elegidas por la misma persona y por nadie más, por lo que estas características aparecen juntas en un extremo. En el lado opuesto aparecen la c1 y c6, que son elegidas por la misma persona, y las c2 y c5 que son elegidas por personas que también eligen la c6. En la segunda dimensión las características extremas son las c1 y c2.

7.2.2 Proyección de las columnas

Podemos aplicar a las columnas de la matriz $\mathbf{F}$ un análisis equivalente al de las filas. Las columnas serán ahora puntos en $\Re^I$. Llamando

$$\mathbf{c} = \mathbf{F}'\mathbf{1}$$

al vector de frecuencias relativas de las columnas y $\mathbf{D}_c$ a la matriz diagonal que contiene estas frecuencias relativas en la diagonal principal, de acuerdo con la sección anterior la mejor representación de los J puntos (columnas) en un espacio de dimensión menor, con la métrica χ^2 conducirá, por simetría, a estudiar la matriz $\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_f^{-1/2}$. Observemos que, si ahora consideramos la matriz $\mathbf{F}'$ y volvemos al problema de representarla por filas (que es equivalente a representar $\mathbf{F}$ por columnas), el problema es idéntico al que hemos resuelto en la sección anterior. Ahora la matriz que contiene las frecuencias relativas de las filas $\mathbf{F}'$ es $\mathbf{D}_c$ y la que contiene la de las columnas es $\mathbf{D}_f$. Intercambiando el papel de estas matrices, las direcciones de proyección son los vectores propios de la matriz

$$\mathbf{Z}\ \mathbf{Z}' = \mathbf{D}_f^{-1/2}\mathbf{F}\mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_f^{-1/2} \tag{7.13}$$

donde $\mathbf{Z}$ es la matriz $I \times J$ definida por (7.10). Como $\mathbf{Z}'\mathbf{Z}$ y $\mathbf{Z}\mathbf{Z}'$ tienen los mismos valores propios no nulos, esa matriz tendrá también un valor propio unidad ligado al vector propio $\mathbf{1}$. Esta solución trivial no se considera. Llamando $\mathbf{b}$ al vector propio ligado al mayor valor

propio distinto de la unidad de $\mathbf{ZZ'}$, la mejor representación de las columnas de la matriz en un espacio de dimensión uno vendrá dada por

$$\mathbf{y}_c(\mathbf{b}) = \mathbf{Y'b} = \mathbf{D}_c^{-1}\mathbf{F'D}_f^{-1/2}\mathbf{b} \tag{7.14}$$

y, análogamente, la mejor representación en dimensión dos de las columnas de la matriz vendrá dada por las coordenadas definidas por las filas de la matriz

$$\mathbf{C}_c = \mathbf{Y'B_2} = \mathbf{D}_c^{-1}\mathbf{F'D}_f^{-1/2}\mathbf{B}_2$$

donde $\mathbf{B}_2 = [\mathbf{b}_1\mathbf{b}_2]$ contiene en columnas los dos vectores propios ligados a los valores propios mayores de $\mathbf{ZZ'}$ y menores que la unidad. La matriz $\mathbf{C}_c$ es $J \times 2$ y cada fila es la mejor representación de las columnas de la matriz $\mathbf{F}$ en un espacio de dos dimensiones.

7.2.3 Análisis Conjunto

Dada la simetría del problema conviene representar conjuntamente las filas y las columnas de la matriz. Observemos que las matrices $\mathbf{Z'\ Z}$ y $\mathbf{Z\ Z'}$ tienen los mismos valores propios no nulos y que los vectores propios de ambas matrices que corresponden al mismo valor propio están relacionados. En efecto, si $\mathbf{a}_i$ es un vector propio de $\mathbf{Z'Z}$ ligado al valor propio λ_i :

$$\mathbf{Z'Za}_i = \lambda_i\mathbf{a}_i$$

entonces, multiplicando por $\mathbf{Z}$

$$\mathbf{ZZ'}(\mathbf{Za}_i) = \lambda_i(\mathbf{Za}_i)$$

y obtenemos que $\mathbf{b}_i = \mathbf{Za}_i$ es un vector propio de $\mathbf{ZZ'}$ ligado al valor propio λ_i. Una manera rápida de obtener estos vectores propios es calcular directamente los vectores propios de la matriz de dimensión más pequeña, $\mathbf{Z'Z}$ o $\mathbf{ZZ'}$, y obtener los otros vectores propios como $\mathbf{Za}_i$ o $\mathbf{Z'b}_i$. Alternativamente podemos utilizar la descomposición en valores singulares de la matriz $\mathbf{Z}$ o $\mathbf{Z'}$, estudiada al introducir los biplots en el capítulo anterior. Esta descomposición aplicada a $\mathbf{Z}$ es

$$\mathbf{Z} = \mathbf{B}_r\mathbf{D}_r\mathbf{A}_r' = \sum_{i=1}^{r} \lambda_i^{1/2}\mathbf{b}_i\mathbf{a}_i'$$

donde $\mathbf{B}_r$ contiene en columnas los vectores propios de $\mathbf{ZZ'}$, $\mathbf{A}_r$ los de $\mathbf{Z'Z}$ y $\mathbf{D}_r$ es digonal y contiene los valores singulares, $\lambda_i^{1/2}$, o raíces de los valores propios no nulos y $r = \min(I, J)$. Entonces la representación de las filas se obtiene con (7.12) y la de las columnas con (7.14). La representación de la matriz $\mathbf{Z}$ con h dimensiones (habitualmente $h = 2$) implica aproximar esta matriz mediante $\widehat{\mathbf{Z}}_h = \mathbf{B}_h\mathbf{D}_h\mathbf{A}_h'$. Esto es equivalente, por (7.10), a una aproximación a la tabla de contingencia observada mediante:

$$\widehat{\mathbf{F}}_h = \mathbf{D}_f^{1/2}\widehat{\mathbf{Z}}_h\mathbf{D}_c^{1/2}, \tag{7.15}$$

y una forma de juzgar la aproximación que estamos utilizando es reconstruir la tabla de contingencia con esta expresión.

Si deseamos eliminar el valor propio unidad desde el principio, dado que no aparta información de interés, podemos reemplazar la matriz $\mathbf{F}$ por $\mathbf{F}-\widehat{\mathbf{F}}_e$, donde $\widehat{\mathbf{F}}_e$ es la matriz de frecuencias esperadas que viene dada por

$$\widehat{\mathbf{F}}_e = \frac{1}{n}\mathbf{r}\mathbf{c}'.$$

Puede comprobarse que la matriz $\mathbf{F}-\widehat{\mathbf{F}}_e$ tiene rango $r-1$, y ya no tiene el valor propio igual a la unidad.

La proporción de variabilidad explicada por cada dimensión se calcula como en componentes principales descartando el valor propio igual a uno y tomando la proporción que representa cada valor propio con relación al resto.

En resumen, el análisis de correspondencias de una tabla de contingencia de dimensiones $I \times J$ se realiza en los pasos siguientes

(1) Se calcula la tabla de frecuencias relativas, $\mathbf{F}$.

(1) Se calcula la tabla estandarizada $\mathbf{Z}$, de frecuencias relativas las mismas dimensiones de la tabla original, $I \times J$, dividiendo cada celda de $\mathbf{F}$ por la raíz de los totales de su fila y columna, $z_{ij} = \left\{ f_{ij}/\sqrt{f_{i.}f_{.j}} \right\}$.

(2) Se calculan los h (normalmente $h = 2$) vectores propios ligados a valores propios mayores, pero distintos de la unidad, de las matriz de menor dimensión de las $\mathbf{Z}\mathbf{Z}'$ y $\mathbf{Z}'\mathbf{Z}$. Si obtenemos lo vectores propios $\mathbf{a}_i$ de $\mathbf{Z}'\mathbf{Z}$, los $\mathbf{b}_i$ de $\mathbf{Z}\mathbf{Z}'$ se obtienen por $\mathbf{b}_i = \mathbf{Z}\mathbf{a}_i$. Analogamente si se obtienen los $\mathbf{b}_i$ de $\mathbf{Z}\mathbf{Z}'$ $\mathbf{a}_i = \mathbf{Z}'\mathbf{b}_i$. Las I filas de la matriz se presentarán como I puntos en $\Re^h$ y las coordenadas de cada fila vienen dadas por

$$\mathbf{C}_f = \mathbf{D}_f^{-1/2}\mathbf{Z}\mathbf{A}_2$$

donde $\mathbf{A}_2$ tiene en columnas los dos vectores propios de $\mathbf{Z}'\mathbf{Z}$. Las J columnas se representarán como J puntos en $\Re^h$ y las coordenadas de cada columna son

$$\mathbf{C}_c = \mathbf{D}_c^{-1/2}\mathbf{Z}'\mathbf{B}_2$$

Ejemplo 7.3 *Vamos a representar conjuntamente las filas y las columnas de la matriz de los colores. La figura 7.3 presenta esta representación. Se observa que el gráfico describe de manera clara la relación entre ambas variables. La dimensión principal gradúa la tonalidad de claro a oscuro y la segunda separa los castaños de los casos más extremos.*

Es importante calcular conjuntamente los vectores propios para evitar problemas de signos, ya sea calculando los vectores propios de una matriz y obteniendo los otros como producto por la matriz Z o bien a través de la descomposición en valores singulares. La razón es que si v es un vector propio también lo es -v y al calcular separadamente las coordenadas y superponerlas podemos obtener un resultado como el que se presenta en la figura 7.4 . En esta figura se han calculado separadamente las dos representaciones y luego se han superpuesto. El lector puede comprobar que si cambiamos de signo las coordenadas del eje de ordenadas se obtiene la representación de la figura (7.3). Estos problemas de signos se evitan calculado los vectores conjuntamente.

Figura 7.3: Representación de los colores de ojos y cabello para los escolares escoceses.

Figura 7.4:

7.3 LA DISTANCIA JI-CUADRADO

El contraste de independencia entre las variables fila y columna en una tabla de contingencia $I \times J$ se realiza con la estadístico

$$X^2 = \sum \frac{(\text{fr. observadas - fr. esperadas})^2}{\text{fr. esperadas}}$$

que, en la hipótesis de independencia, sigue una distribución χ^2 con $(I-1)\times(J-1)$ grados de libertad. De acuerdo con la notación anterior, la frecuencia esperada en cada celda de la fila i, suponiendo independencia de filas y columnas, se obtendrá repartiendo el total de la fila, $nf_{i.}$, porporcionalmente a la frecuencia relativa de cada columna, $f_{.j}$. Por ejemplo, la frecuencia esperada de la primera casilla de la tabla 5.1 se obtendrá multiplicando el número total de elementos de la fila, 1580, por la proporción de personas rubias sobre el total, 1455/5387. Por tanto, el estadístico X^2 para contrastar la independencia puede escribirse:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}} \tag{7.16}$$

donde $f_{i.} = \sum_{j=1}^{J} f_{ij}$ es la frecuencia relativa de la fila i y $f_{j.} = \sum_{i=1}^{I} f_{ij}$ la de columna j. Como

$$\frac{(nf_{ij} - nf_{i.}f_{.j})^2}{nf_{i.}f_{.j}} = \frac{nf_{i.}}{f_{.j}}\frac{(f_{ij} - f_{i.}f_{.j})^2}{{f_{i.}}^2}$$

la expresión del estadístico X^2 puede también escribirse como :

$$X^2 = n\sum_{i=1}^{I} f_{i.}\sum_{j=1}^{J}(\frac{f_{ij}}{f_{i.}} - f_{.j})^2\frac{1}{f_{.j}}. \tag{7.17}$$

En esta representación la distribución condicionada de las frecuencias relativas de cada fila, $\left\{\frac{f_{ij}}{f_{i.}}\right\}$, se compara con la distribución media de las filas $\{f_{.j}\}$, y cada coordenada se pondera inversamente a la frecuencia relativa que existe en esa columna. Se suman luego todas las filas, pero dando a cada fila un peso tanto mayor cuanto mayor es su frecuencia, $nf_{i.}$.

Vamos a ver que esta representación es equivalente a calcular las distancias entre los vectores de la matriz de frecuencias relativas por filas, $\mathbf{R}$, definida en (7.1) si medimos la distancia con la métrica χ^2. Consideremos los vectores $\mathbf{r}_i'$, filas de la matriz $\mathbf{R}$. La media de estos vectores es

$$\overline{\mathbf{r}} = \frac{\sum_{i=1}^{I} w_i\mathbf{r}_i}{\sum_{i=1}^{I} w_i}$$

donde los w_i son coeficientes de ponderacion. La media aritmética se obtiene con $w_i = 1$, dando a todas las filas el mismo peso. Sin embargo, en este caso esta poderación no es

conveniente, porque debemos dar más peso a las filas que contengan más datos. Podemos ponderar por la frecuencia relativa de cada fila, $w_i = f_{i.}$, y entonces $\sum w_i = \sum f_{i.} = 1$. Como las frecuencias relativas de las filas vienen dadas por el vector columna $\mathbf{D}_f\mathbf{1}$, tenemos que

$$\overline{\mathbf{r}} = \mathbf{R}'\mathbf{D}_f\mathbf{1}$$

y utilizando (7.1)

$$\overline{\mathbf{r}} = \mathbf{F}'\mathbf{D}_f^{-1}\mathbf{D}_f\mathbf{1} = \mathbf{F}'\mathbf{1} = \mathbf{c}$$

y el valor medio de las filas viene dado por el vector cuyos componentes son las frecuencias relativas de las columnas. La distancia de cualquier vector de fila, $\mathbf{r}_i$, a su media, $\mathbf{c}$, con la métrica χ^2 será

$$(\mathbf{r}_i - \mathbf{c})'\mathbf{D}_c^{-1}(\mathbf{r}_i - \mathbf{c})$$

donde la matriz $\mathbf{D}_c^{-1}$ se obtuvo en (7.3) para construir la distancia χ^2. La suma de todas estas distancias, ponderadas por su importancia, que se conoce como inercia total de la tabla, es

$$I_T = \sum_{i=1}^{I} f_{i.}(\mathbf{r}_i - \mathbf{c})'\mathbf{D}_c^{-1}(\mathbf{r}_i - \mathbf{c})$$

y esta expresión puede escribirse como

$$I_T = \sum_{i=1}^{I} f_{i.} \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 / f_{.j}$$

y si comparamos con (7.17) vemos que la inercia total es igual a X^2/n.

Se demuestra que la inercia total es la suma de los valores propios de la matriz $\mathbf{Z}'\mathbf{Z}$ eliminado el uno. Por tanto, el análisis de las filas (o de las columnas ya que el problema es simétrico) puede verse como una descomposición de los componentes del estadístico X^2 en sus fuentes de variación.

La distancia χ^2 tiene una propiedad importante que se conoce como el principio de equivalencia distribucional. Esta propiedad es que si dos filas tienen la misma estructura relativa, $f_{ij}/f_{i.}$ y las unimos en una nueva fila única, las distancias entre las restantes filas permanecen invariables. Esta misma propiedad por simetría se aplica a las columnas. Esta propiedad es importante, porque asegura una cierta invarianza del procedimiento ante agregaciones o desagregaciones irrelevantes de las categorías. Para demostrarlo, consideremos la distancia χ^2 entre las filas a y b

$$\sum_{j=1}^{J} (\frac{f_{aj}}{f_{a.}} - \frac{f_{bj}}{f_{b.}})^2 \frac{1}{f_{.j.}}$$

es claro que esta distancia no se modifica si unimos dos filas en una, ya que esta unión no va a afectar a las frecuencias $f_{ij}/f_{i.}$ ni tampoco a $f_{.j.}$. Vamos a comprobar que si unimos dos filas con la misma estructura la distancia de la nueva fila al resto es la misma que las

de las filas originales. En efecto, supongamos que para las filas 1 y 2, se verifica que para $j = 1, ..., J$

$$\frac{f_{1j}}{f_{1.}} = \frac{f_{2j}}{f_{2.}} = g_j$$

entonces, si unimos estas dos filas en una nueva fila, se obtiene que, para la nueva fila

$$\frac{f_{1j} + f_{2j}}{f_{1.} + f_{2.}} = g_j$$

y su distancia a cualquier otra fila permanecerá invariable.

Esta propiedad garantiza que no perdemos nada al agregar categoría homogéneas ni podemos ganar nada por desagregar una categoría homogénea.

Ejemplo 7.4 *Se han contabilizado los pesos y las alturas de 100 estudiantes universitarios y se han formado 4 categorias tanto para el peso como para la altura. Para el peso, las categorias se denotan P1, de 51 a 60 k., P2, de 61 a 70 k., P3, de 71 a 80 k. y P4, de 81 a 90 k. Para la altura se denotan A1, de 151 a 160 cm., A2, de 161 a 170 cm., A3, de 171 a 180 cm. y A4, de 181 a 190 cm. La siguiente tabla de contingencia muestra las frecuencias de cada grupo:*

Peso/Altura	*A1*	*A2*	*A3*	*A4*
P1	*15*	*8*	*3*	*0*
P2	*10*	*15*	*7*	*2*
P3	*2*	*7*	*17*	*3*
P4	*0*	*2*	*3*	*6*

Realizar proyecciones por filas, por columnas y conjunta de filas y columnas. Comprobar como las proyecciones por filas y por columnas separan claramente las categorias, pero que la proyección conjunta asocia claramente cada categoría de un peso con la de una altura.

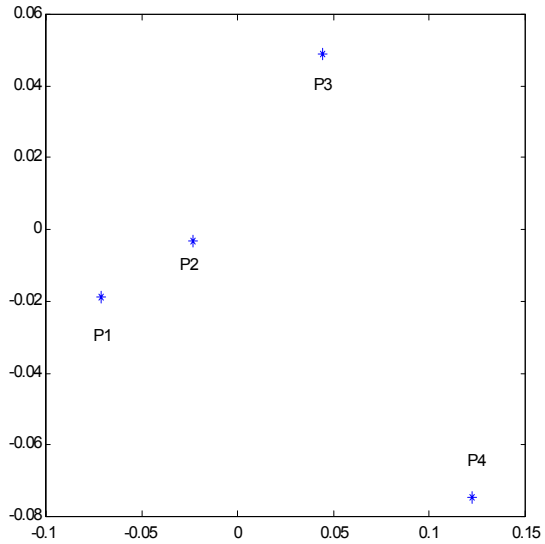
Para la proyección por filas, la variable Y queda:

$$Y = RD_c^{-\frac{1}{2}} = \begin{bmatrix} 0.1110 & 0.0544 & 0.0211 & 0 \\ 0.0566 & 0.0780 & 0.0376 & 0.0177 \\ 0.0133 & 0.0427 & 0.1070 & 0.0312 \\ 0 & 0.0321 & 0.0498 & 0.1645 \end{bmatrix}$$

Los tres valores propios y vectores propios diferentes de uno de esta matriz son:

valor propio	vector propio			
0.3717	-0.6260	-0.1713	0.3673	0.6662
0.1401	-0.2974	-0.0064	0.6890	-0.6610
0.0261	0.4997	-0.8066	0.3007	0.0964

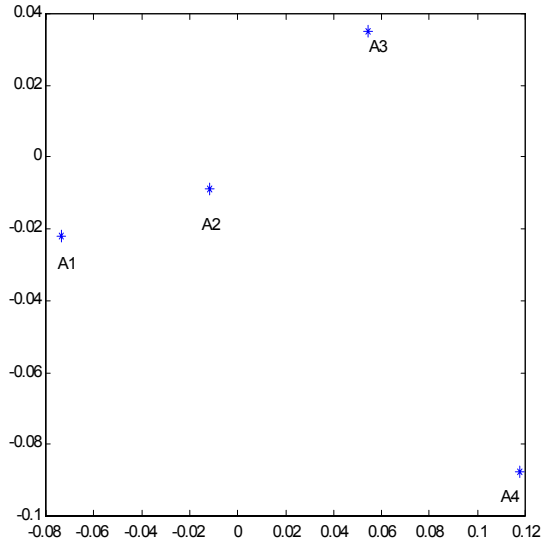
La proyección por filas es:



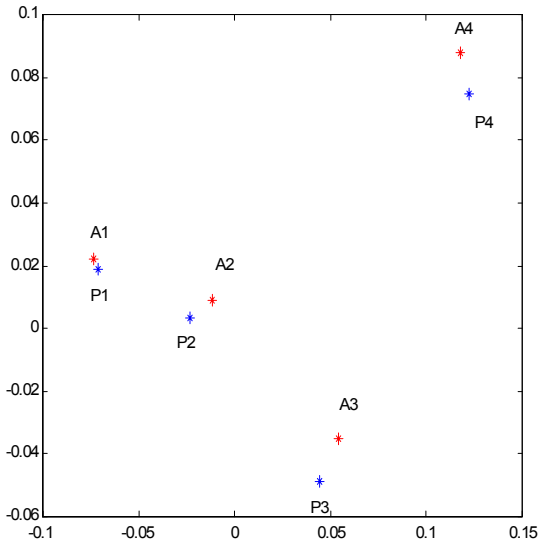
Para las columnas, los tres valores propios y vectores propios diferentes de uno de esta matriz son:

valor propio	vector propio			
0.3717	-0.5945	-0.2216	0.3929	0.6656
0.1401	-0.2568	-0.0492	0.7034	-0.6609
0.0261	0.5662	-0.7801	0.2466	0.1005

La proyección por columnas es:



El resultado de la proyección conjunta es el siguiente donde alturas y pesos quedan asociados:



Ejemplo 7.5 *Del conjunto de datos MUNDODES, se ha tomado la esperanza de vida de hombres y de mujeres. Se han formado 4 categorias tanto para la mujer como para el hombre. Se denotan por M1 y H1, a las esperanzas entre menos de 41 a 50 años, M2 y H2, de 51 a 60 años, M3 y H3, de 61 a 70, y M4 y H4, para entre 71 a más de 80. La siguiente tabla de contingencia muestra las frecuencias de cada grupo:*

Mujer/Hombre	*H1*	*H2*	*H3*	*H4*
M1	*10*	*0*	*0*	*0*
M2	*7*	*12*	*0*	*0*
M3	*0*	*5*	*15*	*0*
M4	*0*	*0*	*23*	*19*

Realizar proyecciones por filas, por columnas y conjunta de filas y columnas. Comprobar que en la proyección por filas las categorias están claramente separadas y que en el caso del hombre, las dos últimas categorías están muy cercanas. Comprobar en la proyección conjunta la cercania de las categorías H3 con M3 y M4.

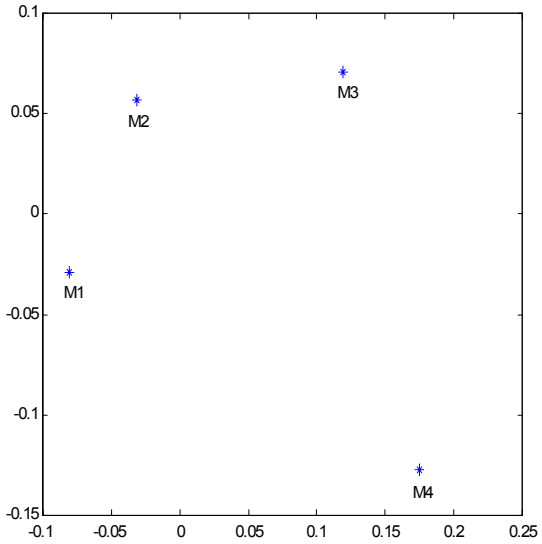
Para la proyección por filas, la variable Y queda:

$$Y = RD_c^{-\frac{1}{2}} = \begin{bmatrix} 0.2425 & 0 & 0 & 0 \\ 0.0894 & 0.1532 & 0 & 0 \\ 0 & 0.0606 & 0.1217 & 0 \\ 0 & 0 & 0.0888 & 0.1038 \end{bmatrix}$$

Los tres valores propios y vectores propios diferentes de uno de esta matriz son:

valor propio	vector propio			
0.8678	0.7221	0.3551	-0.4343	-0.4048
0.3585	-0.5249	0.7699	0.0856	-0.3528
0.1129	-0.1274	0.3072	-0.6217	0.7091

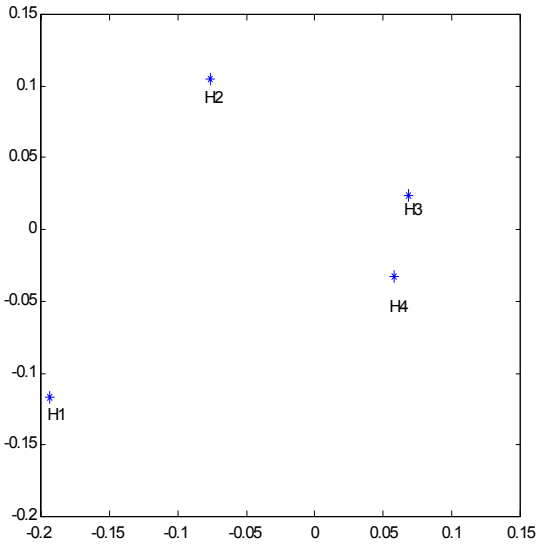
La proyección por filas es:



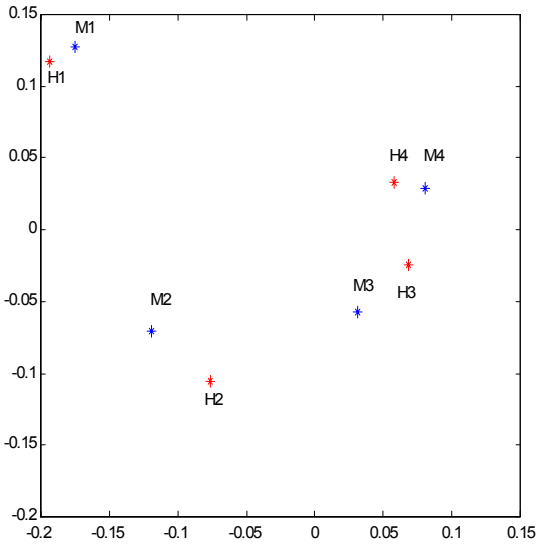
Para las columnas, los tres valores propios y vectores propios diferentes de uno de esta matriz son:

valor propio	vector propio			
0.8678	-0.5945	-0.5564	0.1503	0.5606
0.3585	-0.6723	0.5172	0.4265	-0.3141
0.1129	-0.2908	0.4628	-0.7588	0.3543

La proyección por columnas es:



El resultado de la proyección conjunta es:



7.4 ASIGNACIÓN DE PUNTUACIONES

El análisis de correspondencias puede aplicarse también para resolver el siguiente problema. Supongamos que se desea asignar valores numéricos $y_c(1), ..., y_c(J)$ a las columnas de una matriz $\mathbf{F}$ de observaciones, o, en otros términos, convertir la variable en columnas en una

variable numérica. Por ejemplo, en la tabla (7.3) el color del cabello puede considerarse una variable continúa y es interesante cuantificar las clases de color definidas. Una asignación de valores numéricos a las columnas de la tabla inducirá automáticamente unos valores numéricos para las categorías de la variable en filas. En efecto, podemos asociar a la fila i el promedio de la variable y_c en esa fila, dado por:

$$y_i = \frac{\sum_{j=1}^{J} f_{ij} y_c(j)}{\sum_{j=1}^{J} f_{ij}} = \sum_{j=1}^{J} r_{ij} y_c(j) \tag{7.18}$$

donde $r_{ij} = f_{ij}/f_{i.}$ es la frecuencia relativa condicionada a la fila. El vector de valores así obtenido para todas las filas será un vector $I \times 1$ dado por:

$$\mathbf{y_f} = \mathbf{R}\mathbf{y_c} = \mathbf{D}_r^{-1}\mathbf{F}\mathbf{y_c} \tag{7.19}$$

Análogamente, dadas unas puntuaciones $\mathbf{y_f}$ para las filas, las puntuaciones de las columnas pueden estimarse igualmente por sus valores medios en cada columna, obteniendo el vector $J \times 1$:

$$\mathbf{y_c} = \mathbf{D}_c^{-1}\mathbf{F}'\mathbf{y_f} \tag{7.20}$$

Escribiendo conjuntamente (7.19) y (7.20) resultan las ecuaciones:

$$\mathbf{y_f} = \mathbf{D}_f^{-1}\mathbf{F}\ \mathbf{D}_c^{-1}\mathbf{F}'\mathbf{y_f} \tag{7.21}$$

$$\mathbf{y_c} = \mathbf{D}_c^{-1}\mathbf{F}'\mathbf{D}_f^{-1}\mathbf{F}\mathbf{y_c} \tag{7.22}$$

que indican que las puntuaciones $\mathbf{y_f}$, y $\mathbf{y_c}$ se obtienen como vectores propios de estas matrices. Observemos que estas puntuaciones admiten una solución trivial tomando $\mathbf{y_c} = (1, ..., 1)'_J$, $\mathbf{y_f} = (1, ..., 1)'_I$. En efecto, las matrices $\mathbf{D}_c^{-1}\mathbf{F}'$ y $\mathbf{D}_f^{-1}\mathbf{F}$ suman uno por filas, ya que son de frecuencias relativas. Esta solución equivale en (7.21) y (7.22) al valor propio 1 de la correspondiente matriz. Para encontrar una solución no trivial al problema, vamos a exigir que ambas ecuaciones se cumplan aproximadamente introduciendo un coeficiente de proporcionalidad, $\lambda < 1$, pero que queremos sea tan próximo a uno como sea posible. Multiplicando (7.19) por $\mathbf{D}_f^{1/2}$ y (7.20) por $\mathbf{D}_c^{1/2}$ e introduciendo este coeficiente de proporcionalidad tenemos que

$$\lambda(\mathbf{D}_f^{1/2}\mathbf{y_f}) = \mathbf{D}_f^{-1/2}\mathbf{F}\ \mathbf{D}_c^{-1/2}(\mathbf{D}_c^{1/2}\mathbf{y_c}) \tag{7.23}$$

$$\lambda(\mathbf{D}_c^{1/2}\mathbf{y_c}) = \mathbf{D}_c^{-1/2}\mathbf{F}'\mathbf{D}_f^{-1/2}(\mathbf{D}_f^{1/2}\mathbf{y_f}) \tag{7.24}$$

Para resolver estas ecuaciones, llamemos $\mathbf{b} = \mathbf{D}_f^{1/2}\mathbf{y_f}$, $\mathbf{a} = \mathbf{D}_c^{1/2}\mathbf{y_c}$ y $\mathbf{Z} = \mathbf{D}_f^{-1/2}\mathbf{F}\ \mathbf{D}_c^{-1/2}$. Sustituyendo estas definiciones en (7.23) y (7.24), obtenemos $\lambda\mathbf{b} = \mathbf{Za}$ y $\lambda\mathbf{a} = \mathbf{Z'b}$ y sustiuyendo una de estas ecuacione en la otra se obtiene

$$\lambda^2\mathbf{b} = \mathbf{ZZ'b} \tag{7.25}$$

$$\lambda^2\mathbf{a} = \mathbf{Z'Za} \tag{7.26}$$

Estas ecuaciones muestran que $\mathbf{b}$ y $\mathbf{a}$ son vectores propios ligados al valor propio λ^2 de las matrices $\mathbf{ZZ'}$ y $\mathbf{Z'Z}$. Los vectores de puntuaciones se obtendrán después a partir de la definición de $\mathbf{b} = \mathbf{D}_f^{1/2}\mathbf{y_f}$, con lo que resulta:

$$\mathbf{y_f} = \mathbf{D}_f^{-1/2}\mathbf{b} \tag{7.27}$$

y como $\mathbf{a} = \mathbf{D}_c^{1/2}\mathbf{y_c}$,

$$\mathbf{y_c} = \mathbf{D}_c^{-1/2}\mathbf{a} \tag{7.28}$$

Las matrices $\mathbf{ZZ'}$ o $\mathbf{Z'Z}$ siempre admite el valor propio 1 ligado a un vector propio $(1, ..., 1)'$. Tomando como $\mathbf{a}$ y $\mathbf{b}$ los vectores propios ligados al segundo mayor valor propio, $\lambda < 1$, de estas matrices obtenemos las puntuaciones óptimas de filas y columnas.

Podemos obtener una representación gráfica de las filas y columnas de la matriz de la forma siguiente: si sustituimos las puntuaciones $\mathbf{y_c}$ dadas por (7.28), que se denominan a veces "factores" asociados a las columnas, en la ecuación (7.19) y escribimos

$$\mathbf{y_f}(\mathbf{a}) = \mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}\mathbf{a}$$

obtenemos las proyecciones de las filas encontradas en (7.12). Análogamente, sustituyendo los "factores" $\mathbf{y_f}$ asociados a las filas en (7.20) y escribiendo

$$\mathbf{y_c}(\mathbf{b}) = \mathbf{D}_c^{-1}\mathbf{F'}\mathbf{D}_f^{-1/2}\mathbf{b}$$

encontramos las proyecciones de las columnas de (7.14).

Concluimos que el problema de asignar puntaciones de una forma consistente a las filas y a las columnas de una tabla de contingencia, es equivalente al problema de encontrar una representación óptima en una dimensión de las filas y las columnas de la matriz. En otros términos, el análisis de correspondencia proporciona en la primera coordenada de las filas y columnas una forma consistente de asignar puntuaciones numèricas a las filas y a las columnas de la tabla de contingencia.

Ejemplo 7.6 *La tabla adjunta indica las puntuaciones alta (A), media (M) y baja (B) obtenidas por 4 profesores $P_1, ..., P_4$, que han sido evaluados por un total de 49 estudiantes. ¿Qué puntuaciones habría que asignar a las categorías alta, media y baja? ¿y a los profesores?*

	A	M	B	
P_1	2	6	2	10
P_2	4	4	4	12
P_3	1	10	4	15
P_4	7	5	0	12
	14	25	10	49

Entonces la matriz $\mathbf{Z} = \mathbf{D}_f^{-1/2}\mathbf{F}\ \mathbf{D}_c^{-1/2}$ es

$$\mathbf{Z} = \begin{bmatrix} .169 & .380 & .200 \\ .309 & .230 & .365 \\ .069 & .516 & .327 \\ .540 & .288 & 0 \end{bmatrix}$$

Vamos a obtener la descomposición en valores singulares de esta matriz. Es :

$$\mathbf{Z} = \begin{bmatrix} .452 & .166 & -.249 \\ .495 & .-004 & .869 \\ .553 & .581 & -.317 \\ .495 & -.797 & -.288 \end{bmatrix} \begin{bmatrix} 1 & & \\ & .45 & \\ & & .22 \end{bmatrix} \begin{bmatrix} .534 & -.816 & .221 \\ .714 & .296 & -.634 \\ .452 & .497 & .741 \end{bmatrix}$$

que conduce a las variables

$$\mathbf{y} = \mathbf{D}_f^{-1/2}\mathbf{b}_i = \begin{bmatrix} .143 & .052 & -.079 \\ .143 & -.001 & .251 \\ .143 & .150 & -.082 \\ .143 & -.230 & -.083 \end{bmatrix}$$

$$\mathbf{z} = \mathbf{D}_c^{-1/2}\mathbf{a} = \begin{bmatrix} .143 & -.218 & .059 \\ .143 & .059 & -.127 \\ .143 & .157 & .234 \end{bmatrix}$$

La mejor puntuación -en el sentido de la máxima discriminación- corresponde a (multiplicando por -1 el segundo vector propio para que los números más altos correpondan a puntuaciones altas y favorecer la interpretación) 218, -059, -157 y a los profesores (multiplicando por -1 el segundo vector propio, para ser consistentes con el cambio anterior) 230 -150 001 -052. Si queremos trasladar estas puntuaciones a una escala entre cero y diez, escribiremos

$$y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \times 10$$

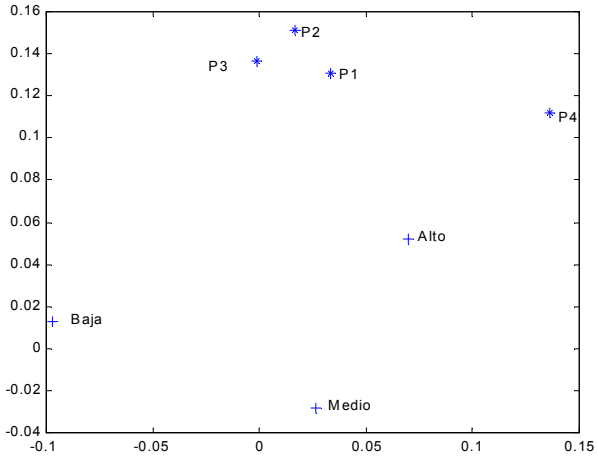


Figura 7.5: Proyección de los profesores y de las puntuaciones

y las puntuaciones se convierten en 10, 7.4 y 0 y. Las evaluaciones de los profesores al pasarlas a la escala de cero a diez se convierten en 10, 0, 3.98, 2.57. La figura 7.5 presenta la proyección de los profesores y de las categorías sobre el plano de mejor representación.

Ejemplo 7.7 *La tabla de contingencia siguiente indica las puntuaciones, muy buena (MB), buena (B), regular (R) o mala (M) obtenidas por las 5 películas nominadas a los Oscars a la mejor película del año 2001 que han sido evaluadas por un total de 100 críticos de cine de todo el mundo. ¿Que puntuaciones habría que asignar a las categorías? ¿y a las películas?*

Películas/Puntuación	*M*	*R*	*B*	*MB*	
P1	*1*	*7*	*2*	*10*	*20*
P2	*0*	*3*	*2*	*15*	*20*
P3	*2*	*7*	*2*	*9*	*20*
P4	*0*	*1*	*3*	*16*	*20*
P5	*1*	*3*	*3*	*13*	*20*
	4	*21*	*12*	*63*	*100*

La matriz P es:

$$\begin{bmatrix} 0.1118 & 0.3416 & 0.1291 & 0.2817 \\ 0 & 0.1464 & 0.1291 & 0.4226 \\ 0.2236 & 0.3416 & 0.1291 & 0.2535 \\ 0 & 0.0488 & 0.1936 & 0.4507 \\ 0.1118 & 0.1464 & 0.1936 & 0.3662 \end{bmatrix}$$

Las variables que se obtienen son:

$$y = Dr^{-\frac{1}{2}}b = \begin{bmatrix} 0.1000 & -0.0934 & -0.1124 & 0.1365 & 0.0000 \\ 0.1000 & 0.0721 & -0.1208 & -0.1234 & 0.0707 \\ 0.1000 & -0.1356 & 0.0707 & -0.1078 & -0.0707 \\ 0.1000 & 0.1304 & 0.0334 & 0.0435 & -0.1414 \\ 0.1000 & 0.0266 & 0.1291 & 0.0512 & 0.1414 \end{bmatrix}$$

$$z = Dc^{-\frac{1}{2}}a = \begin{bmatrix} 0.1000 & -0.2382 & 0.3739 & -0.2085 \\ 0.1000 & -0.1580 & -0.1053 & 0.0396 \\ 0.1000 & 0.0369 & 0.1282 & 0.2357 \\ 0.1000 & 0.0608 & -0.0130 & -0.0448 \end{bmatrix}$$

La mejor puntuación para las categorías corresponde a -0.2382, -0.1580, 0.0369 y 0.0608. Para las películas (multiplicando por -1 el segundo vector propio) a -0.0934, 0.0721, -0.1356, 0.1304 y 0.0266. Si trasladamos todas las puntuaciones entre cero y diez, obtenemos para las categorias los valores 0, 2.6823, 9.2007 y 10. Para las cinco películas tenemos 1.5864, 7.8082, 0, 10 y 6.0977. La proyección conjunta muestra como la pelicula más cercana a la puntuación muy buena (MB) es P4:

7.5 Lecturas complementarias

El análisis de correspondencias puede extenderse para estudiar tablas de cualquier dimensión con el nombre de análisis de correspondencias múltiple. En este enfoque se utiliza la descomposición en valores singulares para aproximar simultaneamente todas las tablas bidimensionales que pueden obtenerse de una tabla multidimensional. Una buena introducción

desde el punto de vista de componentes principales con la métrica ji-cuadrado se encuentra en Gower y Hand (1995). Presentaciones de esta técnica como extensión del análisis de correspondencias presentado en este capítulo se encuentran en Greenacre (1984) y Lebart et al (1984). La literatura sobre análisis de correspondencias está sobre todo en francés, véase Lebart et al (1997) y Saporta (1990). En español Cuadras (1990) y Escofier y Pages (1990). Jackon (1991) contiene una sucinta descripción del método con bastantes referencias históricas y actuales. Lebart, Salem y Bécue (2000) presenta interesantes aplicaciones del análisis de correspondencias para el estudio de textos.

Ejercicios 7

7.1 Demostrar que la traza de las matrices $\mathbf{Z}'\mathbf{Z}$ y $\mathbf{Z}\mathbf{Z}'$ es la misma.

7.2 Demostrar que el centro de los vectores $\mathbf{r}_i$ de las filas, donde cada fila tiene un peso $\mathbf{f}$ es el vector $\mathbf{c}$ de las frecuencias relativas de las columnas (calcule $\overline{\mathbf{r}} = \sum f_{i.}\mathbf{r}_i = \mathbf{R}'\mathbf{D}_f\mathbf{1}$)

7.3 Demostrar que dada una matriz de datos $\mathbf{X}$ donde cada fila tiene un peso $\mathbf{W}$ la operación que convierte a esta matriz en otra de media cero es $\widetilde{\mathbf{X}} = (\mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{W})\mathbf{X}$.

7.4 Demostrar que la suma de las distancias de Mahalanobis ponderadas de las filas es igual a la de las columnas, donde la suma de las filas es $\sum f_{i.}(\mathbf{r}_i - \mathbf{c})'\mathbf{D}_c^{-1}(\mathbf{r}_i - \mathbf{c})$.

7.5 Supongamos que estudiamos dos características en los elementos de un conjunto que pueden darse en los niveles alto, medio y bajo en ambos casos. Si las frecuencias relativas con las que aparecen estos niveles son las mismas para las dos características, indicar la expresión de la representación de las filas y columnas en el plano bidimensional.

7.6 En el ejemplo 7.5 ¿qué podemos decir de la puntuación óptima para cuantificar las filas y columnas?

7.7 Indicar cómo afecta a la representación de filas y columnas que la tabla de contingencias sea simétrica, es decir, $f_{ij} = f_{ji}$.

7.8 Justificar que la variable $\frac{(f_{ij} - r_i c_j/n)}{\sqrt{r_i c_j/n}}$ es aproximadamente una variable normal estándar.

7.9 Demostrar que si definimos la matriz X con elemento genérico $x_{ij} = (f_{ij} - f_{i.}f_{.j})/\sqrt{f_{i.}f_{.j}}$ la matriz $\mathbf{X}'\mathbf{X}$ tiene los mismos valores vectores propios que la $\mathbf{Z}'\mathbf{Z}$, donde $z_{ij} = f_{ij}/\sqrt{f_{i.}f_{.j}}$ salvo el valor propio 1 que aparece en $\mathbf{Z}'\mathbf{Z}$, y no en $\mathbf{X}'\mathbf{X}$.

Capítulo 8

ANÁLISIS DE CONGLOMERADOS

8.1 FUNDAMENTOS

El análisis de conglomerados (clusters) tiene por objeto agrupar elementos en grupos homogéneos en función de las similitudes o similaridades entre ellos. Normalmente se agrupan las observaciones, pero el análisis de conglomerados puede también aplicarse para agrupar variables. Estos métodos se conocen también con el nombre de métodos de clasificación automática o no supervisada, o de reconocimiento de patrones sin supervisión. El nombre de no supervisados se aplica para distinguirlos del análisis discriminante, que estudiaremos en el capítulo 13. El análisis de conglomerados estudia tres tipos de problemas:

Partición de los datos. Disponemos de datos que sospechamos son heterogéneos y se desea dividirlos en un número de grupos prefijado, de manera que:

(1) cada elemento pertenezca a uno y solo uno de los grupos;

(2) todo elemento quede clasificado;

(3) cada grupo sea internamente homogéneo.

Por ejemplo, se dispone de una base de datos de compras de clientes y se desea hacer una tipología de estos clientes en función de sus pautas de consumo.

Construcción de jerarquías. Deseamos estructurar los elementos de un conjunto de forma jerárquica por su similitud. Por ejemplo, tenemos una encuesta de atributos de distintas profesiones y queremos ordenarlas por similitud. Una clasificación jerárquica implica que los datos se ordenan en niveles, de manera que los niveles superiores contienen a los inferiores. Este tipo de clasificación es muy frecuentes en biología, al clasificar animales, plantas etc. Estrictamente, estos métodos no definen grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos. Sin embargo, como veremos, la jerarquía construida permite obtener también una partición de los datos en grupos.

Clasificación de variables. En problemas con muchas variables es interesante hacer un estudio exploratorio inicial para dividir las variables en grupos. Este estudio puede orientarnos para plantear los modelos formales para reducir la dimensión que estudiaremos más adelante. Las variables pueden clasificarse en grupos o estructurarse en una jerarquía.

Los métodos de partición utilizan la matriz de datos, pero los algoritmos jerárquicos utilizan la matriz de distancias o similitudes entre elementos. Para agrupar variables se parte de la matriz de relación entre variables: para variables continuas suele ser la matriz de

correlación, y para variables discretas, se construye, como veremos, a partir de la distancia ji-cuadrado.

Vamos a estudiar en primer lugar los métodos de partición

8.2 MÉTODOS CLÁSICOS DE PARTICIÓN

8.2.1 Fundamentos del algoritmo de k-medias

Supongamos una muestra de n elementos con p variables. El objetivo es dividir esta muestra en un número de grupos prefijado, G. El algoritmo de k-medias (que con nuestra notación debería ser de $G-$medias) requiere las cuatro etapas siguientes :

(1) Seleccionar G puntos como centros de los grupos iniciales. Esto puede hacerse:

   a) asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos así formados;

   b) tomando como centros los G puntos más alejados entre sí ;

   c) construyendo los grupos con información a priori, o bien seleccionando los centros a priori.

(2) Calcular las distancias euclídeas de cada elementoa al centro de los G grupos, y asignar cada elemento al grupo más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.

(3) Definir un criterio de optimalidad y comprobar si reasignando uno a uno cada elemento de un grupo a otro mejora el criterio.

(4) Si no es posible mejorar el criterio de optimalidad, terminar el proceso.

8.2.2 Implementación del algoritmo

El criterio de homogeneidad que se utiliza en el algoritmo de k-medias es *la suma de cuadrados dentro de los grupos (SCDG)* para todas las variables, que es equivalente a la suma ponderada de las varianzas de las variables en los grupos:

$$SCDG=\sum_{g=1}^{G}\sum_{j=1}^{p}\sum_{i=1}^{n_g}(x_{ijg}-\overline{x}_{jg})^2 \tag{8.1}$$

donde x_{ijg} es el valor de la variable j en el elemento i del grupo g y $\overline{x}_{jg}$ la media de esta variable en el grupo. El criterio se escribe

$$\min SCDG = \min \sum_{g=1}^{G}\sum_{j=1}^{p} n_g s_{jg}^2 \tag{8.2}$$

donde n_g es el número de elementos del grupo g y s_{jg}^2 es la varianza de la variable j en dicho grupo. La varianza de cada variable en cada grupo es claramente una medida de la heterogeneidad del grupo y al minimizar las varianzas de todas las variables en los grupos obtendremos grupos más homogéneos. Un posible criterio alternativo de homogeneidad sería minimizar las distancias al cuadrado entre los centros de los grupos y los puntos que pertenecen a ese grupo. Si medimos las distancias con la norma euclídea, este criterio se escribe:

$$\min \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g) = \sum_{g=1}^{G} \sum_{i=1}^{n_g} d^2(i,g)$$

donde $d^2(i,g)$ es el cuadrado de la distancia euclídea entre el elemento i del grupo g y su media de grupo. Es fácil comprobar que ambos criterios son idénticos. Como un escalar es igual a su traza, podemos escribir este último criterio como

$$\min \sum_{g=1}^{G} \sum_{i=1}^{n_g} tr\left[d^2(i,g)\right] = \min tr\left[\sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'\right]$$

y llamando $\mathbf{W}$ a la matriz de suma de cuadrados dentro de los grupos,

$$\mathbf{W} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'$$

teenmos que

$$\min tr(\mathbf{W}) = \min SCDG$$

Como la traza es la suma de los elementos de la diagonal principal ambos criterios coinciden. Este criterio se denomina *criterio de la traza,* y fue propuesto por Ward (1963).

La maximización de este criterio requeriría calcularlo para todas las posibles particiones, labor claramente imposible, salvo para valores de n muy pequeños. El algoritmo de $k-$medias busca la partición óptima con la restricción de que en cada iteración sólo se permite mover un elemento de un grupo a otro. El algoritmo funciona como sigue

(1) Partir de una asignación inicial

(2) Comprobar si moviendo algún elemento se reduce $\mathbf{W}$.

(3) Si es posible reducir $\mathbf{W}$ mover el elemento, recalcular las medias de los dos grupos afectados por el cambio y volver a (2). Si no es posible reducir $\mathbf{W}$ terminar.

En consecuencia, el resultado del algortimo puede depender de la asignación inicial y del orden de los elementos. Conviene siempre repetir el algoritmo desde distintos valores iniciales y permutando los elemento de la muestra. El efecto del orden de las observaciones suele ser pequeño, pero conviene asegurarse en cada caso de que no esta afectando.

El criterio de la traza tiene dos propiedades importantes. La primera es que no es invariante ante cambios de medida en las variables. Cuando las variables vayan en unidades distintas conviene estandarizarlas, para evitar que el resultado del algoritmo de k-medias dependa de cambios irrelevantes en la escala de medida. Cuando vayan en las mismas

unidades suele ser mejor no estandarizar, ya que es posible que una varianza mucho mayor que el resto sea precisamente debida a que existen dos grupos de observaciones en esa variable, y si estandarizamos podemos ocultar la presencia de los grupos. Por ejemplo, la figura 8.1 muestra un ejemplo donde la estandarización puede hacer más difícil la identificación de los grupos.

Figura 8.1: La estandarización puede dificultar la identificación de los grupos.

La segunda propiedad del criterio de la traza es que minimizar la distancia euclídea produce grupos aproximadamente esféricos. Las razones para este hecho se estudiarán en el capítulo 15. Por otro lado este criterio esta pensado para variables cuantitativas y, aunque puede aplicarse si existe un pequeño número de variables binarias, si una parte importante de las variables son atributos, es mejor utilizar los métodos jerárquicos que se describen a continuación.

8.2.3 Número de grupos

En la aplicación habitual del algoritmo de k-medias hay que fijar el número de grupos, G. Es claro que este número no puede estimarse con un criterio de homogeneidad ya que la forma de conseguir grupos muy homogéneos y minimizar la SCDG es hacer tantos grupos como observaciones, con lo que siempre SCDG=0. Se han propuesto distintos métodos para seleccionar el número de grupos. Un procedimiento aproximado que se utiliza bastante, aunque puede no estar justificado en unos datos concretos, es realizar un test F aproximado de reducción de variabilidad, comparando la SCDG con G grupos con la de $G+1$, y calculando la reducción proporcional de variabilidad que se obtiene aumentando un grupo adicional. El test es:

$$F = \frac{SCDG(G) - SCDG(G+1)}{SCDG(G+1)/(n-G-1)} \tag{8.3}$$

y compara la disminución de variabilidad al aumentar un grupo con la varianza promedio. El valor de F suele compararse con una F con $p, p(n-G-1)$ grados de libertad, pero esta regla no esta muy justificada porque los datos no tienen porque verificar las hipótesis necesarias para aplicar la distribución F . Una regla empírica que da resultados razonables, sugerida por Hartigan (1975), e implantada en algunos programas informáticos, es introducir un grupo más si este cociente es mayor que 10.

Ejemplo 8.1 *La figura 8.2 presenta los datos de ruspini (fichero ruspini.dat) que incluye 75 datos de dos variables y que se han utilizado para comparar distintos algoritmos de clasificación. El gráfico muestra claramente cuatro grupos de datos en dos dimensiones.*

Figura 8.2: Datos de Ruspini

La tabla 8.1 muestra el resultado de aplicar el programa de k-medias en Minitab para distinto número de grupos a los datos sin estandarizar. De acuerdo con el criterio F existen tres grupos en los datos. Las figuras 8.3, 8.4, 8.5 y 8.6 muestran los grupos obtenidos con este programa.

La tabla se ha construido a partir de la información proporcionada por el programa. Al pasar de 2 a 3 grupos hay una reducción de variabilidad muy significativa dada por

$$F = \frac{89247 - 51154}{51154/(75-4)} = 52.87$$

Sin embargo al pasar de 3 a 4 grupos la reducción no es significativa

$$F = \frac{51154 - 50017}{50017/(75-5)} = 1.59.$$

El algortimo de k-medias implantado en minitab llevaría a dividir los datos en los tras grupos indicados en la figura 8.4. Si aplicamos el algoritmo a los datos estandarizados se

Número de grupos	tamaño	SCDG(i)	SCDG	F
2	34	43238		
	40	46009	89247	
3	20	3689		
	40	46009		
	15	1456	51154	52.8
4	4	170		
	16	2381		
	15	1456		
	40	46009	50017	1.59
5	4	170		
	5	292		
	11	857		
	40	46009		
	15	1456	48784	

Tabla 8.1: Tabla con la información para seleccionar el número de grupos con el algoritmo de k.medias.

obtienen de nuevo tres grupos, pero distintos: el primero esta formado por los dos conjuntos de puntos situados en la parte superior del gráfico y los otros dos grupos por los dos inferiores.

Figura 8.3: División de los datos de Ruspini en dos grupos con Minitab.

Para estudiar el funcionamiento de distintos programas hemos aplicado el mismo análisis a estos datos con el programa de k-medias de SPSS. La partición en dos grupos es la misma con ambos programas, pero la partición en tres y cuatro grupos es distinta como muestran las figuras 8.7, 8.8 y 8.9. El programa SPSS produce mejores resultados que Minitab. Este ejemplo sugiere que antes de aceptar los resultados de un análisis de conglomerados mediante

Figura 8.4: División de los datos de Ruspini en tres grupos con Minitab

Figura 8.5: División de los datos de Ruspini en cuatro grupos con Minitab

	$G=2$	$G=3$	$G=4$	$G=5$	$G=6$
eh	30	20	14	15	14
em	35	22	13	16	12
mi	509	230	129	76	83
tm	15	11	9	9	9
tn	64	58	37	35	26
Total=MS(G)	653	341	202	151	144
F		82.4	61.5	30.4	6.2

Tabla 8.2: Tabla con la información para seleccionar el número de grupos con el algoritmo de k.medias.

el algoritmo de K-medias conviene probar distintos puntos de partida y distintos algoritmos.

Ejemplo 8.2 *Vamos a aplicar el algoritmo de k-medias a los datos de los países. Se van a utilizar únicamente las 5 variables demográficas de MUNDODES. Comenzaremos comentando los resultados obtenidos al utilizar el programa k-medias con el programa SPSS. Para decidir el número de grupos este programa nos proporciona la varianza promedio dentro de los grupos para cada variable. Por ejemplo, si* $G = 2$*, dos grupos, la segunda columna de la tabla 15.1 indica que la varianza promedio dentro de los dos grupos o no explicada para la variable eh es 30, para la variable em es 35, y así sucesivamente. Este término se calcula como sigue: para cada variable hacemos la descomposición del análisis de la varianza de su suma de cuadrados total* $\sum(x_{ij} - \overline{x})^2$ *en la variabilidad explicada,* $\sum(\overline{x}_i - \overline{x})^2$*, donde* $\overline{x}_i$ *es la media de la variable en cada grupo, y la no explicada,* $\sum(x_{ij} - \overline{x}_i)^2$*. Este último término dividido por sus grados de libertad, que son* $n - G$ *proporciona la varianza promedio dentro de los grupos o no explicada. Según la definición La suma de estas varianzas multiplicada por* $n - G$ *proporciona el estadístico SCDG, como indica la fórmula (15.5). La tabla 15.1 resume esta información*

La tabla muestra que, como es de esperar, las varianzas promedio de las variables disminuyen al hacer más grupos. La tabla muestra que la variable mi tiene mucha más varianza que las demás, y por tanto va a tener un peso muy importante en la construcción de los grupos, que van a hacerse principalmente por los valores de esta variable. La tabla de las varianzas muestra que el número de grupos es cinco, ya que al aumentar a seis la disminución de las varianzas es muy pequeña. Podemos contrastar esta intuición calculando el estadístico F dado por (8.3). Llamando MS(G) a la fila de totales que será igual a SCDG(G)/(n-G), tenemos que este estadistico se calcula como

$$F = \frac{(n-G)MS(G) - (n-G-1)MS(G+1)}{MS(G+1)}$$

donde $n = 91$ *y* G *es el número de grupos indicado por columnas. Por ejemplo, el contraste para ver si conviene pasar de dos grupos a tres será*

$$F = \frac{89.653 - 88.341}{341} = 82.45$$

Figura 8.6: División de los datos de Ruspini en 5 grupos con Minitab

Figura 8.7: División en tres grupos de los datos de Ruspini con SPSS

Figura 8.8: División en cuatro grupos de los datos de Ruspini con SPSS

Figura 8.9: División en cinco grupos de los datos de Ruspini con SPSS

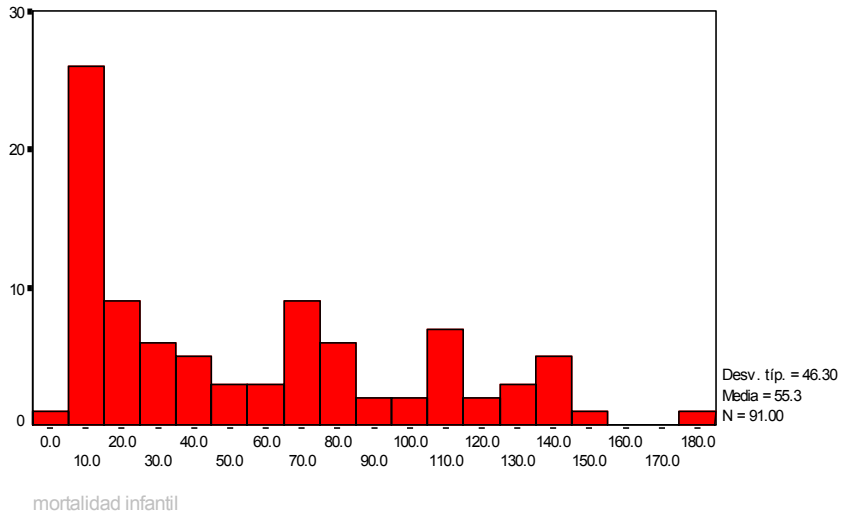


Figura 8.10: Histograma de la variable mortalidad infantil indicando la presencia de entre cuatro y cinco grupos de paises

Así se obtiene la fila de F de la tabla, y, de acuerdo con el criterio de Hartigan, escogeríamos cinco grupos.

Como hemos visto que la variable mi es muy importante en la construcción de los grupos, la figura 15.1 presenta un histograma de esta variable. Se observa que esta variable, que va a tener un peso dominante en la formación de los grupos, indica claramente la heterogeneidad de la muestra. En los grupos construidos el grupo con menor mortalidad infantil es el tres, que incluye los paises de Europa menos Albania, y el de mayor mortalidad, el dos, que incluye a los países más pobres de Africa.

La figura 8.11 ilustra la posición de los 5 grupos en el gráfico de las dos variables más influyentes y la figura 8.12 la composición de los grupos. Se observa que el grupo 3 esta ´formado por la mayoría de los países europeos, japón y norte américa, el grupo1 incluye los países europeos más pobres, los más ricos de latinoamérica y otros países como China y Egipto. El grupo 4 engloba países de desarrollo medio africanos (como suráfrica o Zaire) lationamericanos (Brasil) y de Asia como Arabia Saudita, India e Indonesia. Finalmente los grupos 5 y 2 incluye los países menos desarrollos.

Figura 8.11: Representación de los grupos en el gráfico de dispersión de las variables mortalidad infantil y tasa de natalidad

Figura 8.12: Indicación de los países que pertenecen a cada uno de los grupos.

Hemos repetido el análisis utilizando el programa Minitab para cinco grupos. Este programa propociona la suma de cuadrados dentro de los grupos por clusters (grupos) en lugar de por variables, como se indica:

```
Number of    Within cluster  Average distance Maximum distance
observations  sum of squares   from centroid    from centroid
```

Cluster1	21	10855.985	20.220	58.275
Cluster2	14	833.119	7.357	10.902
Cluster3	28	960.586	5.415	9.925
Cluster4	9	864.347	8.977	15.250
Cluster5	19	3126.014	12.110	21.066

Ejemplo 8.3 *Los resultados para datos sin estandarizar son parecidos, pero no idénticos, como puede verse en la figura 8.13, donde se han representado los grupos en el plano de las dos variables con mayor varianza y que van a tener más peso en la determinación de los grupso. Al estandarizar las variables los resultados cambian sustancialmente, al tener un peso mayor el resto de las variables, los grupos son más homogéneos por continentes y en Europa se separan los países occidentales y los orientales. Los resultados se presentan en la figura 8.14 donde de nuevo se han utilizado las dos variables más importantes.*

Figura 8.13: Resultados de k-medias con minitab para los datos de MUNDODES sin estandarizar. Se forman cinco grupos. En ordenadas la mortalidad infantil(C4) y en abcisas la tasa de natalidad (C2)

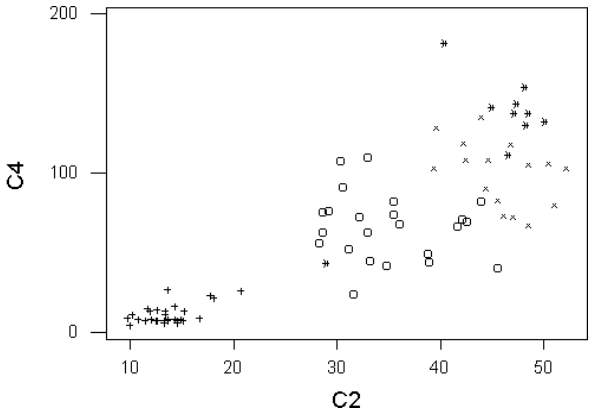


Figura 8.14: Resultados de k-medias para datos estandarizados de MUNDODES con el programa Minitab. En ordenadas la mortalidad infantil(C4) y en abcisas la tasa de natalidad (C2)

8.3 MÉTODOS JERÁRQUICOS

8.3.1 Distancias y Similaridades

Distancias Euclídeas

Los métodos jerárquicos parten de una matriz de distancias o similaridades entre los elementos de la muestra y construyen una jerarquía basada en una distancia. Si todas las variables son continuas, la distancia más utilizada es la distancia euclídea entre las variables estandarizadas. No es, en general, recomendable utilizar las distancias de Mahalanobis, ya que la única matriz de covarianzas disponible es la de toda la muestra, que puede mostrar unas correlaciones muy distintas de las que existen entre las variables dentro de los grupos. Por ejemplo, la figura 8.15 se ha generado con dos grupos de variables normales independientes de medias (0,0) y (5,5) y varianza unidad. La posición de los grupos genera en el conjunto de puntos una correlación positiva fuerte, que desaparece si consideramos cada uno de los grupos por separado.

Figura 8.15: Dos grupos con variables incorreladas pueden dar lugar a alta correlación entre las variables.

Para decidir si estandarizar las variables o no antes del análisis conviene tener en cuenta los comentarios anteriores y el objetivo del estudio. Si no estandarizamos, la distancia euclídea dependerá sobre todo de las variables con valores más grandes, y el resultado del análisis puede cambiar completamente al modificar su escala de medida. Si estandarizamos, estamos dando a priori un peso semejante a las variables, con independencia de su variabilidad original, lo que puede no ser siempre adecuado.

Cuando en la muestra existen variables continuas y atributos el problema se complica. Supongamos que la variable x_1 es binaria. La distancia euclídea entre dos elementos de la muestra en función de esta variable es $(x_{i1} - x_{h1})^2$ que tomará el valor cero si $x_{i1} = x_{h1}$, es decir cuando el atributo está, o no está, en ambos elementos, y uno si el atributo está en un elemento y no en el otro. Sin embargo, la distancia entre dos elementos correspondiente a una variable continua estandarizada, $(x_{i1} - x_{h1})^2/s_1^2$, puede ser mucho mayor que uno, con lo que las variables continuas van en general a pesar mucho más que las binarias. Esto puede ser aceptable en muchos casos, pero cuando, por la naturaleza del problema, esta situación no sea deseable, la solución es trabajar con similaridades.

Similaridades El coeficiente de similaridad según la variable $j = 1, ..., p$ entre dos elementos muestrales (i, h), se define como una función, s_{jih}, no negativa y simétrica:

(1) $s_{jii} = 1$

(2) $0 \leq s_{jih} \leq 1$

(3) $s_{jih} = s_{jhi}$

Si obtenemos las similaridades para cada variable entre dos elementos podemos combinarlas en un coeficiente de similaridad global entre los dos elementos. El coeficiente propuesto

por Gower es

$$s_{ih} = \frac{\sum_{j=1}^{p} w_{jih} s_{jih}}{\sum_{j=1}^{p} w_{jih}} \tag{8.4}$$

donde w_{jih} es una variable ficticia que es igual a uno si la comparación de estos dos elementos mediante la variable j tiene sentido, y será cero si no queremos incluir esa variable en la comparación entre los elementos. Por ejemplo, si la variable x_1 es si una persona ha pedido ($x_1 = 1$) o no ($x_1 = 0$) un crédito y la x_2 si lo ha devuelto o no, si una persona no ha pedido crédito, tiene $x_1 = 0$, no tienen sentido preocuparse de x_2. En este caso al comparar individuos (i, j) si uno cualquiera de los dos tiene un valor cero en x_1, asignaremos a la variable w_{2ij} el valor cero

Las similaridades entre elementos en función de las variables cualitativas pueden construirse individualmente o por bloques. La similaridad entre dos elementos por una variable binaria será uno, si ambos tienen el atributo, y cero en caso contrario. Alternativamente, podemos agrupar las variables binarias en grupos homogéneos y tratarlas conjuntamente. Si suponemos que todos los atributos tienen el mismo peso, podemos construir una medida de similaridad entre dos elementos A y B respecto a todos estos atributos contando el número de atributos que están presentes:

(1) en ambos (a);
(2) en A y no en B, (b);
(3) en B y no en A, (c);
(4) en ninguno de los dos elementos, (d).

Estas cuatro cantidades forman una *tabla de asociación entre elementos,* y servirán para construir medidas de similitud o similaridad entre los dos elementos comparados. En esta tabla se verifica que $n_a = a + b + c + d$, donde n_a es el número de atributos.

Elementos	variables (atributos)						
	x_1	x_2	x_3	x_4	x_5	x_6	x_7
A	0	1	1	0	0	0	1
B	1	0	1	1	1	1	0
C	1	0	0	1	1	1	1
.	.	.	.	.	.	.	

Tabla 8.3: Matriz de datos cuando las variables son atributos binarios

Por ejemplo, la tabla 8.3 presenta una posible matriz de datos con siete atributos binarios y con ella se ha construido la tabla 8.4 de asociación que presenta la distribución conjunta de los valores 0 y 1 para los elementos A y B. El elemento A tiene 3 valores 1 en el conjunto de variables binarias y de estos tres casos, en una ocasión también el elemento B tiene el valor 1, y en otras dos tiene el valor 0. El elemento A toma 4 veces el valor 0, ninguna coincidiendo con B y las cuatro con B tomando el valor uno. La suma de los totales de filas y columnas debe ser siempre el número de atributos binarios considerados. Para calcular un coeficiente de similitud entre dos individuos a partir de su tabla de asociación se utilizan los dos criterios principales siguientes:

		B		
		1	0	
A	1	1 (a)	2 (b)	3
	0	4 (c)	0 (d)	4
Suma		5	2	7

Tabla 8.4: Tabla de asociación correspondiente a los elementos A y B

1. *Proporción de coincidencias.* Se calcula como el número total de coincidencias sobre el número de atributos totales:

$$s_{ij} = \frac{a+d}{n_a}. \tag{8.5}$$

   por ejemplo la similitud de Ay B es 1/7, y la de B y C es 5/7.

2. *Proporción de apariciones.* Cuando la ausencia de un atributo no es relevante, podemos excluir las ausencias y calcular sólo la proporción de veces donde el atributo aparece en ambos elementos. El coeficiente se define por:

$$s_{ij} = \frac{a}{a+b+c} \tag{8.6}$$

Por ejemplo con este criterio en la tabla 8.3 la similitud entre A y B es también 1/7 , y la de B y C es 4/6.

Aunque las dos propuestas anteriores son las más utilizadas puede haber situaciones donde sean recomendables otras medidas. Por ejemplo, podemos querer dar peso doble a las coincidencias, con lo que resulta $s_{ij} = 2(a+d)/(2(a+d)+b+c)$, o tener sólo en cuenta las coincidencias y tomar $s_{ij} = a/(b+c)$. Finalmente los coeficientes de similitud o similaridad para una variable continua se construye mediante

$$s_{jih} = 1 - \frac{|x_{ij} - x_{hj}|}{\text{rango}(x_j)}$$

de esta manera el coeficiente resultante estará siempre entre cero y uno. Cuando tenemos varias variables estos coeficientes pueden combinarse como indica la expresión (8.4).

Una vez obtenida la similaridad global entre los elementos, podemos transformar los coeficientes en distancias. Lo más simple es definir la distancia mediante $d_{ij} = 1 - s_{ij}$, pero está relación puede no verificar la propiedad triangular. Puede demostrarse que si la matriz de similaridades es definida positiva (lo que ocurrirá si calculamos las similitudes por (8.5) o (8.6), y definimos la distancia por:

$$d_{ij} = \sqrt{2(1 - s_{ij})}$$

entonces sí se verifica la propiedad triangular (véase el ejercicio 6.5)

8.3.2 Algoritmos Jerárquicos

Dada una matriz de distancias o de similitudes se desea clasificar los elementos en una jerarquía. Los algoritmos existentes funcionan de manera que los elementos son sucesivamente asignados a los grupos, pero la asignación es irrevocable, es decir, una vez hecha, no se cuestiona nunca más. Los algoritmos son de dos tipos:

1. *De aglomeración.* Parten de los elementos individuales y los van agregando en grupos.
2. *De división.* Parten del conjunto de elementos y lo van dividiendo sucesivamente hasta llegar a los elementos individuales.

Los algoritmos de aglomeración requieren menos tiempo de cálculo y son los más utilizados. El lector puede consultar los algoritmos de división en Seber (1984).

8.3.3 Métodos Aglomerativos

Los algoritmos aglomerativo que se utilizan tienen siempre la misma estructura y sólo se diferencian en la forma de calcular las distancias entre grupos. Su estructura es:

1. Comenzar con tantas clases como elementos, n. Las distancias entre clases son las distancias entre elementos originales.
2. Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos una clase.
3. Sustituir los dos elementos utilizados en (2) para definir la clase por un nuevo elemento que represente la clase construida. Las distancias entre este nuevo elemento y los anteriores se calculan con uno de los criterios que comentamos a continuación.
4. Volver a (2) y repetir (2) y (3) hasta que tengamos todos los elementos agrupados en una clase única.

Criterios para definir distancias entre grupos

Supongamos que tenemos un grupo A con n_a elementos, y un grupo B con n_b elementos, y que ambos se fusionan para crear un grupo (AB) con $n_a + n_b$ elementos. La distancia del nuevo grupo, (AB), a otro grupo C con n_c elementos, se calcula habitualmente por alguna de las cinco reglas siguientes:

1. *Encadenamiento simple o vecino más próximo.* La distancia entre los dos nuevos grupos es la menor de las distancias entre grupos antes de la fusión. Es decir:

$$d(C; AB) = min\,(d_{CA}, d_{CB})$$

Una forma simple de calcular con un ordenador el mínimo entre las dos distancias es utilizar que

$$min\,(d_{CA}, d_{CB}) = 1/2\,(d_{CA} + d_{CB} - |d_{CA} - d_{CB}|)$$

En efecto, si $d_{CB} > d_{CA}$ el término en valor absoluto es $d_{CB} - d_{CA}$ y el resultado de la operación es d_{CA}, la menor de las distancias. Si $d_{CA} > d_{CB}$ el segundo término es $d_{CA} - d_{CB}$ y se obtiene d_{CB}.

Como este criterio sólo depende del orden de las distancias será invariante ante transformaciones monótonas: obtendremos la misma jerarquía aunque las distancias sean numéricamente distintas. Se ha comprobado que este criterio tiende a producir grupos alargados, que pueden incluir elementos muy distintos en los extremos.

2. *Encadenamiento completo o vecino más alejado.* La distancia entre los dos nuevos grupos es la mayor de las distancias entre grupos antes de la fusión. Es decir:

$$d(C; AB) = máx\,(d_{CA}, d_{CB})$$

y puede comprobarse que

$$máx\,(d_{CA}, d_{CB}) = 1/2\,(d_{CA} + d_{CB} + |d_{CA} - d_{CB}|)\,.$$

Este criterio será también invariante ante transformaciones monótonas de las distancias al depender, como el anterior, del orden de las distancias. Tiende a producir grupos esféricos.

3. *Media de grupos.* La distancia entre los dos nuevos grupos es la media ponderada entre las distancias entre grupos antes de la fusión. Es decir:

$$d(C; AB) = \frac{n_a}{n_a + n_b} d_{CA} + \frac{n_b}{n_a + n_b} d_{CB}$$

Como se ponderan los valores de las distancias, este criterio no es invariante ante transformaciones monótonas de las distancias.

4. *Método del centroide.* Se aplica generalmente sólo con variables continuas. La distancia entre dos grupos se hace igual a la distancia euclídea entre sus centros, donde se toman como centros los vectores de medias de las observaciones que pertenecen al grupo. Cuando se unen dos grupos se pueden calcular las nuevas distancias entre ellos sin utilizar los elementos originales. Puede demostrarse (véase ejercicio 8.5) que el cuadrado de la distancia euclídea de un grupo C a la unión de los grupos A, con n_a elementos y B con n_b es

$$d^2(C; AB) = \frac{n_a}{n_a + n_b} d^2_{CA} + \frac{n_b}{n_a + n_b} d^2_{CB} - \frac{n_a n_b}{(n_a + n_b)^2} d^2_{AB}$$

El método de Ward

Un proceso algo diferente de construir el agrupamiento jerárquico ha sido propuesto por Ward y Wishart. La diferencia con los métodos anteriores es que ahora se parte de los elementos directamente, en lugar de utilizar la matriz de distancias, y se define una medida global de la heterogeneidad de una agrupación de observaciones en grupos. Esta medida es

W, ya utilizada en la sección 8.2, la suma de las distancias euclídeas al cuadrado entre cada elemento y la media de su grupo:

$$\mathbf{W} = \sum_{g} \sum_{i \in g} (\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)'(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g) \tag{8.7}$$

donde $\overline{\mathbf{x}}_g$ es la media del grupo g. El criterio comienza suponiendo que cada dato forma un grupo, $g = n$ y por tanto **W** (8.7) es cero. A continuación se unen los elementos que produzcan el incremento minimo de **W**. Obviamente esto implica tomar los más próximos con la distancia euclídea. En la siguiente etapa tenemos $n-1$ grupos, $n-2$ de un elemento y uno de dos elementos. Decidimos de nuevo que dos grupos unir para que **W** crezca lo menos posible, con lo que pasamos a $n-2$ grupos y así sucesivamente hasta tener un único grupo. Los valores de **W** van indicando el crecimiento del criterio al formar grupos y pueden utilizarse para decidir cuantos grupos naturales contienen nuestros datos.

Puede demostrarse que, en cada etapa, los grupos que debe unirse para minimizar **W** son aquellos tales que:

$$\min \frac{n_a n_b}{n_a + n_b} (\overline{\mathbf{x}}_a - \overline{\mathbf{x}}_b)'(\overline{\mathbf{x}}_a - \overline{\mathbf{x}}_b)$$

Comparación

Es difícil dar reglas generales que justifiquen un criterio sobre otro, aunque los más utilizados son los tres últimos. Nuestra recomendación es analizar que criterio es más razonable para los datos que se quieren agrupar y , en caso de duda, probar con varios y comparar los resultados.

El dendrograma

El dendrograma, o árbol jerárquico, es una representación gráfica del resultado del proceso de agrupamiento en forma de árbol. Los criterios para definir distancias que hemos presentado tienen la propiedad de que, si consideramos tres grupos, A, B, C, se verifica que

$$d(A, C) \leq \max \{d(A, B), D(B, C)\}$$

y una medida de distancia que tiene esta propiedad se denomina *ultramétrica.* Esta propiedad es más fuerte que la propiedad triangular, ya que una ultramétrica es siempre una distancia. En efecto si $d^2(A, C)$ es menor o igual que el máximo de $d^2(A, B), d^2(B, C)$ forzosamente será menor o igual que la suma $d^2(A, B) + d^2(B, C)$. El dendrograma es la repreentación de una ultramétrica, y se contruye como sigue:

1. En la parte inferior del gráfico se disponen los n elementos iniciales.
2. Las uniones entre elementos se representan por tres lineas rectas. Dos dirigidas a los elementos que se unen y que son perpendiculares al eje de los elementos y una paralela a este eje que se sitúa al nivel en que se unen.
3. El proceso se repite hasta que todos los elementos están concetados por lineas rectas.

Si cortamos el dendrograma a un nivel de distancia dado, obtenemos una clasificación del número de grupos existentes a ese nivel y los elementos que los forman.

El dendrograma es útil cuando los puntos tienen claramente una estructura jerárquica, pero puede ser engañoso cuando se aplica ciegamente, ya que dos puntos pueden parecer próximos cuando no lo están, y pueden aparecer alejados cuando están próximos.

Ejemplo 8.4 *Aplicaremos los algoritmos estudiados a la siguiente matriz inicial de distancias entre elementos*

$$\begin{array}{c|cccc} & A & B & C & D \\ A & 0 & 1 & 4 & 2,5 \\ B & 1 & 0 & 2 & 3 \\ C & 2 & 2 & 0 & 4 \\ D & 2,5 & 3 & 4 & 0 \end{array} = \begin{array}{|cccc|} 0 & 1 & 4 & 2,5 \\ & 0 & 2 & 3 \\ & & 0 & 4 \\ & & & 0 \end{array}$$

Método 1 encadenamiento simple o vecino más próximo. El valor mínimo fuera de la diagonal de la matriz de distancias es 1, y corresponde a la distancia entre los elementos Ay B. Los unimos para formar un grupo y calcularemos la nueva distancias de un elemento al grupo (AB) como la mínima de las distancias de ese elemento a A y a B. Es decir:

$$\begin{aligned} d(AB, C) &= min(4; 2) = 2; \\ d(AB, D) &= min(2,5; 3) = 2,5. \end{aligned}$$

La nueva tabla de distancias se obtiene de la anterior tachando las filas y columnas de A y B y añadiendo una nueva columna y una nueva fila correspondiente al grupo AB que contiene las nuevas distancias. El resultado es :

	AB	C	D
AB	0	2	2,5
C	2	0	4
D	2,5	4	0

El valor mínimo fuera de la diagonal de la tabla es ahora 2, que corresponde a la distancia entre AB y C. Uniendo estos dos grupos en uno y calculando las distancias al nuevo grupo :

$$d(ABC, D) = min(2,5; 4) = 2,5.$$

y finalmente se unen los dos grupos finales ABC y D. Este proceso se representa en el dendrograma de la figura 8.16

El dendrograma indica que primero se unen los dos elementos A y B a distancia uno, ese grupo se une al C con distancia 2 y el ABC al D a distancia 2,5.

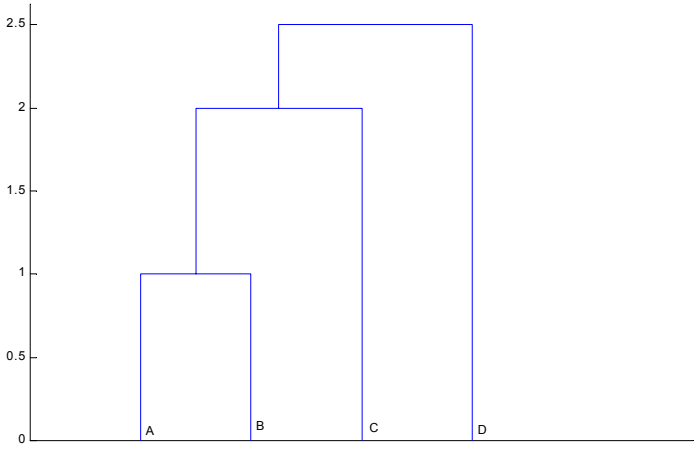


Figura 8.16: Dendrograma del método de encadenamiento simple

Método 2. Encadenamiento completo o vecino más alejado. La primera unión se hace igual que en el caso anterior entre A y B a distancia uno. Sin embargo, ahora las nuevas distancias son:

$$\begin{aligned} d(AB,C) &= máx(4;2) = 4; \\ d(AB,D) &= máx(2,5;3) = 3 \end{aligned}$$

y la siguiente unión será entre AB y D a distancia tres. La distancia de C al grupo ABD es 4 y esa será la siguiente unión. La figura 8.17 resume el proceso.

Método 3 . El inicio es, como en los métodos anteriores, la unión de los elementos más próximos, AB. Las nuevas distancias son d(AB,C)=3; d(AB,D)=2,75. Por tanto, la siguiente unión será entre AB y D a distancia 2,75. Este grupo ABD se unirá a C a su distancia que es d(ABC,D) = 1/2(4+2,75) = 3,375. La figura 8.18 resume el proceso.
Método 4 . El inicio es, como en los métodos anteriores. Las nuevas distancias se calculan como $d^2(C;AB) = \frac{1}{2}d^2_{CA} + \frac{1}{2}d^2_{CB} - \frac{1}{4}d^2_{AB} = 8 + 2 - 0,25 = 9,75$. Análogamente $d^2(D;AB) = 2,5^2/2 + 9/2 - 1/4 = 7,375$. La unión será con D a distancia $\sqrt{7,375} = 2.72$. La distancia de C al nuevo grupo será $d^2(C;ABD) = \frac{1}{3}9,75 + \frac{1}{2}16 - \frac{1}{4}7,375 = 3.16^2$, y C se unirá al grupo a la distancia 3.16. La figura 8.19 presenta el dendograma.

Ejemplo 8.5 *La figura 8.20 presenta el dendrograma hecho con MINITAB para los paises de MUNDODES con el método de la disminución de la suma de cuadrados (Ward). El gráfico sugiere la presencia de cuatro o cinco grupos de paises.*

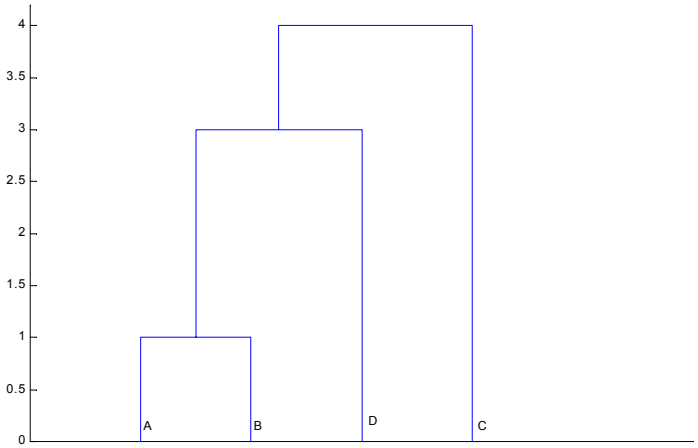


Figura 8.17: Dendrograma del método de encadenamiento completo

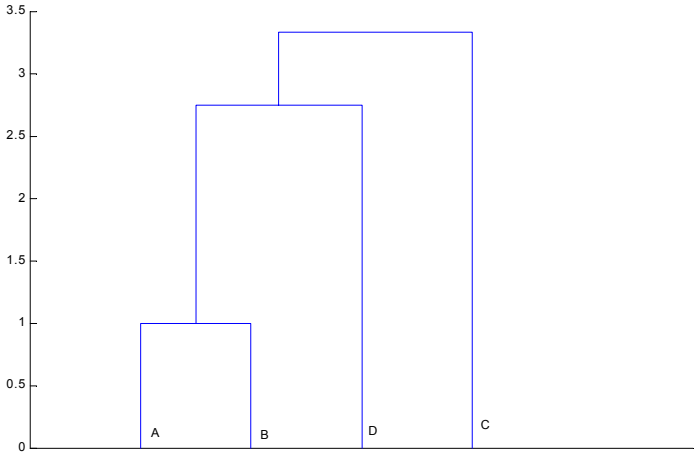


Figura 8.18: Dendrograma del método de la media de los grupos

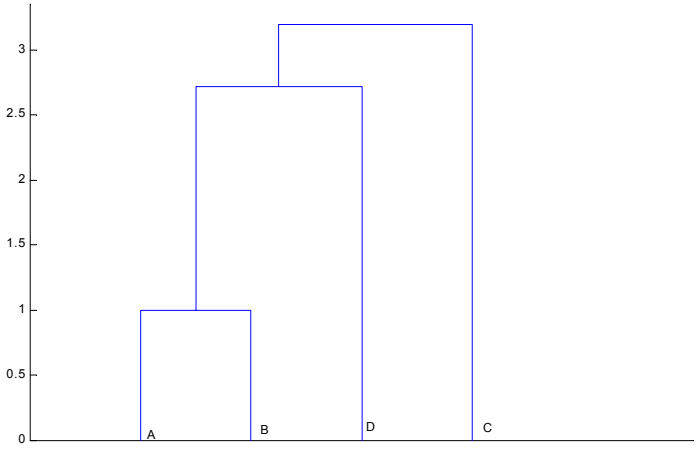


Figura 8.19: Dendrograma del método del centroide.

Figura 8.20: Resultados de un agrupamiento jerárquico de los paises de MUNDODES por las variables de natalidad

La figura muestra el resultado del encadenamiento simple, que es mucho más confuso.

Figura 8.21: Resultados de una aglomeración jerárquica para los paises de MUNDODES con encadenamiento simple.

Para comparar los resultados del agrupamiento jerárquico y el de partición la figura 8.22 presenta los grupos obtenidos para los datos estandarizados y con el criterio de Ward en el gráfico de las variables tasa de natalidad y mortalidad infantil.

Figura 8.22: Resultado del agrupamiento jerárquico cortado en cinco grupos para variables estandarizadas de MUNDODES

8.4 CONGLOMERADOS POR VARIABLES

El análisis de conglomerados de variables es un procedimiento exploratorio que puede sugerir procedimientos de reducción de la dimensión, como el análisis factorial o los métodos de correlación canónica que estudiaremos en la segunda parte del libro. La idea es construir una matriz de distancias o similitudes entre variables y aplicar a esta matriz un algoritmo jerárquico de clasificación.

8.4.1 Medidas de distancia y similitud entre variables

Las medidas habituales de asociación entre variables continuas son la covarianza y la correlación. Estas medidas tienen en cuenta únicamente las relaciones lineales. Alternativamente, podríamos construir una medida de distancia entre dos variables $\mathbf{x}_j$ y $\mathbf{x}_h$ representando cada variable como un punto en $\Re^n$ y calculando la distancia euclídea entre los dos puntos. Esta medida es:

$$d_{jh}^2 = \sum_{i=1}^{n}(x_{ij} - x_{ih})^2 \tag{8.8}$$

$$= \sum x_{ij}^2 + \sum x_{ih}^2 - 2\sum x_{ij}x_{ih}. \tag{8.9}$$

Para que la distancia no dependa de las unidades, las variables deben estar estandarizadas. En otro caso la distancia entre dos variables podría alterarse arbitrariamente mediante transformaciones lineales de éstas. (Por ejemplo, midiendo las estaturas en metros, en lugar de en cm. y en desviaciones respecto a la media poblacional en lugar de con carácter absoluto). Suponiendo, por tanto, que trabajamos con variables estandarizadas de media cero y varianza uno, se obtiene que (8.8) se reduce a:

$$d_{jh}^2 = 2n(1 - r_{jh}).$$

Observemos que:

(a) si $r_{jh} = 1$, la distancia es cero, indicando que las dos variables son idénticas.

(b) si $r_{jh} = 0$, las dos variables están incorreladas y la distancia es $d_{jh} = \sqrt{2n}$.

(c) si $r_{jh} < 0$, las dos variables tienen correlación negativa, y la distancia tomará su valor máximo, $\sqrt{4n}$, cuando las dos variables tengan correlación -1.

Esta medida de distancia puede estandarizarse para que sus valores estén entre cero y uno prescindiendo de la constante n y tomando $d_{jh} = \sqrt{(1 - r_{jh})/2}$.

Para variables cualitativas binarias se puede construir una medida de similitud de forma similar a como se hizo con los elementos construyendo *una tabla de asociación entre variables.* Para ello se cuenta el número de elementos donde están presentes ambas características (a), donde esta sólo una de ellas (b) y (c), y donde no lo están ninguna de las dos (d). En estas

tablas se verifica que si n es el número de individuos $n = a + b + c + d$, y podemos construir coeficientes de similitud como se hizo con los elementos. Alternativamente, esta tabla de asociación entre variables es una tabla de contingencia (véase el capítulo 7) y una medida de distancia es el valor de la ji-cuadrado (veáse el Apéndice 8.1)

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(a+c)(c+d)(b+d)}.$$

Es más habitual definir la distancia por el coeficiente de contingencia

$$d_{ij} = 1 - \sqrt{\frac{\chi^2}{n}}.$$

8.5 Lecturas complementarias

Un libro pionero sobre métodos de agrupamiento en español es Escudero (1977), que presenta una visión muy amplia de distintas técnicas de agrupación. La literatura sobre cluster en inglés es extensa: Anderberg (1973), Everitt (1993), Gordon (1981), Hartigan (1975), Mirkin (1996) , Spath y Bull (1980) y Spath (1985), están dedicados a este tema. La mayoría de los libros generales dedican también un capítulo a estos métodos.

Ejemplo 8.6 *La figura8.23 muestra el dendrograma del agrupamiento de las variables de los datos de EUROSEC. El criterio utilizado es el de Ward. Se observa que la agrupación de las variables coincide con lo esperado: primero se unen minería y energía, sevicios y servicions industriales, e industria y construcción. En un segundo nivel tenemos servicios (que engloba las tres variables servicios, servicios industriales y finanzas), agricultura, que esta sóla e industria, que recoge el resto de las variables industriales.*

Figura 8.23: Agrupamiento por variables de los datos de EUROSEC

Ejemplo 8.7 *El dendrograma de la figura8.24 muestra la agrupación de las variables para las medidas físicas, MEDIFIS. La correlación más estrecha se da entre longitud del pie y estatura, y la variable diámetro del cráneo esta poco relacionada con el resto como obtuvimos anteriormente. Si quisiésemos hacer grupos a un primer nivel tenemos tres grupos de variables, de longitud, con 4 variables, de anchura, con dos, y el diámetro de la cabeza. A un nivel superior quedan todas las variables en un lado y el diámetro de la cabeza en el otro.*

Figura 8.24: Dendrograma de las medidas físicas con el criterio de Ward.

Ejemplo 8.8 *La figura 8.25 presenta los resultados para las variables de INVES. A un nivel bajo tenemos cuatro grupos de variables: química, ingeniería, agricultura y biología y el resto, que incluye 4 variables. A un nivel superior los dos últimos grupos se unen y la distancia mayor se da entre el banco de datos químicos y el resto.*

Figura 8.25: Dendrograma de las variables de INVES

EJERCICIOS

Ejercicio 8.1 *Aplicar el algoritmo de k-medias a los datos de los presupuestos familiares. ¿Cuántos grupos hay en lso datos?*

Ejercicio 8.2 *Aplicar un agrupamiento jerárquico a los datos de los presupuestos familiares. Comparar el resultado con distintos métodos de agrupación. Compararlos con los resultados de k-medias*

Ejercicio 8.3 *Demostrar que el criterio de Hartigan para el algoritmo de k-medias equivale a continuar añadiendo grupos hasta que* $tr(\mathbf{W}_G) < tr(\mathbf{W}_{G+1})(n-G+9)/(n-G-1)$ *(Sugerencia utilizar que* $tr(\mathbf{W}) = SCDG$*, e imponer la condición de que el valor de F sea mayor que 10)*

Ejercicio 8.4 *Demostrar que si definimos* $\mathbf{T} = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(\mathbf{x}_{ig}-\overline{\mathbf{x}})(\mathbf{x}_{ig}-\overline{\mathbf{x}})'$ *a la suma de cuadrados totales podemos escribir* $\mathbf{T} = \mathbf{B} + \mathbf{W}$*, donde W se ha definido en la sección 8.2 y* $\mathbf{B}$ *es la matriz de suma de cuadrados entre grupos.*

Ejercicio 8.5 *Demostrar que las distancias entre grupos con encadenamiento simple, completo y media de grupos pueden calcularse con* $\alpha d_{CA} + \alpha d_{CB} + \beta\,|d_{CA} - d_{CB}|$ *y obtener los valores de* α *y* β *que dan lugar a estas distancias.*

Ejercicio 8.6 *Demostrar que en aglomeramiento jerárquico podemos calcular las distancias euclídeas al cuadrado entre un grupo C a la unión de los grupos A, con* n_a *elementos y B con* n_b *mediante* $d^2(C;AB) = \frac{n_a}{n_a+n_b}d^2_{CA} + \frac{n_b}{n_a+n_b}d^2_{CB} - \frac{n_a n_b}{(n_a+n_b)^2}d^2_{AB}$ *.*

(sugerencia: La media de la unión de los grupos A y B tendrá de coordenadas $\overline{x}_{AB} = \frac{n_a}{n_a+n_b}\overline{x}_A + \frac{n_b}{n_a+n_b}\overline{x}_B$*, sustituir esa expresión en la distancia de C a ese punto* $(\overline{x}_C - \overline{x}_{AB})'(\overline{x}_C - \overline{x}_{AB})$ *y desarrollar.*

APÉNDICE 8.1. CÁLCULO DEL ESTADISTICO JI-CUADRADO EN TABLAS 2×2

En la tabla de contingencia $\{a, b, c, d\}$ las frecuencias esperadas son $\frac{1}{n}\{(a+c)(a+b), (a+b)(b+d), (b+d)(c+d)\}$ y el valor de la χ^2 definida en la seccion 7.3 es:

$$\chi^2 = \left(\frac{ad-bc}{n}\right)^2 \left[\frac{n}{(a+c)(a+b)} + \frac{n}{(a+b)(b+d)} + \frac{n}{(a+c)(c+d)} + \frac{n}{(b+d)(c+d)}\right]$$

En efecto, como la tabla tiene un grado de libertad, las discrepancias entre las frecuencias observadas y esperadas deben de ser iguales, por ejemplo para la primera casilla

$$\left(a - \frac{(a+c)\,(a+b)}{n}\right)^2 = \left(\frac{na - a(a+b+c) - bc}{n}\right)^2 = \left(\frac{ad-bc}{n}\right)^2$$

y lo mismo se obtiene en las restantes. Como:

$$(b+d)(d+c) + (a+c)(c+d) + (a+b)(b+d) + (a+c)(a+b) =$$

$$(b+d)n + (a+c)n = (a+b+c+d)n = n^2$$

resulta finalmente que:

$$\chi^2 = \frac{(ab-bc)^2 n}{(a+b)(a+c)(b+d)(c+d)}.$$

Capítulo 9

DISTRIBUCIONES MULTIVARIANTES

9.1 CONCEPTOS BÁSICOS.

El problema central en la análisis de datos es decidir si las propiedades encontradas en una muestra pueden generalizarse a la población de la que proviene. Para poder realizar esta extrapolación necesitamos construir un modelo del sistema generador de los datos, es decir, suponer una distribución de probabilidad para la variable aleatoria en la población. Este capítulo repasa los conceptos básicos para construir modelos estadísticos multivariantes y presenta las distribuciones que se utilizarán para la inferencia en los capítulos siguientes.

9.1.1 Variables aleatorias vectoriales.

Una variable aleatoria vectorial es el resultado de observar p características en un elemento de una población. Por ejemplo, si observamos la edad y el peso de los estudiantes de una universidad tendremos valores de una variable aleatoria bidimensional; si observamos el número de trabajadores, las ventas y los beneficios de las empresas de un sector, tendremos una variable aleatoria tridimensional.

Diremos que se ha definido la distribución conjunta de una variable aleatoria vectorial cuando se especifique:

1. El espacio muestral o conjunto de sus valores posibles. Representando cada valor por un punto en el espacio de dimensión p, $\Re^p$, de los números reales, el espacio muestral es, en general, un subconjunto de este espacio.

2. Las probabilidades de cada posible resultado (subconjunto de puntos) del espacio muestral.

Diremos que la variable vectorial $p-$dimensional es discreta, cuando lo es cada una de las $p-$variables escalares que la componen. Por ejemplo, el color de los ojos y del cabello forman una variable bidimensional discreta. Análogamente, diremos que la variable es continua si sus componentes lo son. Cuando algunos de sus componentes sean discretos y otros continuos

diremos que la variable vectorial es mixta. Por ejemplo, la variable: género (0=hombre, 1=mujer), estatura y peso de personas, es tridimensional mixta. En este capítulo, para simplificar la exposición, y salvo indicación en otro sentido, supondremos que la variable vectorial es continua.

9.1.2 Distribución conjunta

La función de distribución conjunta de una variable aleatoria vectorial $F(\mathbf{x})$ se define en un punto $\mathbf{x}^0 = (x_1^0, ..., x_p^0)$ mediante:

$$F(\mathbf{x}^0) = P(\mathbf{x} \leq \mathbf{x}^0) = P(x_1 \leq x_1^0, ..., x_p \leq x_p^0)$$

donde $P(\mathbf{x} \leq \mathbf{x}^0)$ representa la probabilidad de que la variable tome valores menores o iguales al valor particular considerado, $\mathbf{x}^0$. Por tanto, la función de distribución acumula las probabilidades de todos los valores menores o iguales al punto considerado, y será no decreciente. Aunque la función de distribución tiene un gran interés teórico, es más cómodo en la práctica trabajar con la función de densidad para variables continuas, o con la función de probabilidades para las discretas. Llamaremos función de probabilidad de una variable discreta a la función $p(\mathbf{x}^0)$ definida por

$$p(\mathbf{x}^0) = P(\mathbf{x} = \mathbf{x}^0) = P(x_1 = x_1^0, ..., x_p = x_p^0).$$

Diremos que el vector $\mathbf{x}$ es absolutamente continuo si existe una función de densidad, $f(\mathbf{x})$, que satisface:

$$F(\mathbf{x}^0) = \int_{-\infty}^{\mathbf{x}^0} f(\mathbf{x}) d\mathbf{x}, \tag{9.1}$$

donde $d\mathbf{x} = dx_1....dx_p$ y la integral es una integral múltiple en dimensión p. La densidad de probabilidad tiene la interpretación habitual de una densidad: masa por unidad de volumen. Por tanto la función de densidad conjunta debe verificar

a) $f(\mathbf{x}) = f(x_1....., x_p) \geq 0$. La densidad es siempre no negativa.

b) $\int_{-\infty}^{\infty} f(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} ..... \int_{-\infty}^{\infty} f(x_1, ...x_p)\ dx_1....dx_p = 1$. Si multiplicamos la densidad en cada punto por el elemento de volumen en p dimensiones (que, si $p = 2$, será el área de un rectángulo, si $p = 3$ el volumen de un paralepípedo, etc) y sumamos (integramos) para todos los puntos con densidad no nula, obtenemos la masa de probabilidad total, que se estandariza al valor unidad.

Las probabilidades de sucesos definidos como subconjuntos del espacio muestral serán iguales a la masa de probabilidad correspondiente al subconjunto. Estas probabilidades se calcularán integrando la función de densidad sobre el subconjunto. Por ejemplo, para una variable bidimensional y sucesos A del tipo $A = (a < x_1 \leq b;\ c < x_2 \leq d)$:

$$P(A) = \int_a^b \int_c^d f(x_1, x_2) dx_1 dx_2$$

mientras que, en general,

$$P(A) = \int_A f(\mathbf{x}) d\mathbf{x}.$$

En este capítulo, y para simplificar la notación, utilizaremos la letra f para referirnos a la función de densidad de cualquier variable e indicaremos la variable por el argumento de la función, de manera que $f(x_1)$ es la función de densidad de la variable x_1 , y $f(x_1, x_2)$ es la función de densidad de la variable bidimensional (x_1, x_2).

9.1.3 Distribuciones marginales y condicionadas

Dada una variable aleatoria vectorial $p-$dimensional $(x_1, ..., x_p)$ llamaremos *distribución marginal* de cada componente x_i a la distribución univariante de dicho componente, considerado individualmente, e ignorando los valores del resto de los componentes. Por ejemplo, para variables bidimensionales continuas las distribuciones marginales se obtienen como:

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2, \tag{9.2}$$

$$f(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1, \tag{9.3}$$

y representan la función de densidad de cada variable ignorando los valores que toma la otra. Como hemos indicado antes, la letra f se refiere genericamente a una función de densidad. Por ejemplo, la ecuación (9.2) indica que si integramos una función de densidad en dos variables, $f(x_1, x_2)$, respecto a la variable x_2 se obtiene una función que es de nuevo una función de densidad, y de ahí el símbolo f, pero que es ahora la función de densidad de la variable x_1. Las funciones $f(x_1)$ y $f(x_1, x_2)$ serán en general totalmente distintas y sólo tienen en común ser ambas funciones de densidad, por tanto $f\ (.) \geq 0$, y

$$\int_{-\infty}^{\infty} f(x_1) dx_1 = 1$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = 1.$$

Para justificar (9.2), calcularemos la probabilidad de que la variable x_1 pertenezca a un intervalo $(a, b]$ a partir de la distribución conjunta. Entonces:

$$P(a < x_1 \leq b) = P(a < x_1 \leq b; -\infty < x_2 \leq \infty) = \int_a^b dx_1 \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 =$$

$$= \int_a^b f(x_1) dx_1$$

que justifica (9.2). Observemos que en esta ecuación x_1 es un valor concreto cualquiera. Supongamos que la precisión de la medida de x_1 es Δx_1, es decir, diremos que ha ocurrido el valor x_1 si se observa un valor en el intervalo $x_1 \pm \Delta x_1/2$. La probabilidad de este valor será el valor de la densidad en el centro del intervalo, $f(x_1)$ por la longitud de la base Δx_1. Si multiplicamos ambos miembros de la ecuación (9.2) por la constante Δx_1, tenemos en el primer miembro $f(x_1)\Delta x_1$, que es la probabilidad de ese valor concreto de x_1 calculada con su distribución univariante. En el segundo miembro tendremos la suma (integral) de todas las probabilidades de los pares de valores posibles (x_1, x_2), cuando x_1 es fijo y x_2 toma todos los valores posibles. En efecto, estas probabilidades vienen dadas por $f(x_1, x_2)dx_2\Delta x_1$, y sumando para todos los valores posibles de x_2 de nuevo obtenemos la probabilidad del valor x_1.

Si $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, donde $\mathbf{x}_1$ y $\mathbf{x}_2$ son a su vez variables vectoriales, se define la distribución condicionada de $\mathbf{x}_1$, para un valor concreto de la variable $\mathbf{x}_2 = \mathbf{x}_2^0$, por:

$$f(\mathbf{x}_1|\mathbf{x}_2^0) = \frac{f(\mathbf{x}_1, \mathbf{x}_2^0)}{f(\mathbf{x}_2^0)} \tag{9.4}$$

supuesto que $f(\mathbf{x}_2^0) \neq 0$. Esta definición es consistente con el concepto de probabilidad condicionada y con el de función de densidad para una variable. En efecto, supongamos para simplificar que ambas variables son escalares. Entonces multiplicando por Δx_1 ambos miembros tendremos

$$f(x_1|x_2^0)\Delta x_1 = \frac{f(x_1, x_2^0)\Delta x_1 \Delta x_2}{f(x_2^0)\Delta x_2}$$

y el primer miembro representa la probabilidad condicionada que se expresa como cociente de la probabilidad conjunta y la marginal. De esta definición se deduce:

$$f(\mathbf{x}_1, \mathbf{x}_2) = f(\mathbf{x}_2|\mathbf{x}_1)f(\mathbf{x}_1). \tag{9.5}$$

La distribución marginal de $\mathbf{x}_2$, puede calcularse en función de (9.3) y (9.5) como:

$$f(\mathbf{x}_2) = \int f(\mathbf{x}_2|\mathbf{x}_1)f(\mathbf{x}_1)d\mathbf{x}_1, \tag{9.6}$$

que tiene una clara interpretación intuitiva. Si multiplicamos ambos miembros por $\Delta\mathbf{x}_2$, el elemento de volumen, tenemos en la izquierda $f(\mathbf{x}_2)\Delta\mathbf{x}_2$, la probabilidad del valor concreto de $\mathbf{x}_2$ considerado. La fórmula (9.6) nos dice que esta probabilidad puede calcularse obteniendo primero la probabilidad del valor $\mathbf{x}_2$ para cada posible valor de $\mathbf{x}_1$, dada por $f(\mathbf{x}_2|\mathbf{x}_1)\Delta\mathbf{x}_2$, y luego multiplicando cada uno de estos valores por las probabilidades de $\mathbf{x}_1$, $f(\mathbf{x}_1)d\mathbf{x}_1$, lo que equivale a promedir las probabilidades condicionadas por $\mathbf{x}_1$ respecto a la distribución de esta variable.

Como resultado de (9.5) y (9.6) la distribución condicionada $f(\mathbf{x}_1|\mathbf{x}_2)$ puede entonces escribirse como:

$$f(\mathbf{x}_1|\mathbf{x}_2) = \frac{f(\mathbf{x}_2|\mathbf{x}_1)f(\mathbf{x}_1)}{\int f(\mathbf{x}_2|\mathbf{x}_1)f(\mathbf{x}_1)d\mathbf{x}_1} \tag{9.7}$$

que es el teorema de Bayes para funciones de densidad, y constituye la herramienta fundamental de la inferencia Bayesiana que estudiaremos en el capítulo 11.

Para variables discretas los conceptos son similares, pero las integrales se sustituyen por sumas, como se indica en el siguiente ejemplo.

Ejemplo 9.1 *La tabla 9.1 presenta al distribución conjunta de las variables aleatorias discretas:* x_1 : *votar a uno de cuatro posibles partidos políticos, que toma los cuatro valores posibles* P_1, P_2, P_3 *y* P_4 *y* x_2 : *nivel de ingresos de los votantes, que toma los tres valores A (alto), M (medio), B (bajo). Calcular las distribuciones marginales, la distribución condicionada de los votos para las personas con ingresos bajos y la distribución condicionada de los ingresos para los votantes del partido* P_4.

	A	M	B
P_1	.1	.05	.01
P_2	.05	.20	.04
P_3	.04	.25	.07
P_4	.01	.1	.08

Tabla 9.1. Distribución conjunta de votos e ingresos en una población

Para calcular la distribución marginal añadimos a la tabla una fila y una columna y colocamos allí el resultado de sumar las filas y las columnas de la tabla. Con esto se obtiene la tabla 9.2. Por ejemplo, la distribución marginal de los ingresos indica que la probabilidad de ingresos altos es .2, de medios .6 y de bajos .2. Observemos que las distribuciones marginales son el resultado que se obtiene en los márgenes de la tabla (lo que justifica su nombre) al sumar las probabilidades conjuntas por filas y por columnas.

	A	M	B	Marginal de votos
P_1	.1	.05	.01	.16
P_2	.05	.20	.04	.29
P_3	.04	.25	.07	.36
P_4	.01	.1	.08	.19
Marginal de ingresos	.2	.6	.2	

Tabla 9.2. Distribución conjunta y marginales de votos e ingresos en una población

Para calcular la distribución condicionada de los votos para las personas de ingresos bajos, dividimos cada casilla de la columna de ingresos bajos por el total de la columna. La distribución resultante se indica en el tabla 9.3

P_1	P_2	P_3	P_4
.05	.20	.35	.40

Tabla 9.3 distribución condicionada de los votos para personas con ingresos medios.

Por ejemplo, el valor .05 es el resultado de dividir .01, la probabilidad conjunta de ingresos bajos y votar a P_1 por la probabilidad marginal de ingresos bajos, .1. Esta tabla indica que el partido preferido para las personas de ingresos bajos es el P_4 con un 40% de los votos, seguido del P_3 con el 35%. La tabla 9.4 indica la distribución condicionada de los ingresos para los votantes del partido P_4. El grupo más numeroso de votantes de este partido es de ingresos medios (52,63%) seguido de ingresos bajos (42,11%) y altos (5,26%).

	A	M	B	Total
P_4	.0526	.5263	.4211	1

Tabla 9.4 distribución condicionada de los ingresos para personas que votan a P_1.

9.1.4 Independencia

El concepto fundamental en el estudio conjunto de varias variables aleatorias es el concepto de independencia. Diremos que dos vectores aleatorios $\mathbf{x}_1$, $\mathbf{x}_2$ son independientes si el conocimiento de uno de ellos no aporta información respecto a los valores del otro. En otros términos, la distribución de valores concretos de $\mathbf{x}_2$ no depende de $\mathbf{x}_1$ y es la misma cualquiera que sea el valor de $\mathbf{x}_1$. Esto se expresa matemáticamente:

$$f(\mathbf{x}_2|\mathbf{x}_1) = f(\mathbf{x}_2) \tag{9.8}$$

que indica que la distribución condicionada es idéntica a la marginal. Utilizando (9.5), una definición equivalente de independencia entre dos vectores aleatorios $\mathbf{x}_1$, $\mathbf{x}_2$ es:

$$f(\mathbf{x}_1, \mathbf{x}_2) = f(\mathbf{x}_1)f(\mathbf{x}_2) \tag{9.9}$$

es decir, dos vectores aleatorios son independientes si su distribución conjunta (su probabilidad conjunta) es el producto de las distribuciones marginales (de las probabilidades individuales). En general, diremos que las variables aleatorias $x_1, ..., x_p$, con densidad conjunta $f(x_1, ..., x_p)$ son independientes, si se verifica:

$$f(x_1, ..., x_p) = f(x_1)f(x_2)....f(x_p) \tag{9.10}$$

La independencia conjunta es una condición muy fuerte: al ser $x_1, ..., x_p$ independientes también lo serán cualquier subconjunto de variables $(x_1, ..., x_h)$ con $h \leq p$, así como cualquier conjunto de funciones de las variables individuales, $g_1(x_1)....g_1(x_p)$, o de conjuntos disjuntos de ellas. Cuando las variables son independientes no ganamos nada con su estudio conjunto y conviene estudiarlas individualmente. Es fácil comprobar que si las variables $\mathbf{x}_1$ y $\mathbf{x}_2$ son independientes y construimos nuevas variables $\mathbf{y}_1 = g_1(\mathbf{x}_1)$, $\mathbf{y}_2 = g_2(\mathbf{x}_2)$, donde la primera variable es sólo función de $\mathbf{x}_1$ y la segunda sólo de $\mathbf{x}_2$, las variables $\mathbf{y}_1$, $\mathbf{y}_2$ son también independientes.

9.1.5 La maldición de la dimensión

La maldición de la dimensión es un término acuñado por el matemático R. Bellman para describir como aumenta la complejidad de un problema al aumentar la dimensión de las variables involucradas. En el análisis estadístico multivariante la maldición de la dimensión se manifiesta de varias formas.

En primer lugar, al aumentar la dimensión, el espacio está cada vez más vacío, haciendo más difícil cualquier proceso de inferencia a partir de los datos. Esto es consecuencia de que, al aumentar la dimensión del espacio aumenta su volumen (o su hipervolumen en general), y como la masa total de probabilidad es la unidad, la densidad de la variable aleatoria

debe disminuir. En consecuencia, la densidad de probabilidad de una variable aleatoria de dimensión alta es muy baja en todo el espacio, o, lo que es equivalente, el espacio esta progresivamente más vacío. Para ilustrar el problema, supongamos que la densidad de una variable $p-$dimensional es uniforme en el hipercubo $[0,1]^p$ y que todos los componentes son independientes. Por ejemplo, pueden generarse muestras de esta variable tomando conjuntos de p números aleatorios entre cero y uno. Consideremos la probabilidad de que un valor al azar de esta variable esté dentro del hipercubo $[0;0,9]^p$. Para $p = 1$, el caso escalar, esta probabilidad es $0,9$, para $p = 10$, este valor baja a $0,9^{10} = 0,35$, y para $p = 30$ es $0,9^{30} = 0,04$. Vemos que, a medida que aumenta la dimensión del espacio, cualquier conjunto va, progresivamente, quedándose vacío.

Un segundo problema es que el número de parámetros necesario para describir los datos aumenta rápidamente con la dimensión. Para representar en dimensión p la media y la matriz de covarianzas necesitamos

$$p + p(p+1)/2 = p(p+3)/2$$

que es de orden p^2. Por tanto, la complejidad de los datos, medida por el número de parámetros necesarios para representarlos, crece, en este caso, con el cuadrado de la dimensión del espacio. Por ejemplo, 100 datos es una muestra grande para una variable unidimensional, pero es muy pequeña para una variable vectorial con $p = 14$: para estimar las medias, varianzas y covarianzas se requieren más de $14(17)/2 = 119$ observaciones. Como norma general, los procedimientos multivariantes necesita un ratio $n/p > 10$ y es deseable que este ratio sea mayor de 20.

La consecuencia del aumento de la dimensión es un aumento de la incertidumbre del problema: la previsión conjunta de los valores de la variable va siendo cada vez más difícil. En la práctica, este problema disminuye si las variables son muy dependientes entre sí, ya que entonces, la densidad de probabilidad se concentra en determinadas zonas del espacio, definidas por la relación de dependencia, en lugar de repartirse por todo el espacio muestral. Esta dependencia puede usarse, extendiendo los métodos que como hemos visto en capítulos anteriores, para reducir la dimensión del espacio de variables y evitar la maldición de la dimensionalidad.

9.2 PROPIEDADES DE VARIABLES VECTORIALES

9.2.1 Vector de medias

Llamaremos esperanza, o vector de medias, $\boldsymbol{\mu}$, de una variable multidimensional, $\mathbf{x}$, al vector cuyos componentes son las esperanzas, o medias, de los componentes de la variable aleatoria. Escribiremos el vector de medias como:

$$\boldsymbol{\mu} = E[\mathbf{x}] \tag{9.11}$$

donde debe entenderse que la esperanza operando sobre un vector o una matriz es el resultado de aplicar este operador (tomar medias) a cada uno de los componentes. Si la variable es

continua:

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int \mathbf{x} f(\mathbf{x}) d\mathbf{x}$$

La esperanza es una función lineal, es decir, para cualquier matriz, $\mathbf{A}$, y vector $\mathbf{b}$, tenemos:

$$E[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}E[\mathbf{x}_1] + \mathbf{b}.$$

Si $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)'$ tenemos también que, para escalares a y b :

$$E[a\mathbf{x}_1 + b\mathbf{x}_2] = aE[\mathbf{x}_1] + bE[\mathbf{x}_2].$$

y si $\mathbf{x}_1$ y $\mathbf{x}_2$ son independientes:

$$E[\mathbf{x}_1\mathbf{x}_2] = E[\mathbf{x}_1]E[\mathbf{x}_2].$$

9.2.2 Esperanza de una función

Generalizando la idea de esperanza, si disponemos de una función escalar $y = g(\mathbf{x})$ de un vector de variables aleatorias, el valor medio de esta función se calcula:

$$E[y] = \int y f(y) dy = \int ... \int g(\mathbf{x}) f(x_1, ..., x_n) dx_1, ..., dx_n \tag{9.12}$$

La primera integral tiene en cuenta que y es escalar y si conocemos su función de densidad, $f(y)$, su esperanza se calcula de la forma habitual. La segunda, muestra que no es necesario calcular $f(y)$ para determinar el valor promedio de $g(\mathbf{x})$: basta ponderar sus valores posibles por las probabilidades que dan lugar a estos valores.

Esta definición es consistente, y es fácil comprobar que ambos métodos conducen al mismo resultado. Si $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)'$, y definimos $y_1 = g_1(\mathbf{x}_1)$, $y_2 = g_2(\mathbf{x}_2)$, si $\mathbf{x}_1$ e $\mathbf{x}_2$ son independientes

$$E[y_1 y_2] = E(g_1(\mathbf{x}_1))E(g_2(\mathbf{x}_2))$$

9.2.3 Matriz de varianzas y covarianzas

Llamaremos matriz de varianzas y covarianzas (o simplemente matriz de covarianzas) de un vector aleatorio $\mathbf{x} = (x_1, ..., x_p)'$, de $\Re^p$, con vector de medias $\boldsymbol{\mu}' = (\mu_1, ..., \mu_p)$, a la matriz cuadrada de orden p obtenida por :

$$\mathbf{V}_x = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'] \tag{9.13}$$

La matriz $\mathbf{V}_x$ contiene en la diagonal las varianzas de los componentes, que representaremos por σ_i^2, y fuera de ella las covarianzas entre los pares de variables, que representaremos por σ_{ij}. La matriz de covarianzas es simétrica y semidefinida positiva. Es decir, dado un vector cualquiera, $\boldsymbol{\omega}$, se verificará:

$$\boldsymbol{\omega}' \mathbf{V}_x \boldsymbol{\omega} \geq 0.$$

Para demostrar esta propiedad definamos una variable unidimensional por:

$$y = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\omega}$$

donde $\boldsymbol{\omega}$ es un vector arbitrario de $\Re^p$. La variable y tiene esperanza cero ya que

$$E(y) = E\left[(\mathbf{x} - \boldsymbol{\mu})\right]'\boldsymbol{\omega} = 0$$

y su varianza debe ser no negativa:

$$var(y) = E\left[y^2\right] = \boldsymbol{\omega}' E\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\right]\boldsymbol{\omega} = \boldsymbol{\omega}'\mathbf{V}_x\boldsymbol{\omega} \geq 0$$

Llamaremos varianza media al promedio de las varianzas dado por $\text{tr}(\mathbf{V}_x)/p$, varianza generalizada a $|\mathbf{V}_x|$ y variabilidad promedio a

$$VP = |\mathbf{V}_x|^{1/p}$$

que es una medida global de la variabilidad conjunta para todas las variables que tiene en cuenta su estructura de dependencia. La interpretación de estas medidas es similar a la estudiada en el capítulo 3 para distribuciones de datos.

9.2.4 Transformaciones de vectores aleatorios.

Al trabajar con funciones de densidad de vectores aleatorios es importante recordar que, como en el caso univariante, la función de densidad tiene dimensiones: si $p = 1$, caso univariante, probabilidad por unidad de longitud, si $p = 2$, probabilidad por unidad de superficie, si $p = 3$ por unidad de volumen y si $p > 3$) de hipervolumen. Por lo tanto, si cambiamos las unidades de medida de las variables, la función de densidad debe modificarse también. En general, sea $\mathbf{x}$ un vector de $\Re^p$ con función de densidad $f_x(\mathbf{x})$ y sea otro vector aleatorio $\mathbf{y}$ de $\Re^p$, definido mediante la transformación uno a uno:

$$\begin{aligned} y_1 &= g_1(x_1, ..., x_p) \\ &\vdots \quad\quad \vdots \\ y_p &= g_p(x_1, ..., x_p), \end{aligned}$$

donde suponemos que existen las funciones inversas $x_1 = h_1(y_1, ..., y_p), ..., x_p = h_p(y_1, ..., y_p)$, y que todas las funciones implicadas son diferenciables. Entonces, puede demostrarse que la función de densidad del vector $\mathbf{y}$ viene dada por:

$$f_y(\mathbf{y}) = f_x(\mathbf{x})\left|\frac{d\mathbf{x}}{d\mathbf{y}}\right|, \tag{9.14}$$

donde aquí hemos utilizado f_y y f_x para representar las funciones de densidad de las variables $\mathbf{y}$, y $\mathbf{x}$, para evitar confusiones. El término $|d\mathbf{x}/d\mathbf{y}|$ representa el jacobiano de la

transformación, (que ajusta la probabilidad por el cambio de escala de medida) dado por el determinante:

$$\left|\frac{d\mathbf{x}}{d\mathbf{y}}\right| = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \cdots & \dfrac{\partial x_1}{\partial y_p} \\ \vdots & & \vdots \\ \dfrac{\partial x_p}{\partial y_1} & \cdots & \dfrac{\partial x_p}{\partial y_p} \end{vmatrix}$$

que suponemos es distinto de cero en el rango de la transformación.

Un caso importante es el de transformaciones lineales de la variable. Si hacemos

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

donde $\mathbf{A}$ es una matriz cuadrada no singular, las derivadas de los componentes de $\mathbf{x}$ respecto a $\mathbf{y}$ se obtendrán de $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$, y serán, por tanto, los elementos de la matriz $\mathbf{A}^{-1}$. El Jacobiano de la transformación será $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$ y la función de densidad de la nueva variable $\mathbf{y}$, será

$$f_y(\mathbf{y}) = f_x(\mathbf{A}^{-1}\mathbf{y})\,|\mathbf{A}|^{-1} \tag{9.15}$$

expresión que indica que para obtener la función de densidad de la variable $\mathbf{y}$ sustituimos en la función de densidad de la variable $\mathbf{x}$ el argumento por $\mathbf{A}^{-1}\mathbf{y}$ y dividimos el resultado por el determinante de la matriz $\mathbf{A}$.

9.2.5 Esperanzas de transformaciones lineales

Sea $\mathbf{x}$ un vector aleatorio de dimensión p y definamos un nuevo vector aleatorio $\mathbf{y}$ de dimensión $m, (m \leq p)$, con

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{9.16}$$

donde $\mathbf{A}$ es una matriz rectangular de dimensiones $m \times p$. Llamando $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$, a sus vectores de medias y $\mathbf{V}_x$, $\mathbf{V}_y$ a las matrices de covarianzas, se verifica la relación:

$$\boldsymbol{\mu}_y = \mathbf{A}\boldsymbol{\mu}_x \tag{9.17}$$

que es inmediata tomando esperanzas en (9.16). Además:

$$\mathbf{V}_y = \mathbf{A}\mathbf{V}_x\mathbf{A}' \tag{9.18}$$

donde $\mathbf{A}'$ es la matriz transpuesta de $\mathbf{A}$. En efecto, aplicando la definición de covarianzas y las relaciones (9.16) y (9.18)

$$\mathbf{V}_y = E\left[(\mathbf{y}-\boldsymbol{\mu}_y)(\mathbf{y}-\boldsymbol{\mu}_y)'\right] = E\left[\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)'\mathbf{A}'\right] = \mathbf{A}\mathbf{V}_x\mathbf{A}'$$

Ejemplo 9.2 *Las valoraciones de los clientes de la puntualidad (x_1), rapidez (x_2) y limpieza (x_3) de un servicio de transporte tienen unas medias, en una escala de cero a diez, de 7, 8 y 8,5 respectivamente con una matriz de varianzas y covarianzas*

$$\mathbf{V}_x = \begin{bmatrix} 1 & .5 & .7 \\ .5 & .64 & .6 \\ .7 & .6 & 1.44 \end{bmatrix}$$

Se construyen dos indicadores de la calidad del servicio. El primero es el promedio de las tres puntuaciones y el segundo es la diferencia entre el promedio de la puntualidad y la rapidez, que indica la fiabilidad del servicio y la limpieza, que indica la comodidad del mismo. Calcular el vector de medias y la matriz de covarianzas para estos dos indicadores.

La expresión del primer indicador es

$$y_1 = \frac{x_1 + x_2 + x_3}{3}$$

y la del segundo

$$y_2 = \frac{x_1 + x_2}{2} - x_3$$

Estas dos ecuaciones pueden escribirse matricialmente

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

El vector de medias será

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & -1 \end{bmatrix} \begin{bmatrix} 7 \\ 8 \\ 8,5 \end{bmatrix} = \begin{bmatrix} 7,83 \\ -1 \end{bmatrix}$$

y el valor 7,83 es una medida global de la calidad promedio del servicio y el menos uno de la relación fiabilidad comodidad. La matriz de varianzas covarianzas es:

$$\begin{aligned} \mathbf{V}_y &= \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/2 & 1/2 & -1 \end{bmatrix} \begin{bmatrix} 1 & .5 & .7 \\ .5 & .64 & .6 \\ .7 & .6 & 1.44 \end{bmatrix} \begin{bmatrix} 1/3 & 1/2 \\ 1/3 & 1/2 \\ 1/3 & -1 \end{bmatrix} = \\ &= \begin{bmatrix} .742\,22 & -.256\,67 \\ -.256\,67 & .8 \end{bmatrix} \end{aligned}$$

que indica que la variabilidad de ambos indicadores es similar y que están relacionados negativamente, ya que la covarianza es negativa.

9.3 Dependencia entre variables aleatorias

9.3.1 Esperanzas condicionadas

La esperanza de un vector $\mathbf{x}_1$ condicionada a un valor concreto de otro vector $\mathbf{x}_2$ es la esperanza de la distribución de $\mathbf{x}_1$ condicionada a $\mathbf{x}_2$ y viene dada por:

$$E\left[\mathbf{x}_1|\mathbf{x}_2\right] = \int \mathbf{x}_1 f\left(\mathbf{x}_1|\mathbf{x}_2\right) d\mathbf{x}_1.$$

En general esta expresión será una función del valor $\mathbf{x}_2$. Cuando $\mathbf{x}_2$ es un valor fijo, la esperanza condicionada será una constante. Si $\mathbf{x}_2$ es una variable aleatoria, la esperanza condicionada será también una variable aleatoria.

La esperanza de un vector aleatorio $\mathbf{x}_1$ puede calcularse a partir de las esperanzas condicionales en dos etapas: en la primera calculamos la esperanza de $\mathbf{x}_1$ condicionada a $\mathbf{x}_2$. El resultado es una función aleatoria que depende de la variable aleatoria $\mathbf{x}_2$. En la segunda, calculamos la esperanza de esta función con relación a la distribución de $\mathbf{x}_2$. Entonces:

$$E(\mathbf{x}_1) = E\left[E(\mathbf{x}_1|\mathbf{x}_2)\right]. \tag{9.19}$$

Esta expresión indica que la esperanza de una variable aleatoria puede obtenerse promediando las esperanzas condicionadas por sus probabilidades de aparición o, en otros términos, que la esperanza de la media condicionada es la esperanza marginal o incondicional.

Demostración

$$\begin{aligned} E(\mathbf{x}_1) &= \int \mathbf{x}_1 f(\mathbf{x}_1) d\mathbf{x}_1 = \iint \mathbf{x}_1 f(\mathbf{x}_1\mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2 = \iint \mathbf{x}_1 f(\mathbf{x}_1|\mathbf{x}_2) f(\mathbf{x}_2) d\mathbf{x}_1 d\mathbf{x}_2 \\ &= \int f(\mathbf{x}_2) \left[\int \mathbf{x}_1 f(\mathbf{x}_1|\mathbf{x}_2) d\mathbf{x}_1\right] d\mathbf{x}_2 = \int E\left[\mathbf{x}_1|\mathbf{x}_2\right] f(\mathbf{x}_2) d\mathbf{x}_2 \\ &= E\left[E(\mathbf{x}_1|\mathbf{x}_2)\right]. \end{aligned}$$

■

9.3.2 Varianzas condicionadas

La varianza de $\mathbf{x}_1$ condicionada a $\mathbf{x}_2$ se define como la varianza de la distribución de $\mathbf{x}_1$ condicionada a $\mathbf{x}_2$. Utilizaremos la notación

$$Var(\mathbf{x}_1|\mathbf{x}_2) = \mathbf{V}_{1/2}$$

y esta matriz tendrá las propiedades ya estudiadas de una matriz de covarianzas.

Si $\mathbf{x}_1$ es escalar, su varianza puede calcularse también a partir de las propiedades de la distribución condicionada. En concreto, puede expresarse como suma de dos términos: el primero asociado a las medias condicionadas y el segundo a las varianzas condicionadas. Para obtener esta expresión partimos de la descomposición:

$$x_1 - \mu_1 = x_1 - E(x_1/\mathbf{x}_2) + E(x_1/\mathbf{x}_2) - \mu_1$$

donde $\mathbf{x}_2$ es un vector aleatorio cualquiera para el que la esperanza condicionada $E(x_1/\mathbf{x}_2)$ es finita. Elevando al cuadrado esta expresión y tomando esperanzas en ambos miembros:

$$var(x_1) = E(x_1 - E(x_1/\mathbf{x}_2))^2 + E(E(x_1/\mathbf{x}_2) - \mu_1)^2 + 2E\left[(x_1 - E(x_1/\mathbf{x}_2)(E(x_1/\mathbf{x}_2) - \mu_1)\right]$$

el doble producto se anula, ya que

$$E\left[(x_1 - E(x_1/\mathbf{x}_2))(E(x_1/\mathbf{x}_2) - \mu_1)\right] =$$

$$= \int (E(x_1/\mathbf{x}_2) - \mu_1) \left[\int (x_1 - E(x_1/\mathbf{x}_2)) f(x_1/\mathbf{x}_2) dx_1 \right] f(\mathbf{x}_2) d\mathbf{x}_2 = 0$$

al ser nula la integral entre corchetes. Por otro lado, como por (9.19):

$$E\left[E(x_1/\mathbf{x}_2)\right] = E(x_1) = \mu_1,$$

el segundo término es la esperanza de la diferencia al cuadrado entre la variable aleatoria $E(x_1/\mathbf{x}_2)$, que dependerá del vector aleatorio $\mathbf{x}_2$ y su media μ_1. Por tanto:

$$var(x_1) = E\left[var(x_1/\mathbf{x}_2)\right] + var\left[E(x_1/\mathbf{x}_2)\right] \tag{9.20}$$

Esta expresión se conoce como *descomposición de la varianza,* ya que descompone la variabilidad de la variable en dos fuentes principales de variación. Por un lado, hay variabilidad porque las varianzas de las distribuciones condicionadas, $var(x_1/\mathbf{x}_2)$, pueden ser distintas, y el primer término promedia estas varianzas. Por otro, hay también variabilidad porque las medias de las distribuciones condicionadas pueden ser distintas, y el segundo término recoge las diferencias entre las medias condicionadas, $E(x_1/\mathbf{x}_2)$, y la media total, μ_1, mediante el término $var\left[E(x_1/\mathbf{x}_2)\right]$. Observemos que la varianza de la variable x_1 es, en general, mayor que el promedio de las varianzas de las distribuciones condicionadas, debido a que en las condicionadas la variabilidad se calcula respecto a las medias condicionadas, $E(x_1/\mathbf{x}_2)$, mientras que $var(x_1)$ mide la variabilidad respecto a la media global, μ_1. Si todas las medias condicionadas son iguales a μ_1, los que ocurrirá por ejemplo si x_1 e $\mathbf{x}_2$ son independientes, entonces el término $var\left[E(x_1/\mathbf{x}_2)\right]$ es cero y la varianza es la media ponderada de las varianzas condicionadas. Si $E(x_1/\mathbf{x}_2)$ no es constante, entonces la varianza de x_1 será tanto mayor cuanto mayor sea la variabilidad de las medias condicionadas.

Esta descomposición de la varianza aparece en el análisis de la varianza de los modelos lineales univariantes:

$$\sum (x_i - \overline{x})^2/n = \sum (x_i - \widehat{x}_i)^2/n + \sum (\widehat{x}_i - \overline{x})^2/n$$

donde, en esta expresión, $\widehat{x}_i$ es la estimación de la media condicionada en el modelo lineal. La variabilidad total, que equivale a $var(x_1)$, se descompone en dos términos incorrelados. Por un lado, el promedio de las estimaciones de $var(x_1/\mathbf{x}_2)$, que se calculan promediando las diferencias entre la variable y la media condicionada. Por el otro, la variabilidad de las esperanzas condicionales respecto a la media global, que se estiman en los modelos lineales por las diferencias $\widehat{x}_i - \overline{x}$.

9.3.3 Matriz de correlación

Se define la matriz de correlación de un vector aleatorio $\mathbf{x}$ con matriz de covarianzas $\mathbf{V}_x$, por

$$\mathbf{R}_x = \mathbf{D}^{-1/2}\mathbf{V}_x\mathbf{D}^{-1/2}$$

donde

$$\mathbf{D} = \mathbf{diag}(\sigma_1^2, ..., \sigma_p^2)$$

es la matriz diagonal que contiene las varianzas de las variables. La matriz de correlación será pues una matriz cuadrada y simétrica con unos en la diagonal y los coeficientes de correlacion entre los pares de variables fuera de la diagonal. Los coeficientes de correlación simple o coeficientes de correlación lineal, vienen dados por

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

La matriz de correlación es también semidefinida positiva. Una medida global de las correlaciones lineales existentes en el conjunto de variables es la dependencia, definida por

$$D_x = 1 - |\mathbf{R}_x|^{1/(p-1)}$$

cuya interpretación para variables aleatorias es análoga a la presentada en el capítulo 3 para variables estadísticas. Para $p = 2$ la matriz $\mathbf{R}_x$ tiene unos en la diagonal y el coeficiente ρ_{12} fuera, $|\mathbf{R}_x| = 1 - \rho_{12}^2$, y la dependencia $D_x = 1 - (1 - \rho_{12}^2) = \rho_{12}^2$ coincide con el coeficiente de determinación. Se demuestra de la misma forma que se hizo en el capítulo 3 que en el caso general, $p > 2$, la dependencia es un promedio geométrico de coeficientes de determinación.

9.3.4 Correlaciones Múltiples

Se denomina correlación múltiple de una variable escalar, y, y un vector de variables $\mathbf{x}$ a una medida de la capacidad de prever y mediante una función lineal de las variables $\mathbf{x}$. Suponiendo, sin pérdida de generalidad, que las variables tienen media cero, definimos la mejor predicción lineal de y como la función $\boldsymbol{\beta}'\mathbf{x}$ que minimiza $E(y - \boldsymbol{\beta}'\mathbf{x})^2$. Puede demostrarse que $\boldsymbol{\beta} = \mathbf{V}_x^{-1}\mathbf{V}_{xy}$ siendo $\mathbf{V}_x$ la matriz de covarianzas de $\mathbf{x}$ y $\mathbf{V}_{xy}$ el vector de covarianzas entre y y $\mathbf{x}$. El coeficiente de correlación simple entre las variables escalares y y $\boldsymbol{\beta}'\mathbf{x}$ se denomina coeficiente de correlación múltiple.

Puede demostrarse que si llamamos σ_{ij} a los términos de la matriz de covarianzas $\mathbf{V}$ de un vector de variables y σ^{ij} a los términos de la matriz $\mathbf{V}^{-1}$, el coeficiente de correlación múltiple, $R_{i.R}$ entre cada variable (i) y todas las demás (R) se calcula como:

$$R_{i.R}^2 = 1 - \frac{1}{\sigma_{ij}\sigma^{ij}}$$

En particular, si $E(y|\mathbf{x})$ es una función lineal de $\mathbf{x}$ entonces $E(y|\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$ y $R_{i.R}^2$ puede también calcularse como $1 - \sigma_{y|x}^2/\sigma_y^2$, donde $\sigma_{y|x}^2$ es la varianza de la distribución condicionada, $y|\mathbf{x}$ y σ_y^2 la varianza marginal de y.

9.3.5 Correlaciones Parciales

Supongamos que obtenemos la mejor aproximación lineal a un vector de variables $\mathbf{x}_1$ de dimensiones $p_1 \times 1$ a partir de otro vector de variables $\mathbf{x}_2$ de dimensiones $p_2 \times 1$. Suponiendo que las variables tienen media cero, esto implica calcular un vector $\mathbf{B}\mathbf{x}_2$ donde $\mathbf{B}$ es una matriz de coeficientes de dimensiones $p_1 \times p_2$ de manera que $\sum_{j=1}^{p_1} E(x_{1j} - \boldsymbol{\beta}_j'\mathbf{x}_2)^2$ sea mínima, donde x_{1j} es el componente j del vector $\mathbf{x}_1$ y $\boldsymbol{\beta}_j'$ la fila j de la matriz $\mathbf{B}$. Llamemos $\mathbf{V}_{1/2}$ a la

matriz de covarianzas de la variable $\mathbf{x}_1 - \mathbf{B}\mathbf{x}_2$. Si estandarizamos esta matriz de covarianzas para pasarla a correlaciones, los coeficientes de correlación resultantes se denominan coeficientes de correlación parcial entre los componentes de $\mathbf{x}_1$ dadas las variables $\mathbf{x}_2$. La matriz cuadrada y simétrica de orden p_1

$$R_{1/2} = \mathbf{D}_{1/2}^{-1/2}\mathbf{V}_{1/2}\mathbf{D}_{1/2}^{-1/2}$$

se denomina matriz de correlaciones parciales entre los componentes del vector $\mathbf{x}_1$ cuando controlamos (o condicionado a) el vector $\mathbf{x}_2$, donde $\mathbf{D}_{1/2}$= $\mathbf{diag}(\sigma^2_{1/2}, ..., \sigma^2_{k/2})$ y $\sigma^2_{j/2}$ es la varianza de la variable $x_{1j} - \boldsymbol{\beta}'_j\mathbf{x}_2$.

En particular si $E(\mathbf{x}_1|\mathbf{x}_2)$ es lineal en $\mathbf{x}_2$, entonces $E(\mathbf{x}_1|\mathbf{x}_2) = \mathbf{B}\mathbf{x}_2$ y $\mathbf{V}_{1/2}$ es la matriz de covarianzas de la distribución condicionada de $\mathbf{x}_1|\mathbf{x}_2$.

9.4 LA DISTRIBUCIÓN MULTINOMIAL

Supongamos que observamos elementos que clasificamos en dos clases, A y $\overline{A}$. Por ejemplo, clasificamos los recién nacidos en un hospital como hombre (A) o mujer ($\overline{A}$), los dias de un mes como lluviosos (A) o no ($\overline{A}$), o los elementos fabricados por una máquina como buenos (A) o defectuosos ($\overline{A}$). Suponemos que el proceso que genera elementos es estable, existiendo un probabilidad constante de aparición de los elementos de cada clase, $P(A) = p = cte$, y que el proceso no tiene memoria, es decir $P(A|A) = P(A|\overline{A})$. Supongamos que observamos elementos al azar de este proceso y definimos la variable

$$x = \left\{ \begin{array}{c} 1, \text{ si la observación pertenece a la clase } A \\ 0, \text{ en otro caso} \end{array} \right\}$$

esta variable sigue una distribución binomial puntual, con $P(x = 1) = p$ y $P(x = 0) = 1 - p$. Si observamos n elementos en lugar de uno y definimos la variable $y = \sum_{i=1}^{n} x_i$, es decir, contamos el número de elementos en n que pertenece a la primera clase, la variable y sigue una distribución binomial con

$$P(y = r) = \frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}.$$

Podemos generalizar esta distribución permitiendo G clases en lugar de dos, y llamamos $\mathbf{p}$ al vector de probabilidades de pertenencia a las clases, $\mathbf{p}$ =$(p_1, ..., p_G)'$, donde $\sum p_j = 1$. Definiremos ahora las G variables aleatorias:

$$x_j = \left\{ \begin{array}{c} 1, \text{ si la observación pertenece al grupo } j \\ 0, \text{en otro caso} \end{array} \right\} \quad j = 1, ..., G$$

y el resultado de una observación es un valor del vector de G-variables $\mathbf{x} = (x_1, ..., x_G)'$, que será siempre de la forma $\mathbf{x} = (0, ..., 1, ...0)'$, ya que sólamente una de las G componentes puede tomar el valor uno, el asociado a la clase observada para ese elemento. En consecuencia, los componentes de esta variable aleatoria no son independientes, ya que están ligadas por la ecuación

$$\sum_{j=1}^{G} x_j = 1.$$

Para describir el resultado de la observación bastaría con definir $G-1$ variables, como se hace en la distribución binomial donde sólo se define una variable cuando hay dos clases, ya que el valor de la última variable queda fijada al conocer las restantes. Sin embargo, con más de dos clases es constumbre trabajar con las G variables y la distribución de la variable multivariante así definida se denomina *multinomial puntual.* Tiene como función de probabilidades

$$P(x_1,..,x_G) = p_1^{x_1}...p_G^{x_G} = \prod p_j^{x_j}$$

En efecto, como sólo una de las x_j es distinta de cero, la probabilidad de que la j-ésima sea uno es precisamente p_j, la probabilidad de que el elemento observado pertenezca a la clase j. Generalizando esta distribución, sea $(\mathbf{x}_1,...,\mathbf{x}_n)$ una muestra de n valores de esta variable multinomial puntual que resulta al clasificar n elementos de una muestra en las G clases. Se denomina *distribución multinomial* a la de la variable suma:

$$\mathbf{y} = \sum_{i=1}^{n} \mathbf{x}_i$$

que indica el número de elementos en la muestra que corresponden a cada una de las clases. Los componentes de esta variable, $\mathbf{y} = (y_1,...,y_G)'$, representan las frecuencias observadas de cada clase y podrán tomar los valores $y_i = 0, 1, ...n$, pero están sujetos a la restricción:

$$\sum y_i = n, \tag{9.21}$$

y su función de probabilidad será:

$$P(y_1 = n_1,..,y_G = n_G) = \frac{n!}{n_1!...n_G!} p_1^{n_1}...p_G^{n_G}$$

donde $\sum n_i = n$. El término combinatorio tiene en cuenta las permutaciones de n elementos cuando hay $n_1,...,n_G$ repetidos. Se comprueba que

$$E(\mathbf{y}) = n\mathbf{p} = \boldsymbol{\mu}_y$$

y

$$Var(\mathbf{y}) = n\left[diag(\mathbf{p}) - \mathbf{p}\mathbf{p}'\right] = diag(\boldsymbol{\mu}_y) - \frac{1}{n}\boldsymbol{\mu}_\mathbf{y}\boldsymbol{\mu}_\mathbf{y}'$$

donde $diag(\mathbf{p})$ es una matriz cuadrada con los elementos de $\mathbf{p}$ en la diagonal y ceros fuera de ella. Esta matriz es singular ya que los elementos de $\mathbf{y}$ están ligados por la ecuación de restricción (9.21). Es fácil comprobar que las distribuciones marginales son binomiales, con:

$$E[y_j] = np_j, \qquad DT\ [y_j] = \sqrt{p_j(1-p_j)}.$$

Además, cualquier distribución condicionada es multinomial. Por ejemplo, la de $G-1$ variables cuando y_G toma el valor fijo n_G es una multinomial en las $G-1$ variables restantes con tamaño muestral $n' = n - n_G$. La distribución condicionada de y_1, y_2 cuando $y_3 = n_3,...,y_G = n_G$ es una binomial, con $n' = n - n_3 - n_4 - ... - n_G$, etc.

Ejemplo 9.3 *En un proceso de control de calidad los elementos pueden tener tres tipos de defectos: leves* (A_1)*, medios* (A_2)*, graves* (A_3) *y se conoce que entre los elementos con defectos la probabilidad de estos errores es* $p_1 = P(A_1) = 0,7$; $p_2 = P(A_2) = 0,2$; *y* $p_3 = P(A_3) = 0,1$. *Calcular la probabilidad de que en los próximos tres elementos defectuosos haya exactamente uno con un defecto grave.*

Los defectos posibles en los tres siguientes elementos son, sin tener en cuenta el orden de aparición :

$$A_1A_1A_3 \ ; \ A_1A_2A_3 \ ; \ A_2A_2A_3$$

y sus probabilidades según la distribución multinomial serán:

$$\begin{aligned} P(x_1 = 2, x_2 = 0, x_3 = 1) &= \frac{3!}{2!0!1!}0,7^2 \cdot 0,2^0 \cdot 0,1 = 0,147 \\ P(x_1 = 1, x_2 = 1, x_3 = 1) &= \frac{3!}{1!1!1!}0,7 \cdot 0,2 \cdot 0,1 = 0,084 \\ P(x_1 = 0, x_2 = 2, x_3 = 1) &= \frac{3!}{0!2!1!}0,7^0 \cdot 0,2^2 \cdot 0,1 = 0,012 \end{aligned}$$

Luego:

$$P(x_3 = 1) = 0,147 + 0,084 + 0,012 = 0,243$$

Este resultado puede también obtenerse considerando la Binomial $(A_3\overline{A}_3)$ con probabilidades $(0,9; 0,1)$ y:

$$P(x_3 = 1) = \left(\begin{array}{c} 3 \\ 1 \end{array} \right) 0,1 + 0,9^2 = 0,243$$

9.5 LA DISTRIBUCIÓN DE DIRICHLET

La distribución de Dirichlet se introduce para representar variables que toman valores entre cero y uno y cuya suma es igual a la unidad. Estos datos se conocen como datos de proporciones (compositional data en inglés). Por ejemplo, supongamos que investigamos el peso relativo que los consumidores asignan a un conjunto de atributos de calidad, y que las evaluaciones de la importancia de los atributos se realizan en una escala de cero a uno. Por ejemplo, con tres atributos un cliente puede dar las valoraciones (0.6, 0,3, 01) indicando que el primer atributo tiene el 60% del peso, el segundo el 30% y el tercero el 10%. Otros ejemplos de este tipo de datos son la proporción de tiempo invertido en ciertas actividades o la composición en % de las distintas sustancias que contienen un grupo de productos. En todos estos casos los datos son vectores de variables continuas $\mathbf{x} = (x_1, ..., x_G)'$ tales que, por construcción, $0 \leq x_j \leq 1$ y existe la ecuación de restricción:

$$\sum_{j=1}^{G} x_j = 1.$$

Una distribución apropiada para representar este tipo de situaciones es la distribución de Dirichlet, cuya función de densidad es:

$$f(x_1, ..., x_G) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)...\Gamma(\alpha_G)} x_1^{\alpha_1 - 1} ... x_G^{\alpha_G - 1}$$

donde $\Gamma(.)$ es la función gamma y $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_G)'$ es el vector de parámetros que caracteriza la distribución, y

$$\alpha_0 = \boldsymbol{\alpha}'\mathbf{1} = \sum_{j=1}^{G} \alpha_j.$$

Se demuestra que

$$E(\mathbf{x}) = \boldsymbol{\alpha}/\alpha_0 = \boldsymbol{\mu}_x,$$

por tanto, los parámetros α_j indican la esperanza relativa de cada componente y

$$Var(\mathbf{x}) = \frac{1}{(\alpha_0 + 1)} (\frac{1}{\alpha_0} diag(\boldsymbol{\alpha}) - \frac{1}{\alpha_0^2} \boldsymbol{\alpha}\boldsymbol{\alpha}').$$

Esta expresión indica que la varianza de cada componente es:

$$var(x_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}.$$

y vemos que el parámetro α_0 determina la varianza de los componentes y que estas varianzas decrecen rápidamente con α_0. Las variables de Dirichlet, al igual que las multinomiales, están ligadas por una ecuación de restricción, con lo que no son linealmente independientes y su matriz de covarianzas será singular. Las covarianzas entre dos componentes son:

$$cov(x_i x_j) = -\frac{\alpha_j \alpha_i}{\alpha_0^2(\alpha_0 + 1)},$$

y las covarianzas también disminuyen con α_0, pero son mayores cuanto mayores sean las esperanzas de las variables.

El lector puede apreciar la similitud entre las fórmulas de las probabilidades, medias y varianzas para la multinomial y la Dirichlet. Esta similitud proviene de que en ambos casos clasificamos el resultado en G grupos. La diferencia es que en el caso multinomial contamos cuantas observaciones de n aparecen de cada grupo, mientras que en el de Dirichlet medimos la proporción que un elemento contiene de la cada clase. En la distribución de Dirichlet el parámetro α_0 tiene un papel similar al tamaño muestral y los cocientes $\alpha_j/\ \alpha_0$ a las probabilidades.

9.6 LA NORMAL k-DIMENSIONAL

La distribución normal escalar tiene como función de densidad:

$$f(x) = (\sigma^2)^{-1/2}(2\pi)^{-1/2} \exp\left\{-(1/2)(x - \mu)^2\sigma^{-2}\right\}.$$

Figura 9.1: Representación de la distribución Normal bivariante y sus marginales.

y escribimos $x \sim N(\mu, \sigma^2)$ para expresar que x tiene distribución normal con media μ y varianza σ^2.

Generalizando esta función, diremos que un vector $\mathbf{x}$ sigue una distribución normal $p-$dimensional si su función de densidad es:

$$f(\mathbf{x}) = |\mathbf{V}|^{-1/2}(2\pi)^{-p/2} \exp\left\{-(1/2)(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{9.22}$$

En la figura 9.1 se muestra el aspecto de una Normal bivariante con $\boldsymbol{\mu} = (0,0)$ y $\mathbf{V} = \begin{bmatrix} 1 & 1/\sqrt{3} \\ 1/\sqrt{3} & 1 \end{bmatrix}$, y sus distribuciones marginales.

Escribiremos que $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \mathbf{V})$. Las propiedades principales de la normal multivariante son:

1. La distribución es simétrica alrededor de $\boldsymbol{\mu}$.

   La simetria se comprueba sustituyendo en la densidad $\mathbf{x}$ por $\boldsymbol{\mu} \pm \mathbf{a}$ y observando que $f(\boldsymbol{\mu} + \mathbf{a}) = f(\boldsymbol{\mu} - \mathbf{a})$.

2. La distribución tiene un único máximo en $\boldsymbol{\mu}$.

   Al ser $\mathbf{V}$ definida positiva el término del exponente $(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ es siempre positivo, y la densidad $f(\mathbf{x})$ será máxima cuando dicho término sea cero, lo que ocurre para $\mathbf{x} = \boldsymbol{\mu}$.

3. La media del vector aleatorio normal es $\boldsymbol{\mu}$ y su matriz de varianzas y covarianzas es $\mathbf{V}$.

   Estas propiedades, que pueden demostrarse rigurosamente, se deducen de la comporación de las densidades univariante y multivariante.

4. Si p variables aleatorias tienen distribución conjunta normal y están incorreladas son independientes.

   La comprobación de esta propiedad consiste en tomar en (9.22) la matriz $\mathbf{V}$ diagonal y comprobar que entonces $f(\mathbf{x}) = f(x_1), ..., f(x_p)$.

5. Cualquier vector $\mathbf{x}$ normal $p-$dimensional con matriz $\mathbf{V}$ no singular puede convertirse mediante una transformación lineal en un vector $\mathbf{z}$ normal $p-$dimensional con vector de medias $\mathbf{0}$ y matriz de varianzas y covarianzas igual a la identidad ($\mathbf{I}$). Llamaremos normal $p-$dimensional estándar a la densidad de $\mathbf{z}$, que vendrá dada por:

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}\mathbf{z}'\mathbf{z}\right\} = \Pi_{i=1}^{p} \frac{1}{(2\pi)^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{z}_i^2\right\} \tag{9.23}$$

La demostración de esta propiedad es la siguiente: al ser $\mathbf{V}$ definida positiva existe una matriz cuadrada $\mathbf{A}$ simétrica que consideramos su raíz cuadrada y verifica:

$$\mathbf{V} = \mathbf{A}\mathbf{A} \tag{9.24}$$

Definiendo una nueva variable:

$$\mathbf{z} = \mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{9.25}$$

entonces $\mathbf{x} = \boldsymbol{\mu} + \mathbf{A}\mathbf{z}$ y según (9.14) la función de densidad de $\mathbf{z}$ es

$$f_z(\mathbf{z}) = f_x(\boldsymbol{\mu} + \mathbf{A}\mathbf{z})\,|\mathbf{A}|$$

y utilizando $\mathbf{A}\mathbf{V}^{-1}\mathbf{A} = \mathbf{I}$, se obtiene (**??**) Por tanto, cualquier vector de variables normales $\mathbf{x}$ en $\Re^p$ puede transformarse en otro vector de $\Re^p$ de variables normales independientes y de varianza unidad.

6. Las distribuciones marginales son normales.

   Si las variables son independientes la comprobación de esta propiedad es inmediata. La demostración general puede verse, por ejemplo, en Mardia et al (1979).

7. Cualquier subconjunto de $h < p$ variables es normal $h-$dimensional.

   Es una extensión del la propiedad anterior y se demuestra analogamente.

8. Si $\mathbf{y}$ es $(k \times 1)$, $k \leq p$, el vector $\mathbf{y} = \mathbf{A}\mathbf{x}$, donde $\mathbf{A}$ es una matriz $(k \times p)$, es normal $k-$dimensional. En particular, cualquier variable escalar $y = \mathbf{a}'\mathbf{x}$, (siendo $\mathbf{a}'$ un vector $1 \times p$ no nulo) tiene distribución normal.

   La demostración puede verse, por ejemplo, en Mardia et al (1979).

9. Al cortar con hiperplanos paralelos al definido por las p variables que forman la variable vectorial, $\mathbf{x}$, se obtienen las curvas de nivel, cuya ecuación es:

$$(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = cte.$$

   Las curvas de nivel son, por tanto, elipsoides, y definen una medida de la distancia de un punto al centro de la distribución. Esta medida ha aparecido ya en la descripción de datos del capítulo 3 donde estudiamos su interpretación. Se denomina distancia de Mahalanobis y la representaremos por :

$$D^2 = (\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \tag{9.26}$$

Como ilustración, consideremos el caso más simple de dos distribuciones univariantes indicado en la figura 13.3. La observación x=3, indicada con una X, en el gráfico, esta con la distancia euclidea más cerca del centro de la distribución A, que es cero, que del centro de la B que es diez. Sin embargo, con la distancia de Mahalanobis la distancia del punto X a la distribución A que tiene desviación típica uno es $(3-0)^2/1$, mientras que la distancia al centro de la B, que tiene desviación típica diez, es $(3-10)^2/10^2 = 0,7^2$ y el punto X está mucho más cerca, con esta distancia, de la distribución B. Esto es consecuencia de que es mucho más probable que este punto provenga de la distribución B que de la A.

10. La distancia de Mahalanobis se distribuye como una χ^2 con p grados de libertad.

    Para comprobarlo, hagamos la transformación (9.25) y como $\mathbf{V}^{-1} = \mathbf{A}^{-1}\mathbf{A}^{-1}$ se obtiene que

$$D^2 = \mathbf{z}'\mathbf{z} = \sum \mathbf{z}_i^2$$

    donde cada $\mathbf{z}_i$ es $N(0,1)$. Por tanto $D^2 \sim \chi_p^2$.

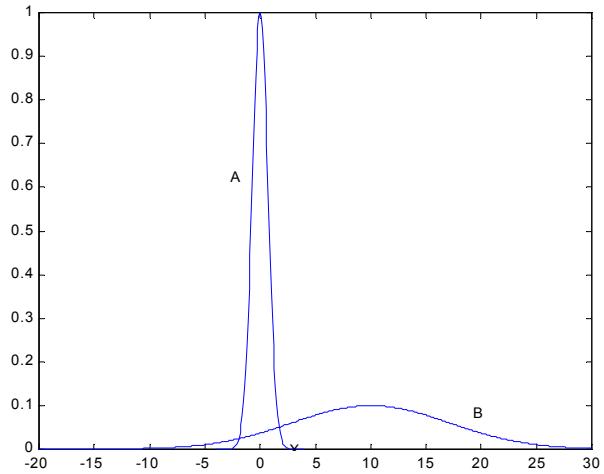


Figura 9.2: El punto X esta más cerca, con la distancia euclidea del centro de la distribución A pero con la distancia de Mahalanobis lo está de la B

9.6.1 Distribuciones condicionadas

Particionemos el vector aleatorio en dos partes, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)'$, donde $\mathbf{x}_1$ es un vector de dimensión p_1 y $\mathbf{x}_2$ de dimensión p_2, siendo $p_1 + p_2 = p$. Particionemos también la matriz de covarianzas del vector $\mathbf{x}$ en bloques asociados a estos dos vectores, como:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \tag{9.27}$$

donde, por ejemplo, $\mathbf{V}_{11}$, la matriz de covarianzas del vector $\mathbf{x}_1$, es cuadrada de orden p_1, $\mathbf{V}_{12}$, la matriz de covarianzas entre los vectores $\mathbf{x}_1$ y $\mathbf{x}_2$ tiene dimensiones $p_1 \times p_2$, y $\mathbf{V}_{22}$, la matriz de covarianzas del vector $\mathbf{x}_2$, es cuadrada de orden p_2. Queremos calcular la distribución condicionada del vector $\mathbf{x}_1$ dados los valores del vector $\mathbf{x}_2$. Vamos a demostrar que esta distribución es normal, con media:

$$E\left[\mathbf{x}_1|\mathbf{x}_2\right] = \boldsymbol{\mu}_1 + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right) \tag{9.28}$$

y matriz de varianzas y covarianzas:

$$Var\left[\mathbf{x}_1|\mathbf{x}_2\right] = \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21} \tag{9.29}$$

Para interpretar estas expresiones supongamos primero el caso bivariante donde ambas variables son escalares de media cero. Entonces la media se reduce a

$$E\left[x_1|x_2\right] = \sigma_{12}\sigma_{22}^{-1}x_2$$

que es la expresión habitual de la recta de regresión con pendiente $\beta = \sigma_{12}/\sigma_{22}$. La expresión de la varianza condicionada alrededor de la recta de regresión es

$$var\left[x_1|x_2\right] = \sigma_{11} - \sigma_{12}^2/\sigma_{22} = \sigma_1^2(1-\rho^2)$$

donde $\rho = \sigma_{12}/\sigma_{22}^{1/2}\sigma_{11}^{1/2}$ es el coeficiente de correlación entre las variables. Esta expresión indica que la variabilidad de la distribución condicionada es siempre menor que la de la marginal y la reducción de variabilidad es tanto mayor cuanto mayor sea ρ^2.

Supongamos ahora que x_1 es escalar pero $\mathbf{x}_2$ es un vector. La expresión de la media condicionada proporciona la ecuación de regresión múltiple

$$E\left[x_1|\mathbf{x}_2\right] = \mu_1 + \boldsymbol{\beta}'\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)$$

donde $\boldsymbol{\beta} = \mathbf{V}_{22}^{-1}\ \mathbf{V}_{21}$ siendo $\mathbf{V}_{21}$ el vector de covarianzas entre x_1 y los componentes de $\mathbf{x}_2$. La varianza de esta distribución condicionada es

$$var\left[x_1|\mathbf{x}_2\right] = \sigma_1^2(1-R^2)$$

donde $R^2 = \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}/\sigma_1^2$ es el coeficiente de correlación múltiple.

En el caso general, estas expresiones corresponden al conjunto de regresiones múltiples de los componentes de $\mathbf{x}_1$ sobre las variables $\mathbf{x}_2$, que se conoce como regresión multivariante.

Demostración La expresión de la distribución condicionada es

$$f\left(\mathbf{x}_1|\mathbf{x}_2\right) = \frac{f\left(\mathbf{x}_1, \mathbf{x}_2\right)}{f\left(\mathbf{x}_2\right)}$$

Como las distribuciones $f\left(\mathbf{x}_1, \mathbf{x}_2\right)$ y $f\left(\mathbf{x}_2\right)$ son normales multivariantes al hacer el cociente quedará un cociente entre determinantes y la diferencia entre los exponentes de las normales. Comencemos calculando el exponente resultante. Será

$$\left(\mathbf{x} - \boldsymbol{\mu}\right)'\mathbf{V}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right) - \left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)'\mathbf{V}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right) \tag{9.30}$$

Vamos a descomponer la primera forma cuadrática en los términos correspondientes a $\mathbf{x}_1$y $\mathbf{x}_2$. Para ello particionaremos $(\mathbf{x}-\boldsymbol{\mu})$ como $(\mathbf{x}_1-\boldsymbol{\mu}_1, \mathbf{x}_2-\boldsymbol{\mu}_2)'$, particionaremos $\mathbf{V}$ como en (9.27), y utilizaremos la expresión de la inversa de una matriz particionada (véase la sección 2.2.3). Realizando el producto se obtiene.

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= (\mathbf{x}_1-\boldsymbol{\mu}_1)'\mathbf{B}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)-(\mathbf{x}_1-\boldsymbol{\mu}_1)'\mathbf{B}^{-1}\mathbf{V}_{12}\mathbf{V}_{12}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)- \\ &-(\mathbf{x}_2-\boldsymbol{\mu}_2)\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{B}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)+(\mathbf{x}_2-\boldsymbol{\mu}_2)\mathbf{V}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)- \\ &+(\mathbf{x}_2-\boldsymbol{\mu}_2)'\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\end{aligned}$$

donde $\mathbf{B}=\left(\mathbf{V}_{11}-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\right)$, que es la expresión utilizada en (9.29). El cuarto término de esta expresión se cancela en la diferencia (9.30), y los otros cuatro pueden agruparse como

$$\left(\mathbf{x}_1-\boldsymbol{\mu}_1-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right)'\mathbf{B}^{-1}\left(\mathbf{x}_1-\boldsymbol{\mu}_1-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right).$$

Esta expresión muestra que el exponente de la distribución corresponde a una variable normal con vector de medias y matriz de covarianzas igules a los indicados en (9.28) y (9.29). Vamos a comprobar que el cociente de determinantes conduce también a la misma matriz de covarianzas. Utilizando que, según la sección 2.3.5, $|\mathbf{V}|=|\mathbf{V}_{22}|\left|\mathbf{V}_{11}-\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\right|=|\mathbf{V}_{22}||\mathbf{B}|$. Como en el denominador tenemos $|\mathbf{V}_{22}|$, el cociente proporciona el término único $|\mathbf{B}|$. Finalmente, quedará en término $(2\pi)^{p/2-p_2/2}=(2\pi)^{p_1/2}$.

En conclusión, la expresión resultante será la de la función de densidad normal multivariante de orden p_1, con vector de medias dado por (9.28) y matriz de covarianzas dada por (9.29). ■

Ejemplo 9.4 *La distribución de los gastos en dos productos (x,y) de un grupo de consumidores sigue una distribución normal bivariante con medias respectivas 2 y 3 euros y matriz de varianzas y covarianzas*

$$\begin{bmatrix} 1 & 0,8 \\ 0,8 & 2 \end{bmatrix}$$

Calcular la distribución condicionada de los gastos en el producto y para los consumidores que gastan 4 euros en el producto x.

La distribución condicionada $f(y/x=4)=f(4,y)/f_x(4)$. La distribución marginal de x es normal, $N(2,1)$. Los términos de la distribución conjunta $f(x,y)$ serán:

$$\begin{aligned}|\mathbf{V}|^{1/2} &= \left(\sigma_1^2\sigma_2^2\left(1-\varrho^2\right)\right)^{1/2}=\sigma_1\sigma_2\sqrt{1-\varrho^2} \\ \mathbf{V}^{-1} &= \frac{1}{\sigma_1^2\sigma_2^2\left(1-\varrho^2\right)}\begin{vmatrix}\sigma_2^2 & -\varrho\sigma_2\sigma_1 \\ -\varrho\sigma_2\sigma_1 & \sigma_1^2\end{vmatrix}\end{aligned}$$

donde en este ejemplo $\sigma_1^2=1$, $\sigma_2^2=2$, y $\varrho=0,8/\sqrt{2}=0,566$. El exponente de la normal bivariante $f(x,y)$ será:

$$-\frac{1}{2\left(1-\varrho^2\right)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\left(\frac{y-\mu_2}{\sigma_2}\right)^2-2\varrho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right\}=-\frac{A}{2}$$

En consecuencia, tendremos:

$$\begin{aligned} f(y \mid x) &= \frac{\left(\sigma_1\sigma_2\sqrt{1-\varrho^2}\right)^{-1}(2\pi)^{-1}\exp\left\{-\frac{A}{2}\right\}}{\sigma_1^{-1}(2\pi)^{-1}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right\}} = \\ &= \frac{1}{\sigma_2\sqrt{1-\varrho^2}}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}B\right\} \end{aligned}$$

donde el término resultante en el exponente, que llamaremos B, será:

$$\begin{aligned} B &= \frac{1}{1-\varrho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2+\left(\frac{y-\mu_2}{\sigma_2}\right)^2-2\varrho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}-\left(\frac{x-\mu_1}{\sigma_1}\right)^2(1-\varrho^2)\right] \\ &= \frac{1}{1-\varrho^2}\left[\left(\frac{y-\mu_2}{\sigma_2}\right)^2-\varrho\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]^2 \\ B &= \frac{1}{\sigma_2^2(1-\varrho^2)}\left[y-\left(\mu_2+\varrho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right)\right]^2 \end{aligned}$$

Este exponente corresponde a una distribución normal con media:

$$E[y \mid x] = \mu_2 + \varrho\frac{\sigma_2}{\sigma_1}(x-\mu_1)$$

que es la recta de regresión, y desviación típica:

$$DT[y \mid x] = \sigma_2\sqrt{1-\varrho^2}$$

Para $x = 4$.

$$E[y \mid 4] = 3 + \left(\frac{0,8}{\sqrt{2}}\right)\cdot\frac{\sqrt{2}}{1}(4-2) = 4,6.$$

Como hay una correlación positiva de 0,566 entre los gastos en ambos productos los consumidores que gastan más en uno también en promedio tienen gastos medios más altos en el otro. La variabilidad de la disribución condicionada será:

$$Var[y \mid 4] = \sigma_2^2\left(1-\varrho^2\right) = 2(1-0,32) = 1,36$$

y será menor que la varianza de la marginal porque cuando condicionamos tenemos más información.

9.7 DISTRIBUCIONES ELÍPTICAS

La distribución normal multivariante es un caso particular de una familia de distribuciones muy utilizadas en el análisis multivariante: las distribuciones elípticas. Para introducirlas, consideremos primero el caso más simple de las distribuciones esféricas

9.7.1 Distribuciones esféricas

Diremos que una variable vectorial $\mathbf{x} = (x_1, ..., x_p)'$ sigue una distribución esférica si su función de densidad depende de la variable sólo por la distancia euclídea $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{p} x_i^2$. Esta propiedad implica que:

1. Los contornos de equiprobabilidad de la distribución son esferas con centro en el origen.

2. La distribución es invariante ante rotaciones. En efecto, si definimos una nueva variables $\mathbf{y} = \mathbf{C}\mathbf{x}$, donde $\mathbf{C}$ es una matriz ortogonal, la densidad de la variable $\mathbf{y}$ es la misma que la de la variable $\mathbf{x}$.

Un ejemplo de distribución esférica, estudiado en la sección anterior, es la función de densidad Normal estándar multivariante, cuya densidad es

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \exp(-\frac{1}{2}\mathbf{x}'\mathbf{x}) = \Pi_{i=1}^{p} \frac{1}{(2\pi)^{1/2}} \exp(-\frac{1}{2}x_i^2)$$

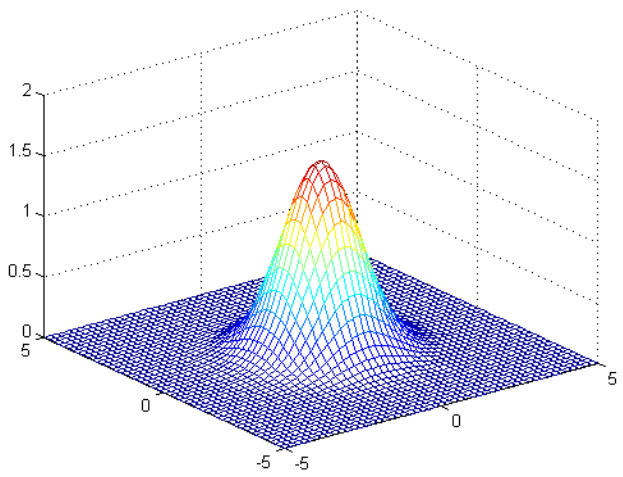


Figura 9.3: Densidad de la normal estandar bivariante

Esta densidad está representada en la figura 9.3, y las dos variables escalares que forman el vector son independientes. Este propiedad es característica de la normal, ya que, habitualmente, los componentes de las distribuciones esféricas son dependientes. Por ejemplo, la

distribución multivariante de Cauchy, dada por

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\pi^{(p+1)/2}}(1+\mathbf{x}'\mathbf{x})^{-(p+1)/2} \tag{9.31}$$

tiene colas más pesadas que la normal, como en el caso univariante, y es fácil comprobar que esta función no puede escribirse como producto de distribuciones univariantes de Cauchy, por lo que sus componentes aunque están incorrelados no son independientes.

Otra distribución esférica importante es la doble exponencial. En el caso bivariante esta distribución tiene función de densidad

$$f(\mathbf{x}) = \frac{1}{2\pi}\exp(-\sqrt{\mathbf{x}'\mathbf{x}})$$

y aunque la función de densidad puede parecer similar a la normal tiene colas mucho más pesadas. La figura 9.4 muestra esta distribución.

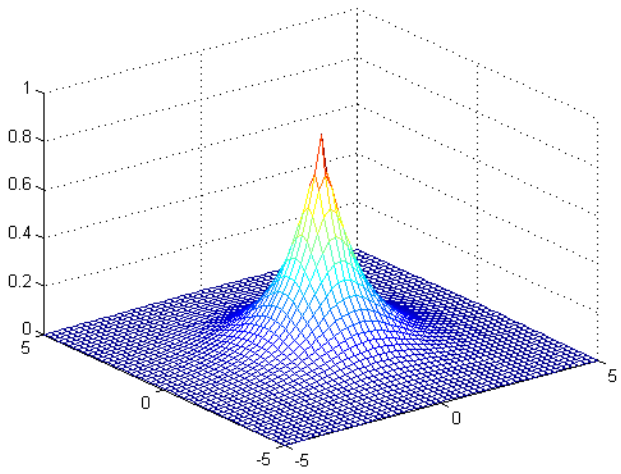


Figura 9.4: Dendidad de la doble exponencial bivariante

9.7.2 Distribuciones elípticas

Si la variable $\mathbf{x}$ sigue una distribución esférica y $\mathbf{A}$ es una matriz cuadrada de dimensión p y $\mathbf{m}$ un vector de dimensión p, la variable

$$\mathbf{y} = \mathbf{m} + \mathbf{A}\mathbf{x} \tag{9.32}$$

se dice que sigue una distribución elíptica. Como una variable esférica tiene media cero y matriz de covarianzas $c\mathbf{I}$, es inmediato que una variable elíptica tiene media $\mathbf{m}$ y matriz de covarianzas $\mathbf{V} = c\mathbf{A}\mathbf{A}'$. Las distribuciones elípticas tienen las propiedades siguientes:

1. Su función de densidad depende de la variable a través de la distancia de Mahalanobis:

$$(\mathbf{y}-\mathbf{m})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{m})$$

2. Los contornos de equiprobabilidad de la distribución son elipsoides con centro en el punto $\mathbf{m}$.

La distribución normal multivariante general es el miembro más conocido de las distribuciones elípticas. Otro miembro de esta familia es la distribución t multivariante. Aunque existen distintas versiones de esta distribución, la más habitual se construye dividiendo cada componente de un vector de variables normales multivariantes $N_p(\mathbf{m},\mathbf{V})$ por la misma variable escalar: la raiz cuadrada de una χ^2 dividida por sus grados de libertad. Es obvio, por construcción, que las marginales serán t de Student, y se obtiene que la función de densidad de la variable multivariante resultante es

$$f(\mathbf{y}) = \frac{\Gamma(\frac{(v+p)}{2})}{(\pi v)^{p/2}\Gamma(\frac{v}{2})}\,|\mathbf{V}|^{-1/2}\left[1+(\mathbf{y}-\mathbf{m})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{m})\right]^{-(v+p)/2} \tag{9.33}$$

donde el escalar v se denomina grados de libertad. Observemos que si hacemos $v=1$, $\mathbf{m}=\mathbf{0}$, $\mathbf{V}=\mathbf{I}$, obtenemos la distribución de Cauchy multivariante (9.31) que tiene simetría esférica. Para $v>2$ la media de la distribución es $\mathbf{m}$ y la varianza $v/(v-2)\mathbf{V}$.

Las distribuciones elípticas comparten muchas propiedades de la normal: las distribuciones marginales y condicionadas son también elípticas, y las medias condicionadas son función lineal de las variables condicionantes. Sin embargo, la normal tiene la propiedad de que es el único miembro de la familia donde si la matriz de covarianzas es diagonal todas las variables componentes son independientes. El lector interesado en la demostración de este resultado puede encontrarlo en Muirhead (1982).

9.8 (*)LA DISTRIBUCIÓN DE WISHART

9.8.1 Concepto

La distribución de Wishart se utiliza para representar la incertidumbre respecto a una matriz de varianzas y covarianzas de variables normales multivariantes. En el caso escalar, la distribución que representa esta incertidumbre es la ji-cuadrado de Pearson, χ^2, y la distribución de Wishart estándar puede considerarse como una generalización multivariante de esta distribución.

Recordemos los resultados univariantes: Si $(x_1,...x_m)$ es un conjunto de variables aleatorias normales independientes $N(0,\sigma^2)$, la suma estandarizada de sus cuadrados, $\sigma^{-2}\sum_{i=1}^m x_i^2$, sigue una distribución χ^2_m. También decimos que $w=\sum_{i=1}^m x_i^2$ sigue una distribución $\sigma^2\chi^2_m$. La densidad de una distribución χ^2_m es un caso particular de la Gamma con parámetros $\left(\frac{1}{2},\frac{m}{2}\right)$ y tiene función de densidad dada por

$$f(\chi^2) = k(\chi^2)^{\frac{m}{2}-1}\exp\left\{-\frac{1}{2}\chi^2\right\}, \tag{9.34}$$

donde k es una constante. Por otro lado, la distribución de la variable $w = \sum_{i=1}^{m} x_i^2$ será la Gamma con parámetros $\left(\frac{1}{2\sigma^2};\ \frac{m}{2}\right)$, y su densidad tendrá la forma:

$$f(w) = k\left(\sigma^2\right)^{-\frac{m}{2}} w^{\frac{m}{2}-1} \exp\left\{-\frac{1}{2}\sigma^{-2}w\right\}. \tag{9.35}$$

Consideremos ahora un conjunto de m vectores aleatorios, $(\mathbf{x}_1, ..., \mathbf{x}_m)$, de dimensión p con la misma distribución $N_p(0, \mathbf{I})$. La estimación de su matriz de varianzas y covarianzas se obtendrá de $\Sigma_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i'/m$, y el numerador de esta expresión

$$\mathbf{W} = \Sigma_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i' \tag{9.36}$$

que es una matriz cuadrada $p \times p$, simétrica y definida positiva, decimos que sigue una distribución Wishart con m grados de libertad. Esta afirmación debe interpretarse en el sentido de que la distribución conjunta de los $\frac{1}{2}p(p+1)$ elementos distintos de $\mathbf{W}$ es

$$f(w_{11}, ..., w_{pp}) = c|\mathbf{W}|^{(m-p-1)/2} \exp\left\{-\frac{1}{2}tr\mathbf{W}\right\} \tag{9.37}$$

donde c es una constante para que la función integre a uno (véase Seber, 1984). Observemos que para $p = 1$ se obtiene (9.34). Escribiremos $\mathbf{W} \sim W_p(m)$, donde p indica que se trata de la distribución de los elementos de una matriz cuadrada y simétrica de orden p, y m son los grados de libertad. Observemos que esta distribución depende únicamente de las dos medidas escalares del tamaño de la matriz: la traza y el determinante. Por tanto, todas las combinaciones de elementos de la matriz que conduzcan a los mismos valores de estas medidas de tamaño tienen la misma probabilidad.

Consideremos ahora m vectores aleatorios $(\mathbf{x}_1, ..., \mathbf{x}_m)$ de una distribución $N_p(0, \Sigma)$, donde hemos utilizado el símbolo Σ en lugar de $\mathbf{V}$ para representar la matriz de covarianzas para evitar confusiones cuando esta distribución se utilice en el análisis bayesiano del capítulo siguiente. La distribución de los elementos de la matriz

$$\mathbf{W} = \Sigma_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i' \tag{9.38}$$

es la distribución Wishart con m grados de libertad y matriz de parámetros Σ, dada por

$$f(w_{11}, ..., w_{pp}) = c|\Sigma|^{-m/2}|\mathbf{W}|^{(m-p-1)/2} \exp\left\{-\frac{1}{2}tr\Sigma^{-1}\mathbf{W}\right\}. \tag{9.39}$$

En general, si una matriz cuadrada y simétrica sigue la distribución (9.39), donde Σ es una matriz simétrica $(p \times p)$ no singular definida positiva de componentes constantes, diremos que sigue la distribución Wishart con m grados de libertad y matriz de parámetros Σ, y escribiremos $\mathbf{W} \sim W_p(m, \Sigma)$. Observemos que para $p = 1$ esta expresión se reduce (9.35), y si hacemos $\Sigma = 1$, la densidad (9.39) se reduce a (9.34). La figura 9.5 presenta un ejemplo de esta distribución

9.8.2 Propiedades de la distribución

La distribución de Wishart tiene las propiedades siguientes:

1. La esperanza de la distribución es:

$$E[\mathbf{W}] = m\Sigma$$

   lo que implica que $\mathbf{W}/m$ tiene esperanza Σ.

2. La suma de dos distribuciones χ^2 independientes es otra distribución χ^2 con grados de libertad la suma de ambas. Análogamente, si $\mathbf{W}_1 \sim W_p(m_1, \Sigma)$ y $\mathbf{W}_2 \sim W_p(m_2, \Sigma)$ son independientes, entonces $\mathbf{W}_1 + \mathbf{W}_2 \sim W_p(m_1 + m_2, \Sigma)$. Este resultado es consecuencia inmediata de la definición de la distribución por (9.34).

3. Si $\mathbf{A}$ es una matriz $h \times p$ de constantes, y $\mathbf{W} \sim \mathbf{W}_p(m, \Sigma)$, la distribución de $\mathbf{AWA}' \sim \mathbf{W}_h(m, \mathbf{A}^{-1}\Sigma\mathbf{A}'^{-1})$.

   En efecto, por (9.38) la variable $\mathbf{AW}'\mathbf{A}$ será

$$\mathbf{A}\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i'\mathbf{A}' = \sum_{i=1}^{m}\mathbf{y}_i\mathbf{y}_i'$$

   donde ahora $\mathbf{y}_i$ es $N(0, \mathbf{A}\Sigma\mathbf{A}')$, y aplicando la definición de la distribución Wishart se obtiene el resultado.

4. Si $\mathbf{S}$ es la matriz de varianzas y covarianzas muestral

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{P}\mathbf{X}$$

   donde $\mathbf{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$ es idempotente, entonces

$$n\mathbf{S} \sim W_p(n-1, \Sigma).$$

   Esta expresión indica que si definimos el estimador

$$\widehat{\mathbf{S}} = \frac{1}{(n-1)}\mathbf{X}'\mathbf{P}\mathbf{X} = \frac{n}{(n-1)}\mathbf{S}$$

   su esperanza será Σ, y $\widehat{\mathbf{S}}$ será un estimador centrado para la matriz de varianzas. Podemos escribir que $(n-1)\widehat{\mathbf{S}} \sim W_p(n-1, \Sigma)$. Este resultado es análogo al del caso escalar: $(n-1)\widehat{s}^2$, donde $\widehat{s}^2$ es el estimador centrado de la varianza, sigue una distribución $\sigma^2\chi^2_{n-1}$.

9.9 LA T^2 DE HOTELLING

Si $\mathbf{x}$ es un vector aleatorio $N_p(\boldsymbol{\mu}, \mathbf{V})$, la variable $(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ es una χ^2 con p grados de libertad. Si sustituimos $\mathbf{V}$ por su estimación $\widehat{\mathbf{S}}$, la matriz de varianzas muestral dividiendo por $n-1$, la distribución que se obtiene se denomina T^2 de Hotelling. En general, si $x \sim N_p(\boldsymbol{\mu}, \mathbf{V})$ y $(n-1)\widehat{\mathbf{S}} \sim W_p(n-1, \mathbf{V})$, la distribución de la variable escalar:

$$T^2 = (\mathbf{x}-\boldsymbol{\mu})'\widehat{\mathbf{S}}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{9.40}$$

que representa la distancia de Mahalanobis entre una variable y su media poblacional, pero calculada con la matriz de covarianzas estimada, se denomina distribución T^2 de Hotelling con p y $n-1$ grados de libertad. Diremos que $T^2 \sim T^2(p, n-1)$. Asintóticamente, como $\widehat{\mathbf{S}} \to \mathbf{V}, T^2$ converge a la distancia de Mahalanobis y la distribución de Hotelling a la distribución χ^2_p. Por tanto, para n grande, la distribución de Hotelling es muy similar a una χ^2_p. Para tamaños muestrales más pequeños tiene una mayor variabilidad que la χ^2_p, como consecuencia de la mayor incertidumbre al utilizar la matriz estimada, $\widehat{\mathbf{S}}$, en lugar de la matriz de covarianzas verdadera, $\mathbf{V}$.

Si $\bar{\mathbf{x}}$ es la media muestral, como $\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \frac{1}{n}\mathbf{V})$, la distribución de

$$(\bar{\mathbf{x}}-\boldsymbol{\mu})'\left(\frac{\widehat{\mathbf{S}}}{n}\right)^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) = n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\widehat{\mathbf{S}}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})$$

es también una T^2 de Hotelling. Observemos que si $p=1$, la T^2 se reduce a:

$$T^2 = \frac{n(\overline{x}-\mu)^2}{\widehat{s^2}} = t^2 \tag{9.41}$$

y coincide con el estadístico t de Student. Por tanto $T^2(1,m) = t^2_m$.

La distribución de Hotelling no se tabula, ya que con una simple transformación se reduce a la distribución F del análisis de la varianza de Fisher. Se demuestra (véase Seber, 1984 o Muirhead, 1982) que:

$$F_{p,n-p} = \frac{n-p}{p(n-1)}T^2(p, n-1) \tag{9.42}$$

lo que permite calcular la distribución de T^2 en función de las tablas de la distribución F. Este resultado es consistente con (9.42), ya que, asintóticamente, $pF_{p,n-p}$ tiende a una distribución χ^2_p. La figura (**??**) muestra un ejemplo de la distribución de Hotelling comparada con la χ^2_p. Vemos que para tamaño muestral muy pequeño, $n=15$, las colas de la distribución son más planas que las de la ji-cuadrado indicando la mayor incertidumbre existente, pero para n=50, ambas son ya muy similares. La aproximación depende del cociente n/p, y si este es grande, mayor de 25, podemos aproximar bien la distribución de Hotelling mediante la ji-cuadrado.

Figura 9.5: Distribución Wishart dibujada en función de la traza y el determinante

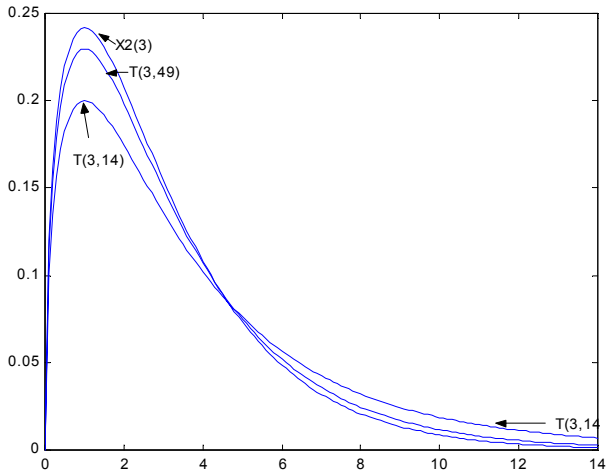


Figura 9.6: La distribución de Hotelling para dos valores del tamaño muestral y la distribución ji-cuadrado.

9.10 DISTRIBUCIONES MEZCLADAS

Los datos multivariantes son frecuentemente heterogéneos. Por ejemplo, si observamos el gasto en distintos productos en una muestra de consumidores, es esperable que haya grupos de consumidores con patrones de gasto distintos: los consumidores sin hijos respecto a que las que los tienen, o los jóvenes respecto a los ancianos. En general, si una población donde hemos definido una variable aleatoria vectorial, $\mathbf{x}$, puede subdividirse en G estratos más homogéneos y llamamos π_i a la proporción de elementos en el estrato i ($\sum_{i=1}^{G} \pi_i = 1$) y $f_i(\mathbf{x})$ a la función de densidad de la variable en el estrato i, la función de densidad en toda la población vendrá dada por la mezcla de densidades

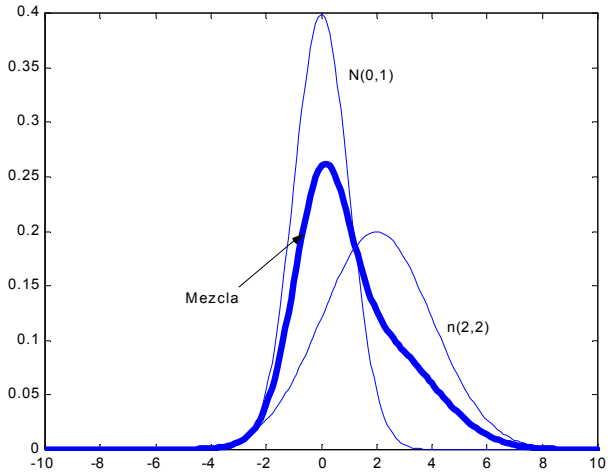
$$f(\mathbf{x}) = \sum_{i=1}^{G} \pi_i f_i(\mathbf{x}). \tag{9.43}$$

Para justificar esta distribución, nótese que observar un elemento al azar de esa población puede plantearse en dos etapas. En la primera, seleccionamos el estrato al azar mediante una variable escalar, g, que toma los valores $1, ..., G$ con probabilidades $\pi_1, ..., \pi_G$. En la segunda, seleccionamos aleatoriamente el elemento de la población seleccionada, $f_i(\mathbf{x})$. La probabilidad de que el elemento seleccionado tome un valor $\mathbf{x} \in \mathbf{A}$ será

$$P(\mathbf{x} \in \mathbf{A}) = \sum_{i=1}^{G} P(\mathbf{x} \in \mathbf{A}/g = i)P(g = i)$$

y llamando $\pi_i = P(g = i)$, la distribución marginal de la variable $\mathbf{x}$ cuando no se conoce la variable g viene dada por (9.43).

Las figuras 9.7 y **??** presentan ejemplos de distribuciones obtenidas como mezclas de dos distribuciones univariantes con proporcion de mezcla 50% ($\pi_1 = \pi_2 = .5$). En la figura 9.7 las dos distribuciones de partida son normales con la misma media y diferente varianza. La distribución resultante tiene la misma media y una varianza que es el promedio de las varianzas de las distribuciones. Observemos que la distribución mezclada no es normal. En la figura **??** las distribuciones tienen distinta media y varianza. Como comprobaremos ahora la media es en este caso el promedio de las medias pero la varianza tiene una expresión más complicada porque además de la variabilidad de las distribuciones con respecto a sus medias se añade la variabilidad debida a las diferencias entre las medias.



Mezcla al 50% de dos distribuciones normales con distinta media y varianza

Los parámetros de la distribución de la variable mezclada $(\boldsymbol{\mu}, \mathbf{V})$, o marginal, se obtienen fácilmente conocidos las medias $\boldsymbol{\mu}_i$ y matrices de varianzas $\mathbf{V}_i$ de las distribuciones que generan la mezcla, o condicionadas.

1. La media de la distribución mezclada o media de la distribución marginal es

$$\boldsymbol{\mu} = \sum_{i=1}^{G} \pi_i \boldsymbol{\mu}_i \tag{9.44}$$

La demostración de este resultado es inmediato aplicando las propiedades de la esperanza condicional (9.19). Introduciendo la variable de clasificación g, tenemos que, como $E(\mathbf{x}/g{=}i) = \boldsymbol{\mu}_i$

$$E(\mathbf{x}) = E_g E_{x/g}(\mathbf{x}) = E_g(\boldsymbol{\mu}_i) = \sum_{i=1}^{G} \pi_i \boldsymbol{\mu}_i$$

2. La matriz de varianzas y covarianzas de la distribución marginal viene dada por

$$\mathbf{V} = \sum_{i=1}^{G} \pi_i \mathbf{V}_i + \sum_{i=1}^{G} \pi_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' \tag{9.45}$$

Para demostrar este resultado, introduciendo que

$$\mathbf{V} = E\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\right] = E\left[(\mathbf{x} - \boldsymbol{\mu}_i + \boldsymbol{\mu}_i - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu}_i + \boldsymbol{\mu}_i - \boldsymbol{\mu})'\right]$$

y aplicando de nuevo las propiedades de la esperanza condicional (9.19), obtenemos que

$$E_{x/y}\left[(\mathbf{x} - \boldsymbol{\mu}_i + \boldsymbol{\mu}_i - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu}_i + \boldsymbol{\mu}_i - \boldsymbol{\mu})'\right] = \mathbf{V}_i + (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'$$

y tomando ahora la esperanza de esta cantidad respecto a la distribución de g se obtiene el resultado deseado.

La expresión (9.45) puede interpretarse como una descomposición de la variabilidad similar a la del análisis de la varianza. La variabilidad total, que es la matriz de varianzas y covarianzas de la marginal, $\mathbf{V}$, se descompone en una variabilidad explicada, $\sum_{i=1}^{G} \pi_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'$, que tiene en cuenta las diferencias entre las medias de las distribuciones condicionadas $\boldsymbol{\mu}_i$ y la marginal, $\boldsymbol{\mu}$, y una variabilidad no explicada $\sum_{i=1}^{G} \pi_i \mathbf{V}_i$, que es la variabilidad con respecto a las distribuciones condicionadas. Por ejemplo, en el caso escalar representado en la figura **??**, esta expresión se reduce a :

$$\sigma^2 = \sum_{i=1}^{G} \pi_i \sigma_i^2 + \sum_{i=1}^{G} \pi_i (\mu_i - \mu)^2$$

y descompone la varianza de los datos en sus fuentes de variabilidad. En la figura **??** las medias son cero y dos y las varianzas uno y cuatro, y tenemos que

$$\sigma^2 = .5(1) + .5(4) + .5(0-1)^2 + .5(2-1)^2 = 3.5$$

que corresponde a una desviación típica de 1.87, que está de acuerdo con la distribución de la figura **??**.

En el caso multivariante las mezclas de distribuciones normales pueden representar una gama muy amplia de distribuciones. La figura 9.8 presenta un ejemplo.

9.11 Lecturas complementarias

El lector puede encontrar exposiciones más detalladas y más ejemplos de la teoría aquí expuesta en la mayoría de los textos básicos de estadística y en los primeros capítulos de la mayoría de los textos multivariantes. En inglés, Flury (1997), Johnson y Wichern (1998) y Mardia et al (1979) son buenas exposiciones en orden creciente de complejidad matemática.

Existen otras distribuciones más flexibles que la Dirichlet para modelar datos multivariantes de proporciones. Aitchinson (1986) es una buena referencia de distintas distribuciones que pueden usarse para este objetivo. El lector interesado en ampliar las propiedades de las propiedades elípticas puede acudir a Flury (1997), que es una excelente introducción, y a Muirhead (1982). Otras buenas referencias sobre las distribuciones aquí expuestas son Anderson (1984), Seber (1984) y Johnson y Kotz (1970). Patel y Read (1982) se concentran en la distribución normal.

Las distribuciones mezcladas han ido teniendo un papel cada vez mayor en Estadística, tanto desde el punto de vista clásico como Bayesiano. Un referencia básica es Titterington at al (1987). Muchos de los textos de cluster, que comentaremos en el capítulo 14, incluyen el estudio de estas distribuciones.

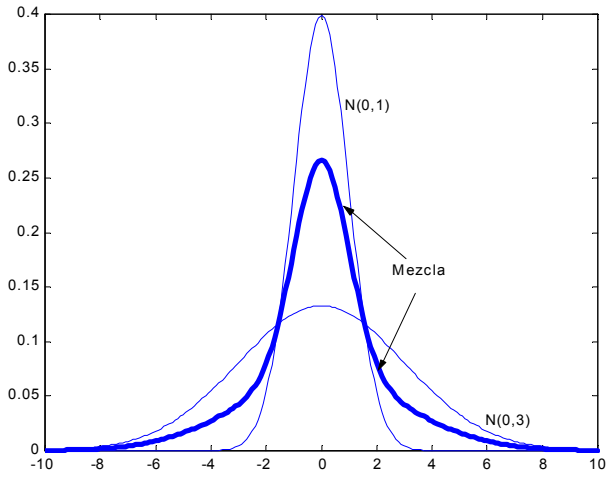


Figura 9.7: Mezcla al 50% de dos distribuciones normales con la misma media y distinta varianza

Figura 9.8: Mezcla de dos normales bivariantes en la proporción 50% con medias (0,0) y (3,3) y distintas matrices de covarianzas.

Ejercicios

Ejercicio 9.1 *Dada la función de densidad conjunta* $f(x,y) = 6x$ *definida en* $0 < x < 1$, $0 < y < 1 - x$, *comprobar que las densidades marginales de ambas variables son* $f(x) = 6x(1-x)$, $0 < x < 1$ *y* $f(y) = 3(1-y)^2$, $0 < y < 1$.

Ejercicio 9.2 *Comprobar que las densidades condicionadas en el ejemplo anterior son* $f(y|x) = \frac{1}{(1-x)}$, $0 < y < 1 - x$ *y* $f(x|y) = \frac{2x}{(1-y)^2}$, $0 < x < 1 - y$.

Ejercicio 9.3 *Utilizar la fórmula de transformaciones lineales de variables vectoriales para demostrar que si definimos la variable normal estándar como la que tiene función de densidad* $f(\mathbf{z}) = (2\pi)^{-p/2} \exp(-\mathbf{z}'\mathbf{z}/2)$ *y hacemos la transformación* $\mathbf{x} = \mathbf{m} + \mathbf{A}\mathbf{z}$ *se obtiene la expresión de la normal general.*

Ejercicio 9.4 *Obtener las distribuciones condicionadas en la normal bivariante de media cero y matriz de covarianzas* $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

Ejercicio 9.5 *Demostrar en el ejercicio anterior que si* $\rho > 1$ *tanto la matriz de covarianzas como la de correlación no son definidas positivas*

Ejercicio 9.6 *Comprobar las fórmulas (9.19) y (9.20) para las esperanzas y varianzas globales de las variables del ejercicio 4.3*

Ejercicio 9.7 *Sea* $\binom{x}{y}$ *un vector bidimensional de variables aleatorias normales incorreladas. Escribir la función de densidad conjunta del vector de variables* $a = \sum_{i=1}^{m} x_i^2$, $b \sum_{i=1}^{m} y_i^2$, $c = \sum_{i=1}^{m} x_i y_i$.

Ejercicio 9.8 *Calcular en el ejercicio anterior la densidad condicionada* $f(c|ab)$.

Ejercicio 9.9 *Demostrar que la distancia de Mahalanobis entre la variable multinomial y su media,* $(\mathbf{y} - n\mathbf{p})'\mathbf{Var}(\mathbf{y})^{-1}(\mathbf{y} - n\mathbf{p})$ *es la distancia ji-cuadrado* $\sum (y_i - np_i)^2/np_i$.

Ejercicio 9.10 *En la normal bivariante , demostrar que existe una matriz triangular (descomposición de Cholesky)* $L = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix}$ *tal que* $LL' = V$. *Encontrar los parámetros* l_{11}, l_{21}, l_{22} *como función de las varianzas y covarianzas de las variables. Interpretar el resultado como parámetros de las distribuciones marginales y condicionadas de las variables.*

Ejercicio 9.11 *Aplicar la descomposición de Cholesky del ejercicio anterior a la matriz de covarianzas* $\begin{bmatrix} 9 & 3 \\ 3 & 4 \end{bmatrix}$

Ejercicio 9.12 *Generar muestras de una distribución normal bivariante por el método siguiente: (1) generar un valor al azar de la distribución marginal de la primera variable; (2) generar un valor al azar de la distribución univariante de la segunda variable dada la primera. Aplicarlo para generar valores al azar de una variable aleatoria con vector de medias* $\boldsymbol{\mu} = (0,5)'$ *desviaciones típicas (2,3) y correlación 0,5.*

Ejercicio 9.13 *Demostrar que el método anterior es equivalente a generar dos variables aleatorias independientes de media cero y desviación típica unidad , $\mathbf{z} = (z_1, z_2)'$, y obtener los valores al azar de las variables mediante la transformación $\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z}$, donde $\mathbf{L}$ es al matriz triangular de la descomposición de Cholesky.*

Ejercicio 9.14 *Demostrar que si particionamos el vector de variables y la matriz de covarianzas como $\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}$, y llamamos L_{11}, L_{12}, L_{22} a las matrices correspondientes a la descomposición de Cholesky de esta matriz se verifica que $L_{11}^2 = V_{11}, L_{12} = V_{11}^{-1/2}V_{12}, L_{22}^2 = V_{22} - V_{21}V_{11}^{-1}V_{12}$ e interpretar estos resultados de acuerdo con la sección 5.3.1*

Ejercicio 9.15 *Demostrar que si $\mathbf{x}_1, ..., \mathbf{x}_h$ son vectores con medias $\boldsymbol{\mu}_i$ y matrices de covarianzas $\mathbf{V}_i$ la variable $\mathbf{y} = \sum_{I=1}^{h} c\mathbf{x}_i$ tiene media $\sum_{I=1}^{h} c\boldsymbol{\mu}_i$ y covarianza $\sum_{I=1}^{h} c^2\mathbf{V}_i$.*

Ejercicio 9.16 *Cuando aumenta la dimensión del vector de datos la maldición de la dimensión se manifiesta en que cada vez hay menos densidad en una región del espacio. Para ilustrar este problema, considere la normal estándar y calcule con tablas de la χ^2 la probabilidad de encontrar un valor en la esfera unidad definida por la relación $\mathbf{x}\prime\mathbf{x} \leq 1$, cuando la dimensión de $\mathbf{x}$, es $p = 2, 4, 8, 16$. ¿Qué pasará al aumentar p?*

Ejercicio 9.17 *Considere una variable normal $N_p(0, \mathbf{I})$, donde $p = 10$. Tomemos un valor al azar, x_0 y construyamos la dirección que une ese punto con el centro de la distribución. ¿Cuál es el valor esperado de la distancia entre ese punto y el centro de la distribución?. Supongamos que ahora tomamos una muestra de 100 valores de la variable aleatoria y los proyectamos sobre la dirección anterior. ¿Cuál será la distribución que observamos? ¿Cuál será la distancia esperada entre el centro de esos datos y el punto x_0 ?*

Ejercicio 9.18 *La función generatriz de momentos de una variable aleatoria multivariante viene dada por $\varphi(\mathbf{t}) = E(e^{\mathbf{t}'\mathbf{x}})$, donde $\mathbf{t}$ es un vector de parámetros. Comprobar que para una variable normal multivariante esta función es $\varphi(\mathbf{t}) = exp(\mathbf{t}'\boldsymbol{\mu} + \mathbf{t}'\mathbf{V}\mathbf{t})$.*

APÉNDICE 9.1 La distribución Wishart invertida

Si $\mathbf{W}$ es $W_p(m, \Sigma)$, la distribución de $\mathbf{U} = \mathbf{W}^{-1}$ se denomina distribución Wishart invertida, escribiremos $\mathbf{U} \sim IW_p(m, \Sigma)$. Su función de densidad es

$$f(\mathbf{U}) = C|\Sigma|^{-m/2}|\mathbf{U}|^{-(m+p+1)/2}\exp(-1/2\ tr\Sigma^{-1}\mathbf{U}^{-1})$$

y se verifica que

$$E[\mathbf{U}] = \frac{\Sigma^{-1}}{m-p-1}.$$

La distribución Wishart invertida es utilizada por muchos autores para la estimación bayesiana de matrices de covarianzas. Como es equivalente decir que si $\mathbf{U}$ es Wishart invertida, $\mathbf{U} \sim IW_p(m, \Sigma)$ y que $\mathbf{U}^{-1} = \mathbf{W}$ sigue una distribución Wishart, $\mathbf{U}^{-1} \sim W_p(m, \Sigma)$, en este libro para simplificar, hemos optado por no utilizarla y se incluye aquí únicamente como referencia para el lector que consulte otra bibliografía.

Capítulo 10

INFERENCIA CON DATOS MULTIVARIANTES

10.1 INTRODUCCIÓN

En este capítulo vamos a presentar una introducción a la inferencia en modelos multivariantes. Suponemos al lector familiarizado con los conceptos básicos de inferencia al nivel de Peña (2001). El objetivo de este capítulo es repasar los resultados de estimación y contrastes principales que serán necesarios en los temas posteriores. El lector puede encontrar en Anderson (1984), Mardia et al. (1979) o Seber (1983) presentaciones más completas de lo aquí expuesto.

Se estudia primero la estimación de los parámetros en modelos normales multivariantes por máxima verosimilitud. En segundo lugar se presenta el método de la razón de verosimilitudes, como procedimiento general para obtener contrastes con buenas propiedades en muestras grandes. Existen otros procedimientos para construir contrastes multivariantes que no revisaremos aquí, y que el lector puede encontrar en Anderson (1984). A continuación, se presenta un contraste sobre el valor del vector de medias en una población normal multivariante. Este contraste se generaliza para comprobar la igualdad de los vectores de medias de varias poblaciones normales multivariantes con la misma matriz de covarianzas, que es la herramienta principal del análisis de la varianza multivariante. Un caso particular de este contraste es el test de valores atípicos, que puede formularse como una prueba de que una observación proviene de una distribución con media distinta a la del resto de los datos. Finalmente, se presentan los contrastes de normalidad conjunta de los datos y transformaciones posibles para llevarlos a la normalidad.

10.2 Fundamentos de la Estimación Máximo Verosimil

El método de máxima verosimilitud, debido a Fisher, escoge como estimadores de los parámetros aquellos valores que hacen máxima la probabilidad de que el modelo a estimar genere la muestra observada. Para precisar esta idea, supongamos que se dispone de una muestra

aleatoria simple de n elementos de una variable aleatoria $p-$dimensional, $\mathbf{x}$, con función de densidad $f(\mathbf{x} \mid \boldsymbol{\theta})$, donde $\boldsymbol{\theta} = (\theta_1, ..., \theta_r)'$ es un vector de parámetros que supondremos tiene dimensión $r \leq pn$. Llamando $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, a los datos muestrales, la función de densidad conjunta de la muestra será, por la independencia de las observaciones:

$$f(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i \mid \boldsymbol{\theta}).$$

Cuando el parámetro $\boldsymbol{\theta}$ es conocido, esta función determina la probabilidad de aparición de cada muestra. En el problema de estimación se dispone de la muestra, pero $\boldsymbol{\theta}$ es desconocido. Considerando en la expresión de la densidad conjunta a $\boldsymbol{\theta}$ como una variable y particularizando esta función para los datos observados, se obtiene una función que llamaremos *función de verosimilitud*, $\ell(\boldsymbol{\theta}|\mathbf{X})$, o $\ell(\boldsymbol{\theta})$:

$$\ell(\boldsymbol{\theta}|\mathbf{X}) = \ell(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i \mid \boldsymbol{\theta}) \qquad \mathbf{X}\ fijo; \boldsymbol{\theta}\ variable \tag{10.1}$$

El estimador de máxima verosimilitud, o estimador MV, es el valor de $\boldsymbol{\theta}$ que hace máxima la probabilidad de aparición de los valores muestrales efectivamente observados y se obtiene calculando el valor máximo de la función $\ell(\boldsymbol{\theta})$. Suponiendo que esta función es diferenciable y que su máximo no ocurre en un extremo de su dominio de definición, el máximo se obtendrá resolviendo el sistema de ecuaciones:

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} &= 0 \\ &\vdots \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_r} &= 0 \end{aligned}$$

El vector $\widehat{\boldsymbol{\theta}}$ que satisface este sistema de ecuaciones corresponderá a un máximo si la matriz hessiana de segundas derivadas H, evaluada en $\widehat{\boldsymbol{\theta}}$, es definida negativa:

$$H(\widehat{\boldsymbol{\theta}}) = \left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right)_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}} \quad \text{definida negativa.}$$

En ese caso $\widehat{\boldsymbol{\theta}}$ es el estimador de máxima verosimilitud o estimador MV de $\boldsymbol{\theta}$. En la práctica suele ser más cómodo obtener el máximo del logaritmo de la función de verosimilitud:

$$L(\boldsymbol{\theta}) = \ln \ell(\boldsymbol{\theta}) \tag{10.2}$$

que llamaremos *función soporte*. Como el logaritmo es una transformación monótona, ambas funciones tienen el mismo máximo, pero trabajar con el soporte tiene dos ventajas principales. En primer lugar pasamos del producto de densidades (10.1) a la suma de sus logaritmos y la expresión resultante suele ser más simple que la verosimilitud, con lo que resulta más cómodo

obtener el máximo. En segundo lugar, al tomar logaritmos las constantes multiplicativas de la función de densidad, que son irrelevante para el máximo, se hacen aditivas y desaparecen al derivar, con lo que la derivada del soporte tiene siempre la misma expresión y no depende de constantes arbitrarias. En tercer lugar, el doble de la función soporte cambiada de signo proporciona un método general para juzgar el ajuste de un modelo a los datos que se denomina *desviación*:

$$D(\boldsymbol{\theta}) = -\mathbf{2}L\left(\boldsymbol{\theta}\right)$$

y la desviación $D(\boldsymbol{\theta})$ mide la discrepancia entre el modelo y los datos. Cuanto mayor sea el soporte, $L\left(\boldsymbol{\theta}\right)$, mayor es la concordancia entre el valor del parámetro y los datos y menor la desviación. La desviación aparecerá de manera natural en el contraste de hipótesis y es una medida global de ajuste de un modelo a los datos.

Para distribuciones cuyo rango de valores posibles es conocido a priori y no depende de ningún parámetro, puede demostrarse (veáse por ejemplo Casella y Berger, 1990) que, en condiciones muy generales respecto al modelo de distribución de probabilidad, el método de máxima verosimilitud (MV) proporciona estimadores que son:

1. Asintóticamente centrados.
2. Con distribución asintóticamente normal.
3. Asintóticamente de varianza mínima (eficientes).
4. Si existe un estadístico suficiente para el parámetro, el estimador MV es suficiente.
5. Invariantes en el sentido siguiente: si $\widehat{\boldsymbol{\theta}}$ es el estimador MV de $\boldsymbol{\theta}$, y $g(\boldsymbol{\theta})$ es una función cualquiera del vector de parámetros, entonces $g(\widehat{\boldsymbol{\theta}})$ es el estimador MV de $g(\boldsymbol{\theta})$.

10.3 Estimación de los parámetros de variables normales p-dimensionales.

Sea $\mathbf{x}_1, ..., \mathbf{x}_n$ una muestra aleatoria simple donde $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \mathbf{V})$. Vamos a obtener los estimadores MV de los parámetros desconocidos $\boldsymbol{\mu}$ y $\mathbf{V}$. El primer paso es construir la función de densidad conjunta de las observaciones, que es, utilizando la expresión de la normal multivariante estudiada en el capítulo 8:

$$f(\mathbf{X} \mid \boldsymbol{\mu}, \mathbf{V}) = \prod_{i=1}^{n} |\mathbf{V}|^{-1/2}(2\pi)^{-p/2} \exp\left\{-(1/2)(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

y la función soporte será, despreciando las constantes:

$$L(\boldsymbol{\mu}, \mathbf{V}|\mathbf{X}) = -\frac{n}{2}\log|\mathbf{V}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}).$$

Observemos que la función soporte así escrita es siempre negativa, ya que tanto el determinante como la forma cuadrática son positivos por ser definida positiva la matriz $\mathbf{V}$. Ésta función nos indica el apoyo o soporte que reciben los posibles valores de los parámetros dados los valores muestrales observados. Cuanto mayor sea esta función (menos negativa) para unos valores de los parámetros, mayor será la concordancia entre estos parámetros y los datos. Vamos a expresar esta función de una forma más conveniente. Llamando $\overline{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{x}_i/n$ al vector de medias muestral y escribiendo $(\mathbf{x}_i - \boldsymbol{\mu}) = (\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu})$ y desarrollando la forma cuadrática

$$\sum_{i=1}^{n}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \sum_{i=1}^{n}(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) + n(\bar{\mathbf{x}}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})$$

ya que $\sum_{i=1}^{n}(\mathbf{x}-\bar{\mathbf{x}}) = 0$. Concentrandonos en el primer término de esta descomposición, como un escalar es igual a su traza:

$$tr\left(\sum_{i=1}^{n}(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}-\bar{\mathbf{x}})\right) = \sum_{i=1}^{n} tr\left[(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{V}^{-1}(\mathbf{x}-\bar{\mathbf{x}})\right] =$$

$$= \sum_{i=1}^{n} tr\left[\mathbf{V}^{-1}(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})'\right] = tr\left(\mathbf{V}^{-1}\sum_{i=1}^{n}(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})'\right),$$

y llamando:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})', \tag{10.3}$$

a la matriz de covarianzas muestral, y sustituyendo en la función soporte:

$$L(\boldsymbol{\mu},\mathbf{V}|\mathbf{x}) = -\frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}tr\mathbf{V}^{-1}\mathbf{S} - \frac{n}{2}(\bar{\mathbf{x}}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \tag{10.4}$$

Esta es la expresión que utilizaremos para el soporte de los parámetros en muestras de una normal multivariante. Observemos que esta función sólo depende de la muestra a través de los valores $\bar{\mathbf{x}}$ y $\mathbf{S}$, que serán, por tanto, estimadores suficientes de $\boldsymbol{\mu}$ y $\mathbf{V}$. Todas la muestras que proporcionen los mismos valores de $\bar{\mathbf{x}}$ y $\mathbf{S}$ darán lugar a las mismas inferencias respecto a los parámetros.

Para obtener el estimador del vector de medias en la población, utilizamos que, por ser $\mathbf{V}^{-1}$ definida positiva, $(\bar{\mathbf{x}}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \geq 0$. Como este término aparece con signo menos, el valor de $\boldsymbol{\mu}$ que maximiza la función soporte es aquel que hace este término lo menor posible, y se hará cero tomando:

$$\widehat{\boldsymbol{\mu}} = \bar{\mathbf{x}} \tag{10.5}$$

por lo que concluimos que $\bar{\mathbf{x}}$ es el estimador máximo verosímil de $\boldsymbol{\mu}$. Sustituyendo este estimador en la función soporte este término desaparece. Para obtener el máximo de la función respecto a $\mathbf{V}$, sumaremos la constante $\frac{n}{2}\log|\mathbf{S}|$, y escribiremos el soporte como:

$$L(\mathbf{V}|\mathbf{X}) = \frac{n}{2}\log|\mathbf{V}^{-1}\mathbf{S}| - \frac{n}{2}tr\mathbf{V}^{-1}\mathbf{S} \tag{10.6}$$

Esta expresión es útil porque el valor de la verosimilitud escrita de esta forma no depende de las unidades de medida de las variables. También es fácil comprobar (veáse ejercicio 10.1) que el valor de la verosimilitud es invariante ante transformaciones lineales no singulares de las variables. Llamemos λ_i a los valores propios de la matriz $\mathbf{V}^{-1}\mathbf{S}$, entonces:

$$L(\mathbf{V}|\mathbf{X}) = \frac{n}{2}\sum \log \lambda_i - \frac{n}{2}\sum \lambda_i = \frac{n}{2}\sum(\log \lambda_i - \lambda_i).$$

Esta expresión indica que la verosimilitud es una suma de funciones del tipo $\log x - x$. Derivando respecto a x es inmediato que una función de este tipo tiene un máximo para $x = 1$. Por tanto, $L(\mathbf{V}|\mathbf{X})$ será máxima si todos los valores propios de $\mathbf{V}^{-1}\mathbf{S}$ son iguales a la unidad, lo que implica que $\mathbf{V^{-1}S} = \mathbf{I}$. Esto se consigue tomando como estimador de máxima verosimilitud de $\mathbf{V}$:

$$\widehat{\mathbf{V}} = \mathbf{S} \tag{10.7}$$

Los estimadores MV de $\boldsymbol{\mu}$ y $\mathbf{V}$ son pues $\bar{\mathbf{x}}$ y $\mathbf{S}$. Se demuestra, como en el caso univariante, que $\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, 1/n\mathbf{V})$. Además $n\mathbf{S}$ se distribuye como la distribución de Wishart, $W_p(n-1, \mathbf{V})$. El estimador $\mathbf{S}$ es sesgado, pero $\frac{n}{n-1}\mathbf{S}$ es un estimador centrado de $\mathbf{V}$. Estos estimadores tienen las buenas propiedades asintóticas de los estimadores de máxima verosimilitud: consistencia, eficiencia y normalidad asintótica. En el ejercicio 10.2 se presenta una deducción alternativa, más clásica, de estos estimadores derivando la función soporte.

10.4 El método de la razón de verosimilitudes

En esta sección repasamos la metodología general para construir contrastes utilizando la razón de verosimilitides y la aplicaremos al caso de poblaciones normales. Con frecuencia se desea comprobar si una muestra dada puede provenir de una distribución con ciertos parámetros conocidos. Por ejemplo, en el control de calidad de ciertos procesos se toman muestras de elementos, se mide una variable multivariante y se desea contrastar si el proceso está en estado de control, lo que supone que las muestras provienen de una población normal con ciertos valores de los parámetros. En otros casos, interesa comprobar si varias muestras multivariantes provienen o no de la misma población. Por ejemplo, queremos comprobar si ciertos mercados son igualmente rentables o si varios medicamentos producen efectos similares. Finalmente, si hemos basado nuestra inferencia en la hipótesis de normalidad conviene realizar un contraste para ver si esta hipótesis no es rechazada por los datos observados.

Para realizar contrastes de parámetros vectoriales podemos aplicar la teoría del contraste de verosimilitudes. Esta teoría proporciona pruebas estadísticas que, como veremos, tienen ciertas propiedades óptimas para tamaños muestrales grandes. Dado un parámetro vectorial, $\boldsymbol{\theta}$, $p-$dimensional, que toma valores en Ω (donde Ω es un subconjunto de $\Re^p$), suponemos que se desea contrastar la hipótesis:

$$H_0 : \boldsymbol{\theta} \in \Omega_0,$$

que establece que $\boldsymbol{\theta}$ está contenido en una región Ω_0 del espacio paramétrico, frente a una hipótesis alternativa:

$$H_1 : \boldsymbol{\theta} \in \Omega - \Omega_0,$$

que supone que $\boldsymbol{\theta}$ no está restringida a la región Ω_0. Para comparar estas hipótesis, analizaremos su capacidad de prever los datos observados, y, para ello, compararemos las probabilidades de obtenerlos bajo ambas hipótesis. Calcular estas probabilidades requiere el valor del vector de parámetros, que es desconocido. El método de razón de verosimilitudes resuelve este problema tomando el valor que hace más probable obtener la muestra observada y que es compatible con la hipótesis. En concreto:

1. La máxima probabilidad de obtener la muestra observada bajo H_0 se obtiene como sigue. Si Ω_0 determina un valor único para los parámetros, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, entonces se calcula la probabilidad de los datos supuesto $\boldsymbol{\theta}_0$. Si Ω_0 permite muchos valores, elegiremos entre ellos el valor del parámetro que haga máxima la probabilidad de obtener la muestra. Como la probabilidad de la muestra observada es proporcional a la distribución conjunta de las observaciones, sustituyendo en esta función los datos disponibles resulta la función de verosimilitud. Calculando el máximo de esta función en Ω_0, se obtiene el máximo valor de la verosimilitud compatible con H_0, que representaremos por $f(H_0)$.

2. La máxima probabilidad de obtener la muestra observada bajo H_1 se calcula obteniendo el máximo absoluto de la función sobre todo el espacio paramétrico. Estrictamente debería calcularse en el conjunto $\Omega - \Omega_0$, pero es más simple hacerlo sobre todo el espacio, ya que en general se obtiene el mismo resultado. La razón es que, habitualmente, H_0 impone restricciones en el espacio paramétrico mientras que H_1 supone que estas restricciones no existen. Particularizando la función de verosimilitud en su máximo, que corresponde al estimador MV de los parámetros, se obtiene una cantidad que representaremos como $f(H_1)$.

A continuación compararemos $f(H_0)$ y $f(H_1)$. Para eliminar las constantes y hacer la comparación invariante ante cambios de escala de las variables, construimos su cociente, que llamaremos razón de verosimilitudes (RV) :

$$\boxed{RV = \frac{f(H_0)}{f(H_1)}} \tag{10.8}$$

Por construcción $RV \leq 1$ y rechazaremos H_0 cuando RV sea suficientemente pequeño. La región de rechazo de H_0 vendrá, en consecuencia, definida por:

$$RV \leq a,$$

donde a se determinará imponiendo que el nivel de significación del test sea α. Para calcular el valor a es necesario conocer la distribución de RV cuando H_0 es cierta, lo que suele ser difícil en la práctica. Sin embargo, cuando el tamaño muestral es grande, el doble de la diferencia de soportes entre la alternativa y la nula, cuando H_0 es cierta, definida por:

$$\lambda = -2 \ \ln RV = 2\left(L(H_1) - L(H_0)\right),$$

donde $L(H_i) = \log f(H_i)$, $i = 0, 1$. se distribuye asintóticamente como una χ^2 con un número de grados de libertad igual a la diferencia de dimensión entre los espacios Ω, y Ω_0. Intuitivamente rechazamos H_0 cuando el soporte de los datos para H_1 es significativamente mayor

que para H_0. La diferencia se juzga, para muestras grandes, con la distribución χ^2. Utilizando la definición de la desviación, este contraste puede interpretarse como la diferencia entre las desviaciónes para H_0 y para H_1 :

$$\lambda = D(H_0) - D(H_1)$$

Es frecuente que la dimensión de Ω sea p y la dimensión de Ω_0 sea $p-r$, siendo r el número de restricciones lineales sobre el vector de parámetros. Entonces, el número de grados de libertad de la diferencia de soportes, λ, es:

$$g = gl(\lambda) \ = \dim(\Omega) - \dim(\Omega_0) = p - (p-r) = r$$

igual al número de restricciones lineales impuestas por H_0.

10.5 Contraste sobre la media de una población normal

Consideremos una muestra $(\mathbf{x}_1, ..., \mathbf{x}_n)$ de una población $N_p(\boldsymbol{\mu}, \mathbf{V})$. Se desea realizar el contraste de la hipótesis:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0, \mathbf{V} = cualquiera$$

frente a la alternativa:

$$H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \mathbf{V} = cualquiera.$$

Para construir un contraste de razón de verosimilitudes, calcularemos el máximo de la función de verosimilitud bajo H_0 y bajo H_1. La función soporte es:

$$L(\boldsymbol{\mu}, \mathbf{V}|\mathbf{X}) = -\frac{n}{2}\log|\mathbf{V}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}).$$

Se requiere obtener los estimadores MV de $\boldsymbol{\mu}$ y $\mathbf{V}$ bajo H_0 y bajo H_1. Por la sección 10.2 sabemos que, bajo H_1, estos estimadores son $\bar{\mathbf{x}}$ y $\mathbf{S}$, y sustituyendo en (10.4) tenemos que el soporte para H_1 es:

$$L(H_1) = -\frac{n}{2}\log|\mathbf{S}| - \frac{np}{2}$$

Bajo H_0 el estimador de $\boldsymbol{\mu}$ es directamente $\boldsymbol{\mu}_0$, y operando en la forma cuadratica como vimos en la sección 10.2.2 (tomando trazas y utilizando las propiedades lineales de la traza) podemos escribir esta función como:

$$L(\mathbf{V}|\mathbf{X}) = -\frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}tr\mathbf{V}^{-1}\mathbf{S}_0 \tag{10.9}$$

donde

$$\mathbf{S}_0 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu}_0)(\mathbf{x}_i-\boldsymbol{\mu}_0)'. \tag{10.10}$$

Si sumamos en la expresión (10.9) la constante $\frac{n}{2}\log|\mathbf{S}_0|$ obtenemos una expresión análoga a (10.6), con lo que concluimos que $\mathbf{S}_0$ es el estimador MV de $\mathbf{V}$ bajo H_0. Sustituyendo $\mathbf{V}$ por $\mathbf{S}_0$ en (10.9) el soporte para H_0 será

$$L(H_0) = -\frac{n}{2}\log|\mathbf{S}_0| - \frac{np}{2}$$

y la diferencia de soportes será

$$\lambda = 2(L(H_1) - L(H_0)) = n\log\frac{|\mathbf{S}_0|}{|\mathbf{S}|} \tag{10.11}$$

y rechazaremos H_0 cuando el soporte para H_1 sea significativamente mayor que para H_0. Esta condición equivale a que la varianza generalizada bajo H_0, ($|\mathbf{S}_0|$) sea significativamente mayor que bajo H_1. La distribución de λ es una χ^2, con grados de libertad igual a la diferencia de las dimensiones del espacio en que se mueven los parámetros bajo ambas hipótesis. La dimensión del espacio paramétrico bajo H_0 es $p + p(p-1)/2 = p(p+1)/2$, el número de términos distintos en $\mathbf{V}$, y la dimensión del espacio parámetro bajo H_1 es $p + p(p+1)/2$. La diferencia es p que serán los grados de libertad del estadístico χ^2.

En este caso, podemos obtener la distribución exacta del ratio de verosimilitudes, no siendo necesaria la distribución asintótica. Se demuestra en el apéndice 10.2 que:

$$\frac{|\mathbf{S}_0|}{|\mathbf{S}|} = 1 + \frac{T^2}{n-1} \tag{10.12}$$

donde el estadístico

$$T^2 = (n-1)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0),$$

sigue la distribución T^2 de Hotelling con p y $n-1$ grados de libertad. Utilizando la relación entre el estadístico T^2 y la distribución F, podemos calcular los percentiles de T^2. Como la diferencia de soportes es una función monótona de T^2, podemos utilizar directamente este estadístico en lugar de la razón de verosimilitudes, y rechazaremos H_0 cuando T^2 sea suficientemente grande. Observemos que de (10.11) y (10.12) podemos escribir

$$\lambda = n\log(1 + \frac{T^2}{n-1})$$

que es consistente con la distribución asintótica, ya que, para n grande, $\log(1 + a/n) \approx a/n$, y entonces $\lambda \approx T^2$, que sabemos tiene una distribución asintótica χ^2_p.

Ejemplo 10.1 *Un proceso industrial fabrica elementos cuyas características de calidad se miden por un vector de tres variables,* $\mathbf{x}$. *Cuando el proceso está en estado de control, los valores medios de las variables deben ser* $(12, 4, 2)$. *Para comprobar si el proceso funciona adecuadamente, se toma una muestra de 20 elementos y se miden las tres características. La media muestral es*

$$\bar{\mathbf{x}} = (11.5,\ 4.3,\ 1.2)$$

y la matriz de covarianzas entre estas tres variables es

$$S = \begin{bmatrix} 10 & 4 & -5 \\ 4 & 12 & -3 \\ -5 & -3 & 4 \end{bmatrix}$$

(Los valores numéricos se han simplificado para facilitar los cálculos) Observemos que si miramos cada variable aisladamente como

$$t = (\overline{x} - \mu)\sqrt{n}/\widehat{s}$$

es una t de Student con $n-1$ grados de libertad, obtendríamos unos valores de las t para cada variable de $t_1 = (11.5-12)\sqrt{20}/\sqrt{20 \times 10/19} = -.68$; $t_2 = (4.3-4)\sqrt{20}/\sqrt{20 \times 12/19} = .88$; y $t_3 = (1.2-2)\sqrt{20}/\sqrt{20 \times 4/19} = .85$. Aparentemente, mirando cada variable separadamente no hay diferencias significativas entre las medias muestrales y las del proceso bajo control y concluiríamos que no hay evidencia de que el proceso esté fuera de control. Si calculamos ahora el estadístico de Hotelling

$$T^2 = 19(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) = 14.52$$

Para juzgar el tamaño de esta discrepancia lo llevamos a la distribución F

$$F_{3,17} = ((20-3)/3)(T^2/19) = 4.33$$

y como el valor $F_{3,17}(.001) = 3.4$, rechazamos sin ninguna duda que el proceso esta en estado de control.

Para entender la razones de esta discrepancia entre el contraste multivariante y los univariantes, observemos que el contraste multivariante tiene en cuenta las correlaciones entre las discrepancias individuales. La matriz de correlaciones de los datos muestrales obtenida a partir de la matriz de covarianzas es

$$\mathbf{R} = \begin{bmatrix} 1 & .37 & -0.79 \\ .37 & 1 & -0.43 \\ -0.79 & -0.43 & 1 \end{bmatrix}$$

la correlación entre la primera variable y la tercera es negativa. Esto quiere decir que si observamos un valor por debajo de la media en la primera variable, esperamos que aparezca un valor por encima de la media en la tercera. En la muestra ocurre lo contrario, y esto contribuye a sugerir un desplazamiento de la media del proceso.

10.6 Contrastes sobre la matriz de varianzas de una población normal

El contraste de la razón de verosimilitudes se aplica para hacer contrastes de matrices de varianzas de forma similar a la estudiada para vectores de medias en la sección anterior. Vamos a ver cuatro contrastes sobre la matriz de covarianzas de variables normales. En el

primer caso la hipótesis nula es que esta matriz toma un valor fijo dado. En el segundo, que la matriz es diagonal y las variables están incorreladas. En el tercero las variables además tienen la misma varianza, es el contraste de esfericidad donde suponemos que la matriz de covarianzas es $\sigma^2\mathbf{I}$. En el cuarto caso suponemos una esfericidad parcial: la matriz de covarianzas puede descomponerse como una matriz de rango $m < p$ más $\sigma^2\mathbf{I}$. Si $m = 0$ este contraste se reduce al de esfericidad.

10.6.1 Contraste de un valor particular

Supongamos que se desea realizar el contraste de la hipótesis:

$$H_0 : \mathbf{V} = \mathbf{V}_0, \qquad \boldsymbol{\mu} \text{ cualquiera}$$

frente a la alternativa:

$$H_1 : \boldsymbol{\mu}, \text{ y } \mathbf{V} = \text{cualquiera.}$$

Para construir un contraste de razón de verosimilitudes, calcularemos el máximo de la función de soporte bajo H_0 y bajo H_1. Utilizando la expresión del soporte:

$$L(\boldsymbol{\mu}, \mathbf{V}|\mathbf{x}) = -\frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}tr\mathbf{V}^{-1}\mathbf{S} - \frac{n}{2}(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$$

Bajo H_0, el valor de $\mathbf{V}$ queda especificado, $\mathbf{V}_0$, y $\boldsymbol{\mu}$ se estimará mediante $\bar{\mathbf{x}}$, con lo que :

$$L(H_0) = -\frac{n}{2}\log|\mathbf{V}_0| - \frac{n}{2}tr\mathbf{V}_0^{-1}\mathbf{S}$$

mientras que bajo H_1, los estimadores son $\bar{\mathbf{x}}$ y $\mathbf{S}$, con lo que, como vimos en la sección anterior:

$$L(H_1) = -\frac{n}{2}\log|\mathbf{S}| - \frac{np}{2}$$

y la diferencia de soportes será

$$\lambda = 2(L(H_1) - L(H_0)) = n\log\frac{|\mathbf{V}_0|}{|\mathbf{S}|} + ntr\mathbf{V}_0^{-1}\mathbf{S} - np \tag{10.13}$$

Vemos que el contraste consiste en comparar $\mathbf{V}_0$, el valor teórico y $\mathbf{S}$, el estimado con la métrica del determinante y con la de la traza. La distribución del estadístico λ es una χ^2, con grados de libertad igual a la diferencia de las dimensiones del espacio en que se mueven los parámetros bajo ambas hipótesis que es $p(p+1)/2$, el número de términos distintos en $\mathbf{V}$.

En particular este test sirve para contrastar si $\mathbf{V}_0 = \mathbf{I}$. Entonces el estadístico (10.13) se reduce a

$$\lambda = -n\log|\mathbf{S}| + ntr\mathbf{S} - np.$$

10.6.2 Contraste de independencia

Otro contraste de interés es el de independencia, donde suponemos que la matriz $\mathbf{V}_0$ es diagonal. Es decir:

$$H_0 : \mathbf{V} = \text{diagonal} \quad \boldsymbol{\mu} \text{ cualquiera}$$

frente a la alternativa:

$$H_1 : \boldsymbol{\mu}, \text{ y } \mathbf{V} = \text{cualquiera.}$$

Entonces la estimación máximo verosímil de $\mathbf{V}_0$ es $\widehat{\mathbf{V}_0} = \text{diag}(\mathbf{S})$, donde $\text{diag}(\mathbf{S})$ es una matriz diagonal con términos s_{ii} iguales a los de $\mathbf{S}$, y el estadístico (10.13) se reduce a

$$\lambda = n \log \frac{\prod s_{ii}}{|\mathbf{S}|} + ntr\widehat{\mathbf{V}}_0^{-1}\mathbf{S} - np$$

y como $tr\widehat{\mathbf{V}}_0^{-1}\mathbf{S} = tr\widehat{\mathbf{V}}_0^{-1/2}\mathbf{S}\widehat{\mathbf{V}}_0^{-1/2} = tr\mathbf{R} = p$, el contraste se reduce a:

$$\lambda = -n \log |\mathbf{R}| \tag{10.14}$$

que suele escribirse en términos de los valores propios de $\mathbf{R}$, llamando λ_i a estos valores propios una forma equivalente del contraste es

$$\lambda = -n \sum_{i=1}^{p} \log \lambda_i$$

y su distribución asintótica será una χ^2, con grados de libertad igual $p(p+1)/2 - p = p(p-1)/2$.

10.6.3 Contraste de esfericidad

Un caso particular importante del contraste anterior es suponer que todas las variables tienen la misma varianza y están incorreladas. En este caso no ganamos nada por analizarlas conjuntamente, ya que no hay información común. Este contraste equivale a suponer que la matriz $\mathbf{V}_0$ es escalar, es decir $\mathbf{V} = \sigma^2\mathbf{I}$, y se denomina de esfericidad, ya que entonces la distribución de las variables tiene curvas de nivel que son esferas: hay una total simetría en todas las direcciones en el espacio. El contraste es

$$H_0 : \mathbf{V} = \sigma^2\mathbf{I}, \quad \boldsymbol{\mu} \text{ cualquiera}$$

frente a:

$$H_1 : \boldsymbol{\mu}, \text{ y } \mathbf{V} = \text{cualquiera}$$

Sustituyendo $\mathbf{V}_0 = \sigma^2\mathbf{I}$ en (10.13), la función soporte bajo H_0 es

$$L(H_0) = -\frac{np}{2} \log \sigma^2 - \frac{n}{2\sigma^2} tr\mathbf{S}$$

y derivando respecto a σ^2 es inmediato comprobar que el estimador MV es $\widehat{\sigma}^2 = tr\mathbf{S}/p$, .el promedio de las varianzas. La función soporte $L(H_1)$ es la misma que en el contraste anterior y la diferencia de soportes es

$$\lambda = n \log \frac{\widehat{\sigma}^{2p}}{|\mathbf{S}|} + ntr\mathbf{S}/\widehat{\sigma}^2 - np \tag{10.15}$$

y sustituyendo $\widehat{\sigma}^2 = tr\mathbf{S}/p$ el contraste se reduce a:

$$\lambda = np \log \widehat{\sigma}^2 - n \log |\mathbf{S}|$$

y se distribuirá asintoticamente como una χ^2 con $p(p+1)/2 - 1 = (p+2)(p-1)/2$ grados de libertad.

10.6.4 (*)Contraste de esfericidad parcial

El cuarto contraste que estudiaremos se denomina de esfericidad parcial porque supone que la matriz de covarianzas tiene dependencias en un espacio de dimensión m, pero en el espacio complementario de dimensión $p-m$ se da la situación de esfericidad. Esto supone que toda la estructura de dependencias entre las variables puede explicarse en función de m variables, como veremos al estudiar el modelo factorial. Observemos que no tiene sentido contrastar que una matriz cuadrada de orden p tiene rango $m < p$, porque, en este caso, la matriz debe tener exactamente $p - m$ valores propios nulos. Si es así, lo comprobaremos al calcular sus valores propios, ya que si se da esta condición en la población tiene que darse también en todas las muestras. Sin embargo, sí tiene sentido contrastar que la matriz tiene m valores propios relativamente grandes, que coresponden a m direcciones informativas, y $p-m$ valores propios pequeños e iguales, que corresponden a las no informativas. Esta es la esfericidad parcial. El contraste será:

$$H_0 : \mathbf{V} = \mathbf{B} + \sigma^2\mathbf{I}, \qquad \boldsymbol{\mu} \text{ cualquiera y rango}(\mathbf{B}) = m$$

frente a:

$$H_1 : \boldsymbol{\mu}, \text{ y } \mathbf{V} = \text{cualquiera}$$

Puede demostrarse utilizando los mismos principios (véase Anderson, 1963) que, llamando λ_i a los valores propios de $\mathbf{S}$, el contraste es:

$$\lambda = -n \sum_{i=m+1}^{p} \log \lambda_i + n(p-m) \log \frac{\sum_{j=m+1}^{p} \lambda_j}{p-m} \tag{10.16}$$

y sigue asintóticamente una distribución χ^2 con $(p-m+2)(p-m-1)/2$ grados de libertad. Observemos que si las variables están estandarizadas y $m = 0$, este contraste se reduce a (10.15). También si las variables estan estandarizadas $\sum_{j=1}^{p} \lambda_j = p$, el segundo término se anula y este contraste se reduce (10.14). Concluimos que cuando $m = 0$ este contraste coincide con el general de esfericidad presentado en la sección anterior.

10.6.5 Ajustes en la distribución

La aproximación de la distribución del estadístico λ a la χ^2 cuando el tamaño muestral no es muy grande puede mejorarse introduciendo factores de corrección. Box (1949) y Bartlett (1954) han demostrado que las aproximaciones mejoran si sutituimos en los estadísticos anteriores n por n_c donde n_c es menor que n y dependen de p y del contraste. Por ejemplo, Box (1949) demostró que el contraste de independencia mejora si sutituimos n por $n_c = n-(2p+11)/2$. Estas correciones pueden ser importantes si el tamaño muestral es pequeño, p es grande y el estadístico obtenido esta cerca del valor crítico, pero no van a ser importantes si p/n es pequeño y el estadístico resultante es claramente conluyente en cualquiera de las direcciones. El lector interesado puede acudir a Muirhead (1982).

Ejemplo 10.2 *Contrastar si podemos admitir que la matriz de covarianzas de las medidas de calidad del ejercicio 10.1 es de la forma $\sigma^2 I$. Si no es así contrastar si las variables aunque tengan distinta varianza son independientes.*

La estimación de $\widehat{\sigma}^2$ bajo la nula es $tr\mathbf{S}/p = (10+12+4)/3 = 8,67$. Por otro lado se comprueba que $|\mathbf{S}| = 146$. Entonces

$$\lambda = 60\log 8,67 - 20\log 146 = 29.92$$

que debe compararse con una χ^2 con $(3+2)(3-1)/2 = 5$ grados de libertad, y el valor obtenido es claramente significativo, por lo que rechazamos que las variables tengan la misma varianza y estén incorreladas.

Para realizar el contraste de independencia transformemos las variables dividiendo cada una de ellas por su varianza. Es decir, pasamos a nuevas variables $z_1 = x_1/\sqrt{10}$, $z_2 = x_2/\sqrt{12}, z_3 = x_3/\sqrt{4}$, que tendrán matriz de covarianzas, llamando $\mathbf{D}$ a la matriz diagonal con elementos $(1/\sqrt{10}, 1/\sqrt{12}, 1/\sqrt{4})$, tendremos:

$$\mathbf{V}_z = \mathbf{D}V_x\mathbf{D}' = \begin{bmatrix} 1 & 0.3651 & -0.7906 \\ 0.3651 & 1 & -0.4330 \\ -0.7906 & -0.4330 & 1 \end{bmatrix} = R_x$$

y el contraste ahora es

$$\lambda = -20\log 0,304 = 23.8$$

que debe compararse ahora con χ^2 con 3 grados de libertad, con lo que se rechaza sin duda la hipótesis de independencia.

10.7 Contraste de igualdad de varias medias: el Análisis de la Varianza Multivariante

Supongamos que hemos observado una muestra de tamaño n de una variable p dimensional que puede estratificarse en G clases o grupos, de manera que existen n_1 observaciones del grupo $1, ..., n_G$ del grupo G. Un problema importante es contrastar que las medias de las

G clases o grupos son iguales. Vamos a resolverlo aplicando el contraste de la razón de verosimilitudes. La hipótesis a contrastar es:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = ... = \boldsymbol{\mu}_G = \boldsymbol{\mu};$$

donde, además, $\mathbf{V}$ es definida positiva, e idéntica en los grupos. La hipótesis alternativa es:

$$H_1 : \text{no todas las } \boldsymbol{\mu}_i \text{ son iguales;}$$

con las mismas condiciones para $\mathbf{V}$.

La función de verosimilitud bajo H_0 de una muestra normal homogénea se ha calculado en la sección 10.2 y sabemos que su máximo se alcanza para $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$ y $\widehat{\mathbf{V}} = \mathbf{S}$. Sustituyendo estas estimaciones en la función soporte tenemos que

$$L(H_0) = -\frac{n}{2}\log|\mathbf{S}| - \frac{np}{2}. \tag{10.17}$$

Bajo H_1, los n vectores de observaciones se subdividen en n_1 del grupo $1, ..., n_G$ del grupo G. La función de verosimilitud bajo H_1 será:

$$f(\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_p, V|X) = |\mathbf{V}|^{-n/2}(2\pi)^{-np/2}\exp\left\{-\frac{1}{2}\sum_{g=1}^{G}\sum_{h=1}^{n_g}(\mathbf{x}_{hg} - \boldsymbol{\mu}_g)'\mathbf{V}^{-1}(\mathbf{x}_{hg} - \boldsymbol{\mu}_g)\right\},$$

donde $\mathbf{x}_{hg}$ es el h vector de variables del grupo g, y $\boldsymbol{\mu}_g$ su media. La maximización de esta función en el espacio paramétrico definido por H_1 se realiza por el procedimiento estudiado en 10.2. La estimación de la media de cada grupo será la media muestral, $\widehat{\boldsymbol{\mu}}_g = \overline{\mathbf{x}}_g$, y la estimación de la matriz de covarianzas común se obtiene utilizando que:

$$\sum_{g=1}^{G}\sum_{h=1}^{n_g}(\mathbf{x}_{hg} - \overline{\mathbf{x}}_g)'\mathbf{V}^{-1}(\mathbf{x}_{hg} - \overline{\mathbf{x}}_g) = tr\left(\sum_{g=1}^{G}\sum_{h=1}^{n_g}(\mathbf{x}_{hg} - \overline{\mathbf{x}}_g)'\mathbf{V}^{-1}(\mathbf{x}_{hg} - \overline{\mathbf{x}}_g)\right)$$

$$\sum_{g=1}^{G}\sum_{h=1}^{n_g} tr\left(\mathbf{V}^{-1}(\mathbf{x}_{hg} - \bar{\mathbf{x}}_g)(\mathbf{x}_{hg} - \bar{\mathbf{x}}_g)'\right) = tr\ \left(\mathbf{V}^{-1}\mathbf{W}\right)$$

donde

$$\mathbf{W} = \sum_{g=1}^{G}\sum_{h=1}^{n_g}(\mathbf{x}_{hg} - \bar{\mathbf{x}}_g)(\mathbf{x}_{hg} - \bar{\mathbf{x}}_g)' \tag{10.18}$$

es la matriz de suma de cuadrados dentro de los grupos. Sustituyendo en la función de verosimilitud y tomando logaritmos se obtiene

$$L(\mathbf{V}|\mathbf{X}) = \frac{n}{2}\log|\mathbf{V}^{-1}| - \frac{n}{2}tr\mathbf{V}^{-1}\mathbf{W}/n$$

y, según los resultados de 10.2, la varianza común a los grupos cuando estos tienen distinta media se estima por:

$$\widehat{\mathbf{V}} = \mathbf{S}_w = \frac{1}{n}\mathbf{W} \tag{10.19}$$

donde $\mathbf{W}$ está dada por (10.18). Sustituyendo estas expresiones en la función soporte tendremos que

$$L(H_1) = -\frac{n}{2}\log|\mathbf{S}_w| - \frac{np}{2}. \tag{10.20}$$

La diferencia de soportes será:

$$\lambda = n\log\frac{|\mathbf{S}|}{|\mathbf{S}_w|} \tag{10.21}$$

y rechazaremos H_0 cuando esta diferencia sera grande, es decir, cuando la variabilidad suponiendo H_0 cierta, medida por $|\mathbf{S}|$, sea mucho mayor que la variabilidad cuando permitimos que las medias de los grupos sean distintas, medida por $|\mathbf{S}_w|$. Su distribución es, asintoticamente, una χ^2_g donde los grados de libertad, g, se obtienen como por la diferencia entre ambos espacio paramétricos. H_0 determina una región Ω_0 donde hay que estimar los p componentes del vector de medias común y la matriz de covarianzas, en total $p + p(p+1)/2$ parámetros. Bajo la hipótesis H_1 hay que estimar G vectores de medias más la matriz de covarianzas lo que supone $Gp + p(p+1)/2$ parámetros. La diferencia es g :

$$g = \dim(\Omega) - \dim(\Omega_0) = p(G-1) \tag{10.22}$$

que serán los grados de libertad de la distribución asintótica.

La aproximación a la distribución χ^2_g del cociente de verosimilitudes puede mejorarse para tamaños muestrales pequeños. Se demuestra que el estadístico :

$$\lambda_0 = m\log\log\frac{|\mathbf{S}|}{|\mathbf{S}_w|}, \tag{10.23}$$

donde

$$m = (n-1) - (p+G)/2,$$

sigue asintóticamente una distribución χ^2_g , donde g viene dada por (10.22), y la aproximación es mejor que tomando $m = n$ en pequeñas muestras.

El análisis de la varianza multivariante

Este contraste es la generalización multivariante del análisis de la varianza y puede deducirse alternativamente como sigue. Llamemos variabilidad total de los datos a:

$$\mathbf{T} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})', \tag{10.24}$$

que mide las desviaciones respecto a una media común. Vamos a descomponer la matriz $\mathbf{T}$ como suma de dos matrices. La primera, $\mathbf{W}$, es la matriz de las desviaciones respecto a las medias de cada grupo y viene dada por (10.18). La segunda medirá la variabilidad explicada por las diferencias entre las medias y la llamaremos $\mathbf{B}$. Esta descomposición generaliza al caso vectorial la descomposición clásica de análisis de la varianza. Para obtenerla sumaremos y restaremos las medias de grupo en la expresión de $\mathbf{T}$, como:

$$\mathbf{T} = \sum_{g=1}^{G} \sum_{h=1}^{n_g} (\mathbf{x}_{gh} - \bar{\mathbf{x}} + \bar{\mathbf{x}}_g - \bar{\mathbf{x}}_g)(\mathbf{x}_{gh} - \bar{\mathbf{x}} + \bar{\mathbf{x}}_g - \bar{\mathbf{x}}_g)'$$

y desarrollando se comprueba que el doble producto se anula y resulta:

$$\boxed{\mathbf{T} = \mathbf{B} + \mathbf{W},} \tag{10.25}$$

donde $\mathbf{T}$ viene dado por (10.24), $\mathbf{W}$ por (10.18) y $\mathbf{B}$, la matriz de variabilidad explicada o de sumas de cuadrados entre grupos, se calcula por:

$$\mathbf{B} = \sum_{g=1}^{G} n_g (\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})'.$$

La descomposición (10.25) puede expresarse como

$$\text{Variabilidad Total } (\mathbf{T}) = \text{Variabilidad Explicada } (\mathbf{B}) + \text{Variabilidad Residual } (\mathbf{W})$$

que es la descomposición habitual del análisis de la varianza.

Para hacer un contraste de que las medias son iguales podemos comparar el tamaño de las matrices $\mathbf{T}$ y $\mathbf{B}$. La medida de tamaño adecuada es el determinante, con lo que concluimos que el contraste debe basarse en el cociente $|\mathbf{T}|/|\mathbf{W}|$. La distribución exacta de este cociente fue estudiada por Wilks. Para tamaños moderados el contraste es similar al de la razón de verosimilitudes (10.23), que puede escribirse también como:

$$\lambda_0 = m \log \frac{|\mathbf{T}|}{|\mathbf{W}|} = m \log \frac{|\mathbf{W} + \mathbf{B}|}{|\mathbf{W}|} = m \log |\mathbf{I} + \mathbf{W}^{-1}\mathbf{B}| \tag{10.26}$$

Desde el punto de vista del cálculo de (10.26) como $|\mathbf{I} + \mathbf{A}| = \Pi(1 + \lambda_i)$ donde λ_i son los vectores propios de $\mathbf{A}$, este estadístico se reduce a

$$\lambda_0 = m \sum \log(1 + \lambda_i)$$

donde λ_i son los vectores propios de la matriz $\mathbf{W}^{-1}\mathbf{B}$.

Ejemplo 10.3 *Vamos a aplicar este contraste para ver si se observan diferencias detectables en pequeñas muestras en los datos de Medifis, entre las medidas físicas de hombres y mujeres de la tabla A.5. En la muestra hay 15 mujeres (variable sexo = 0) y 12 hombres (sexo = 1). El primer paso del análisis es calcular las medias y matrices de covarianzas en cada grupo,*

por separado, y para el conjunto de los datos. En la tabla siguiente se presentan las medias para cada variable, para toda la muestra, y para los grupos de mujeres y hombres

	est	*pes*	*pie*	*lbr*	*aes*	*dcr*	*lrt*
total	168.78	63.89	38.98	73.46	45.85	57.24	43.09
mujeres	161.73	55.60	36.83	70.03	43.33	56.63	41.06
hom*bres*	177.58	74.25	41.67	77.75	49.00	58.00	45.62

Las matrices de covarianzas dividiendo por $n-1$ *para toda la muestra, mujeres y hombres son*

Para las mujeres:

$$\widehat{S}_M = \begin{bmatrix} 37.64 & & & & & & \\ 22.10 & 80.40 & & & & & \\ 6.38 & 7.36 & 1.92 & & & & \\ 15.65 & 12.94 & 3.06 & 7.41 & & & \\ 9.49 & 14.39 & 1.49 & 3.99 & 9.42 & & \\ 2.75 & 7.20 & 0.76 & 1.17 & 2.559 & 2.94 & \\ 9.02 & 9.31 & 1.98 & 4.53 & 1.12 & 0.95 & 3.78 \end{bmatrix}$$

Para los hombres

$$\widehat{S}_H = \begin{bmatrix} 45.53 & & & & & & \\ 48.84 & 74.20 & & & & & \\ 9.48 & 9.63 & 2.79 & & & & \\ 14.34 & 19.34 & 2.09 & 12.57 & & & \\ 14.86 & 19.77 & 3.23 & 6.18 & 6.77 & & \\ 9.45 & 9.90 & 1.86 & 2.36 & 3.02 & 3.13 & \\ 8.92 & 5.23 & 2.31 & 1.21 & 1.84 & 2.63 & 6.14 \end{bmatrix}$$

y para el conjunto de hombres y mujeres, se calcula como media ponderada de estas dos matrices

$$\widehat{S}_T = (14\widehat{S_M} + 11\widehat{S_H})/25$$

con lo que se obtiene

$$\widehat{S}_T = \begin{bmatrix} 41.11 & & & & & & \\ 33.86 & 77.67 & & & & & \\ 7.476 & 8.36 & 2.30 & & & & \\ 15.07 & 15.76 & 2.63 & 9.68 & & & \\ 11.85 & 16.76 & 2.25 & 4.95 & 8.25 & & \\ 5.70 & 8.390 & 1.24 & 1.70 & 2.76 & 3.03 & \\ 8.98 & 7.52 & 2.13 & 3.07 & 1.44 & 1.70 & 4.82 \end{bmatrix}$$

Vamos a calcular el ratio de verosimilitudes como cociente de las variabilidades promedio ante ambas hipótesis. Bajo H_0 *se obtiene que la matriz de varianzas y covarianzas cuando suponemos la misma media,* $\mathbf{S}$ *conduce a la variabilidad promedio*

$$VP(H_0) = |\mathbf{S}|^{1/7} = 5.77$$

mientras que

$$VP(H_1) = |\mathbf{S}_w|^{1/7} = 4.67$$

con lo que el contraste es

$$27((27-1)-(7+7)/2)\log(5.77/4.67) = 108.5$$

que debe compararse con una χ^2 con 7 grados de libertad, y no hay ninguna duda de que las diferencias son significativas.

10.8 Contrastes de datos atípicos

El contraste de igualdad de medias puede aplicarse, como caso particular, para contrastar si una observación de una muestra de datos normales es atípica. La hipótesis nula será que todos los datos vienen de la misma población normal. La hipótesis alternativa será que el dato sospechoso ha sido generado por otra población desconocida. Para caracterizar la población alternativa podemos suponer que la media es distinta y la varianza la misma, o que la media es la misma y la varianza distinta. Si supusiesemos que tanto la media como la matriz de covarianzas tendríamos un problema de identificación, porque no es posible con un solo dato estimar la media y la variabilidad. Puede comprobarse que los contrastes suponiendo la media distinta o la varianza distinta son similares (véase Peña y Guttman, 1993) y aquí consideraremos el caso más simple de media distinta pero misma matriz de covarianzas. Para aplicar este contraste a un dato sospechoso, $\mathbf{x}_i$, estableceremos:

$$H_0 : E(\mathbf{x}_i) = \boldsymbol{\mu};$$

frente a

$$H_1 : E(\mathbf{x}_i) = \boldsymbol{\mu}_i \neq \boldsymbol{\mu};$$

La función de verosimilitud bajo H_0 es (10.17). Bajo H_1, como la estimación $\boldsymbol{\mu}_i$ es $\mathbf{x}_i$, la estimación de la varianza será

$$\mathbf{S}_{(i)} = \frac{1}{n-1}\mathbf{W}_{(i)},$$

donde

$$\mathbf{W}_{(i)} = \sum_{h=1, j\neq i}^{n} (\mathbf{x}_h - \bar{\mathbf{x}}_{(i)})(\mathbf{x}_h - \bar{\mathbf{x}}_{(i)})',$$

es la estimación de la suma de cuadrados de los residuos, y $\bar{\mathbf{x}}_{(i)}$ es la media de las observaciones, en ambos casos eliminando la observación $\mathbf{x}_i$. La diferencia de soportes es, particularizando (10.26):

$$\lambda = n \log \frac{|\mathbf{T}|}{|\mathbf{W}_{(i)}|}$$

y, se demuestra en el apéndice 10.3, que se verifica la relación:

$$\frac{|\mathbf{T}|}{|\mathbf{W}_{(i)}|} = 1 + \frac{1}{n} D^2(\mathbf{x}_i, \bar{\mathbf{x}}_{(i)})$$

donde $D^2(\mathbf{x}_i, \bar{\mathbf{x}}_{(i)})$ es:

$$D^2(\mathbf{x}_i, \bar{\mathbf{x}}_{(i)}) = (\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})' \mathbf{S}_{(i)}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{(i)}). \tag{10.27}$$

la distancia de Mahalanobis entre el dato y la media sin incluirle. Por tanto, para realizar el test calcularemos la distancia de Mahalanobis (10.27), que se distribuirá, si H_0 es cierta, para muestras grandes como una χ_p^2 .

En la práctica, para detectar atípicos se calcula el máximo de las distancias $D^2(\mathbf{x}_i, \bar{\mathbf{x}}_{(i)})$ y este valor se compara con el percentil 0,95 o 0,99 de las tablas de percentiles del máximo de una χ_p^2. El problema, entonces, es que si existe más de un atípico, la potencia del contraste puede ser muy baja, al estar contaminadas las estimaciones de los parámetros. Un procedimiento más recomendable siempre que se trabaje con muestras que pueden ser heterogéneas es identificar primero todas las observaciones sospechosas, con los procedimientos indicados en el capítulo 3, y después ir contrastando una por una si las observaciones se aceptan. Es decir, ordenamos todos los datos sospechosos por $D^2(\mathbf{x}_i, \bar{\mathbf{x}}_{(i)})$ y contrastamos si el más próximo puede incorporarse a la muestra. Si se rechaza esta incorporación el procedimiento termina y todos los datos sospechosos son declarados atípicos. En caso contrario, el dato se incorpora a la muestra y se recalculan los parámetros y las distancias de Mahalanobis, y se vuelve a aplicar el procedimiento a las restantes excluidas.

10.9 Contrastes de Normalidad

Los métodos más utilizados en análisis multivariante suponen normalidad conjunta de las observaciones y conviene, cuando dispongamos de datos suficientes, contrastar esta hipótesis.

Normalidad unidimensional

La normalidad de las distribuciones univariantes puede contrastarse con los contrastes χ^2, Kolmogorov-Smirnov, Shapiro y Wilks, o con los contrastes basados en coeficientes de asimetría y curtosis, que pueden consultarse en Peña (2001). Llamando

$$A = \frac{m_3}{m_2^{3/2}}; \quad K = \frac{m_4}{m_2^2},$$

donde

$$m_h = \frac{1}{n} \sum (x_i - \overline{x})^h.$$

Se demuestra que, asintóticamente, con datos normales:

$$A \sim N(0; 6/n); \quad K \sim N(3; 24/n)$$

y por tanto la variable

$$X^2 = \frac{nA^2}{6} + \frac{n\,(K-3)^2}{24}$$

se distribuirá, si la hipótesis de normalidad es cierta, como una χ^2 con 2 grados de libertad. Rechazaremos la hipótesis de normalidad si $X^2 > \chi_2^2(\alpha)$.

Normalidad multivariante

La normalidad multivariante implica la normalidad de distribuciones marginales unidimensionales, pero la existencia de esta propiedad no garantiza la normalidad multivariante de los datos. Para contrastar la normalidad conjunta existen varios contrastes posibles, y aquí sólo comentaremos la generalización multivariante de los contrastes de asimetría y curtosis. (Véase Justel, Peña y Zamar (1997) para una generalización del contraste de Kolmogorov-Smirnov al caso multivariante).

Definiendo los coeficientes de asimetría y curtosis multivariantes como en la sección 3.6:

$$A_p = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^3$$

$$K_p = \frac{1}{n}\sum_{i=1}^{n} d_{ii}^2$$

donde $d_{ij} = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})$, se verifica asintóticamente:

$$nA_p/6 \sim \chi_f^2 \quad \text{con} \quad f = \frac{1}{6}p(p+1)(p+2)$$

$$K_p \sim N\left(p(p+2); 8p(p+2)/n\right)$$

La potencia de este contraste no es muy alta a no ser que tengamos una muestra muy grande. Dos casos frecuentes en la práctica en que se rechaza la hipótesis de normalidad conjunta son:

(1) Las distribuciones marginales son aproximadamente simétricas, y las relaciones entre las variables son lineales, pero existen valores atípicos que no pueden explicarse por la hipótesis de normalidad. En este caso si eliminamos (o descontamos con un estimador robusto) los valores atípicos, la normalidad conjunta no se rechaza y los métodos basados en la normalidad suelen dar buenos resultados.

(2) Algunas (o todas) las distribuciones marginales son asimétricas y existen relaciones no lineales entre las variables. Una solución simple y que funciona bien en muchos casos es transformar las variables para conseguir simetría y relaciones lineales.

10.9.1 Transformaciones

Para variables escalares Box y Cox (1964) han sugerido la siguiente familia de transformaciones para conseguir la normalidad:

$$x^{(\lambda)} = \begin{array}{lll} \frac{(x+m)^\lambda - 1}{\lambda} & (\lambda \neq 0) & (x > -m) \\ \ln(x+m) & (\lambda = 0) & (m > 0) \end{array}$$

donde λ es el parámetro de la transformación que se estima a partir de los datos y la constante m se elige de forma que $x + m$ sea siempre positiva. Por lo tanto, m será cero si

trabajamos con datos positivos e igual en valor absoluto al valor más negativo observado, en otro caso. Suponiendo $m = 0$ esta familia incluye como casos particulares la transformación logarítmica, la raíz cuadrada y la inversa. Cuando $\lambda > 1$, la transformación produce una mayor separación o dispersión de los valores grandes de x, tanto más acusada cuanto mayor sea el valor de λ mientras que cuando $\lambda < 1$ el efecto es el contrario: los valores de x grandes tienden a concentrarse y los valores pequeños ($x < 1$) a dispersarse.

Estas transformaciones son muy útiles para las distribuciones marginales. Para estudiar cómo determinar el valor del parámetro con una variable escalar, supongamos que $m = 0$ y que existe un valor de λ que transforma la variable en normal. La relación entre el modelo para los datos originales, x, y para los transformados, $x^{(\lambda)}$, será:

$$f(x) = f(x^{(\lambda)}) \left| \frac{dx^{(\lambda)}}{dx} \right|, \tag{10.28}$$

y como:

$$\frac{dx^{(\lambda)}}{dx} = \frac{\lambda x^{\lambda-1}}{\lambda} = x^{\lambda-1}$$

y suponiendo que $x^{(\lambda)}$ es $N(\mu,\ \sigma^2)$, para cierto valor de λ, la función de densidad de las variables originales será:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left(\frac{x^\lambda - 1}{\lambda} - \mu\right)^2} x^{\lambda-1}$$

Por tanto, la función de densidad conjunta de $\mathbf{X} = (x_i, ..., x_n)$ será, por la independencia de las observaciones:

$$f(\mathbf{X}) = \frac{1}{\sigma^n \left(\sqrt{2\pi}\right)^n} \left(\prod_{i=1}^{n} x_i^{\lambda-1} \right) e^{-\frac{1}{2\sigma^2}\sum\left(\frac{x_i^\lambda - 1}{\lambda} - \mu\right)^2} \tag{10.29}$$

y la función soporte es:

$$L\left(\lambda; \mu, \sigma^2\right) = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi + (\lambda - 1)\sum \ln x_i - \frac{1}{2\sigma^2}\sum\left(\frac{x_i^\lambda - 1}{\lambda} - \mu\right)^2.$$

Para obtener el máximo de esta función utilizaremos que, para λ fijo, los valores de σ^2 y μ que maximizan la verosimilitud (o el soporte) son, derivando e igualando a cero:

$$\begin{aligned} \widehat{\sigma}^2(\lambda) &= \frac{1}{n}\sum\left(x^{(\lambda)} - \widehat{\mu}(\lambda)\right)^2, \\ \widehat{\mu}(\lambda) &= \overline{x}^{(\lambda)} = \sum \frac{x_i^{(\lambda)}}{n} = \frac{1}{n}\sum\left(\frac{x_i^\lambda - 1}{\lambda}\right). \end{aligned}$$

Al sustituir estos valores en la verosimilitud obtenemos lo que se denomina la función de verosimilitud *concentrada* en λ. Su expresión es, prescindiendo de constantes:

$$L(\lambda) = -\frac{n}{2}\ln\widehat{\sigma}(\lambda)^2 + (\lambda - 1)\sum \ln x_i \tag{10.30}$$

El procedimiento para obtener $\widehat{\lambda}$ consiste en calcular $L(\lambda)$ para distintos valores de λ. El valor que maximice esta función es el estimador MV de la transformación.

Para conseguir normalidad multivariante supondremos que existe un vector de parámetros $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p)$ que produce normalidad multivariante, donde λ_j es la transformación aplicada al componente j del vector. Aplicando un análisis similar al caso univariante, la función soporte multivariante concentrada en el vector de parámetros de la transformación es:

$$L(\boldsymbol{\lambda}) = -\frac{n}{2} \ln \left| \widehat{\Sigma} \right| + \sum_{j=1}^{p} \left[(\lambda_j - 1) \sum_{i=1}^{n} \ln x_{ij} \right],$$

donde los parámetros se han estimado aplicando las formulas habituales a los datos transformados:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^{(\lambda)},$$

y

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^{(\lambda)} - \widehat{\boldsymbol{\mu}})(\mathbf{x}_i^{(\lambda)} - \widehat{\boldsymbol{\mu}})'.$$

La estimación MV del parámetro vectorial $\boldsymbol{\lambda}$ no suele aportar mejoras importantes respecto a transformar individualmente cada variable para que las marginales sean normales. Véase Johnson y Wichern (1998).

10.10 Lecturas recomendadas

En este capítulo hemos resumido un tema muy amplio sobre el que existe abundante bibliografía. El lector puede encontrar una buena introducción en inglés a los métodos de estimación y contraste basados en la función de verosimilitud en Casella y Berger (1990). En español vease Peña (2001) y las referencias allì indicadas. La estimación MV de la normal multivariante se trata con detalle en Anderson (1984), Mardia et al (1979) y Muirhead (1982). Los contrastes de matrices de covarianzas se presentan claramente en Mardia et al (1979) y Rechner (1998). El análisis de la varianza multivariante es un tema muy amplio y puede ampliarse en Johnson y Wichern (1998), Morrison (1976), Seber (1984) y Rechner (1998). Métodos tradicionales para el contraste de datos atípicos pueden encontrarse en Barnett y Lewis (1994), y referencias a métodos más recientes en Peña y Prieto (2001). Para la transformación multivariante de Box-Cox véase Gnanadesikan (1997) y Velilla (1993).

Por razones de espacio no hemos incluido la aplicación de nuevos métodos de estimación, como la estimación autosuficiente de Efron, al caso multivariante. El lector puede consultarla en Efron (1982) and Efron y Tibshirani (1993).

Ejercicios

10.1 Demostrar que la verosimilitud $L(\mathbf{V}|\mathbf{X}) = \frac{n}{2} \log |\mathbf{V}^{-1}\mathbf{S}| - \frac{n}{2} tr\mathbf{V}^{-1}\mathbf{S}$ es invariante ante trasnformaciones de las variables $\mathbf{y} = \mathbf{A}\mathbf{x}$, con $\mathbf{A}$ cuadrada no singular.

10.2 Obtener los estimadores máximos verosímiles de los parámetros en la normal multivariante derivando en la función de verosimilitud (10.4). Para la varianza escribir la función como función de $\mathbf{V}^{-1}$ y utilizar que $\frac{\partial log\left|\mathbf{V}^{-1}\right|}{\partial \mathbf{V}^{-1}} = \mathbf{V}$, y $\frac{\partial tr\mathbf{V}^{-1}\mathbf{S}}{\partial \mathbf{V}^{-1}} = \mathbf{S}$. Comprobar entonces que $\frac{\partial L(\mathbf{V}^{-1}|\mathbf{X})}{\partial \mathbf{V}^{-1}} = \frac{n}{2}\left(\frac{\partial log\left|\mathbf{V}^{-1}\right|}{\partial \mathbf{V}^{-1}} - \frac{\partial tr(\mathbf{V}^{-1}\mathbf{S})}{\partial \mathbf{V}^{-1}}\right) = \frac{n}{2}(\mathbf{V}-\mathbf{S}) = 0$.

10.3 Demostrar que la función soporte de la sección 10.2 puede escribirse como $L(\mathbf{V}|\mathbf{X}) = -\frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}tr\mathbf{V}^{-1}\mathbf{S}(\boldsymbol{\mu})$, donde $\mathbf{S}(\boldsymbol{\mu}) = \sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})'/n$ y utilizar esta expresión para demostrar que el estimador MV de $\mathbf{V}$ cuando restringimos los valores de $\boldsymbol{\mu}$ a una región A es es $\mathbf{S}(\widehat{\boldsymbol{\mu}})$, donde $\widehat{\boldsymbol{\mu}}$ es el valor que maximiza $\sum_{i=1}^{n}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ sobre A.

10.4 Demostrar que la función de verosimilitud del ejercicio anterior 10.1 puede escribirse como $L(\mathbf{V}|\mathbf{X}) = \frac{np}{2}(\log\overline{\lambda}_g - \overline{\lambda})$, donde $\overline{\lambda}_g$ y $\overline{\lambda}$ son la media geométrica y aritmética de los valores propios de la matriz $\mathbf{V}^{-1}\mathbf{S}$.

10.5 Demostrar que el contraste del análisis de la varianza multivariante equivale a comparar las medias geométricas de los valores propios de las matrices de variabilidad total y no explicada.

10.6 Demostrar que el contraste del análisis de la varianza multivariante no se modifica si en lugar de trabajar con las variables originales lo hacemos con las componentes principales.

10.7 Demostrar que el contraste multivariante de que una muestra viene de una población es invariante ante transformaciones lineales no singulares de las variables. ¿Cómo sería el contraste si en lugar de las variables utilizamo sus componentes principales?

10.8 Demostrar que el estimador MV del parámetro σ^2 en el modelo $\mathbf{x} \sim \mathbf{N}_p(\boldsymbol{\mu},\sigma^2\mathbf{I})$ es $\widehat{\sigma}^2 = tr\mathbf{S}/p$.

10.9 Demostrar que el contraste $H_0 : \mathbf{x} \sim \mathbf{N}_p(\boldsymbol{\mu},\sigma^2\mathbf{V}_0)$ frente a $H_1 : \mathbf{x} \sim \mathbf{N}_p(\boldsymbol{\mu},\mathbf{V})$ depende sólo de la media aritmética y geométrica de los valores propios $\mathbf{V}_0^{-1}\mathbf{S}$.

APÉNDICE 10.1: Inadmisibilidad de la media muestral para $p \geq 3$

Stein (1956) demostró que para $p \geq 3$ la media muestral no es necesariamente el estimador óptimo de la media poblacional de una normal multivariante. Este resultado es consecuencia de que si tomamos como criterio seleccionar como estimador de $\boldsymbol{\mu}$ el que minimice el error cuadrático medio de estimación, dado por

$$E\left[(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})'\mathbf{M}^{-1}(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}})\right],$$

donde $\mathbf{M}$ es una cierta matriz que sirve para definir como se mide la distancia entre el estimador $\widehat{\boldsymbol{\mu}}$ y el parámetro $\boldsymbol{\mu}$. Eligiendo $\mathbf{M} = \mathbf{I}$, el estimador:

$$\widehat{\boldsymbol{\mu}}_s = \left(1 - \frac{(p-2)}{\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}}\right)\bar{\mathbf{x}},$$

es sesgado para $\boldsymbol{\mu}$, pero puede demostrarse que tiene un error cuadrático medio menor que $\bar{\mathbf{x}}$. Por tanto, con este criterio, $\widehat{\boldsymbol{\mu}}_s$ es mejor estimador que la media muestral y en consecuencia la media muestral se dice que es un estimador inadmisible si $p \geq 3$, ya que el estimador $\widehat{\boldsymbol{\mu}}_s$ tiene siempre menor error cuadrático medio. Observemos que $\widehat{\boldsymbol{\mu}}_s$ "contrae" (shrinkage) el valor de $\bar{\mathbf{x}}$, ya que $|\widehat{\boldsymbol{\mu}}_s| \leq |\overline{\mathbf{x}}|$. Este resultado ha despertado un gran interés por los estimadores "shrinkage" que mejoran el error cuadrático medio de los estimadores MV tradicionales.

APÉNDICE 10.2: RAZÓN DE VEROSIMILITUDES Y LA T DE HOTELLING

Para demostrar la relación entre la razón de verosimilitudes y la T de Hotelling utilizaremos el siguiente

Lemma 1 *Si* $\mathbf{A}$ *es una matriz no singular y* b *un vector,* $|\mathbf{I} + \mathbf{A}bb'| = 1 + b'\mathbf{A}b$. *En efecto, observemos que la matriz* $\mathbf{bb}'$ *tiene rango uno y también tendrá rango uno* $\mathbf{Abb}'$. *Por tanto,* $\mathbf{Abb}'$ *tiene un único valor propio no nulo. Llamando* λ *a este valor propio no nulo y* v *al vector propio, como* $\mathbf{Abb}'v = \lambda v$, *multiplicando por* b' *se obtiene que* $\lambda = b'\mathbf{A}b$. *Entonces la matriz* $\mathbf{I} + \mathbf{A}bb'$ *tendrá un valor propio igual a 1+*λ *y el resto serán la unidad . Como el determinante es el producto de los valores propios, queda demostrado el lema.*

Partiendo ahora de

$$n\mathbf{S}_0 = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu}_0)',$$

y desarrollando en los términos $(\mathbf{x}_i - \bar{\mathbf{x}})$ y $(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$, resulta:

$$n\mathbf{S}_0 = n\mathbf{S} + n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'.$$

Por tanto

$$\frac{|\mathbf{S}_0|}{|\mathbf{S}|} = \frac{|\mathbf{S} + (\overline{\mathbf{x}} - \boldsymbol{\mu}_0)(\overline{\mathbf{x}} - \boldsymbol{\mu}_0)'|}{|\mathbf{S}|},$$

que puede escribirse

$$\frac{|\mathbf{S}_0|}{|\mathbf{S}|} = |\mathbf{S}^{-1}||\mathbf{S} + (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'| = |\mathbf{I} + \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'|,$$

y aplicando el lema anterior tenemos que:

$$|\mathbf{I} + \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'| = 1 + (\overline{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}_0),$$

con lo que, tenemos finalmente que :

$$\frac{|\mathbf{S}_0|}{|\mathbf{S}|} = 1 + (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) = 1 + \frac{T^2}{n-1}.$$

APÉNDICE 10.3: CONTRASTE DE VALORES ATIPICOS

La relación entre $\mathbf{T}$ y $\mathbf{W}_{(i)}$ se obtiene restando y sumando $\bar{\mathbf{x}}_{(i)}$:

$$\mathbf{T} = \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}}_{(i)} + \bar{\mathbf{x}}_{(i)} - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}}_{(i)} + \bar{\mathbf{x}}_{(i)} - \bar{\mathbf{x}})',$$

que resulta en

$$\mathbf{T} = \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}}_{(i)})(\mathbf{x}_j - \bar{\mathbf{x}}_{(i)})' + n(\bar{\mathbf{x}}_{(i)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{(i)} - \bar{\mathbf{x}})' + F + F'$$

donde $F = \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}}_{(i)})(\bar{\mathbf{x}}_{(i)} - \bar{\mathbf{x}})'$. El primer término puede escribirse

$$\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}}_{(i)})(\mathbf{x}_j - \bar{\mathbf{x}}_{(i)})' = \mathbf{W}_{(i)} + (\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})'$$

y utilizando que

$$\bar{\mathbf{x}}_{(i)} - \bar{\mathbf{x}} = \bar{\mathbf{x}}_{(i)} - \frac{(n-1)\bar{\mathbf{x}}_{(i)} + \mathbf{x}_i}{n} = \frac{1}{n}(\bar{\mathbf{x}}_{(i)} - \mathbf{x}_i)$$

y reemplazando en todos los términos $(\bar{\mathbf{x}}_{(i)} - \bar{\mathbf{x}})$ por $(\bar{\mathbf{x}}_{(i)} - \mathbf{x}_i)/n$ se obtiene finalmente que

$$\mathbf{T} = \mathbf{W}_{(i)} + \frac{n-1}{n}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})'$$

y, por tanto,

$$|\mathbf{T}| = \left|\mathbf{W}_{(i)}\right| \left|1 + \frac{n-1}{n}\mathbf{W}_{(i)}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})'\right|$$

y aplicando el lema del apéndice 10.2

$$\frac{|\mathbf{T}|}{|\mathbf{W}_{(i)}|} = 1 + \frac{1}{n}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})'\mathbf{S}_{(i)}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})$$

donde $\mathbf{S}_{(i)}^{-1} = (n-1)\mathbf{W}_{(i)}^{-1}$. Finalmente

$$DS = n\log(1 + \frac{1}{n}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})'\mathbf{S}_{(i)}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)}))$$

y para n grande como $\log(1 + x/n) \simeq x/n$, tenemos que la distancia de Mahalanobis

$$D^2(\mathbf{x}_i, \bar{\mathbf{x}}_{(i)}) = (\mathbf{x}_i - \bar{\mathbf{x}}_{(i)})'\mathbf{S}_{(i)}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{(i)}) \sim \chi_p^2.$$

Capítulo 11

METODOS DE INFERENCIA AVANZADA MULTIVARIANTE

11.1 INTRODUCCIÓN

En este capítulo vamos a presentar métodos más avanzados de inferencia para datos multivariantes. En primer lugar presentamos un algoritmo para estimar por máxima verosimilitud muestras con datos incompletos. Este algoritmo, el EM, es muy útil para estimar distribuciones mezcladas, que utilizaremos en el capítulo 14 en problemas de clasificación y también es útil en la estimación del modelo factorial que se presenta en el capítulo 11. Además este algoritmo tiene un interés general por sí mismo para resolver la estimación de valores ausentes en cualquier problema multivariante. A continuación se presentan los métodos robustos clásicos de estimación, que pueden también considerarse como métodos de estimación de mezclas en un caso especial: hay una distribución central, que genera la mayoría de las observaciones, y una distribución contaminante de forma desconocida que introduce una pequeña proporción de atípicos en la muestra. Se presentan brevemente los métodos clásicos y se introduce un método reciente basado en proyecciones que es fácil de implementar y que puede evitar el efecto perturbador de los datos atípicos en la estimación de los parámetros. Se presenta también una breve introducción a la inferencia bayesiana. Además de su atractivo metodológico, la inferencia bayesiana permite incorporar información a priori, que puede ser importante en problemas de clasificación (capítulo 12, análisis discriminante) y construcción de conglomerados (capítulo 14, clasificación mediante mezclas). Los métodos bayesianos son también útiles en análisis factorial (capítulo 11). Finalmente, los métodos bayesianos de estimación por Montecarlo son muy eficaces para la estimación de mezclas, como veremos en el capítulo 14. En este capítulo se revisa brevemente el enfoque Bayesiano para la estimación y el contraste de hipótesis y se deduce un criterio de comparación de modelos a partir de este enfoque. Finalmente se presentan algunos métodos clasicos y bayesianos para selección de modelos.

Este capítulo es más avanzado que los anteriores y puede saltarse en una primera lectura sin pérdida de continuidad, ya que la comprensión básica de los métodos que se presentan en los capítulos siguientes no requiere el material de este capítulo. Sin embargo este capítulo será necesario para el lector interesado en los detalles de aplicación de los métodos, y en

la comprensión de los algoritmos de estimación actuales para el análisis multivariante y los métodos de data mining, que están adquiriendo una popularidad creciente.

11.2 ESTIMACIÓN MV CON DATOS FALTANTES

La estimación máximo verosímil con datos faltantes es importante por dos razones principales. En primer lugar, es posible que la muestra tenga observaciones faltantes en algunas variables. Por ejemplo, si tomamos una muestra de personas desempleadas y queremos relacionar sus características físicas con la duración de desempleo, es posible que para algunas personas no se consiga este dato. (En otros casos podemos tener información parcial, por ejemplo un valor superior o inferior de la duración, y en estos casos decimos que el dato está censurado o truncado, no consideraremos estos casos en este libro). Como segundo ejemplo, si hacemos una encuesta de opinión, y representamos por $\mathbf{x}$ el vector de respuestas de un individuo, es posible que determinadas preguntas del cuestionario no sean respondidas por algunos individuos, dando lugar a un problema de datos faltantes. Si los datos faltantes ocurren en unos pocos elementos de la muestra, podemos eliminar las observaciones incompletas, pero si ocurren en una proporción importante de observaciones, podemos mejorar la precisión de las estimaciones utilizando todos los datos, con el coste de un mayor esfuerzo computacional.

En segundo lugar, la estimación MV de muchos modelos de análisis multivariante puede realizarse más facilmente con este algoritmo. Por ejemplo, en el modelo factorial, que estudiaremos en el capítulo 12, o en la estimación de distribuciones mezcladas para clasificación, que estudiaremos en el capítulo 15. En el primer caso, podemos suponer que los factores son variables ausentes y en el segundo, que faltan los valores de las variables de clasificación que nos indican de que población proviene cada elemento.

Intuitivamente, el procedimiento para estimar los parámetros de un modelo con una muestra que contiene datos faltantes podría ser:

(1) estimar los parámetros del modelo con los datos que están completos, maximizando la verosimilitud de la forma habitual;

(2) Utilizar los parámetros estimados en (1) para predecir los valores ausentes;

(3) Sustituir los datos ausentes por sus predicciones y obtener nuevos valores de los parámetros maximizando la verosimilitud de la muestra completada.

Adicionalmente podríamos iterar entre (2) y (3) hasta que se obtenga convergencia, es decir hasta que el valor de los parámetros no cambie de una iteración a la siguiente. Veremos en la sección siguiente que este procedimiento intuitivo es óptimo en muchos casos, pero no siempre. La razón es que no tiene en cuenta cómo se utilizan los datos ausentes para estimar los parámetros a partir de la verosimilitud. Por ejemplo, supongamos el caso más simple de una variable escalar, x, y un único parámetro a estimar, θ. Supongamos que la función soporte para θ es de la forma: $\theta^2 - 2\theta \sum x_i^2$. Entonces, el estimador MV de θ es, derivando respecto a θ e igualando a cero, $\widehat{\theta}_{MV} = \sum x_i^2$. Supongamos ahora que la observación x_1 falta. Para obtener entonces el estimador $\widehat{\theta}_{MV}$ tendriamos que estimar el valor esperado de x_1^2 a la vista de la información disponible y utilizarlo en $\widehat{\theta}_{MV} = \sum x_i^2$. Si en lugar del valor esperado de x_1^2 calculamos el valor esperado de x_1 y lo sustituimos en esta ecuación

elevado al cuadrado, como en general $E(x_1^2) \neq [E(x_1)]^2$, este segundo procedimiento no es necesariamente óptimo. Por ejemplo, si la variable x_1 tiene media cero dada la información disponible lo que necesitamos es calcular su varianza, $E(x_1^2)$, y sustituirlo en la ecuación del parámetro. Esto no es lo mismo que calcular $E(x_1)$, que es cero, y sustituirlo como x_1, con lo que x_1^2 será cero. Un procedimiento eficiente y general para maximizar la verosimilitud cuando tenemos datos faltantes es el algoritmo EM (Dempster, Laird y Rubin, 1977), que extiende el procedimiento intuitivo anterior, como describimos a continuación

11.2.1 Estimación MV con el algoritmo EM

Supongamos que tenemos una muestra de tamaño n de una variable vectorial, $\mathbf{x}$, pero en algunos de los n elementos observados faltan los valores de algunas variables. Por ejemplo, observamos el peso y la altura de personas y en algunos casos tenemos sólo el dato de la estatura o sólo el dato del peso. Vamos a suponer que estos datos ausentes aparecen al azar, es decir, que en el ejemplo anterior la falta del dato del peso no aparece con más frecuencia en individuos de peso alto o bajo o con estatura más alta o más baja, sino que falta ese dato por razones no relacionadas con los valores de las variables. Para cada elemento no hay una relación entre los valores observados y la aparición o no de un dato ausente. Un ejemplo donde no se cumple esta condición es una encuesta de opinión donde las personas que manifiestan desacuerdo en un punto, por ejemplo con la pregunta diez, dejan de responder al cuestionario a continuación. En este caso, los valores ausentes en la pregunta once no aparecen al azar, sino que son consecuencia del desacuerdo con la pregunta diez.

Los dos casos más importantes de aparición de valores ausentes son:

(1) Algunos elementos tienen datos faltantes: los elementos de la muestra $\mathbf{x}_1, ..., \mathbf{x}_{n_1}$, están completos, pero los restantes, $\mathbf{x}_{n_1+1}, ..., \mathbf{x}_n$, carecen de los valores de algunas variables, o de todas ellas;

(2) Algunas variables tienen datos faltantes: si dividimos el vector de variables en dos grupos y escribimos $\mathbf{x} = (\mathbf{y}', \mathbf{z}')'$, las variables $\mathbf{y}$ están completas pero las $\mathbf{z}$ tienen datos ausentes.

Para plantear el problema de manera que englobe estos dos casos, supondremos que tenemos una muestra con una matriz de datos observados $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$, donde $\mathbf{y}_i$ es un vector $p_1 \times 1$, y un conjunto de datos ausentes $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_m)$, donde $\mathbf{z}_i$ es un vector $p_2 \times 1$. Esta formulación cubre los dos casos anteriores, ya que en el primero tomamos $\mathbf{z}_1 = \mathbf{x}_{n_1+1}, ..., \mathbf{z}_m = \mathbf{x}_n$ y $m = n - n_1 + 1$. En el segundo $m = n$.

Este conjunto de variables proviene de un modelo con parámetros $\boldsymbol{\theta}$, y se desea estimar el vector de parámetros con la información disponible.La función de densidad conjunta de todas las variables $(\mathbf{Y}, \mathbf{Z})$ puede escribirse

$$f(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\theta}) = f(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}) f(\mathbf{Y}|\boldsymbol{\theta})$$

que implica

$$\log f(\mathbf{Y}|\boldsymbol{\theta}) = \log f(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\theta}) - \log f(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}).$$

En la estimación MV el primer miembro de esta expresión, $\log f(\mathbf{Y}|\boldsymbol{\theta})$, es la función soporte de los datos observados, cuya maximización sobre $\boldsymbol{\theta}$ nos proporcionará el estimador

MV de los parámetros. El término $\log f(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\theta})$ es la función soporte si hubiésemos observado la muestra completa, y el término $\log f(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta})$ proporciona la densidad de los datos ausentes conocida la muestra y los parámetros. Podemos escribir

$$L(\boldsymbol{\theta}|\mathbf{Y}) = L_C(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}) - \log f(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}). \tag{11.1}$$

Es frecuente que la maximización del soporte supuesta la muestra completa, $L_C(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})$, sea fácil, mientras que la maximización del soporte con los datos observados, $L(\boldsymbol{\theta}|\mathbf{Y})$, sea complicada. El algoritmo EM es un procedimiento iterativo para encontrar el estimador MV de $\boldsymbol{\theta}$ trabajando siempre con la función más simple, $L_C(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})$, en lugar de la compleja, $L(\boldsymbol{\theta}|\mathbf{Y})$. La estimación se obtiene iterando en los dos pasos siguientes:

1. Partiendo de un estimador inicial, $\widehat{\boldsymbol{\theta}}^{(i)}$, (en la primera iteración $i = 1$) se calcula la esperanza de las funciones de los valores ausentes que aparecen en la función de verosimilitud completa, $L_C(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})$, con respecto a la distribución de $\mathbf{Z}$ dados el valor $\widehat{\boldsymbol{\theta}}^{(i)}$ y los datos observados $\mathbf{Y}$. Sea:

   $$L_C^*(\boldsymbol{\theta}|\mathbf{Y}) = \mathbf{E}_{Z/\widehat{\boldsymbol{\theta}}^{(i)}}\left[L_C(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})\right]$$

   el resultado de esta operación que se denomina el paso E (de tomar valores esperados) del algoritmo. Observemos que cuando $L_C(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})$ sea una función lineal de $\mathbf{Z}$, este paso llevará a sustituir en esta función los valores ausentes por sus esperanzas dados los parámetros. Sin embargo, cuando en la verosimilitud aparezcan funciones $g(\mathbf{Z})$ calcularemos la esperanza de estas funciones dados el resto de los datos y la estimación disponible de los parámetros.

2. A continuación se maximiza la función $L_C^*(\boldsymbol{\theta}|\mathbf{Y})$ con respecto a $\boldsymbol{\theta}$. Este es el paso M (maximización) del algoritmo. Este paso M equivale a maximizar la verosimilitud completa donde se han sustituido las observaciones faltantes por ciertas estimaciones de sus valores.

3. Sea $\widehat{\boldsymbol{\theta}}^{(i+1)}$ el estimador obtenido en el paso M. Con este valor volvemos al paso E. Se itera entre ellos hasta obtener convergencia, es decir hasta que la diferencia $\left\|\widehat{\boldsymbol{\theta}}^{(i+1)} - \widehat{\boldsymbol{\theta}}^{(i)}\right\|$ sea suficientemente pequeña.

Puede demostrarse (Dempster, Laird y Rubin, 1977), véase el apéndice 11.1, que este algoritmo maximiza $L(\boldsymbol{\theta}|\mathbf{Y})$. Además, la verosimilitud aumenta en cada iteración, aunque la convergencia puede ser muy lenta.

A continuación presentamos dos ejemplos de utilización del algoritmo. En el primero, la función de verosimilitud completa es lineal en los datos ausentes, con lo que les sustituimos por sus estimaciones. En el segundo, los valores ausentes aparecen de forma no lineal en la verosimilitud, y sustituiremos estas funciones por sus estimaciones.

11.2.2 Estimación MV de mezclas

Para ilustrar el comportamiento del algoritmo EM vamos a considerar un problema simple de estimación de mezclas que abordaremos con más generalidad en el capítulo 14. Supondremos que los datos de una muestra, $\mathbf{x}_1, ..., \mathbf{x}_n$ se generan mediante la distribución

$$\pi_1 f_1(\mathbf{x}) + (1-\pi_1) f_2(\mathbf{x})$$

donde $f_i(\mathbf{x})$ es $\mathrm{N}_p(\boldsymbol{\mu}_i, \mathbf{V}_i)$, $i = 1, 2$. La función soporte para la muestra es

$$L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^{n} \log(\pi_1 f_1(\mathbf{x}_i) + (1-\pi_1) f_2(\mathbf{x}_i))$$

donde $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{V}_1, \mathbf{V}_2, \pi_1)$ es el vector de parámetros. La estimación MV de los parámetros es complicada, porque tenemos que resolver las ecuaciones:

$$\frac{\partial L(\boldsymbol{\theta}|\mathbf{X})}{\partial \pi_1} = \sum_{i=1}^{n} \frac{f_1(\mathbf{x}_i) - f_2(\mathbf{x}_i)}{\pi_1 f_1(\mathbf{x}_i) + (1-\pi_1) f_2(\mathbf{x}_i)} = 0 \tag{11.2}$$

Para interpretar esta ecuación llamemos

$$\pi_{1i} = \frac{\pi_1 f_1(\mathbf{x}_i)}{\pi_1 f_1(\mathbf{x}_i) + (1-\pi_1) f_2(\mathbf{x}_i)}$$

a la probabilidad a posteriori de que la observación i sea generada por la primera población. Entonces

$$1 - \pi_{1i} = \frac{(1-\pi_1) f_2(\mathbf{x}_i)}{\pi_1 f_1(\mathbf{x}_i) + (1-\pi_1) f_2(\mathbf{x}_i)}$$

y la ecuación (11.2) puede escribirse:

$$\sum_{i=1}^{n} \left(\frac{\pi_{1i}}{\pi_1} - \frac{1-\pi_{1i}}{(1-\pi_1)}\right) = 0$$

que equivale a

$$\sum_{i=1}^{n} (\pi_{1i} - \pi_1) = 0$$

Es decir

$$\widehat{\pi}_1 = \frac{\sum_{i=1}^{n} \widehat{\pi}_{1i}}{n}$$

Esta ecuación indica que la probabilidad estimada de pertenencia a la primera población debe ser igual al promedio de las probabilidades estimadas de que cada observación pertenezca a esa población. Desgraciadamente no puede aplicarse directamente porque para calcular las

$\widehat{\pi}_{1i}$ necesitamos todos los parámetros del modelo. Derivando la función soporte respecto a $\boldsymbol{\mu}_1$:

$$\frac{\partial L(\boldsymbol{\theta}|\mathbf{X})}{\partial \boldsymbol{\mu}_1} = \sum_{i=1}^{n} \frac{\pi_1 f_1(\mathbf{x}_i)\mathbf{V}_1^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1)}{\pi_1 f_1(\mathbf{x}_i) + (1-\pi_1) f_2(\mathbf{x}_i)} = 0$$

que puede escribirse como

$$\sum_{i=1}^{n} \widehat{\pi}_{1i}(\mathbf{x}_i - \boldsymbol{\mu}_1) = 0$$

de donde obtenemos:

$$\boxed{\widehat{\boldsymbol{\mu}}_1 = \sum_{i=1}^{n} \frac{\widehat{\pi}_{1g}}{\sum_{i=1}^{n} \widehat{\pi}_{1i}} \mathbf{x}_i.} \tag{11.3}$$

que indica que la media de la primera población se estima dando un peso a cada observación proporcional a la probabilidad relativa de pertenecer a esta población. El mismo resultado se obtiene por simetría para $\widehat{\boldsymbol{\mu}_2}$ intercambiando $\widehat{\pi}_{1g}$ por $\widehat{\pi}_{2g} = 1 - \widehat{\pi}_{1g}$. Análogamente, derivando respecto a $\mathbf{V}_1$ puede demostrarse que el estimador es:

$$\boxed{\widehat{\mathbf{V}}_1 = \sum_{i=1}^{n} \frac{\widehat{\pi}_{1g}}{\sum_{i=1}^{n} \widehat{\pi}_{1g}} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_1)'} \tag{11.4}$$

que tiene un interpretación similar, como promedio de desviaciones de los datos respecto a sus medias, con pesos proporcionales a las probabilidades a posteriori.

Para resolver estas ecuaciones y obtener los estimadores necesitamos las probabilidades $\widehat{\pi}_{1i}$, y para calcular estas probabilidades con (15.10) necesitamos los parámetros del modelo. Por otro lado vemos que si las observaciones estuviesen clasificadas como viniendo de una u otra población el problema es muy simple, porque entonces $\widehat{\pi}_{1i}$ es uno, si la i proviene de la primera población o cero, si viene de la segunda, y las fórmulas (11.3) y (11.4) se reducen a aplicar las fórmulas de estimación habituales a las observaciones de cada grupo. Intuitivamente, podríamos partir de una asignación, estimar los parámetros y calcular las probabilidades $\widehat{\pi}_{1i}$ e iterar entre ambas etapas y esta es la solución que se obtiene con el algoritmo EM.

Como la estimación es muy simple si tenemos bien clasificadas las observaciones vamos a aplicar el algoritmo EM introduciendo $2n$ variables de clasificación que van a indicar de que población proviene cada dato muestral y que consideraremos como datos ausentes. Las primeras n variables z_{1i}, $i = 1, ..., n$ se definen mediante :

$$\begin{aligned} z_{1i} &= 1, \quad \text{si } \mathbf{x}_i \text{ proviene de } f_1 \\ &= 0, \quad \text{si } \mathbf{x}_i \text{ proviene de } f_2 \end{aligned}$$

y analogamente z_{2i} se define para que tome el valor uno si $\mathbf{x}_i$ proviene de f_2, de manera que $z_{1i} + z_{2i} = 1$. Para escribir la verosimilitud completa de las variables $\mathbf{x}$ y de las $\mathbf{z}$ tenemos en cuenta que, llamando $\boldsymbol{\theta}$ al vector de parámetros, para una observación cualquiera:

$$f(\mathbf{x}_1, z_{11}, z_{21}|\boldsymbol{\theta}) = f(\mathbf{x}_1/z_{11}, z_{21}\boldsymbol{\theta}) p(z_{11}, z_{21}|\boldsymbol{\theta})$$

donde

$$f(\mathbf{x}_1|z_{11}, z_{21}\boldsymbol{\theta}) = f_1(\mathbf{x}_1)^{z_{11}} f_2(\mathbf{x}_1)^{z_{21}}$$

ya que si $z_{11} = 1$ la dato $\mathbf{x}_1$ proviene de f_1 y entonces forzosamente $z_{21} = 0$ y viceversa. La probabilidad de los valores z es:

$$p(z_{11}, z_{21}|\boldsymbol{\theta}) = \pi_1^{z_{11}}(1-\pi_1)^{z_{21}}$$

ya que la probabilidad de $z_{11} = 1$ (en cuyo caso $z_{21} = 0$) es π_1. Uniendo ambas ecuaciones podemos escribir

$$\log f(\mathbf{x}_1, z_{11}, z_{21}|\boldsymbol{\theta}) = z_{11}\log\pi_1 + z_{11}\log f_1(\mathbf{x}_1) + z_{21}\log(1-\pi_1) + z_{21}\log f_2(\mathbf{x}_1)$$

y, para toda la muestra, llamando $\mathbf{X} = (\mathbf{x_1}, ..., \mathbf{x_n})$ e incluyendo las variables de clasificación $\mathbf{Z} = (z_{11}, ..., z_{1n}, z_{21}, ..., z_{2n})$ es

$$L(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \sum z_{1i}\log\pi_1 + \sum z_{1i}\log f_1(\mathbf{x}_i) + \sum z_{2i}\log(1-\pi_1) + \sum z_{2i}\log f_2(\mathbf{x}_i) \quad (11.5)$$

Para aplicar el algoritmo EM primero necesitamos una estimación inicial de los parámetros. Esto puede hacerse representando gráficamente los datos en un diagrama de dispersión bivariante y diviendo en función de ese gráfico los datos en dos grupos. Supongamos que tomamos como grupo 1 el de menor dispersión aparente. Entonces definimos unos valores iniciales para las variables z, que llamaremos $z^{(1)}$ de manera que $z_{1i}^{(1)} = 1$ si la observación $\mathbf{x}_i$ se clasifica en la primera problación (entonces $z_{2i}^{(1)} = 0$) y cero si se clasifica en la segunda (entonces $z_{2i}^{(1)} = 1$). Una vez definidas estas variables de clasificación estimaremos las medias mediante

$$\widehat{\boldsymbol{\mu}}_1^{(1)} = \frac{\sum z_{1i}^{(1)}\mathbf{x}_i}{\sum z_{1i}^{(1)}} \quad (11.6)$$

y lo mismo para $\widehat{\boldsymbol{\mu}}_2^{(1)}$. Analogamente, estimaremos la matriz de varianzas y covarianzas mediante

$$\widehat{\mathbf{V}}_1^{(1)} = \frac{1}{\sum z_{1i}^{(1)}} \sum_{i=1}^{n} z_{1i}^{(1)}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_1^{(1)})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_1^{(1)})', \quad (11.7)$$

que es simplemente la matriz de covarianzas muestrales de las observaciones de ese grupo. Finalmente, estimaremos la probabilidad de que un dato pertenezca al grupo uno por la proporción de datos en ese grupo:

$$\widehat{\pi}_1^{(1)} = \frac{\sum z_{1i}^{(1)}}{n} \quad (11.8)$$

A continuación tomamos esperanzas en la distribución conjunta (11.5) respecto a la distribución de los z supuesto este valor inicial $\widehat{\boldsymbol{\theta}}^{(1)}$ de los parámetros. Como las variables z aparecen linealmente, esto se reduce a calcular sus esperanzas y sustituirlas en la ecuación (11.5). Las

1	2	3	4	5	6	7	8	9	10
0.837	-0.722	-0.722	-0.201	-0.021	0.279	2.143	4.382	4.219	0.337
-0.655	-1.081	-0.048	0.379	-0.330	-0.500	3.530	5.355	2.324	1.623
11	12	13	14	15	16	17	18	19	20
2.408	0.595	6.925	3.680	-1.265	-0.538	6.351	5.289	4.868	-2.191
2.992	1.310	4.634	3.565	0.264	1.052	3.896	2.549	2.556	-0.414

Tabla 11.1: Datos simulados, los 6 primeros de una distribución y los 14 siguientes de otra

variables z son binomiales puntuales, y su esperanza coincide con la probabilidad de que tomen el valor uno. Por tanto:

$$\widehat{z_{1i}} = E(z_{1i}|\widehat{\boldsymbol{\theta}}^{(1)}, \mathbf{x}_i) = P(z_{1i} = 1|\widehat{\boldsymbol{\theta}}^{(1)}, \mathbf{x}_i)$$

y esta probabilidad se calcula mediante el teorema de Bayes

$$P(z_{1i} = |1\widehat{\boldsymbol{\theta}}^{(1)}, \mathbf{x}_i) = \frac{\widehat{\pi}_1^{(1)} f_1(\mathbf{x}_1|\widehat{\boldsymbol{\theta}}^{(1)})}{\widehat{\pi}_1^{(1)} f_1(\mathbf{x}_1|\widehat{\boldsymbol{\theta}}^{(1)}) + (1-\widehat{\pi}_1^{(1)}) f_2(\mathbf{x}_1|\widehat{\boldsymbol{\theta}}^{(1)})} \tag{11.9}$$

Una vez obtenidos los valores $\widehat{z_{ji}}$ los sustituiremos en la función de verosimilitud (11.5) y la maximizaremos respecto a los parámetros. Esto conduce a resolver las ecuaciones (11.6), (11.7) y (11.8) pero sustituyendo ahora las $z_{ij}^{(1)}$ por las estimaciones $\widehat{z_{ji}}$. Observemos que ahora las $\widehat{z_{ji}}$ ya no serán valores cero o uno, y la fórmula (11.6) ya no calcula la media de las observaciones de un grupo sino que hace una media ponderada de todas las observaciones con peso proporcional a la probabilidad de pertenecer al grupo. Esto propocionará otro nuevo estimador $\widehat{\boldsymbol{\theta}}^{(2)}$ que, mediante (11.9) conducirá a nuevos valores de las $\widehat{z_{ji}}$, y el proceso se itera hasta la convergencia.

Ejemplo 11.1 *Vamos a ilustrar el funcionamiento del algoritmo EM para estimar distribuciones normales con datos simulados. Hemos generado 20 observaciones de una variable bidimensional de acuerdo con el modelo* $.3N(0, I) + .7N(\mu, V)$ *donde* $\mu = (2, 2)'$ *y* $V = \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}$. *Los datos generados, donde los seis primeros provienen de la primera mezcla y los 14 siguiente de la segunda se presentan en la tabla 11.1*

Para obtener una estimación inicial de los parámetros, consideramos el histograma de cada variable.

Figura 11.1: Histograma de la primera variable de la mezcla de normales

El histograma de la primera variable representado en la figura (11.1) indica que los datos parecen ser mezcla de dos poblaciones con medias (0, 4) y similar variabilidad. Las desviaciones típicas de las poblaciones sobre esta variable son del orden de uno. El histograma de la segunda, figura (11.2), parece de nuevo también una mezcla, aunque dadas las pocas observaciones no es muy claro. Las medias parecen ser (0,3) pero ahora parece haber mas variabilidad en la segunda variables que en la primera. El diagrama de dispersión de las variables de la figura (**??**) indica dos grupos y correlación entre las variables. A primera vista este gráfico de dispersión sugiere dos poblaciones, la primera con 11 elementos y media próxima al punto (0,0) y la segunda con nueve elementos y media alrededor del punto (4,3). Uniendo la información univariante y bivariante vamos a tomar como estimación inicial $\widehat{\boldsymbol{\mu}}_1^{(1)} = (0,0)'$, $\widehat{\boldsymbol{\mu}}_2^{(1)} = (4,3)'$ y matrices de covarianzas $\widehat{\mathbf{V}}_1^{(1)} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$ y $\widehat{\mathbf{V}}_2^{(1)} = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$. Para las proporciones tomaremos la estimación inicial más simple $\pi_1 = \pi_2 = .5$.

Una asignación mejor sería clasificar las varaibles en los grupos y estimar a partir de esa clasificación los parámetros pero vamos a tomar una estimación inicial rápida para ilustrar como funciona el algoritmo con estimaciones iniciales no muy precisas.

Figura 11.2: Histograma de la segunda variable en la mezcla de normales

Diagrama de dispersión de las dos variables.

La aplicación del algoritmo EM se resume en la tabla siguiente. Se indican las iteraciones, el valor de π_1, y las medias estimadas de cada variable en cada iteración.

$iter$	$\widehat{\pi}_1$	$\widehat{\mu}_{11}$	$\widehat{\mu}_{12}$	$\widehat{\mu}_{21}$	$\widehat{\mu}_{22}$
1	0.5551	−0.3021	0.1780	4.4961	3.4869
2	0.5488	−0.3289	0.1435	4.4614	3.4826
3	0.5479	−0.3308	0.1408	4.4551	3.4798
4	0.5478	−0.3311	0.1404	4.4536	3.4791
5	0.5477	−0.3311	0.1403	4.4533	3.4790
6	0.5477	−0.3311	0.1403	4.4532	3.4789

Se observa que la convergencia se alcanza bastante rápido, y que los resultados obtenidos son consistentes con los datos del gráfico de dispersión de las variables. En efecto, once

observaciones son clasificadas en el primer grupo y nueve en el segundo. Las probabilidades a posterior de cada observación de pertenecer al grupo 1 son (0.9999 1.0000 1.0000 0.9998 1.0000 1.0000 0.0003 0.0000 0.0000 0.9725 0.0008 0.9828 0.0000 0.0000 1.0000 0.9989 0.0000 0.0000 0.0000 1.0000). Esto es consecuencia de que algunas observaciones generadas por el grupo 2 han aparecido muy próximas a las del grupo uno y, en consecuencia, se han marcado como provenientes del grupo 1. La estimación final de las matrices de covarianzas es

$V_1 = \begin{bmatrix} 0.7131 & 0.1717 \\ 0.1717 & 0.6841 \end{bmatrix}$ y $V_2 = \begin{bmatrix} 2.3784 & 0.4209 \\ 0.4209 & 0.9414 \end{bmatrix}$.

Se ha comprobado que esta solución no parece depender de los valores iniciales. Comenzando con $V_1 = V_2 = I$ se obtiene la misma solución y si tomamos como valores iniciales los exactos utilizados para generar los datos se obtiene de nuevo este resultado. El problema es que esta estimación es consistente con los datos, y dado el pequeño tamaño muestral la precisión de los estimadores es baja. Si repetimos el problema con $n = 100$ los parámetros obtenidos se aproximan mucho más a los verdaderos, pero la convergencia es muy lenta y hacen falta más de 50 iteraciones para alcanzarla.

11.2.3 Estimación de poblaciones normales con datos ausentes

Vamos a aplicar el algoritmo EM para estimar los parámetros de una distribución normal multivariante cuando disponemos de observaciones ausentes. La función de verosimilitud para una muestra sin valores ausentes puede escribirse, según (10.4), en función de las observaciones

$$L(\boldsymbol{\mu}, \mathbf{V}|\mathbf{X}) = -\frac{n}{2}\log|\mathbf{V}| - \frac{1}{2}tr(\mathbf{V}^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i') - \frac{n}{2}\boldsymbol{\mu}'\mathbf{V}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}'\mathbf{V}^{-1}\sum_{i=1}^{n}\mathbf{x}_i \qquad (11.10)$$

y sabemos que la estimación MV cuando tenemos toda la muestra es $\widehat{\boldsymbol{\mu}} = \sum_{i=1}^{n}\mathbf{x}_i/n$ y $\widehat{\mathbf{V}} = \mathbf{S} = \sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{x_i}\mathbf{x_i'}/\mathbf{n} - \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}'$.

Supongamos ahora que los vectores de observaciones $\mathbf{x}_1, ..., \mathbf{x}_m$ $(m < n)$ están completos, pero que los vectores $\mathbf{x}_{m+1} = \mathbf{z}_1, ..., \mathbf{x}_n = \mathbf{z}_{n-m}$ carecen de los valores de algunas variables (o de todas ellas). Con la notación de la sección anterior, sea $\mathbf{Y}$ el conjunto de datos disponibles y $\mathbf{Z}$ las variables ausentes. La función de verosimilitud completa viene dada por (11.10). Para aplicar el algoritmo EM, comenzaremos calculando un estimador inicial con los datos disponibles, y sean $\widehat{\boldsymbol{\mu}}^{(0)}$ y $\widehat{\mathbf{V}}^{(0)}$ estos estimadores iniciales. Tomamos $\widehat{\boldsymbol{\mu}}^{(i)} = \widehat{\boldsymbol{\mu}}^{(0)}$ y $\widehat{\mathbf{V}}^{(i)} = \widehat{\mathbf{V}}^{(0)}$ y iteraremos entre los dos pasos siguientes:

1. *Paso E.* Hay que calcular la esperanza de la función de verosimilitud completa (11.10) respecto a la distribución de los datos faltantes $\mathbf{Z}$, dados los parámetros $\widehat{\boldsymbol{\theta}}^{(i)}$=($\widehat{\boldsymbol{\mu}}^{(i)}$, $\widehat{\mathbf{V}}^{(i)}$) y los datos observados $\mathbf{Y}$. En esta función los datos faltantes aparecen en dos términos. El primero es $\boldsymbol{\mu}'\mathbf{V}^{-1}\sum_{i=1}^{n}\mathbf{x}_i$, y allí aparecen de forma lineal, por lo que tendremos simplemente que sustituir los datos ausentes por sus estimaciones. El segundo es $tr(\mathbf{V}^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i')$, y aquí tendremos que sustituir las expresiones $\mathbf{x}_i\mathbf{x}_i'$ por sus estimaciones. Comencemos con el primer término, tomar esperanzas de dados los parámetros y los datos conocidos implica sustituir $\mathbf{x}_i$ para $i > m$ por $E(\mathbf{x}_i/\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(i)})$. El calculo de esta esperanza se realiza como sigue:

(a) Si el vector $\mathbf{x}_i$ es completamente inobservado, es decir, no se ha observado ninguna variable para ese elemento, entonces $E(\mathbf{x}_i/\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(i)}) = \widehat{\boldsymbol{\mu}}^{(i)}$, no depende de los datos observados. Puede comprobarse que, finalmente, esta sustitución es equivalente a desechar completamente esta observación, lo que resulta intuitivo. Si no observamos en un elemento ninguna variable es equivalente a no tomarlo en la muestra.

(b) Si el vector $\mathbf{x}_i = [\mathbf{x}'_{1i}\ \mathbf{x}'_{2i}]'$ se observa parcialmente, de manera que no conocemos los valores de ciertas variables $\mathbf{x}_{1i}$, pero si hemos observado los valores de otras $\mathbf{x}_{2i}$,entonces $E(\mathbf{x}_i/\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(i)})$ depende de los valores observados de $\mathbf{x}_{2i}$ y será igual a la esperanza condicionada $E(\mathbf{x}_{1i}/\mathbf{x}_{2i}, \widehat{\boldsymbol{\theta}}^{(i)})$. Esta esperanza se calcula, según la sección 8.5.1, por regresión mediante :

$$E(\mathbf{x}_{1i}/\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(i)}) = E(\mathbf{x}_{1i}/\mathbf{x}_{2i}, \widehat{\boldsymbol{\theta}}^{(i)}) = \widehat{\mathbf{x}}^{(i)}_{1i.2} = \widehat{\boldsymbol{\mu}_1}^{(i)} + \widehat{\mathbf{V}}^{(i)}_{12}\widehat{\mathbf{V}}^{(i)}_{22}(\mathbf{x}_{2i} - \widehat{\boldsymbol{\mu}_2}^{(i)}) \tag{11.11}$$

donde hemos particionado el vector de medias y la matriz de covarianzas con relación a los dos bloques de variables.

Para calcular la esperanza del segundo termino, observemos primero que $E\left[tr(\mathbf{V}^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}'_i)\right] = tr\left[E(\mathbf{V}^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}'_i)\right] = tr\left[\mathbf{V}^{-1}\sum_{i=1}^{n}E(\mathbf{x}_i\mathbf{x}'_i)\right]$. Por tanto tenemos que obtener las esperanzas $E(\mathbf{x}_i\mathbf{x}'_i/\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(i)})$ para $i > m$. Consideremos, como antes, los dos casos siguientes:

(a) Si el vector es completamente inobservado, $E(\mathbf{x}_i\mathbf{x}'_i/\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(i)}) = \widehat{\mathbf{V}}^{(i)} - \widehat{\boldsymbol{\mu}}^{(i)}\widehat{\boldsymbol{\mu}}^{(i)\prime}$, y, de nuevo, puede comprobarse que esto va a ser equivalente a desechar completamente esta observación.

(b) Si el vector $\mathbf{x}_i$ se observa parcialmente y no conocemos los valores de $\mathbf{x}_{1i}$ pero si los de $\mathbf{x}_{2i}$, utilizaremos la relación

$$E(\mathbf{x}_{1i}\mathbf{x}'_{1i}/\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(i)}) = E(\mathbf{x}_{1i}\mathbf{x}'_{1i}/\mathbf{x}_{2i}, \widehat{\boldsymbol{\theta}}^{(i)}) = \widehat{\mathbf{V}}^{(i)}_{11.2} + \widehat{\mathbf{x}}^{(i)}_{1i.2}\widehat{\mathbf{x}}^{(i)\prime}_{1i.2} \tag{11.12}$$

donde $\widehat{\mathbf{V}}^{(i)}_{11.2}$ es la matriz de varianzas de la variable $\mathbf{x}_{1i}$ dado $\mathbf{x}_{2i}$. Según la sección 8.5.1 esta varianza condicionada viene dada por

$$\widehat{\mathbf{V}}^{(i)}_{11.2} = \widehat{\mathbf{V}}^{(i)}_{11} - \widehat{\mathbf{V}}^{(i)}_{12}\widehat{\mathbf{V}}^{(i)-1}_{22}\widehat{\mathbf{V}}^{(i)}_{21}. \tag{11.13}$$

que podemos calcular a partir de $\widehat{\mathbf{V}}^{(i)}$ y sustituir en (11.12).

2. *Paso M.* En la función de verosimilitud completa (11.10) reemplazamos las funciones de los valores ausentes por sus estimaciones (11.11) y (11.12) y calculamos los nuevos estimadores de máxima verosimilitud, que vendrán dados por

$$\widehat{\boldsymbol{\mu}}^{(i+1)} = \frac{\sum_{i=1}^{n}\widehat{\mathbf{x}}^{(i)}_i}{n}$$

donde en $\widehat{\mathbf{x}}^{(i)}_i$ los valores observados no se modifican y los no observados se han sustituido por sus esperanzas condicionales (11.11). La estimación de $\widehat{\mathbf{V}}^{(i+1)}$ será

$$\widehat{\mathbf{V}}^{(i+1)} = \sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{E}(\mathbf{x_i}\mathbf{x_i'}/\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(\mathbf{i})})/\mathbf{n} - \widehat{\boldsymbol{\mu}}^{(i+1)}\widehat{\boldsymbol{\mu}}^{(i+1)\prime}$$

donde las esperanzas de los valores observados son ellos mismos y las de los faltantes vienen dadas por (11.12) y (11.13).

Con los valores estimados en el paso M volvemos al E, haciendo $\widehat{\boldsymbol{\mu}}^{(i)} = \widehat{\boldsymbol{\mu}}^{(i+1)}$ y $\widehat{\mathbf{V}}^{(i)} = \widehat{\mathbf{V}}^{(i+1)}$. El algoritmo finaliza cuando el cambio en los parámetros de una iteración a la siguiente es menor que un valor pequeño, como .001. A continuación, presentamos un ejemplo de su funcionamiento.

Ejemplo 11.2 *Vamos a ilustrar el funcionamiento del algoritmo EM con los diez primeros datos de las variables estatura y peso de la base de datos MEDIFIS. Supondremos que en las tres primeras personas en la muestra no se ha observado la variable peso. Llamando x_1 a esta variable, la muestra es:* $x_1 = (*, *, *, 52, 51, 67, 48, 74, 74, 50)$, *donde el signo * indica que el valor está ausente. Sin embargo, suponemos que se han observado los diez valores de la variable estatura:* $x_2 =$ *(159, 164, 172, 167, 164, 161, 168, 181, 183, 158)*

Ejercicio 11.1 *Para comenzar el algoritmo obtenemos una estimación inicial del vector de medias con los diez datos de x_2 y los siete de x_1. Este vector es $\widehat{\boldsymbol{\mu}}^{(0)} =$ (59.43, 167.7). Con las siete parejas de datos completos calculamos la matriz de covarianzas* $\widehat{\mathbf{V}}^{(0)} = \begin{bmatrix} 118.24 & 70.06 \\ 70.06 & 79.26 \end{bmatrix}$. *Con estos parámetros iniciamos el paso E, calculo de las esperanzas condicionadas. La esperanza condicionada (regresión) de la primera variable en la segunda es*

$$E(x_1/x_2) = 59.43 + 70.06/79.26(x_2 - 167.7)$$

es decir, el peso se prevé con la recta de regresión entre peso y estatura cuyo coeficiente de regresión es $70.06/79.26 = .8839$. *Aplicándolo a los valores faltantes*

$$E\begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \end{bmatrix} = 59.43 + 70.06/79.26 \begin{bmatrix} 159 - 167.7 \\ 164 - 167.7 \\ 172 - 167.7 \end{bmatrix} = \begin{bmatrix} 51.738 \\ 56.158 \\ 63.229 \end{bmatrix}$$

Después de esta primera estimación de los valores ausentes, estimaremos los productos cruzados. Los productos de la primera variable por la segunda son :

$$E(x_{1i}x_{2i}/\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(i)}) = x_{2i}E(x_{1i}/x_{2i}, \widehat{\boldsymbol{\theta}}^{(i)})$$

que serán 159×51.738, *para i=1,* 164×56.158, *para i=2, y* 172×63.229, *para i=3. Los cuadrados de la variable ausente se estima por*

$$E(x_{1i}^2/\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(i)}) = \widehat{\mathbf{V}}_{11.2}^{(i)} + \widehat{\mathbf{x}}_{1i.2}^{(i)2}$$

donde $\widehat{\mathbf{V}}_{11.2}^{(i)}$ *es la varianza residual de la regresión entre el peso y la estatura dada por* $118.24 - 70.06^2/79.26 = 56.31$. *Por tanto para* $i = 1, 2, 3$ *los valores serán* $56.31 + 51.738^2 = 2733.13$, $56.31 + 56.158^2 = 3210.03$, $56.31 + 63.229 = 4.0542$.

Con estas estimaciones pasamos al paso M. En el calculo de la media, la nueva estimación será $\widehat{\boldsymbol{\mu}}^{(i)} =$ *(58.71, 167.7) donde ahora la primera componente se calcula con diez datos,*

N^a	$\widehat{x}_{11}$	$\widehat{x}_{12}$	$\widehat{x}_{13}$	$\widehat{\mu}_1$	s_{12}	s_1^2	ss_1^2
1	51.739	56.159	63.229	58.713	58.577	107.582	90.686
2	52.283	55.978	61.890	58.615	57.594	108.300	89.012
3	52.294	55.927	61.740	58.596	57.540	108.865	88.929
4	52.281	55.910	61.717	58.591	57.547	109.070	88.941
5	52.275	55.905	61.713	58.589	57.553	109.135	88.948
6	52.272	55.903	61.711	58.589	57.555	109.156	88.9515
7	52.272	55.902	61.711	58.588	57.556	109.162	88.9525
8	52.271	55.902	61.711	58.588	57.556	109.164	88.953
9	52.271	55.902	61.711	58.588	57.556	109.164	88.953

Tabla 11.2: Estimaciones del algoritmo EM en las distintas iteraciones

sustituyendo los ausentes por sus estimaciones. Para calcular la matriz de covarianzas se utiliza la expresión

$$\widehat{\mathbf{V}}^{(i)} = \frac{1}{10}\begin{bmatrix} \sum \widehat{x}_{1i}^2 & \sum \widehat{x}_{1i}x_{2i} \\ \sum \widehat{x}_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} - \begin{bmatrix} 58.71^2 & 58.71 \times 167.7 \\ 58.71 \times 167.7 & 167.7^2 \end{bmatrix}$$

donde los valores $\widehat{x}_{1i}$ son los observados (para i=4,...,10) o las estimaciones de los ausentes (para i=1,2,3) y análogamente para $\widehat{x}_{1i}^2$. Observemos que la estimación de $\widehat{x}_{1i}^2$ NO es la estimación de $\widehat{x}_{1i}$ elevada al cuadrado sino que además se le añade el valor de la varianza residual como hemos visto.

La tabla 11.2siguiente indica la evolución de las estimaciones proporcionadas por el algoritmo hasta la convergencia para los valores ausentes, la media de la primera variable, la covarianza y la varianza de la primera variable. El algoritmo converge en nueve iteraciones. Se ha añadido una columna adicional , ss_1^2,varianza de la primera variable, para ilustrar las estimaciones que se obtendrían si en lugar de utilizar el algoritmo EM utlizamos el método más simple de modificar el paso E sustituyendo cada observación faltante por su media condicionada e iterando después. Se observa que este segundo procedimiento al no tener en cuenta toda la incertidumbre subestima las varianzas: la varianza con este método es siempre menor que la estimada por el algoritmo EM.

11.3 ESTIMACIÓN ROBUSTA

La estimación MV depende de la hipótesis de normalidad en los datos. Esta es una hipótesis fuerte, y difícil de comprobar con muestras de tamaño mediano. En particular, la estimación MV de los parámetros suponiendo normalidad puede ser muy mala cuando los datos provengan de distribuciones con colas pesadas, que pueden generar valores atípicos. Supongamos por ejemplo que los datos provienen de una normal contaminada

$$\pi_1 N_1(\boldsymbol{\mu}, \mathbf{V}) + (1-\pi_1)N_2(\boldsymbol{\mu}, c\mathbf{V})$$

donde la mayoría de los datos, por ejemplo, $\pi_1 = .95$, se hab generado con la distribución central, $N_1(\boldsymbol{\mu}, \mathbf{V})$, pero una pequeña proporción $(1-\pi_1)$, por ejemplo el 5%, provienen de la

distribución alternativa, que tiene mayor variabilidad, tanto más cuanto mayor sea c, que es siempre mayor que uno de manera que los elementos generados por ella pueden ser atípicos y estar mucho más alejados del centro que los de la primera.

Hemos visto en la sección 11.2 , estimación de mezclas, que los estimadores MV de los parámetros se calculan como:

$$\widehat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} \widehat{z}_{1i}\mathbf{x}_i}{\sum_{i=1}^{n} \widehat{z}_{1i}}$$

y

$$\widehat{\mathbf{V}} = \frac{1}{\sum \widehat{z}_{1i}} \sum_{i=1}^{n} \widehat{z}_{1i}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})',$$

donde las variables $\widehat{z}_{1i}$ son estimaciones de la probabilidad de que la observación provenga de la primera población (el dato no sea atípico). Los métodos robustos parten de estas ecuaciones pero eligen los pesos $\widehat{z}_{1i}$ que se aplican a las observaciones de manera que el estimador resultante tenga buenas propiedades de robustez ante un conjunto amplio de distribuciones contaminantes, no necesariamente normales. Maronna (1976) propuso estimar iterativamente los parametros de la normal multivariante con estas ecuaciones pero sustituyendo los $\widehat{z}_{1i}$ por pesos $w_i(D_i)$ convenientemente elegidos en función de la distancia de Mahalanobis del punto al centro de los datos. Por ejemplo, $w_i(D_i)$ se toma igual a uno si la distancia es menor que una cierta constante y tiene a cero cuando la distancia crece. El proceso es iterativo y recuerda el algoritmo EM. Se parte de una estimación inicial de los parámetros, con ella se calculan las distancias de Mahalanobis y los pesos $w_i(D_i)$. Con estas estimaciones se reestiman los parámetros con los nuevos pesos, lo que conducirá a nuevas distancias de Mahalanobis, que determinarán nuevos pesos y así sucesivamente. Este método de modificar las ecuaciones de verosimilitud mediante pesos se conoce como M-estimación.

Aunque este procedimiento es atractivo, no funciona bien en dimensiones altas. Puede demostrarse que el punto de ruptura de un M-estimador como el que hemos presentado, que descuenta las observaciones extremas, es, como máximo $1/(p+1)$. Esta propiedad implica que en alta dimensión es necesario buscar un enfoque alternativo a los estimadores clásicos robustos. Hay dos enfoques al problema. El primero, buscar un estimador que se base sólo en una fracción de los datos, presumiblemente no contaminados. El segundo es eliminar los atípicos, y construir el estimador a partir de los datos limpios de atípicos.

Con el primer enfoque un procedimiento simple es el introducido por Rousseeuw (1985), que propone calcular el elipsoide de mínimo volumen, o de mínimo determinante, que engloba al menos el 50% de los datos. La justificación intuitiva del método es la siguiente. Los datos atípicos estarán en los extremos de la distribución, por lo que podemos buscar una zona de alta concentración de puntos y determinar con ellos el centro de los datos y la matriz de covarianzas, ya que los puntos de esa zona serán presumiblemente puntos buenos. Para encontrar ese núcleo central con alta densidad de datos, exigimos que el ellipsoide que cubre al menos el 50% de los datos tenga volumen mínimo. Esta idea es una generalización de los resultados univariantes, donde se obtienen estimadores muy robustos a partir de la idea de mediana. Por ejemplo, la mediana es una medida de centralización que se ve poco afectada

por una alta contaminación de los datos. Analogamente, para dispersión podemos utilizar la meda, o mediana de las desviaciones de los datos respcto a la mediana, que tiene también buenas propiedades. Generalizando estas ideas, podemos buscar el centro de la distribución de los datos multivariantes y su variabilidad construyendo el intervalo mínimo alrededor de un punto central que englobe el 50% de los datos. El centro de este intervalo será una estimación de la media y la matriz de covarianzas estimada en este intervalo, convenientemente escalada, estimará la matriz de varianzas de la problación.

Para obtener este intervalo, el proceso se implementa como sigue. Tomamos una muestra mínima de tamaño $p+1$ y calculamos su media, $\overline{\mathbf{x}}^{(1)}$, y su matriz de covarianzas, $\widehat{\mathbf{V}}^{(1)}$. A esta muestra se la llama mínima, porque tiene el número exacto de elementos que necesitamos para calcular un valor del vector de medias y de la matriz de covarianzas, donde suponemos que la matriz de covarianzas estimada resulta no singular (en otro caso se tomaría otra muestra mínima). A continuación, calculamos las distancias de Mahalanobis al centro de esta muestra mínima para todos los puntos de la muestra completa de n puntos:

$$D_i = (\mathbf{x}_i - \overline{\mathbf{x}}^{(1)})' \widehat{\mathbf{V}}^{(1)-1} (\mathbf{x}_i - \overline{\mathbf{x}}^{(1)}),$$

y tomamos la mediana, $m^{(1)}$, de estas n distancias. Entonces, por construcción, el elipsoide definido por $(\mathbf{x} - \overline{\mathbf{x}}^{(1)})' \widehat{\mathbf{V}}^{(1)-1} (\mathbf{x} - \overline{\mathbf{x}}^{(1)}) \leq m^{(1)}$ contiene el 50% de los datos, o, lo que es equivalente, el elipsoide definido por $(\mathbf{x} - \overline{\mathbf{x}}^{(1)})' (m^{(1)} \widehat{\mathbf{V}}^{(1)})^{-1} (\mathbf{x} - \overline{\mathbf{x}}^{(1)}) \leq 1$ contiene el 50% de los datos. El volumen de un elipsoide de este tipo es proporcional a

$$\left| m^{(1)} \widehat{\mathbf{V}}^{(1)} \right|^{1/2} = \left| \widehat{\mathbf{V}}^{(1)} \right|^{1/2} (m^{(1)})^{p/2}. \tag{11.14}$$

El procedimiento de calcular el elipsoide de volumen mínimo que engloba el 50% de los datos es tomar N muestras mínimas, obteniendo centros, $\overline{\mathbf{x}}^{(1)}, ..., \overline{\mathbf{x}}^{(N)}$, matrices de covarianzas, $\widehat{\mathbf{V}}^{(1)}, ..., \widehat{\mathbf{V}}^{(N)}$, y medianas, $m^{(1)}, ..., m^{(N)}$, y calcular en cada muestra mínima el volumen (11.14). La muestra mínima que proporcione un menor valor del criterio (11.14) se utiliza para calcular los estimadores robustos como sigue. Supongamos que la muestra de volumen mínimo es la muestra J. Entonces, el estimador robusto de la media de los datos es $\overline{\mathbf{x}}^{(J)}$, y la estimación de la matriz de covarianzas $m^{(J)} \widehat{\mathbf{V}}^{(J)}$se expande para que corresponda a una estimación de la matriz en la población. Como la distancia de Mahalanobis con respecto al centro de la problación con la matriz de la población es una χ^2_p , tenemos que, para muestras grandes $(\mathbf{x} - \overline{\mathbf{x}}^{(J)})' (m^{(J)} \widehat{\mathbf{V}}^{(J)})^{-1} (\mathbf{x} - \overline{\mathbf{x}}^{(J)})$ que contiene el 50% de los datos debe ser próximo a la mediana de la χ^2_p, que representaremos por $\chi^2_{p.50}$. Una estimación consistente de la matriz de covarianzas para datos normales es

$$\widehat{\mathbf{V}} = (\chi^2_{p.50})^{-1} m^{(J)} \widehat{\mathbf{V}}^{(J)}.$$

Un procedimiento alternativo, más rápido y eficiente que el método anterior, ha sido propuesto por Peña y Prieto (2001) basado en las ideas de proyecciones expuestas en el capítulo 3. El método consiste en tres etapas. En la primera se identifican los posibles atípicos como datos extremos de las proyecciones de la muestra sobre las direcciones que maximizan o minimizan la kurtosis de los puntos proyectados. En segundo lugar, se eliminan todos los

atípicos potenciales o puntos sospechosos, y llamando U al conjunto de observaciones no sospechosas, la estimación incial robusta de los parámetros se realiza con:

$$\begin{aligned}\tilde{\mathbf{m}} &= \frac{1}{|U|}\sum_{i\in U}\mathbf{x}_i,\\ \tilde{\mathbf{S}} &= \frac{1}{|U|-1}\sum_{i\in U}(\mathbf{x}_i-\tilde{\mathbf{m}})(\mathbf{x}_i-\tilde{\mathbf{m}})',\end{aligned}$$

En tercer lugar, utilizando estos estimadores robustos se contrastan una por una las observaciones sospechosas para ver si son atípicas. Como vimos en la sección 10.8 el contraste utiliza la distancia de Mahalanobis:

$$v_i = (\mathbf{x}_i-\tilde{\mathbf{m}})^T\tilde{\mathbf{S}}^{-1}(\mathbf{x}_i-\tilde{\mathbf{m}}), \qquad \forall i \notin U.$$

y aquellas observaciones $i \notin U$ tales que $v_i < T^2_{0.99}(p, n-1)$, donde $T^2_{0.99}(p, n-1)$ es el percentil .99 de la distribución de Hotelling las consideramos como aceptables y las incluimos en U. Cuando una nueva observación se incluye en U los parámetros se recalculan y el proceso se repite hasta que no se encuentran nuevas observaciones. Finalmente, una vez contrastados todos los puntos se estiman los parámetros utilizando los elementos que no se han considerado atípicos, y estos serán los estimadores robustos finales.

Este método se basa en los resultados de la sección 4.5 donde justificamos que los atípicos aislados van a identificarse buscando la dirección de máxima kurtosis y los grupos numerosos de atípicos van a aparecer en las direcciones de mínima kurtosis asociada a distribuciones bimodales.

11.4 ESTIMACIÓN BAYESIANA

11.4.1 Concepto

En el enfoque bayesiano un parámetro es una variable aleatoria y la inferencia respecto a sus posibles valores se obtiene aplicando el cálculo de probabilidades (teorema de Bayes) para obtener la distribución del parámetro condicionada a la información disponible. Si se desea un estimador puntual, se tomará la media o la moda de dicha distribución; si se desea un intervalo de confianza, se tomará la zona que encierre una probabilidad fijada en dicha distribución. En consecuencia, una vez obtenida la distribución de probabilidad del parámetro, los problemas habituales de inferencia quedan resueltos con la distribución a posteriori de manera automática y simple.

El enfoque bayesiano tiene dos ventajas principales. La primera es su generalidad y coherencia: conceptualmente todos los problemas de estimación se resuelven con los principios del cálculo de probabilidades. La segunda es la capacidad de incorporar información a priori con respecto al parámetro adicional a la muestral. Esta fortaleza es, sin embargo, también su debilidad, porque exige siempre representar la información inicial respecto al vector de parámetros mediante una *distribución inicial* o *a priori*, $p(\boldsymbol{\theta})$. Este es el aspecto más controvertido del método, ya que algunos científicos rechazan que la información inicial -que

puede incluir los prejuicios del investigador- se incluya en un proceso de inferencia científica. En principio esto podría evitarse estableciendo una distribución neutra, de referencia o no informativa para el problema, pero, aunque esto es factible en casos simples, puede ser en si mismo un problema complejo en el caso multivariante, como veremos a continuación.

La *distribución final* o *a posteriori* se obtiene mediante el teorema de Bayes. Si llamamos $\mathbf{X}$ a la matriz de datos, con distribución conjunta $f(\mathbf{X}|\boldsymbol{\theta})$, que proporciona las probabilidades de los valores muestrales conocido el vector de parámetros, la distribución a posteriori $p(\boldsymbol{\theta}|\mathbf{X})$ será:

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{f(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int f(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d(\boldsymbol{\theta})}. \tag{11.15}$$

El denominador de esta expresión puede escribirse como $m(\mathbf{X})$ la distribución marginal de los datos. Esta distribución se denomina *distribución predictiva* y se obtiene ponderando las distribuciones $f(\mathbf{X}|\boldsymbol{\theta})$ para cada posible valor del parámetro por las probabilidades que la distribución a priori asigna a estos valores.

En la práctica, el cálculo de (11.15) se simplifica observando que el denominador no depende de $\boldsymbol{\theta}$, y actúa únicamente como una constante normalizadora para que la integral de $p(\boldsymbol{\theta}|\mathbf{X})$ sea la unidad. Por tanto, podemos calcular la distribución posterior escribiendo:

$$p(\boldsymbol{\theta}|\mathbf{X}) = k\ell(\boldsymbol{\theta}|\mathbf{X})p(\boldsymbol{\theta}), \tag{11.16}$$

ya que, dada la muestra, $\mathbf{X}$ es constante y al considerar $f(\mathbf{X}|\boldsymbol{\theta})$ como función de $\boldsymbol{\theta}$ se convierte en la función de verosimilitud $\ell(\boldsymbol{\theta}|\mathbf{x})$. Multiplicando para cada valor de $\boldsymbol{\theta}$ las ordenadas de $\ell(\boldsymbol{\theta}|\mathbf{X})$ y $p(\boldsymbol{\theta})$ resulta la distribución posterior. Para la forma de la posterior la constante k es irrelevante, y siempre puede determinarse al final con la condición de que $p(\boldsymbol{\theta}|\mathbf{X})$ sea una función de densidad e integre a uno. Esta regla se resume en:

$$\boxed{\text{Posterior} \propto \text{Prior} \times \text{Verosimilitud}}$$

donde $\propto$ indica proporcional. La distribución a posteriori es un compromiso entre la prior y la verosimilitud. Cuando $p(\boldsymbol{\theta})$ es aproximadamente constante sobre el rango de valores en los que la verosimilitud no es nula, diremos que $p(\boldsymbol{\theta})$ es localmente uniforme o no informativa, y la posterior vendrá determinada por la función de verosimilitud.

Una ventaja adicional del enfoque bayesiano es su facilidad para procesar información secuencialmente. Supongamos que después de calcular (11.16) observamos una nueva muestra de la misma población $\mathbf{Y}$, independiente de la primera. Entonces, la distribución inicial será ahora $p(\boldsymbol{\theta}|\mathbf{X})$ y la distribución final será :

$$p(\boldsymbol{\theta}|\mathbf{XY}) = k\ell(\boldsymbol{\theta}|\mathbf{Y})p(\boldsymbol{\theta}|\mathbf{X}).$$

Naturalmente este mismo resultado se obtendría considerando una muestra ampliada $(\mathbf{X}, \mathbf{Y})$ y aplicando el teorema de Bayes sobre dicha muestra, ya que por la independencia de $\mathbf{X}$ e $\mathbf{Y}$:

$$p(\boldsymbol{\theta}|\mathbf{XY}) = k\ell(\boldsymbol{\theta}|\mathbf{XY})p(\boldsymbol{\theta}) = k\ell(\boldsymbol{\theta}|\mathbf{X})p(\boldsymbol{\theta}|\mathbf{Y})p(\boldsymbol{\theta})$$

La estimación bayesiana proporciona estimadores (la media de la distribución a posteriori) que son admisibles con criterios clásicos.

11.4.2 Distribuciones a priori

Una manera simple de introducir la información a priori en el análisis es utilizar distribuciones a *priori conjugadas,* que se combinan con la verosimilitud para producir distribuciones a posteriori simples, como veremos en la sección siguiente.

Si no se dispone de información a priori, o se desea que los datos hablen por sí mismos, se debe establecer una distribución a priori no informativa o de referencia. Intuitivamente, una distribución a priori no informativa para un vector de parámetros de localización es aquella que es localmente uniforme sobre la zona relevante del espacio paramétrico, y escribiremos $p(\boldsymbol{\theta}) = c$. Sin embargo, esta elección tiene el problema de que si el vector de parámetros puede tomar cualquier valor real $\int_{-\infty}^{\infty} p(\theta)d\theta = \infty$, y la prior no puede interpretarse como una distribución de probabilidad, sino como una herramienta para calcular la posterior. En efecto, si podemos suponer que a priori un parámetro escalar debe estar en el intervalo $(-h, h)$, donde h puede ser muy grande pero es un valor fijo, la distribución a priori $p(\theta) = 1/2h$ es propia, ya que integra a uno. La distribución $p(\boldsymbol{\theta}) = c$ debe pues considerarse como una herramienta simple para obtener la posterior. Estas distribuciones se denominan impropias. En problemas simples trabajar con distribuciones a priori impropias no produce problemas, (aunque puede dar lugar a paradojas, véase por ejemplo Bernardo y Smith, 1994), pero en situaciones un poco más complicadas la distribución a posteriori correspondiente puede no existir.

Las distribuciones constantes están sujetas a una dificultad conceptual adicional: si suponemos que la distribución a priori para un parámetro escalar θ es del tipo $p(\theta) = c$ y hacemos una transformación uno a uno del parámetro $\varphi = g(\theta)$, como

$$p(\varphi) = p(\theta) \left| \frac{d\theta}{d\varphi} \right|$$

si la distribución es constante para el parámetro θ, no puede ser constante para el parámetro φ. Por ejemplo, si $p(\theta) = c$, y $\varphi = 1/\theta$, entonces $|d\theta/d\varphi| = \varphi^{-2}$ y $p(\varphi) = c\varphi^{-2}$, que no es uniforme. Nos encontramos con la paradoja de que si no sabemos nada sobre θ y $\theta > 0$, no podemos decir que no sabemos nada (en el sentido de una distribución uniforme) sobre $\log\theta$ 0 θ^2. Una solución es utilizar las propiedades de invarianza del problema para elegir que transformación del parámetro es razonable suponer con distribución constante, pero aunque esto suele ser claro en casos simples (para las medias y para los logartimos de las varianzas) , no es inmediato cómo hacerlo para parámetros más complejos.

Jeffreys (1961), Box y Tiao (1973), Bernardo (1979) y Bernardo y Smith (1994), entre otros, han estudiado el problema de establecer distribuciones de referencia con propiedades razonables. Para distribuciones normales, y para los casos simples considerados en este libro, la distribución de referencia para un vector de parámetros de localización podemos tomarla como localmente uniforme y suponer que en la zona relevante para la inferencia $p(\boldsymbol{\theta}) = c$. Para matrices de covarianza, Jeffreys, por consideraciones de invarianza ante transformaciones, propuso tomar la distribución de referencia proporcional al determinante de la matriz de covarianzas elevado a $-(p+1)/2$, donde p es la dimensión de la matriz.

Señalaremos por último que el problema de la distribución a priori, aunque de gran importancia conceptual, no es tan crucial en la práctica como puede parecer a primera vista

ya que :

(1) Si tenemos muchos datos, la verosimilitud será muy apuntada, y la posterior vendrá determinada por la verosimilitud, ya que entonces cualquier priori razonable será casi constante sobre la zona relevante para la inferencia.

(2) Si tenemos poca información muestral, cualquier procedimiento estadístico va a ser muy sensible a las hipótesis que hagamos sobre el modelo de distribución de probabilidad, que van a afectar tanto o más que la prior al análisis. Sin embargo, estas hipótesis no podremos comprobarlas con eficacia con muestras pequeñas. Conviene en estos casos, sea cual sea la prior elegida, estudiar la sensibilidad de la solución a cambios en el modelo y en la prior.

11.4.3 Cálculo de la Posterior

Distribuciones Conjugadas

El cálculo de la distribución posterior puede ser complicado y requerir métodos numéricos. El problema se simplifica si podemos expresar aproximadamente nuestra información a priori con una distribución que simplifique el análisis. Una familia de distribuciones a priori adecuada para este objetivo es aquella con la misma forma que la verosimilitud, de manera que la posterior pueda calcularse fácilmente y pertenezca a la misma familia que la priori. A estas familias se las denomina *conjugadas*.

Una clase $\mathcal{C}$ de distribuciones a priori para un parámetro vectorial, $\boldsymbol{\theta}$, es conjugada, si cuando la prior pertenece a esa clase, $p(\boldsymbol{\theta}) \in \mathcal{C}$ entonces la posterior también pertence a la clase, $p(\boldsymbol{\theta}|X) \in \mathcal{C}$. La distribución conjugada puede elegirse tomando la distribución a priori con la forma de la verosimilitud. Por ejemplo, supongamos que queremos hacer inferencia respecto al parámetro θ en un modelo de la familia exponencial

$$f(X|\theta) = g(X)h(\theta)\exp\left\{t(X)g(\theta)\right\}.$$

La verosimilitud de la muestra será

$$l(\theta|X) = k \times h(\theta)^n \exp\left\{g(\theta)\sum t(X)\right\}$$

y podemos tomar como familia conjugada :

$$p(\theta) = k \times h(\theta)^\nu \exp\left\{g(\theta)m\right\},$$

con lo que se obtiene inmediatamente la posterior:

$$p(\theta|X) = k \times h(\theta)^{\nu+n} \exp\left\{g(X)\left[m + \sum t(X)\right]\right\}.$$

En la sección siguiente veremos ejemplos de su utilización para estimar los parámetros de una normal multivariante.

Métodos de Monte Carlo con Cadenas de Markov (MC^2)

Cuando no sea posible utilizar una distribución a priori conjugada y el cálculo de la posterior sea complejo, podemos utilizar el ordenador para generar muestras de la distribución posterior. Existe una variedad de métodos para realizar esta simulación, que se conocen bajo el nombre común de métodos de Monte Carlo con Cadenas de Markov (o métodos MC^2) y el lector interesado puede acudir a Robert y Casella(1999) , Carlin y Louis (1996) y Gaberman (1997). En este libro sólo presentaremos uno de estos métodos, el muestreo de Gibbs o Gibbs sampling, que es especialmente útil para la estimación de las distribuciones mezcladas consideradas en el capítulo 14.

El muestro de Gibbs es apropiado para obtener muestras de una distribución conjunta cuando es fácil muestrear de las distribuciones condicionadas. Supongamos que estamos interesados en obtener muestras de la distribución conjunta de dos variables aleatorias, $f(x, y)$, y supongamos que conocemos las distribuciones condicionadas $f(x/y)$, y $f(y/x)$. Este método se implementa como sigue :

1. Fijar un valor arbitrario inicial $y^{(0)}$ y obtener un valor al azar para x de la distribución $f(x/y^{(0)})$. Sea $x^{(0)}$ este valor.

2. Obtener un valor al azar para y de la distribución $f(y/x^{(0)})$. Sea $y^{(1)}$ este valor.

3. Volver a 1 con $y^{(1)}$ en lugar de $y^{(0)}$ y alternar entre 1 y 2 para obtener parejas de valores $(x^{(i)}, y^{(i)})$, para $i = 1, ..., N$.

Se demuestra que, para N suficientemente grande, la pareja $(x^{(N)}, y^{(N)})$ es un valor al azar de la distribución conjunta $f(x, y)$.

Un problema importante es investigar la convergencia de la secuencia. Puede demostrarse que, bajo ciertas condiciones generales, el algoritmo converge, pero la convergencia puede requerir un número enorme de iteraciones en algunos problemas (véase por ejemplo Justel y Peña, 1996).

11.4.4 Estimación Bayesiana de referencia en el modelo normal

Supongamos que se desea estimar los parámetros de una normal multivariante sin introducir información a priori. Es más simple tomar como parámetros $\boldsymbol{\mu}, \mathbf{V}^{-1}$, donde $\mathbf{V}^{-1}$ es la matriz de precisión. La estimación de referencia para este problema supone que, a priori, $p(\boldsymbol{\mu}, \mathbf{V}^{-1}) = p(\boldsymbol{\mu})p(\mathbf{V}^{-1})$, donde $p(\boldsymbol{\mu})$ es constante en la región donde la verosimilitud es no nula y $p(\mathbf{V}^{-1})$ se elige como no informativa en el sentido de Jeffreys. Puede demostrarse que, entonces, una prior conveniente es proporcional a $|\mathbf{V}^{-1}|^{-1/2(p+1)}$, con lo que la prior resultante es

$$p(\boldsymbol{\mu}, \mathbf{V}^{-1}) \propto |\mathbf{V}|^{1/2(p+1)}. \tag{11.17}$$

La expresión de la verosimilitud es, según lo expuesto en la sección 10.2.2, y expresándola ahora en función de $\mathbf{V}^{-1}$

$$f(\mathbf{X}|\boldsymbol{\mu}, \mathbf{V}^{-1}) = C|\mathbf{V}^{-1}|^{n/2} \exp\left\{-\frac{n}{2} tr\mathbf{V}^{-1}\mathbf{S}(\boldsymbol{\mu})\right\}, \tag{11.18}$$

y multiplicando estas dos ecuaciones, (11.17) y (11.18), resulta la posterior

$$p(\boldsymbol{\mu}, \mathbf{V}^{-1}|\mathbf{X}) = C_1|\mathbf{V}^{-1}|^{(n-p-1)/2} \exp\left\{-\frac{1}{2} tr\mathbf{V}^{-1}\mathbf{S}(\boldsymbol{\mu})n\right\}, \tag{11.19}$$

donde C_1 es la constante necesaria para que la densidad integre a uno. Observemos que si el objetivo es obtener la moda de la posterior no necesitamos calcular esta constante.

La marginal de $\boldsymbol{\mu}$ se obtiene integrando respecto a $\mathbf{V}$. Para ello, observamos que en esta integración la matriz $\mathbf{S}(\boldsymbol{\mu})$ es una constante, ya que no depende de $\mathbf{V}$, y la función a integrar es similar a la distribución Wishart, siendo ahora $\mathbf{V}^{-1}$ la variable, en lugar de $\mathbf{W}$, $n = m$, y $\mathbf{S}(\boldsymbol{\mu})n$ igual a la matriz de constantes Σ. El término que falta para tener la distribución completa es $|\Sigma|^{-m/2}$, que equivale a $|\mathbf{S}(\boldsymbol{\mu})n|^{n/2}$. Introduciendo esta constante, multiplicando y dividiendo para completar la integral y prescindiendo de constantes, obtenemos que la posterior será

$$p(\boldsymbol{\mu}|\mathbf{X}) \propto |\mathbf{S}(\boldsymbol{\mu})|^{-n/2} \tag{11.20}$$

y, se demuestra en el Apéndice 11.2, que este determinante puede escribirse como

$$p(\boldsymbol{\mu}|\mathbf{X}) \propto |1 + (\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})|^{-n/2}$$

Esta expresión indica que la densidad marginal del vector de medias es una t multivariante con $n-p$ grados de libertad (véase la sección 9.6.2). La moda de la densidad se alcanza para $\boldsymbol{\mu} = \bar{\mathbf{x}}$, resultado análogo al obtenido por MV. De la forma de la densidad (11.20) concluimos que este valor de $\boldsymbol{\mu}$ minimiza el determinante de la matriz de sumas de cuadrados $\mathbf{S}(\boldsymbol{\mu})$. Este criterio, minimizar el determinante de las sumas de cuadrados residuales, suele llamarse de mínimos cuadrados multivariante.

La posterior de $\mathbf{V}$ se obtiene integrando (11.19) con respecto a $\boldsymbol{\mu}$. Se demuestra en el Apéndice 11.2 que la distribución a posteriori para $\mathbf{V}^{-1}$ es una distribución Wishart $W_p(n-1, \mathbf{S}^{-1}/n)$. Puede comprobarse que la media de la distribución a posteriori de $\mathbf{V}$ es $n\mathbf{S}/(n-p-2)$, por lo que si tomamos este valor como estimador de $\mathbf{V}$ obtendremos un valor distinto que con el método MV.

11.4.5 Estimación con información a priori

Supongamos que disponemos de información a priori para estimar los parámetros de una distribución $N_p(\boldsymbol{\mu}, \mathbf{V})$. La forma de la verosimilitud (11.18) sugiere (véase el apéndice 11.2.) las siguientes distribuciones a priori. Para $\boldsymbol{\mu}$ dada $\mathbf{V}^{-1}$ estableceremos que

$$p(\boldsymbol{\mu}|\mathbf{V}^{-1}) \sim N_p(\boldsymbol{\mu}_0, \mathbf{V}|n_0),$$

y esta distribución indica que, conocida $\mathbf{V}$, la mejor estimación a priori que podemos dar del valor de $\boldsymbol{\mu}$ es $\boldsymbol{\mu}_0$, y la incertidumbre que asignamos a esta estimación a priori es $\mathbf{V}|n_0$. En principio podríamos reflejar nuestra incertidumbre respecto $\boldsymbol{\mu}$ con cualquier matriz de covarianzas, pero el análisis se simplifica si suponemos que esta incertidumbre es una fracción de la incertidumbre del muestreo. Observemos que ésta es la distribución a priori para $\boldsymbol{\mu}$

condicionada a $\mathbf{V}$, por lo que tiene sentido expresar la incertidumbre en función de la varianza muestral. Una vez que hayamos visto el papel que juega el parámetro n_0 comentaremos cómo fijarlo. Para la matriz de precisión estableceremos que:

$$p(\mathbf{V}^{-1}) \sim W_p(m_0, \mathbf{M}|m_0)$$

que escribimos de esta forma para que los parámetros tengan una interpretación más sencilla. Así, a priori, el valor esperado de la matriz de precisión es $\mathbf{M}$, y, como veremos, el parámetro m_0 controla la precisión que queremos dar a esta estimación inicial. Utilizando estas dos distribuciones, la distribución a priori conjunta resultante es

$$p(\boldsymbol{\mu}, \mathbf{V}^{-1}) = p(\boldsymbol{\mu}|\mathbf{V}^{-1})p(\mathbf{V}^{-1}).$$

El apéndice 11.2. calcula la distribución a posteriori mediante

$$p(\boldsymbol{\mu}, \mathbf{V}|\mathbf{X}) \propto \ell(\mathbf{X}|\mu, \mathbf{V})p(\mu, \mathbf{V})$$

y allí se obtiene que, a posteriori, la distribución de la media condicionada a la varianza es también normal:

$$p(\boldsymbol{\mu}|\mathbf{V}^{-1}, \mathbf{X}) \sim N_p(\boldsymbol{\mu}_p, \mathbf{V}_p)$$

donde la media a posteriori, que puede tomarse como el estimador bayesiano de $\boldsymbol{\mu}$, es:

$$\boldsymbol{\mu}_p = \frac{n_0\boldsymbol{\mu}_0 + n\bar{\mathbf{x}}}{n_0 + n}$$

y la incertidumbre en esta estimación es

$$\mathbf{V}_p = \frac{\mathbf{V}}{n_0 + n}$$

La media a posteriori es una media ponderada de la información a priori y la proporcionada por la muestra, y los coeficientes de ponderación son n_0 y n. El parámetro n_0 representa pues el peso que queremos dar a nuestra estimación prior con relación a la muestral. Vemos también que la incertidumbre asociada equivale a la de una muestra de tamaño $n_0 + n$. Podemos interpretar n_0 como el número de observaciones equivalentes que asignamos a la información contenida en la prior. Por ejemplo, si $n_0 = 10$ y tomamos una muestra de tamaño 90, queremos que nuestra prior tenga un peso del 10% en el cálculo de la posterior.

La distribución a posteriori de la matriz de precisión es

$$p(\mathbf{V}^{-1}|\mathbf{X}) \sim W_p(n + m_0, \mathbf{M}_p),$$

donde la matriz de la Wishart es

$$\mathbf{M}_p^{-1} = m_0\mathbf{M}^{-1} + n\mathbf{S} + \frac{nn_0}{n + n_0}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'.$$

Para interpretar este resultado, recordemos que la media de una distribución de Wishart $W_p(n+m_0,M_p)$ es

$$E(\mathbf{V}^{-1}|\mathbf{X}) = \left(\frac{m_0}{n+m_0}\mathbf{M}^{-1} + \frac{n}{n+m_0}\mathbf{S} + \frac{nn_0}{(n+n_0)^2}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)'\right)^{-1},$$

el término entre paréntesis juega el papel de la matriz de varianzas y vemos que suma tres fuentes de incertidumbre: las dos que vienen de la distribución prior y la muestral. El primer término es la matriz de covarianzas a priori, el segundo la matriz de covarianzas muestral y el tercero el incremento de covarianzas debido a la discordancia entre la media a priori y la muestral. El término m_0 controla el peso que queremos dar a la estimación prior de la varianza, frente a la varianza muestral, y el término n_0 el peso de la discrepancia entre la media a prioiri y la estimada. Observemos también que si la información proporcionada por la muestra es grande con relación a la prior, es decir, n es grande con relación a m_0 y n_0, de manera que m_0/n y n_0/n sean pequeños, la esperanza de la precisión posterior es, aproximadamente, la precisión muestral, $\mathbf{S}^{-1}$.

11.5 CONTRASTES BAYESIANOS

11.5.1 Conceptos básicos

En el enfoque bayesiano, la hipótesis nula no se acepta o rechaza, como en el enfoque clásico, sino que se determina su probabilidad a posteriori dados los datos. Supongamos el contraste general considerado en el capítulo anterior: dado un parámetro vectorial, $\boldsymbol{\theta}$, $p-$dimensional, que toma valores en Ω se desea contrastar la hipótesis:

$$H_0 : \boldsymbol{\theta} \in \Omega_0,$$

frente a la hipótesis alternativa

$$H_1 : \boldsymbol{\theta} \in \Omega - \Omega_0.$$

Suponemos que existen probabilidades a priori para cada una de las dos hipótesis. Estas probabilidades quedan automáticamente determinadas si establecemos una distribución a priori sobre $\boldsymbol{\theta}$, ya que entonces:

$$p_0 = P(H_0) = P(\boldsymbol{\theta} \in \Omega_0) = \int_{\Omega_0} p(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}$$

y

$$p_1 = P(H_1) = P(\boldsymbol{\theta} \in \Omega - \Omega_0) = \int_{\Omega-\Omega_0} p(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}.$$

Las probabilidades a posteriori de las hipótesis las calcularemos mediante el teorema de Bayes

$$P(H_i|\mathbf{X}) = \frac{P(\mathbf{X}|H_i)P(H_i)}{P(\mathbf{X})} \quad i = 0, 1$$

$\log_{10} B_{01}$	B_{01}	$P(H_0)$ para $(p_0/p_1 = 1)$	Interpretación.
0	1	$0,5$	indecisión
-1	10^{-1}	$0,1$	débil rechazo de H_0
-2	10^{-2}	$0,01$	rechazo de H_0
-3	10^{-3}	$0,001$	rechazo sin duda de H_0

Tabla 11.3: Interpretación del factor de Bayes según Jefreys

y de aquí se obtiene el resultado fundamental:

$$\frac{P(H_0|\mathbf{X})}{P(H_1|\mathbf{X})} = \frac{f(\mathbf{X}|H_0)}{f(\mathbf{X}|H_1)}.\frac{P(H_0)}{P(H_1)} \tag{11.21}$$

que puede expresarse como

Ratio de posteriores = R.Verosimilitudes x R. Prioris

Esta expresión indica que la evidencia respecto a la hipótesis nula se obtiene multiplicando la evidencia proporcionada por los datos, con la evidencia a priori. Al cociente entre las verosimilitudes se denomina factor de Bayes, B, y si las probabilidades a prioiri de ambas hipótesis son las mismas, determina las probabilidades a posteriori de las hipótesis. Expresando las probabilidades a posteriori en términos del parámetro, se obtiene

$$P(H_i|\mathbf{X}) = P(\boldsymbol{\theta} \in \Omega_i/\mathbf{X}) = \int_{\Omega_i} p(\boldsymbol{\theta}|\mathbf{X})d\boldsymbol{\theta} \quad i = 0, 1$$

donde $\Omega_1 = \Omega - \Omega_0$, y $p(\boldsymbol{\theta}|\mathbf{X})$ es la distribución a posteriori para el vector de parámetros de interés dada por (11.15). Por tanto

$$p(H_i|\mathbf{X}) = \frac{1}{f(\mathbf{X})} \int_{\Omega_i} f(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \quad i = 0, 1$$

donde $\Omega_1 = \Omega - \Omega_0$. Sustituyendo en (11.21) se obtiene que el factor de Bayes de la primera hipótesis respecto a la segunda, B_{01}, es

$$B_{01} = \frac{p_1}{p_0} \frac{\int_{\Omega_0} f(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\Omega-\Omega_0} f(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

Jeffreys ha dado la escala de evidencia para el factor de Bayes que se presenta en la Tabla 11.3. La primera columna presenta el factor de Bayes en una escala logaritmica, la segunda el factor de Bayes, la tercera la probabilidad de la hipótesis nula supuesto que las probabilidades a priori para las dos hipótesis son las mismas. La última columna propone la decisión a tomar respecto a H_0.

11.5.2 Comparación entre los contraste bayesianos y los clásicos

Si suponemos que, a priori, las probabilidades de ambas hipótesis son las mismas, el factor de Bayes es comparable al ratio del contraste de verosimilitudes, pero existe una diferencia fundamental: en el constraste de verosimilitudes se toma el máximo de la verosimilitud, mientras que en el enfoque bayesiano se toma el promedio sobre la región relevante, promediando con la distribución a priori. Por tanto, el contraste tiene en cuenta al calcular la integral el tamaño del espacio definido por Ω_0 y por Ω_1. Por ejemplo, supongamos que $\boldsymbol{\theta}$ es un parámetro escalar $0 \leq \theta \leq 1$ y que contrastamos:

$$H_0 : \theta = \theta_0$$

frente a

$$H_1 : \theta \neq \theta_0.$$

Para que las probabilidades a priori de ambas hipótesis sean las mismas, supongamos que fijamos $p(\theta = \theta_0) = 1/2$ y que $p(\theta) = 1/2$ si $\theta \neq \theta_0$. Entonces, el factor de Bayes compara $f(\mathbf{X}|\theta_0)$ con el valor promedio de la verosimilitud cuando $\theta \neq \theta_0$, mientras que el contraste de verosimilitudes compara $f(\mathbf{X}|\theta_0)$ con el valor máximo de la verosimilitud. Si el valor $\theta = \theta_0$ no es exactamente cierto, sino sólo aproximadamente cierto, y el tamaño de la muestra es muy grande:

1. Con el enfoque bayesiano, los valores alejados de θ_0 tendrán una verosimilitud muy pequeña con muestras grandes y al promediar sobre todos los valores se tenderá a favorecer a H_0. Al aumentar n puede hacer muy difícil rechazar H_0.

2. Con el enfoque clásico, comparamos $f(\mathbf{X}|\theta_0)$ con $f(\mathbf{X}|\theta_{MV})$, donde θ_{MV} es el estimador MV que estará próximo al verdadero valor para muestras grandes, y esta diferencia aumentará con el tamaño muestral, por lo que terminaremos siempre rechazando H_0.

En resumen, con el enfoque clásico, cuando $n \to \infty$ se rechaza H_0 en la práctica, mientras que con el enfoque bayesiano cuando $n \to \infty$ es más difícil rechazar H_0 en la práctica. Esto es consecuencia de que el enfoque bayesiano tiene en cuenta la verosimilitud de H_0 *y de* H_1, mientras que el enfoque clásico mira sólo a H_0.

Es importante señalar que esta contradicción desaparece en el momento en que reformulamos el problema como uno de estimación. Entonces ambos métodos coincidirán con muestras grandes en la estimación del parámetro.

11.6 Selección de Modelos

11.6.1 El Criterio de Akaike

El método de máxima verosimilitud supone que la forma del modelo es conocida y sólo falta estimar los parámetros. Cuando no es así debe aplicarse con cuidado. Por ejemplo, supongamos que se desea estimar un vector de parámetros $\boldsymbol{\theta} = (\theta_1, ...\theta_p)'$ y admitimos en lugar de un modelo único la secuencia de modelos $\mathrm{M}_1 = (\theta_1, 0, ..., 0)$, ..., $\mathrm{M}_i = (\theta_1, ..., \theta_i, 0, .., 0)$, ...,$\mathrm{M}_p = (\theta_1, ..., \theta_p)$, es decir el modelo M_i $(i = 1, ..., p)$ indica que los primeros i parámetros son

distintos de cero y los restantes cero. Es claro que si estimamos los parámetros bajo cada modelo y calculamos el valor máximo del soporte sustituyendo los parámetros por sus estimaciones MV, el modelo con mayor soporte de los datos será el modelo M_p con todos los parámetros libres. Este resultado es general: el método de máxima verosimilitud siempre da mayor soporte al modelo con más parámetros, ya que la verosimilud sólo puedo aumentar si introduzco más parámetros para explicar los datos.

Esta limitación del método de máxima verosimilitud fue percibida por Fisher, que propuso el método en 1936 para estimar los parámetros de un modelo, indicando sus limitaciones para comparar modelos distintos. La solución habitula para seleccionar entre los modelos es hacer un contraste de hipótesis utilizando el contraste de verosimilitudes y elegiendo el modelo M_i frente al M_p mediante

$$\lambda = 2(L(M_p) - L(M_i)) = D(M_i) - D(M_p)$$

donde $L(M_p)$ es el soporte del modelo M_p al sustituir en el función soporte el parámetro $\boldsymbol{\theta}$ por su estimación MV y $L(M_i)$ el soporte del modelo M_i al estimar los parámetros con la restriccion $\theta_{i+1} = .. = \theta_p = 0$, y $D(M_j) = -2L(M_j)$ es al desviación. Suponiendo que el modelo más simple, M_i, es correcto, el estadístico λ se distribuye como una χ^2 con $p - i$ grados de libertad.

Akaike propuso un enfoque alternativo para resolver el problema de seleccionar el modelo suponiendo que el objetivo es hacer predicciones tan precisas como sea posible. Sea $f(\mathbf{y}|M_i)$ la densidad de una nueva observacion bajo el modelo M_i y sea $f(\mathbf{y})$ la verdadera función de densidad que puede o no ser una de las consideradas, es decir, el modelo verdadero puede o no ser uno de los M_i. Queremos seleccionar el modelo de manera que $f(\mathbf{y}|M_i)$ sea tan próxima como sea posible a $f(\mathbf{y})$. Una manera razonable de medir la distancia entre estas dos funciones de densidad es mediante la distancia de Kullback-Leibler entre las dos densidades, que se calcula:

$$KL(f(\mathbf{y}|M_i), f(\mathbf{y})) = \int \log \frac{f(\mathbf{y}|M_i)}{f(\mathbf{y})} f(\mathbf{y}) d\mathbf{y} \tag{11.22}$$

Para interpretar esta medida observemos que la diferencia de logaritmos equivale, cuando los valores de ambas funciones son similares, a la diferencia relativa, ya que

$$\log \frac{f(\mathbf{y}|M_i)}{f(\mathbf{y})} = \log(1 + \frac{f(\mathbf{y}|M_i) - f(\mathbf{y})}{f(\mathbf{y})}) \cong \frac{f(\mathbf{y}|M_i) - f(\mathbf{y})}{f(\mathbf{y})}$$

y cuando las diferencias son grandes, el logaritmo es mejor medida de discrepancia que la diferencia relativa. Las discrepancias se promedian respecto a la verdadera distribución de la observación y la medida (11.22) puede demostrarse que es siempre positiva. Una manera alternativa de escribir esta medida es

$$KL(f(\mathbf{y}|M_i), f(\mathbf{y})) = E_y \log f(\mathbf{y}|M_i) - E_y \log f(\mathbf{y})$$

donde $E_{\mathbf{y}}$ indica obtener la esperanza bajo la verdadera distribución de y. Como esta cantidad es siempre positiva, minimizaremos la distancia entre la verdadera distribución y

$f(\mathbf{y}|M_i)$ haciendo el primer término lo más pequeño posible. Puede demostrarse que (Akaike, 1985) que esto equivale a minimizar

$$AIC = -2L(M_i) + 2i = D(M_i) + 2i \tag{11.23}$$

es decir, minimizamos la suma de la desviación del modelo, que disminuirá si introducimos más parámetros, y el número de parámetros en el modelo, que tiende a corregir por este efecto.

11.6.2 El criterio BIC

Una ventaja del enfoque bayesiano es que el problema de selección de modelos puede abordarse con los mismos principios que el contraste de hipótesis. Supongamos que en un problema estadístico dudamos entre un conjunto de m modelos posibles para los datos observados $M_1, \ldots, M_m$. Si consideramos los modelos como posibles hipótesis sobre los datos, calcularemos sus probabilidades a posteriori, $M_1, \ldots, M_m$ y seleccionaremos el modelo con máxima probabilidad a posteriori. Estas probabilidades vienen dadas por :

$$P(M_j|\mathbf{X}) = \frac{f(\mathbf{X}|M_j)}{f(\mathbf{X})} P(M_j) \quad j = 1, ..., m \tag{11.24}$$

donde $P(M_j)$ es la probabilidad a priori del modelo j. Esta ecuación indica cómo pasamos de la probabilidad a priori a la posteriori para cada modelo: se calcula las verosimilitud marginal de los datos para ese modelo, $f(\mathbf{X}|M_j)$,donde el nombre marginal proviene de que esta función no depende de los valores de los parámetros, y se compara con la verosimilitud marginal promedio para todos los modelos, $f(\mathbf{X})$. En efecto, llamemos $\boldsymbol{\theta}_j$ a los parámetros del modelo M_j. La distribución $f(\mathbf{X}|M_j)$ viene dada por

$$\begin{aligned} f(\mathbf{X}|M_j) &= \int f(\mathbf{X}|\boldsymbol{\theta}_j, M_j)\, p(\boldsymbol{\theta}_j|M_j)\, d\boldsymbol{\theta}_j \\ &= \int L_j(\mathbf{X}|\boldsymbol{\theta}_j)\, p(\boldsymbol{\theta}_j|M_j)\, d\boldsymbol{\theta}_j \end{aligned}$$

es decir, se obtiene promediando la verosimilitud del modelo, $L_j(\mathbf{X}|\boldsymbol{\theta})$, por las probabilidades a priori de los parámetros, $p(\boldsymbol{\theta}_j|M_j)$. Por lo tanto, esta función expresa la verosimilitud de los datos dado el modelo, sea cual sea el valor de los parámetros, lo que justifica el nombre de verosimilitud marginal. El denominador de (11.24) es

$$f(\mathbf{X}) = \sum f(\mathbf{X}|M_j)\, P(M_j)$$

y puede interpretarse como una media ponderada de las verosimilitudes marginales, siendo los coeficientes de la ponderación las probabilidades a priori.

La conclusión que se desprende de (11.24) es que seleccionar el modelo con mayor probabilidad a posteriori equivale a seleccionar el modelo donde el producto de la verosimilitud marginal $f(\mathbf{X}|M_j)$ y de la prior del modelo $P(M_j)$ sea máxima.

Las expresiones anteriores se derivan de las reglas del cálculo de probabilidades y son exactas. Es posible obtener una expresión aproximada de $f(\mathbf{X}|M_j)$ si suponemos que la distribución a posteriori del vector de parámetros es asintóticamente normal multivariante. Supongamos que para el modelo j esta distribución a posteriori es :

$$p(\boldsymbol{\theta}_j|\mathbf{X}, M_j) = (2\pi)^{-p_j/2} |S_j|^{-1/2} \exp\left\{-1/2 \left(\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j\right)' S_j^{-1} \left(\boldsymbol{\theta}_j - \widehat{\boldsymbol{\theta}}_j\right)\right\}$$

donde p_j es la dimensión del vector de parámetros del modelo M_j, y $\widehat{\boldsymbol{\theta}}_j$ es el estimador MV de $\boldsymbol{\theta}_j$ y S_j la matriz de covarianzas de este estimador. Por el teorema de Bayes:

$$p(\boldsymbol{\theta}_j|\mathbf{X}, M_j) = \frac{l_j(\boldsymbol{\theta}_j|\mathbf{X})\, p(\boldsymbol{\theta}_j|M_j)}{f(\mathbf{X}|M_j)}$$

donde $l_j(\boldsymbol{\theta}_j|\mathbf{X})$ es la verosimilitud, $p(\boldsymbol{\theta}_j|M_j)$ la probabilidad a priori para los parámetros y $f(\mathbf{X}|M_j)$ la verosimilitud marginal. Esta expresión es cierta para cualquier valor del parámetro y en particular para $\boldsymbol{\theta}_j = \widehat{\boldsymbol{\theta}}_j$. Tomando logaritmos y particularizando esta expresión para $\widehat{\boldsymbol{\theta}}_j$, podemos escribir

$$\log f(\mathbf{X}|M_j) = L_j\left(\widehat{\boldsymbol{\theta}}_j|\mathbf{X}\right) + \log p\left(\widehat{\boldsymbol{\theta}}_j|M_j\right) - \left[-(p_j/2)\log 2\pi - \frac{1}{2}\log|S_j|\right] \tag{11.25}$$

La matriz S_j de covarianzas del estimador de los parámetros tiene términos habitualmente del tipo a/n. Escribiendo

$$S_j \equiv \frac{1}{n} R_j$$

entonces $|S_j| = n^{-p_j}|R_j|$ y sustituyendo en (11.25):

$$\log f(\mathbf{X}|M_j) = L_j\left(\widehat{\boldsymbol{\theta}}_j|\mathbf{X}\right) + \log p\left(\widehat{\boldsymbol{\theta}}_j|M_j\right) + \frac{p_j}{2}\log 2\pi - \frac{p_j}{2}\log n + \frac{1}{2}\log|R_j|\,.$$

Vamos a aproximar esta expresión para n grande. Para ello vamos a mantener en esta expresión únicamente los términos que crecen con n y despreciar los que tomen un valor acotado que no crece con n. El primer término es el valor del soporte en el máximo que es la suma de n términos para las n observaciones y será de orden n. El segundo es el valor de la prior y, para n grande, podemos suponer que va a ser aproximadamente constante con relación a la verosimilitud. El tercer término, $(p_j/2)\log 2\pi$ es de orden constante. El cuarto crece con n y el último, por construcción, esta acotado. En consecuencia, para n grande podemos escribir:

$$\log f(\mathbf{X}|M_j) \simeq L_j\left(\widehat{\boldsymbol{\theta}}_j|\mathbf{X}\right) - \frac{p_j}{2}\log n.$$

Esta expresión fue obtenida por primera vez por Schwarz (1978), que propuso escoger el modelo que conduzca a un valor máximo de esta cantidad. Una forma equivalente de

este criterio, llamada criterio BIC (Bayesian Information Criterion), es xalcular para cada modelo la cantidad:

$$\boxed{\text{BIC}(M_j) = -2L_j\left(\widehat{\boldsymbol{\theta}}_j|\mathbf{X}\right) + p_j \log n}$$

y seleccionar aquel modelo para el cual esta cantidad es mínima. De esta manera este criterio pondera la desviación del modelo, medida por $-2L_j\left(\widehat{\boldsymbol{\theta}}_j|\mathbf{X}\right)$, con el número de parámetros. Si introducimos más parámetros en el modelo mejorará el ajuste, con lo que aumentará el soporte o disminuirá la desviación, y este efecto queda compensado por el aumento del número de parámetros que aparece en $p_j \log n$.

11.6.3 Relación entre el BIC y EL AIC

La forma general de estos dos criterios de selección es

$$D(M_j) + p_j g(n)$$

donde $D(M_j)$ es la desviación del modelo medida por $-2L_j\left(\widehat{\boldsymbol{\theta}}_j|\mathbf{X}\right)$, y p_j el número de parámetros. La constante que multiplica al número de parámetros es distinta en ambos criterios En el criterio BIC esta constante es $\log n$, mientras que en el AIC es 2. Por tanto, el criterio BIC seleccionará modelos más parsimoniosos, es decir, con menor número de parámetros que el AIC. Otros autores han propuesto otros criterios que corresponden a distintas funciones $g(n)$

La diferencia entre estos criterios se explica por su distinto objetivo El criterio BIC trata de seleccionar el modelo correcto, con máxima probabilidad a posteriori, y puede demostrarse que es un criterio consitente, de manera que la probabilidad de seleccionar el modelo correcto tiende a uno cuando crece el tamaño muestral. El criterio AIC no pretende seleccionar el modelo correcto, ya que admite que este modelo verdadero puede no estar entre los estimados, y trata de obtener el modelo que proporcione mejores predicciones entre los existentes. Puede demostrarse que, en condiciones generales de que el modelo verdadero puede aproximarse arbitariamente bien con los estimados al crecer el tamaño muestral, el criterio AIC es eficiente, en el sentido de escoger el modelo que proporciona, en promedio, mejores predicciones. Sin embargo, en muestras pequeñas o medianas, el criterio AIC tiende a seleccionar modelos con más parámetros de los necesarios.

11.7 Lecturas complementarias

Una buena introducción al algoritmo EM se encuentra en Tanner (1991) y con ejemplos multivariantes en Flury (1997). Versiones más amplias se encuentran en Gelman et al (1995) y Little y Rubin (1987). El libro de Schafer (1997) contiene numerosos ejemplos de su aplicación con datos multivariantes.

La estimación de mezclas se estudia con detalle en Titterington at al (1987), y varios de los textos de cluster, que comentaremos en el capítulo 15, incluyen el estudio de estas distribuciones. La estimación robusta puede consultarse en Hampel at al (1986) and

Rousseew and Leroy (1987). La estimación Bayesiana multivariante en Bernardo y Smith (1994), O'Hagan (1994) y Press (1989). Los algoritmos de cadenas de Markov (métodos MC^2) en Gamerman (1997), Carlin y Louis (1996) y Robert y Casella (1999). Los contrastes bayesianos en Berger (1985). La literatura de selección de modelos es muy amplia. Algunas referencias básicas son Akaike(1974), Miller (1990) y McQuarrie y Tsai (1998), Chow (1981) y Lanterman (2001).

APÉNDICE 11.1.CONVERGENCIA DEL ALGORITMO EM

Sea

$$L_C^*(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(i)}) = E\left[L_C(\boldsymbol{\theta}|\mathbf{Y},\mathbf{Z})|\widehat{\boldsymbol{\theta}}_{(i)},\mathbf{Y}\right]$$

la función que maximizamos en el paso M del algoritmo. Vamos a demostrar que cuando $\widehat{\boldsymbol{\theta}}_{(i)} = \widehat{\boldsymbol{\theta}}_{(i+1)} = \widehat{\boldsymbol{\theta}}_F$ entonces

$$\left[\frac{\partial L(\boldsymbol{\theta}|\mathbf{Y})}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathbf{F}}} = 0$$

y $\widehat{\boldsymbol{\theta}}_F$ es el estimador MV. Para ello observemos que

$$L_C^*(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(i)}) = \int \log f(\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}) f(\mathbf{Z}|\mathbf{Y},\widehat{\boldsymbol{\theta}}_{(i)})\mathbf{dZ} + L(\boldsymbol{\theta}|\mathbf{Y})$$

y si maximizamos esta expresión derivando e igualando a cero se obtiene:

$$\frac{\partial L_C^*(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{(i)})}{\partial \boldsymbol{\theta}} = \int \frac{\partial f(\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{f(\mathbf{Z}|\mathbf{Y},\widehat{\boldsymbol{\theta}}_{(\mathbf{i})})}{f(\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta})}\mathbf{dZ} + L'(\boldsymbol{\theta}|\mathbf{Y}) = 0,$$

con lo que tendremos que $\widehat{\boldsymbol{\theta}}_{(i+1)}$ verifica

$$\int \left[\frac{\partial f(\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\widehat{\boldsymbol{\theta}}_{(i+1)}} \frac{f(\mathbf{Z}|\mathbf{Y},\widehat{\boldsymbol{\theta}}_{(\mathbf{i})})}{f(\mathbf{Z}|\mathbf{Y},\widehat{\boldsymbol{\theta}}_{(\mathbf{i+1})})}\mathbf{dZ} + L'(\widehat{\boldsymbol{\theta}}_{(\mathbf{i+1})}|\mathbf{Y}) = 0.$$

Cuando $\widehat{\boldsymbol{\theta}}_{(i)} = \widehat{\boldsymbol{\theta}}_{(i+1)} = \widehat{\boldsymbol{\theta}}_F$ el primer miembro es cero, ya que se reduce a

$$\int \left[\frac{\partial f(\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\widehat{\boldsymbol{\theta}}_{(i+1)}} \mathbf{dZ}$$

que es siempre cero, como se comprueba derivando en la ecuación $\int f(\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta})\mathbf{dZ} = 1$. Por tanto tendrá que verificarse que

$$L'(\widehat{\boldsymbol{\theta}}_{(\mathbf{i+1})}|\mathbf{Y}) = 0$$

que implica que $\widehat{\boldsymbol{\theta}}_{(\mathbf{i+1})}$ es el estimador MV.

APENDICE 11.2: ESTIMACIÓN BAYESIANA

Demostraremos primero que la distribución marginal a posteriori de $\mathbf{V}^{-1}$con la prior de referencia es una Wishart invertida. Integrando en la conjunta

$$p(\mathbf{V}^{-1}|\mathbf{X}) = \int C_1|\mathbf{V}^{-1}|^{(n-p-1)/2}\exp\left\{-\frac{1}{2}tr\mathbf{V}^{-1}\mathbf{S}(\boldsymbol{\mu})n\right\}d\boldsymbol{\mu},$$

y utilizando $\mathbf{S}(\boldsymbol{\mu})n = n\mathbf{S} + n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})'$, podemos escribir

$$p(\mathbf{V}^{-1}|\mathbf{X}) = C_1|\mathbf{V}^{-1}|^{(n-p-2)/2}\exp\left\{-\frac{1}{2}tr\mathbf{V}^{-1}n\mathbf{S}\right\}A$$

donde

$$A = \int |\mathbf{V}|^{-1/2}\exp\left\{-\frac{n}{2}(\bar{\mathbf{x}}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\right\}d\boldsymbol{\mu},$$

que con las constantes adecuadas integra a uno. Por tanto, podemos concluir que la distribución a posteriori para $\mathbf{V}^{-1}$ es una distribución Wishart $W_p(n-1,\mathbf{S}^{-1}/n)$.

Obtendremos ahora las distribuciones en el caso de prioris informativas. La verosimilitud de los parámetros de la normal tiene la forma del producto de una normal por una Wishart, con lo que la prior conjugada a este problema debe ser de la forma:

$$p(\boldsymbol{\mu},\mathbf{V}^{-1}) \propto |\mathbf{V}^{-1}|^{(m_0-p)/2}\exp\left\{-(1/2)\left[tr\mathbf{V}^{-1}\mathbf{M}^{-1}m_0 + n_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\mathbf{V}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right]\right\}.$$

De acuerdo con esta distribución a priori, $p(\boldsymbol{\mu}|\mathbf{V}^{-1})$ es una normal multivariante con media $\boldsymbol{\mu}_0$ y varianza $\mathbf{V}|n_0$,

$$p(\boldsymbol{\mu}|\mathbf{V}^{-1}) \propto |\mathbf{V}^{-1}|^{1/2}\exp\left\{-1/2\left[n_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\mathbf{V}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right]\right\}$$

mientras que $p(\mathbf{V}^{-1})$ sigue una distribución Wishart $W_p(m_0,M/m_0)$

$$p(\mathbf{V}^{-1}) \propto |\mathbf{V}^{-1}|^{(m_0-p-1)/2}\exp\left\{-(1/2)tr\mathbf{V}^{-1}\mathbf{M}^{-1}m_0\right\}.$$

La distribución posterior será

$$p(\boldsymbol{\mu},\mathbf{V}^{-1}|\mathbf{X}) = C|\mathbf{V}^{-1}|^{(n+m_0-p)/2}\exp\left\{-E/2\right\},$$

donde el exponente, E, puede escribirse:

$$E = tr(\mathbf{V}^{-1}(\mathbf{M}^{-1}m_0 + n\mathbf{S})) + n(\bar{\mathbf{x}}-\boldsymbol{\mu})\mathbf{V}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})' + n_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\mathbf{V}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0).$$

Vamos a expresar de otra forma las formas cuadráticas. Para ello utilizaremos el siguiente resultado general:

Lemma 2 *Si* $\mathbf{A}$ *y* $\mathbf{B}$ *son matrices no singulares, se verifica que*

$$(\mathbf{z}-\mathbf{a})'\mathbf{A}(\mathbf{z}-\mathbf{a}) + (\mathbf{z}-\mathbf{b})'\mathbf{B}(\mathbf{z}-\mathbf{b}) = (\mathbf{z}-\mathbf{c})'\mathbf{D}(\mathbf{z}-\mathbf{c}) + (\mathbf{a}-\mathbf{b})'\mathbf{H}(\mathbf{a}-\mathbf{b})$$

donde $\mathbf{c} = (\mathbf{A}+\mathbf{B})^{-1}(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b})$, $D = (A+B)$ *y* $\mathbf{H} = (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}$. *Además se verifica*

$$|\mathbf{A}|^{\frac{1}{2}}|\mathbf{B}|^{\frac{1}{2}} = |\mathbf{A}+\mathbf{B}|^{\frac{1}{2}}|\mathbf{A}^{-1}+\mathbf{B}^{-1}|^{-\frac{1}{2}}$$

Comencemos demostrando que los dos miembros de las formas cuadráticas son idénticos. El primer miembro puede escribirse

$$\mathbf{z}'(\mathbf{A}+\mathbf{B})\mathbf{z}-2\mathbf{z}'(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b})+\mathbf{a}'\mathbf{A}\mathbf{a}+\mathbf{b}'\mathbf{B}\mathbf{b}$$

y llamando $\mathbf{c}=(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b})$,*también puede escribirse*

$$\mathbf{z}'(\mathbf{A}+\mathbf{B})\mathbf{z}-2\mathbf{z}'(\mathbf{A}+\mathbf{B})\mathbf{c}+\mathbf{c}'(\mathbf{A}+\mathbf{B})\mathbf{c}-\mathbf{c}'(\mathbf{A}+\mathbf{B})\mathbf{c}+\mathbf{a}'\mathbf{A}\mathbf{a}+\mathbf{b}'\mathbf{B}\mathbf{b}$$

que es igual a

$$(\mathbf{z}-\mathbf{c})'(\mathbf{A}+\mathbf{B})(\mathbf{z}-\mathbf{c})+\mathbf{a}'\mathbf{A}\mathbf{a}+\mathbf{b}'\mathbf{B}\mathbf{b}-(\mathbf{a}'\mathbf{A}+\mathbf{b}'\mathbf{B})(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b})$$

La primera parte de esta expresión es la primera forma cuadrática del segundo miembro del Lemma. Operando en la segunda parte, resulta

$$\mathbf{a}'(\mathbf{A}-\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A})\mathbf{a}+\mathbf{b}'(\mathbf{B}-\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B})\mathbf{b}-2\mathbf{b}'\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}\mathbf{a}$$

y utilizando que, según la sección 2.3.4:

$$(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}=\mathbf{A}-\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}=\mathbf{B}-\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}=\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}$$

resulta que la segunda forma cuadrática es

$$(\mathbf{a}-\mathbf{b})'(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{a}-\mathbf{b})$$

Para comprobar la segunda parte, como $(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}=\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A}$, *tenemos que* $|\mathbf{A}^{-1}+\mathbf{B}^{-1}|^{-1}=|\mathbf{B}||\mathbf{A}+\mathbf{B}|^{-1}|\mathbf{A}|$ *con lo que* $|\mathbf{A}||\mathbf{B}|=|\mathbf{A}+\mathbf{B}||\mathbf{A}^{-1}+\mathbf{B}^{-1}|^{-1}$

Utilizando este lema, la suma de $(\bar{\mathbf{x}}-\boldsymbol{\mu})\mathbf{V}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})'$ y $n_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\mathbf{V}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)$ puede escribirse como:

$$(n+n_0)(\boldsymbol{\mu}-\boldsymbol{\mu}_p)'\mathbf{V}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_p)+\frac{nn_0}{n+n_0}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)\mathbf{V}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)$$

donde

$$\boldsymbol{\mu}_p=\frac{n_0\boldsymbol{\mu}_0+n\bar{\mathbf{x}}}{n_0+n}$$

Con estos resultados la posterior puede descomponerse como producto de $p(\boldsymbol{\mu}|\mathbf{V}^{-1}\mathbf{X})$ por $p(\mathbf{V}^{-1}|\mathbf{X})$. La primera distribución es la de la media a posteriori dada la varianza, que es normal multivariante

$$p(\boldsymbol{\mu}|\mathbf{V}^{-1}\mathbf{X})=c|\mathbf{V}^{-1}|^{1/2}\exp\left\{-1/2\left[(n+n_0)(\boldsymbol{\mu}-\boldsymbol{\mu}_p)'\mathbf{V}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_p)\right]\right\}.$$

y la segunda es la distribución marginal a posteriori de la matriz de precisión, $p(\mathbf{V}^{-1}|\mathbf{X})$, dada por

$$p(\mathbf{V}^{-1}|\mathbf{X})=C|\mathbf{V}^{-1}|^{(n+m_0-p-1)/2}\exp\left\{-1/2(tr\mathbf{V}^{-1}\mathbf{M}_p^{-1})\right\}$$

donde

$$\mathbf{M}_p^{-1}=\mathbf{M}^{-1}m_0+n\mathbf{S}+\frac{nn_0}{n+n_0}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)'$$

y representa una distribución de Wishart $W_p(n+m_0,M_p)$.

Capítulo 12

ANÁLISIS FACTORIAL

12.1 INTRODUCCIÓN

El análisis factorial tiene por objeto explicar si un conjunto de variables observadas por un pequeño número de variables *latentes,* o no observadas, que llamaremos factores. Por ejemplo, supongamos que hemos tomado veinte medidas físicas del cuerpo de una persona: estatura, longitud del tronco y de las extremidades, anchura de hombros, peso, etc. Es intuitivo que todas estas medidas no son independientes entre sí, y que conocidas algunas de ellas podemos prever con poco error las restantes. Una explicación de este hecho es que las dimensiones del cuerpo humano dependen de ciertos factores, y si estos fuesen conocidos podríamos prever con pequeño error los valores de las variables observadas. Como segundo ejemplo, supongamos que estamos interesados en estudiar el desarrollo humano en los países del mundo, y que disponemos de muchas variables económicas, sociales y demográficas, en general dependientes entre sí, que están relacionadas con el desarrollo. Podemos preguntarnos si el desarrollo de un país depende de un pequeño número de factores tales que, conocidos sus valores, podríamos prever el conjunto de las variables de cada país. Como tercer ejemplo, supongamos que medimos con distintas pruebas la capacidad mental de un individuo para procesar información y resolver problemas. Podemos preguntarnos si existen unos factores, no directamente observables, que explican el conjunto de resultados observados. El conjunto de estos factores será lo que llamamos inteligencia y es importante conocer cuántas dimensiones distintas tiene este concepto y cómo caracterizarlas y medirlas. El análisis factorial surge impulsado por el interés de Karl Pearson y Charles Sperman en comprender las dimensiones de la inteligencia humana en los años 30, y muchos de sus avances se han producido en el área de la psicometría.

El análisis factorial esta relacionado con los componentes principales, pero existen ciertas diferencias. En primer lugar, los componentes principales se construyen para explicar las varianzas, mientras que los factores se construyen para explicar las covarianzas o correlaciones entre las variables. En segundo lugar, componentes principales es un herramienta descriptiva, mientras que el análisis factorial presupone un modelo estadístico formal de generación de la muestra dada.

12.2 EL MODELO FACTORIAL

12.2.1 Hipótesis básicas

Supondremos que observamos un vector de variables $\mathbf{x}$, de dimensiones $(p \times 1)$, en elementos de una población. El modelo de análisis factorial establece que este vector de datos observados se genera mediante la relación:

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda \mathbf{f} + \mathbf{u} \tag{12.1}$$

donde:

1. $\mathbf{f}$ es un vector $(m \times 1)$ de variables latentes o factores no observadas. Supondremos que sigue una distribución $N_m(\mathbf{0}, \mathbf{I})$, es decir los factores son variables de media cero e independientes entre sí y con distribución normal.

2. Λ es una matriz $(p \times m)$ de constantes desconocidas $(m < p)$. Contiene los coeficientes que describen como los factores, $\mathbf{f}$, afectan a las variables observadas, $\mathbf{x}$, y se denomina matriz de carga.

3. $\mathbf{u}$ es un vector $(p \times 1)$ de perturbaciones no observadas. Recoge el efecto de todas las variables distintas de los factores que influyen sobre $\mathbf{x}$. Supondremos que $\mathbf{u}$ tiene distribución $N_p(\mathbf{0}, \boldsymbol{\psi})$ donde $\boldsymbol{\psi}$ es diagonal, y que las perturbaciones están incorreladas con los factores $\mathbf{f}$.

Con estas tres hipótesis deducimos que:
(a) $\boldsymbol{\mu}$ es la media de las variables $\mathbf{x}$, ya que tanto los factores como las perturbaciones tienen media cero;
(b) $\mathbf{x}$ tiene distribución normal, al ser suma de variables normales, y llamando $\mathbf{V}$ a su matriz de covarianzas

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \mathbf{V}).$$

La ecuación (12.1) implica que dada una muestra aleatoria simple de n elementos generada por el modelo factorial, cada dato x_{ij} puede escribirse como:

$$x_{ij} = \mu_j + \lambda_{j1} f_{1i} + ... + \lambda_{jm} f_{mi} + u_{ij} \quad i = 1, ..., n \quad j = 1, ..., p$$

que descompone x_{ij}, el valor observado en el individuo i de la variable j, como suma de $m + 2$ términos. El primero es la media de la variable j, μ_j, del segundo al $m + 1$ recogen el efecto de los m factores, y el último es una perturbación específica de cada observación, u_{ij}. Los efectos de los factores sobre x_{ij} son el producto de los coeficientes $\lambda_{j1}, ..., \lambda_{jm}$, que dependen de la relación entre cada factor y la variable j, (y que son los mismos para todos los elementos de la muestra), por los valores de los m factores en el elemento muestral i, $f_{1i}, ..., f_{mi}$. Poniendo juntas las ecuaciones para todas las observaciones, la matriz de datos, $\mathbf{X}$, $(n \times p)$, puede escribirse como:

$$\mathbf{X} = \mathbf{1}\boldsymbol{\mu}' + \mathbf{F}\Lambda' + \mathbf{U}$$

donde $\mathbf{1}$ es un vector $n \times 1$ de unos, $\mathbf{F}$ es una matriz $(n \times m)$ que contiene los m factores para los n elementos de la población, Λ' es la transpuesta de la matriz de carga $(m \times p)$ cuyos coeficientes constantes relacionan las variables y los factores y $\mathbf{U}$ es una matriz $(n \times p)$ de perturbaciones.

12.2.2 Propiedades

La matriz de carga Λ contiene las covarianzas entre los factores y las variables observadas. En efecto, la matriz de covarianzas $(p \times m)$ entre las variables y los factores se obtiene multiplicando (12.1) por $\mathbf{f}'$ por la derecha y tomando esperanzas:

$$E\left[(\mathbf{x}-\boldsymbol{\mu})\mathbf{f}'\right] = \Lambda E\left[\mathbf{f}\mathbf{f}'\right] + E\left[\mathbf{u}\mathbf{f}'\right] = \Lambda$$

ya que, por hipótesis, los factores están incorrelados ($E\left[\mathbf{f}\mathbf{f}'\right] = \mathbf{I}$) y tienen media cero y están incorrelados con las perturbaciones ($E\left[\mathbf{u}\mathbf{f}'\right] = 0$). Esta ecuación indica que los términos λ_{ij} de la matriz de carga, Λ, representan la covarianza entre la variable x_i y el factor f_j, y, al tener los factores varianza unidad, son los coeficientes de regresión cuando explicamos las variables observadas por los factores. En el caso particular en que las variables $\mathbf{x}$ estén estandarizadas, los términos λ_{ij} coeficientes son también las correlaciones entre las variables y los factores.

La matriz de covarianzas entre las observaciones verifica, según (12.1):

$$\mathbf{V} = E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'\right] = \Lambda E\left[\mathbf{f}\mathbf{f}'\right]\Lambda' + E\left[\mathbf{u}\mathbf{u}'\right]$$

ya que $E[\mathbf{f}\mathbf{u}'] = 0$ al estar incorrelados los factores y el ruido. Entonces, se obtiene la propiedad fundamental:

$$\boxed{\mathbf{V} = \Lambda\Lambda' + \boldsymbol{\psi},} \tag{12.2}$$

que establece que la matriz de covarianzas de los datos observados admite una descomposición como suma de dos matrices:

(1) La primera, $\Lambda\Lambda'$, es una matriz simétrica de rango $m < p$. Esta matriz contiene la parte común al conjunto de las variables y depende de las covarianzas entre las variables y los factores.

(2) La segunda, $\boldsymbol{\psi}$, es diagonal, y contiene la parte específica de cada variable, que es independiente del resto.

Esta descomposición implica que las varianzas de las variables observadas pueden descomponerse como:

$$\sigma_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2 + \psi_i^2, \qquad i = 1, \ldots, p.$$

donde el primer término es la suma de los efectos de los factores y el segundo el efecto de la perturbación. Llamando

$$h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2,$$

a la suma de los efectos de los factores que llamaremos *comunalidad,* tenemos que

$$\sigma_i^2 = h_i^2 + \psi_i^2, \qquad i = 1, \ldots, p. \tag{12.3}$$

Esta igualdad puede interpretarse como una descomposición de la varianza en:

$$\text{Varianza observada} = \underset{\text{(Comunalidad)}}{\text{Variabilidad común}} + \text{Variabilidad específica}$$

que es análoga a la descomposición clásica de la variabilidad de los datos en una parte explicada y otra no explicada que se realiza en el análisis de la varianza. En el modelo factorial la parte explicada es debida a los factores y la no explicada al ruido o componente aleatorio. Esta relación es la base del análisis que presentamos a continuación.

Ejemplo 12.1 *Supongamos que tenemos tres variables generadas por dos factores. La matriz de covarianzas debe verificar*

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \end{bmatrix} \begin{bmatrix} \lambda_{11} & \lambda_{21} & \lambda_{31} \\ \lambda_{12} & \lambda_{22} & \lambda_{32} \end{bmatrix} + \begin{bmatrix} \psi_{11} & 0 & 0 \\ 0 & \psi_{22} & 0 \\ 0 & 0 & \psi_{33} \end{bmatrix}$$

Esta igualdad proporciona 6 ecuaciones distintas (recordemos que al ser $\mathbf{V}$ *simétrica sólo tiene 6 términos distintos). La primera será:*

$$\sigma_{11} = \lambda_{11}^2 + \lambda_{12}^2 + \psi_{11}$$

Llamando $h_1^2 = \lambda_{11}^2 + \lambda_{12}^2$ *a la contribución de los dos factores en la variable* 1*. Las seis ecuaciones son :*

$$\begin{array}{ll} \sigma_{ii} = h_i^2 + \psi_i^2 & i = 1, 2, 3 \\ \sigma_{ij} = \lambda_{i1}\lambda_{j1} + \lambda_{i2}\lambda_{j2} & i = 1, 2, 3 \\ & i \neq j \end{array}$$

12.2.3 Unicidad del modelo

En el modelo factorial ni la matriz de carga, Λ, ni los factores, $\mathbf{f}$, son observables. Esto plantea un problema de indeterminación: dos representaciones $(\Lambda, \mathbf{f})$ y $(\Lambda^*, \mathbf{f}^*)$ serán equivalentes si

$$\Lambda\mathbf{f} = \Lambda^*\mathbf{f}^*$$

Esta situación conduce a dos tipos de indeterminación.

(1) Un conjunto de datos puede explicarse con la misma precisión con factores incorrelados o correlados.

(2) Los factores no quedan determinados de manera única.

Vamos a analizar estas dos indeterminaciones. Para mostrar la primera, si $\mathbf{H}$ es cualquier matriz no singular, la representación (12.1) puede también escribirse como

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{H}\mathbf{H}^{-1}\mathbf{f} + \mathbf{u} \tag{12.4}$$

y llamando $\Lambda^* = \Lambda\mathbf{H}$ a la nueva matriz de carga, y $\mathbf{f}^* = \mathbf{H}^{-1}\mathbf{f}$ a los nuevos factores:

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda^*\mathbf{f}^* + \mathbf{u}, \tag{12.5}$$

donde los nuevos factores $\mathbf{f}^*$ tienen ahora una distribución $N\left(\mathbf{0}, \mathbf{H^{-1}}(\mathbf{H^{-1}})'\right)$ y, por lo tanto, están correlados. Análogamente, partiendo de factores correlados, $\mathbf{f} \sim N(\mathbf{0}, \mathbf{V_f})$, siempre podemos encontrar una expresión equivalente de las variables mediante un modelo con factores incorrelados. En efecto, sea $\mathbf{A}$ una matriz tal que $\mathbf{V_f} = \mathbf{AA}'$. (Esta matriz siempre existe si $\mathbf{V_f}$ es definida positiva), entonces $\mathbf{A^{-1}V_f}(\mathbf{A^{-1}})' = \mathbf{I}$, y escribiendo

$$\mathbf{x} = \boldsymbol{\mu} + \Lambda(\mathbf{A})(\mathbf{A^{-1}})\mathbf{f} + \mathbf{u},$$

y tomando $\Lambda^* = \Lambda\mathbf{A}$ como la nueva matriz de coeficientes de los factores y $\mathbf{f}^* = \mathbf{A^{-1}f}$ como los nuevos factores, el modelo es equivalente a otro con factores incorrelados. Esta indeterminación se ha resuelto en las hipótesis del modelo tomando siempre los factores como incorrelados.

En segundo lugar, si $\mathbf{H}$ es ortogonal, el modelo $\mathbf{x} = \boldsymbol{\mu} + \Lambda\mathbf{f} + \mathbf{u}$ y el $\mathbf{x} = \boldsymbol{\mu} + (\Lambda\mathbf{H})(\mathbf{H}'\mathbf{f}) + \mathbf{u}$ son indistinguibles. Ambos contienen factores incorrelados, con matriz de covarianzas la identidad. En este sentido, decimos que el modelo factorial está indeterminado ante rotaciones. Esta indeterminación se resuelve imponiendo restricciones sobre los componentes de la matriz de carga, como veremos en la sección siguiente.

Ejemplo 12.2 *Supongamos* $\mathbf{x} = (x_1, x_2, x_3)'$ *y el modelo factorial* M_1 *siguiente:*

$$\mathbf{x} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

y los factores están incorrelados. Vamos a escribirlo como otro modelo equivalente de factores también incorrelados. Tomando $\mathbf{H} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, *esta matriz es ortogonal, ya que* $\mathbf{H}^{-1} = \mathbf{H}' = \mathbf{H}$. *Entonces*

$$\mathbf{x} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + [\mathbf{u}].$$

Llamando a este modelo, M_2, *puede escribirse como:*

$$\mathbf{x} = \begin{bmatrix} \frac{2}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} + [\mathbf{u}]$$

y los nuevos factores, $\mathbf{g}$, *están relacionados con los anteriores,* $\mathbf{f}$, *por:*

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \left(\sqrt{2}\right)^{-1} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$$

y son por lo tanto una rotación de los iniciales. Comprobemos que estos nuevos factores están también incorrelados. Su matriz de varianzas es:

$$\mathbf{V}_g = \left(\sqrt{2}\right)^{-1} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{V}_f \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \left(\sqrt{2}\right)^{-1}$$

y si $\mathbf{V}_f = \mathbf{I} \Rightarrow \mathbf{V}_g = \mathbf{I}$*, de donde se deduce que los modelos* $M1$ *y* $M2$ *son indistinguibles.*

12.2.4 Normalización del modelo factorial

Como el modelo factorial esta indeterminado ante rotaciones la matriz Λ no está identificada. Esto implica que aunque observemos toda la población, y $\boldsymbol{\mu}$, y $\mathbf{V}$ sean conocidos, no podemos determinar Λ de manera única. La solución para poder estimar esta matriz es imponer restricciones sobre sus términos. Los dos métodos principales de estimación que vamos a estudiar utilizan alguna de las dos siguientes normalizaciones:

Criterio 1:

Exigir:

$$\Lambda'_{m\times p}\Lambda_{p\times m} = \mathbf{D} = \text{Diagonal} \tag{12.6}$$

Con esta normalización los vectores que definen el efecto de cada factor sobre las p variables observadas son ortogonales. De esta manera, los factores además de estar incorrelados producen efectos lo más distintos posibles en las variables. Vamos a comprobar que esta normalización define una matriz de carga de manera única. Supongamos primero que tenemos una matriz Λ tal que el producto $\Lambda'\Lambda$ no es diagonal. Transformamos los factores con $\Lambda^* = \Lambda\mathbf{H}$, donde $\mathbf{H}$ es la matriz que contiene en columnas los vectores propios de $\Lambda'\Lambda$. Entonces:

$$\Lambda'^*\Lambda^* = \mathbf{H}'\Lambda'\Lambda\mathbf{H} \tag{12.7}$$

y como $\mathbf{H}$ diagonaliza $\Lambda'\Lambda$ la matriz Λ^* verifica la condición (12.6). Veamos ahora que esta es la única matriz que lo verifica. Supongamos que rotamos esta matriz y sea $\Lambda^{**} = \Lambda\mathbf{C}$ donde $\mathbf{C}$ es ortogonal. Entonces la matriz $\Lambda^{**\prime}\Lambda^{**} = \mathbf{C}'\Lambda'^*\Lambda^*\mathbf{C}$ no será diagonal. Analogamente, si partimos de una matriz que verifica (12.6) si la rotamos dejará de verificar esta condición.

Cuando se verifica esta normalización, postmultiplicando la ecuación (12.2) por Λ, podemos escribir

$$(\mathbf{V} - \boldsymbol{\psi})\,\Lambda = \Lambda\mathbf{D},$$

que implica que las columnas de Λ son vectores propios de la matriz $\mathbf{V} - \boldsymbol{\psi}$, que tiene como valores propios los términos diagonales de $\mathbf{D}$. Esta propiedad se utiliza en la estimación mediante el método del factor principal.

Criterio 2:

Exigir:

$$\Lambda'\psi^{-1}\Lambda = \mathbf{D} = \text{Diagonal} \tag{12.8}$$

En esta normalización los efectos de los factores sobre las variables, ponderados por las varianzas de las perturbaciones de cada ecuación, se hacen incorrelados. Como la anterior, esta normalización define una matriz de carga de manera única. En efecto, supongamos que $\Lambda\psi^{-1}\Lambda$ no es diagonal, y transformamos con $\Lambda^* = \Lambda\mathbf{H}$. Entonces:

$$\Lambda'^*\psi^{-1}\Lambda^* = \mathbf{H}'\left(\Lambda'\psi^{-1}\Lambda\right)\mathbf{H} \tag{12.9}$$

y como $\Lambda'\psi^{-1}\Lambda$ es una matriz simétrica y definida no negativa, siempre puede diagonalizarse si escogemos como $\mathbf{H}$ la matriz que contiene en columnas los vectores propios de $\Lambda'\psi^{-1}\Lambda$. Analogamente, si se verifica de partida (12.8) y rotamos la matriz de carga esta condición dejará de verificarse. Esta es la normalización que utiliza la estimación máximo verosímil. Su justificación es que de esta manera los factores son condicionalmente independientes dados los datos, como veremos en el apéndice 12.4.

Con esta normalización, postmultiplicando la ecuación (12.2) por $\boldsymbol{\psi}^{-1}\Lambda$, tenemos que

$$\mathbf{V}\boldsymbol{\psi}^{-1}\Lambda - \Lambda = \Lambda\,\mathbf{D}$$

y premultiplicando por $\boldsymbol{\psi}^{-1/2}$**,** resulta:

$$\boldsymbol{\psi}^{-1/2}\mathbf{V}\boldsymbol{\psi}^{-1}\Lambda - \boldsymbol{\psi}^{-\mathbf{1/2}}\Lambda = \boldsymbol{\psi}^{-1/2}\Lambda\mathbf{D}$$

que implica

$$\boldsymbol{\psi}^{-1/2}\mathbf{V}\boldsymbol{\psi}^{-1/2}\boldsymbol{\psi}^{-1/2}\Lambda = \boldsymbol{\psi}^{-1/2}\Lambda\left(\mathbf{D}+\mathbf{I}\right)$$

y concluimos que la matriz $\boldsymbol{\psi}^{-1/2}\mathbf{V}\boldsymbol{\psi}^{-1/2}$ tiene vectores propios $\boldsymbol{\psi}^{-1/2}\Lambda$ con valores propios $\mathbf{D}+\mathbf{I}$. Esta propiedad se utiliza en la estimación máximo verosímil.

12.2.5 Número máximo de factores

Si sustituimos en (12.2) la matriz teórica de covarianzas, $\mathbf{V}$, por la matriz muestral, $\mathbf{S}$, el sistema estará identificado si es posible resolverlo de manera única. Para ello existe una restricción en el número de factores posibles. El número de ecuaciones que obtenemos de (12.2) es igual al conjunto de términos de $\mathbf{S}$, que es $p+p(p-1)/2 = p(p+1)/2$. El número de incógnitas en el segundo término es pm, los coeficientes de la matriz Λ, más los p términos de la diagonal de $\boldsymbol{\psi}$, menos las restricciones impuestas para identificar la matriz Λ. Suponiendo que $\Lambda'\boldsymbol{\psi}^{-1}\Lambda$ debe ser diagonal, esto supone $m(m-1)/2$ restricciones sobre los términos de Λ, .

Para que el sistema este determinado debe haber un número de ecuaciones igual o mayor que el de incógnitas. En efecto, si existen menos ecuaciones que incógnitas no es posible

encontrar una solución única y el modelo no está identificado. Si el número de ecuaciones es exactamente igual al de incógnitas existirá una solución única. Si existen más ecuaciones que incógnitas, podremos resolver el sistema en el sentido de los mínimos cuadrados y encontrar unos valores de los parámetros que minimicen los errores de estimación. Por lo tanto:

$$p + pm - \frac{m(m-1)}{2} \leq \frac{p(p+1)}{2}$$

que supone:

$$p + m \leq p^2 - 2pm + m^2,$$

es decir

$$(p-m)^2 \geqslant p + m.$$

El lector puede comprobar que esta ecuación implica que, cuando p no es muy grande (menor de 10) aproximadamente el número máximo de factores debe ser menor que la mitad del número de variables menos uno. Por ejemplo, el número máximo de factores con 7 variables es 3. Esta es la regla que se obtiene si escribimos la desigualdad anterior despreciando el término de las restricciones sobre los elementos de Λ.

12.3 EL MÉTODO DEL FACTOR PRINCIPAL

El método del factor principal es un método para estimar la matriz de carga basado en componentes principales. Evita tener que resolver las ecuaciones de máxima verosimilitud, que son más complejas. Tiene la ventaja de que la dimensión del sistema puede identificarse de forma aproximada. Se utiliza en muchos programas de ordenador por su simplicidad. Su base es la siguiente: supongamos que podemos obtener una estimación inicial de la matriz de varianzas de las perturbaciones $\widehat{\boldsymbol{\psi}}$. Entonces, podemos escribir

$$\mathbf{S} - \widehat{\boldsymbol{\psi}} = \Lambda\Lambda', \tag{12.10}$$

y como $\mathbf{S} - \hat{\boldsymbol{\psi}}$ es simétrica, siempre puede descomponerse como:

$$\mathbf{S} - \widehat{\boldsymbol{\psi}} = \mathbf{HGH}' = (\mathbf{HG}^{1/2})(\mathbf{HG}^{1/2})' \tag{12.11}$$

donde $\mathbf{H}$ es cuadrada de orden p y ortogonal, $\mathbf{G}$ es también de orden p, diagonal y contiene las raíces características de $\mathbf{S} - \widehat{\boldsymbol{\psi}}$. El modelo factorial establece que $\mathbf{G}$ debe ser diagonal del tipo:

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{1m\times m} & \mathbf{O}_{m\times(p-m)} \\ \mathbf{O}_{(p-m)\times m} & \mathbf{O}_{(p-m)\times(p-m)} \end{bmatrix}$$

ya que $\mathbf{S} - \widehat{\boldsymbol{\psi}}$ tiene rango m. Por tanto, si llamamos $\mathbf{H}_1$ a la matriz $p \times m$ que contiene los vectores propios asociados a los valores propios no nulos de $\mathbf{G}_1$ podemos tomar como estimador de Λ la matriz $p \times m$:

$$\widehat{\Lambda} = \mathbf{H}_1\mathbf{G}_1^{1/2} \tag{12.12}$$

con lo que resolvemos el problema. Observemos que la normalización resultante es:

$$\widehat{\Lambda}'\widehat{\Lambda} = \mathbf{G}_1^{1/2}\mathbf{H}_1'\mathbf{H}_1\mathbf{G}_1^{1/2} = \mathbf{G}_1 = \text{Diagonal} \tag{12.13}$$

ya que los vectores propios de matrices simétricas son ortogonales, por lo que $\mathbf{H}_1'\mathbf{H}_1 = \mathbf{I}_m$. Por tanto, con este método se obtienen estimadores de la matriz $\widehat{\Lambda}$ con columnas ortogonales entre sí.

En la práctica la estimación se lleva a cabo de forma iterativa como sigue:

1. Partir de una estimación inicial de $\widehat{\Lambda}_i$ o de $\widehat{\boldsymbol{\psi}}_i$ mediante $\widehat{\boldsymbol{\psi}}_i = diag\left(\mathbf{S} - \widehat{\Lambda}\widehat{\Lambda}'\right)$.

2. Calcular la matriz cuadrada y simétrica $\mathbf{Q}_i = \mathbf{S} - \widehat{\boldsymbol{\psi}}_i$.

3. Obtener la descomposición espectral de $\mathbf{Q}_i$ de forma

$$\mathbf{Q}_i = \mathbf{H}_{1i}\mathbf{G}_{1i}\mathbf{H}_{1i}' + \mathbf{H}_{2i}\mathbf{G}_{2i}\mathbf{H}_{2i}'$$

   donde $\mathbf{G}_{1i}$ contiene los m mayores valores propios de $\mathbf{Q}_i$ y $\mathbf{H}_{1i}$ sus valores propios. Elegiremos m de manera que los restantes vectores propios contenidos en $\mathbf{G}_{2i}$ sean todos pequeños y de tamaño similar. La matriz $\mathbf{Q}_i$ puede no ser definida positiva y algunos de sus valores propios pueden ser negativos. Esto no es un problema grave si estos valores propios son muy pequeños y podemos suponerlos próximos a cero.

4. Tomar $\widehat{\Lambda}_{i+1} = \mathbf{H}_{1i}\mathbf{G}_{1i}^{1/2}$ y volver a (1). Iterar hasta convergencia, es decir hasta que $\|\Lambda_{n+1} - \Lambda_n\| < \epsilon$.

Los estimadores obtenidos serán consistentes pero no eficientes, como en el caso de Máxima verosimilitud. Tampoco son invariantes ante transformaciones lineales, como los MV, es decir, no se obtiene necesariamente el mismo resultado con la matriz de covarianzas y con la de correlaciones.

Para llevar a la práctica esta idea, debemos especificar cómo obtener el estimador inicial $\widehat{\boldsymbol{\psi}}$, problema que se conoce como la estimación de las comunalidades.

12.3.1 Estimación de las comunalidades

Estimar los términos ψ_i^2 equivale a definir valores para los términos diagonales, h_i^2, de $\Lambda\Lambda'$, ya que $h_i^2 = s_i^2 - \widehat{\psi}_i^2$. Existen las siguientes alternativas:

1. tomar $\widehat{\psi}_i = 0$. Esto equivale a extraer los componentes principales de $\mathbf{S}$. Supone tomar $\widehat{h}_i^2 = s_i^2$ (en el caso de correlaciones $\widehat{h}_i^2 = 1$), que es claramente su valor máximo, por lo que podemos comenzar con un sesgo importante.

2. tomar $\widehat{\psi}_j^2 = 1/s_{jj}^*$, donde s_{jj}^* es el elemento diagonal j-ésimo de la matriz de precisión $\mathbf{S}^{-1}$. Según el apéndice 3.2 esto equivale a tomar h_j^2 como:

$$\widehat{h}_j^2 = s_j^2 - s_j^2(1 - R_j^2) = s_j^2 R_j^2, \tag{12.14}$$

donde R_j^2 es el coeficiente de correlación múltiple entre x_j y el resto de las variables. Intuitivamente, cuanto mayor sea R_j^2 mayor será la comunalidad $\widehat{h}_j^2$. Con este método comenzamos con una estimación sesgada a la baja de h_i^2, ya que $\widehat{h}_i^2 \leq h_i^2$. En efecto, por ejemplo, suponemos que para la variable x_1 el modelo verdadero es

$$x_1 = \sum_{j=1}^{m} \lambda_{1j} f_j + u_1 \tag{12.15}$$

que está asociado a la descomposición $\sigma_1^2 = h_1^2 + \psi_1^2$. La proporción de varianza explicada es h_1^2/σ_1^2. Si escribimos la ecuación de regresión

$$x_1 = b_2 x_2 + \ldots + b_p x_p + \epsilon_1$$

sustituyendo cada variable por su expresión en términos de los factores tenemos que:

$$x_1 = b_2 \left(\sum \lambda_{2j} f_j + u_2 \right) + \ldots + b_p \left(\sum \lambda_{pj} f_j + u_p \right) + \epsilon. \tag{12.16}$$

que conducirá a una descomposición de la varianza $\sigma_1^2 = \widehat{h}_1^2 + \widehat{\psi}_1^2$. Claramente $\widehat{h}_1^2 \leq h_1^2$, ya que en (12.16) forzamos a que aparezcan como regresores además de los factores, como en (12.15) los ruidos $u_1, \ldots, u_p$ de cada ecuación. Además, es posible que un factor afecte a x_1 pero no al resto, con lo que no aparecerá en la ecuación (12.16). En resumen, la comunalidad estimada en (12.16) será una cota inferior del valor real de la comunalidad.

Ejemplo 12.3 *En este ejemplo mostraremos las iteraciones del algoritmo del factor principal de forma detallada para los datos de ACCIONES del Anexo I. La matriz de varianzas covarianzas de estos datos en logaritmos es,*

$$S = \begin{bmatrix} 0.13 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix}$$

Para estimar la matriz de cargas realizamos los pasos del algoritmo del factor principal descritos anteriormente. Antes de empezar el algoritmo tenemos que fijar la cota para decidir la convergencia . Fijaremos un ε grande, 0.05, de forma que en pocas iteraciones el algoritmo converja a pesar de los errores acumulados por el redondeo.

Paso 1. Tomando la segunda alternativa para la estimación inicial de las comunalidades $diag(\widehat{\psi}_i^2) = 1/s_{jj}^*$. donde s_{jj}^* es el elemento j-ésimo de la matriz S^{-1}

$$S^{-1} = \begin{bmatrix} 52.094 & -47.906 & 52.88 \\ -47.906 & 52.094 & -47.12 \\ 52.88 & -47.12 & 60.209 \end{bmatrix}$$

$$\widehat{\psi}_i^2 = \begin{bmatrix} 1/52.094 & 0 & 0 \\ 0 & 1/52.094 & 0 \\ 0 & 0 & 1/60.209 \end{bmatrix} = \begin{bmatrix} 0.019 & 0 & 0 \\ 0 & 0.019 & 0 \\ 0 & 0 & 0.017 \end{bmatrix}$$

Paso 2. Calculamos la matriz cuadrada y simétrica $\mathbf{Q}_i = \mathbf{S} - \widehat{\boldsymbol{\psi}}_i$

$$\mathbf{Q}_i = \begin{bmatrix} 0.13 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} - \begin{bmatrix} 0.019 & 0 & 0 \\ 0 & 0.019 & 0 \\ 0 & 0 & 0.017 \end{bmatrix} = \begin{bmatrix} 0.111 & 0.15 & -0.19 \\ 0.15 & 0.111 & -0.03 \\ -0.19 & -0.03 & 0.143 \end{bmatrix}$$

Paso 3. Descomposición espectral de $\mathbf{Q}_i$ *y separación en dos términos* $\mathbf{H}_{1i}\mathbf{G}_{1i}\mathbf{H}'_{1i}$ *y* $\mathbf{H}_{2i}\mathbf{G}_{2i}\mathbf{H}'_{2i}$. *Los valores propios de* $\mathbf{Q}_i$ *son* 0.379, 0.094, *y* -0.108. *Observemos que uno de ellos es negativo, con lo que la matriz no es definida positiva. Como hay un valor propio mucho mayor que los demás tomaremos un único factor. Esto supone la descomposición*

$$\begin{bmatrix} 0.111 & 0.15 & -0.19 \\ 0.15 & 0.111 & -0.03 \\ -0.19 & -0.03 & 0.143 \end{bmatrix} = \begin{bmatrix} -0.670 \\ -0.442 \\ 0.596 \end{bmatrix} \times 0.379 \times \begin{bmatrix} -0.670 \\ -0.442 \\ 0.596 \end{bmatrix}' +$$
$$+ \begin{bmatrix} -0.036 & 0.741 \\ -0.783 & -0.438 \\ -0.621 & 0.508 \end{bmatrix} \begin{bmatrix} 0.094 & 0 \\ 0 & -0.108 \end{bmatrix} \begin{bmatrix} -0.036 & 0.741 \\ -0.783 & -0.438 \\ -0.621 & 0.508 \end{bmatrix}'$$

Paso 4. Calculamos $\widehat{\Lambda}_{i+1} = \mathbf{H}_{1i}\mathbf{G}_{1i}^{1/2}$

$$\widehat{\Lambda}_{i+1} = \begin{bmatrix} -0.670 \\ -0.442 \\ 0.596 \end{bmatrix} \times \sqrt{0.379} = \begin{bmatrix} -0.412 \\ -0.272 \\ 0.367 \end{bmatrix}$$

Esta es la primera estimación de la matriz de carga. Vamos a iterar para mejorar esta estimación. Para ello volvemos al paso 1.

Paso 1. Estimamos los términos de la diagonal de $\widehat{\boldsymbol{\psi}}_i$ *mediante* $\widehat{\boldsymbol{\psi}}_i = diag\left(\mathbf{S} - \widehat{\Lambda}\widehat{\Lambda}'\right)$

$$\widehat{\boldsymbol{\psi}}_i = diag\left\{ \begin{bmatrix} 0.13 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} - \begin{bmatrix} -0.412 \\ -0.272 \\ 0.367 \end{bmatrix} \begin{bmatrix} -0.412 & -0.272 & 0.367 \end{bmatrix} \right\}$$
$$= \begin{bmatrix} 0.180 & 0 & 0 \\ 0 & 0.056 & 0 \\ 0 & 0 & 0.0253 \end{bmatrix}$$

Paso 2. Calculamos la matriz cuadrada y simétrica $\mathbf{Q}_i = \mathbf{S} - \widehat{\boldsymbol{\psi}}_i$

$$\mathbf{Q}_i = \begin{bmatrix} 0.13 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} - \begin{bmatrix} 0.180 & 0 & 0 \\ 0 & 0.056 & 0 \\ 0 & 0 & 0.0253 \end{bmatrix} = \begin{bmatrix} -0.05 & 0.15 & -0.19 \\ 0.15 & 0.074 & -0.03 \\ -0.19 & -0.03 & 0.135 \end{bmatrix}$$

Paso 3. Descomposición espectral de $\mathbf{Q}_i = \mathbf{H}_{1i}\mathbf{G}_{1i}\mathbf{H}'_{1i} + \mathbf{H}_{2i}\mathbf{G}_{2i}\mathbf{H}'_{2i}$

$$\begin{bmatrix} -0.05 & 0.15 & -0.19 \\ 0.15 & 0.074 & -0.03 \\ -0.19 & -0.03 & 0.135 \end{bmatrix} = \begin{bmatrix} -0.559 \\ -0.450 \\ 0.696 \end{bmatrix} \times 0.307 \times \begin{bmatrix} -0.559 \\ -0.450 \\ 0.696 \end{bmatrix}' +$$
$$+ \begin{bmatrix} 0.081 & 0.825 \\ 0.806 & -0.385 \\ 0.586 & 0.414 \end{bmatrix} \begin{bmatrix} 0.067 & 0 \\ 0 & -0.215 \end{bmatrix} \begin{bmatrix} 0.081 & 0.825 \\ 0.806 & -0.385 \\ 0.586 & 0.414 \end{bmatrix}'$$

Paso 4. Calculamos $\widehat{\Lambda}_{i+1} = \mathbf{H}_{1i}\mathbf{G}_{1i}^{1/2}$

$$\widehat{\Lambda}_{i+1} = \begin{bmatrix} -0.559 \\ -0.450 \\ 0.696 \end{bmatrix} \times \sqrt{0.307} = \begin{bmatrix} -0.310 \\ -0.249 \\ 0.386 \end{bmatrix}$$

comprobamos si se cumple el criterio de convergencia $\|\Lambda_{n+1} - \Lambda_n\| < \epsilon$.

$$\left\| \begin{bmatrix} -0.310 \\ -0.249 \\ 0.386 \end{bmatrix} - \begin{bmatrix} -0.412 \\ -0.272 \\ 0.367 \end{bmatrix} \right\| = 0.106 \geq \epsilon = 0.05$$

volvemos al paso 1 hasta que se cumpla el criterio.

Paso 1. Volvemos a estimar $\widehat{\boldsymbol{\psi}}_i = diag\left(\mathbf{S} - \widehat{\Lambda}\widehat{\Lambda}'\right)$

$$\begin{aligned} \widehat{\boldsymbol{\psi}}_i &= diag\left\{ \begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} - \begin{bmatrix} -0.310 \\ -0.249 \\ 0.386 \end{bmatrix} \begin{bmatrix} -.31 & -.249 & .386 \end{bmatrix} \right\} \\ &= \begin{bmatrix} 0.254 & 0 & 0 \\ 0 & 0.068 & 0 \\ 0 & 0 & 0.011 \end{bmatrix} \end{aligned}$$

Paso 2. Calculamos la matriz cuadrada y simétrica $\mathbf{Q}_i = \mathbf{S} - \widehat{\boldsymbol{\psi}}_i$

$$\mathbf{Q}_i = \begin{bmatrix} 0.13 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} - \begin{bmatrix} 0.254 & 0 & 0 \\ 0 & 0.068 & 0 \\ 0 & 0 & 0.011 \end{bmatrix} = \begin{bmatrix} -0.124 & 0.15 & -0.19 \\ 0.15 & 0.062 & -0.03 \\ -0.19 & -0.03 & 0.149 \end{bmatrix}$$

Paso 3. Descomposición espectral de $\mathbf{Q}_i$. *Indicaremos sólo el primer vector y valor propio*

$$\begin{bmatrix} -0.124 & 0.15 & -0.19 \\ 0.15 & 0.062 & -0.03 \\ -0.19 & -0.03 & 0.149 \end{bmatrix} = \begin{bmatrix} -0499 \\ -0.425 \\ 0.755 \end{bmatrix} \times 0.291 \times \begin{bmatrix} -0499 \\ -0.425 \\ 0.755 \end{bmatrix}' + \mathbf{H}_{2i}\mathbf{G}_{2i}\mathbf{H}'_{2i}$$

Paso 4. Calculamos $\widehat{\Lambda}_{i+1} = \mathbf{H}_{1i}\mathbf{G}_{1i}^{1/2}$

$$\widehat{\Lambda}_{i+1} = \begin{bmatrix} -0499 \\ -0.425 \\ 0.755 \end{bmatrix} \times \sqrt{0.291} = \begin{bmatrix} -0.269 \\ -0.229 \\ 0.407 \end{bmatrix}$$

comprobamos si se cumple el criterio de convergencia $\left\|\hat{\Lambda}_{n+1} - \hat{\Lambda}_n\right\| < \epsilon$.

$$\left\| \begin{bmatrix} -0.269 \\ -0.229 \\ 0.407 \end{bmatrix} - \begin{bmatrix} -0.310 \\ -0.249 \\ 0.386 \end{bmatrix} \right\| = 0.05 \geq \epsilon = 0.05$$

El criterio de convergencia se ha cumplido y el modelo con los parámetros estimados es:

$$\mathbf{x} = \begin{bmatrix} -0.269 \\ -0.229 \\ 0.407 \end{bmatrix} f_1 + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \sim N_3 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.254 & 0 & 0 \\ 0 & 0.068 & 0 \\ 0 & 0 & 0.011 \end{bmatrix} \right)$$

Observemos que la expresión del factor obtenido es bastante distinta a la del primer componente principal que se obtuvo en el ejercicio 5.1

Ejemplo 12.4 *Para la base de datos de INVEST se realizó un análisis descriptivo en el capítulo 4 en el que se propuso una transformación logarítmica en todas las variables y la eliminación de EEUU. Sobre este conjunto de datos, una vez estandarizados, vamos a ilustrar el cálculo de un único factor mediante el método del factor principal (en el ejemplo siguiente se consideran 2 factores). Vamos a comparar los dos métodos propuestos para inicializar el algortimo con los datos estandarizados. En el primer caso comenzamos las iteraciones con*

$$\widehat{\psi}_j = 0 \Longrightarrow \widehat{h}_{(0)}^2 = 1,$$

y el número de iteraciones antes de converger es 6. El criterio de parada en el paso k del algoritmo es, en este caso, que la diferencia máxima entre las comunalidades en k y k-1 sea menor de 0.0001. En la siguiente tabla se presentan las estimaciones de las comunalidades para los pasos i=0,1,2,3,6.

	$\hat{h}_{(0)}^2$	$\hat{h}_{(1)}^2$	$\hat{h}_{(2)}^2$	$\hat{h}_{(3)}^2$	$\hat{h}_{(6)}^2$
INTER.A	*1*	*0.96*	*0.96*	*0.96*	***0.96***
INTER.B	*1*	*0.79*	*0.76*	*0.75*	***0.75***
AGRIC.	*1*	*0.94*	*0.94*	*0.94*	***0.94***
BIOLO.	*1*	*0.92*	*0.91*	*0.91*	***0.91***
MEDIC.	*1*	*0.97*	*0.97*	*0.97*	***0.97***
QUIMI.	*1*	*0.85*	*0.83*	*0.82*	***0.82***
INGEN.	*1*	*0.9*	*0.88*	*0.88*	***0.88***
FÍSICA	*1*	*0.94*	*0.93*	*0.93*	***0.93***

En negrilla figura el resultado final una vez que ha convergido el algoritmo.

Si inicializamos el algoritmo con el segundo método,

$$\widehat{\psi}_j = 1 - R_j^2 \Longrightarrow \widehat{h}_{(0)}^2 = R_j^2,$$

el número de iteraciones antes de converger es 5. En la siguiente tabla se presentan cómo varían la estimaciones de las comunalidades para los pasos i=0,1,2,3,5.

	$\hat{h}^2_{(0)}$	$\hat{h}^2_{(1)}$	$\hat{h}^2_{(2)}$	$\hat{h}^2_{(3)}$	$\hat{h}^2_{(5)}$
INTER.A	*0.98*	*0.96*	*0.96*	*0.96*	***0.96***
INTER.B	*0.82*	*0.76*	*0.75*	*0.75*	***0.75***
AGRIC.	*0.95*	*0.94*	*0.94*	*0.94*	***0.94***
BIOLO.	*0.97*	*0.92*	*0.91*	*0.91*	***0.91***
MEDIC.	*0.98*	*0.97*	*0.97*	*0.97*	***0.97***
QUIMI.	*0.85*	*0.82*	*0.82*	*0.82*	***0.82***
INGEN.	*0.93*	*0.89*	*0.88*	*0.88*	***0.88***
FÍSICA	*0.97*	*0.94*	*0.93*	*0.93*	***0.93***

En negrilla figura el resultado final una vez que ha convergido el algoritmo. Al haber inicializado el algoritmo en un punto más próximo al final, la convergencia ha sido más rápida, y ya en la segunda iteración el resultado es muy próximo al final. Se observa como la estimación inicial de las comunalidades, $\hat{h}^2_{(0)}$, es cota superior de la estimación final, $\hat{h}^2_{(5)}$. En la siguiente tabla presentamos la estimación de $\hat{\Lambda}_{(0)}$ de las que partimos en ambos métodos y la estimación de las cargas finales obtenidas.

	$\widehat{\psi}_j = 0$	$\widehat{\psi}_j = 1 - R_j^2$	*Final*
	Factor1	*Factor1*	*Factor1*
INTER.A	*0.97*	*0.97*	*0.98*
INTER.B	*0.89*	*0.87*	*0.87*
AGRIC.	*0.97*	*0.97*	*0.97*
BIOLO.	*0.96*	*0.96*	*0.95*
MEDIC.	*0.98*	*0.98*	*0.99*
QUIMI.	*0.92*	*0.90*	*0.91*
INGEN.	*0.94*	*0.94*	*0.94*
FÍSICA	*0.96*	*0.97*	*0.97*

El segundo método proporciona un $\hat{\Lambda}_{(0)}$ más próximo al resultado final, sobre todo para aquellas variables donde la variabilidad específica es mayor.

12.3.2 Generalizaciones

El método de estimación del factor principal es un procedimiento de minimizar la función:

$$F = tr\left(\mathbf{S} - \Lambda\Lambda' - \boldsymbol{\psi}\right)^2. \tag{12.17}$$

En efecto, esta función puede escribirse

$$F = \sum_{i=1}^{p} \sum_{j=1}^{p} (s_{ij} - v_{ij})^2 \tag{12.18}$$

donde v_{ij} son los elementos de la matriz $\mathbf{V} = \Lambda\Lambda' + \boldsymbol{\psi}$. Ahora bien, por la descomposición espectral, dada una matriz $\mathbf{S}$ cuadrada simétrica y no negativa la mejor aproximación en el sentido de mínimos cuadrados (12.18) mediante una matriz de rango m, $\mathbf{AA}'$ se obtiene tomando $\mathbf{A} = \mathbf{HD}^{1/2}$, donde $\mathbf{H}$ contiene los vectores propios y $\mathbf{D}^{1/2}$ las raíces de los valores propios de $\mathbf{S}$ (véase el apéndice 5.2), que es lo que hace el método del factor principal.

Harman (1976) ha desarrollado el algoritmo MINRES que minimiza (12.17) más eficientemente que el método del factor principal y Joreskog (1976) ha propuesto el algoritmo USL (unweighted least squares), que se basa en derivar en (12.17), obtener $\widehat{\Lambda}$ como función de $\boldsymbol{\psi}$ y luego minimizar la función resultante por un algoritmo no lineal tipo Newton-Raphson.

Ejemplo 12.5 *Con los datos de INVEST, utilizados en el ejemplo anterior, presentamos el análisis factorial para dos factores realizado con un programa de ordenador con el método del factor principal. La tabla 12.1 indica la variabilidad de ambos factores. El segundo factor explica poca variabilidad (2%) pero ha sido incluido por tener una clara interpretación.*

	Factor1	Factor2
Variabilidad	7.18	0.17
P_h	0.89	0.02
$\sum_{i=1}^{h} P_h$	0.89	0.91

Tabla 12.1: Variabilidad explicada por los dos primeros factores estimados por el método del factor principal.

El algoritmo del factor principal se inicia con $\widehat{\psi}_j = 1 - R_j^2$, *y se han realizado 14 iteraciones antes de converger a los pesos que se presentan en la tabla 12.2.*

	Factor1	Factor2	ψ_i^2
INTER.A	0.97	-0.06	0.04
INTER.B	0.87	0.16	0.22
AGRIC.	0.97	-0.03	0.06
BIOLO.	0.95	-0.24	0.02
MEDIC.	0.99	-0.10	0.02
QUIMI.	0.91	-0.09	0.17
INGEN.	0.94	0.21	0.06
FÍSICA	0.97	0.17	0.03

Tabla 12.2: Matriz de cargas de los factores y comunalidades

El primer factor es la suma de las publicaciones en todas las bases, nos da una idea de volumen. Según este factor los países quedarían ordenados en función de su producción

científica. El segundo factor contrapone la investigación en biomedicina con la investigación en tecnología. Este segundo componente separa a Japón y Reino Unido, países con una gran producción científica.

En la figura 12.1 se presenta un gráfico de los paises sobre estos dos factores. El lector debe comparar estos resultados con los obtenidos en el capítulo 5 (ejercicios 5.6 y 5.10) con componentes principales.

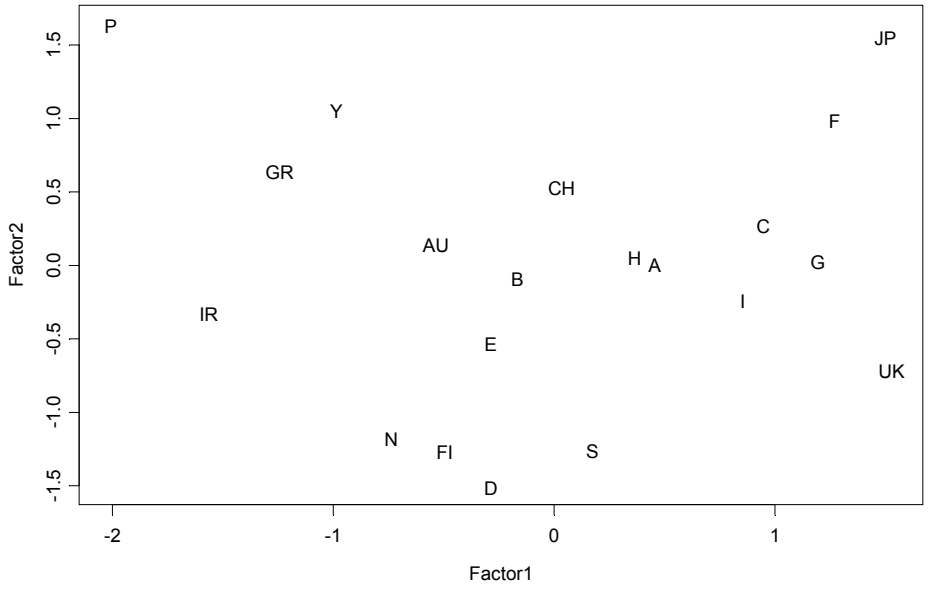


Figura 12.1: Representación de los países en el plano formado por los dos primeros factores.

12.4 ESTIMACIÓN MÁXIMO VEROSÍMIL

12.4.1 Estimación MV de los parámetros

Enfoque directo

Las matrices de parámetros pueden estimarse formalmente mediante máxima verosimilitud. La función de densidad de las observaciones originales es $N_p(\boldsymbol{\mu}, \mathbf{V})$. Por tanto la verosimilitud es la estudiada en el capítulo 10. Sustituyendo $\boldsymbol{\mu}$ por su estimador, $\overline{\mathbf{x}}$, la función soporte para $\mathbf{V}$ es:

$$\log(\mathbf{V}|\mathbf{X}) = -\frac{n}{2}\log|\mathbf{V}| - \frac{n}{2}\, tr\left(\mathbf{S}\mathbf{V}^{-1}\right), \tag{12.19}$$

y sustituyendo $\mathbf{V}$ por (12.2) la función soporte de Λ y $\boldsymbol{\psi}$ es :

$$L(\Lambda, \boldsymbol{\psi}) = -\frac{n}{2}\left(\log|\Lambda\Lambda' + \boldsymbol{\psi}| + tr(\mathbf{S}(\Lambda\Lambda' + \boldsymbol{\psi})^{-1}\right). \tag{12.20}$$

Los estimadores de máxima verosimilitud se obtienen maximizando (12.20) respecto a las matrices Λ y $\boldsymbol{\psi}$. Derivando con respecto a estas matrices y tras ciertas manipulaciones algebraicas que se resumen en el Apéndice 12.1, (vease Anderson, 1984, pp. 557-562) o Lawley y Maxwell, 1971), se obtienen las ecuaciones:

$$\widehat{\boldsymbol{\psi}} = diag\ (\mathbf{S} - \hat{\Lambda}\hat{\Lambda}') \tag{12.21}$$

$$\left(\widehat{\boldsymbol{\psi}}^{-1/2}(\mathbf{S}-\mathbf{I})\,\widehat{\boldsymbol{\psi}}^{-1/2}\right)\left(\widehat{\boldsymbol{\psi}}^{-1/2}\widehat{\Lambda}\right) = \left(\widehat{\boldsymbol{\psi}}^{-1/2}\widehat{\Lambda}\right)\mathbf{D} \tag{12.22}$$

donde $\mathbf{D}$ es la matriz resultado de la normalización

$$\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda} = \mathbf{D} = \text{diagonal.} \tag{12.23}$$

Estas tres ecuaciones permiten resolver el sistema utilizando un algoritmo iterativo tipo Newton-Raphson. La solución numérica es a veces difícil porque puede no haber una solución en la cual $\widehat{\boldsymbol{\psi}}$ sea definida positiva, y es necesario entonces acudir a la estimación con restricciones. Observemos que (12.22) conduce a una ecuación de valores propios: nos dice que $\widehat{\boldsymbol{\psi}}^{-1/2}\widehat{\Lambda}$ contienen los vectores propios de la matriz simétrica $\left(\widehat{\boldsymbol{\psi}}^{-1/2}(\mathbf{S}-\mathbf{I})\,\widehat{\boldsymbol{\psi}}^{-1/2}\right)$ y que $\mathbf{D}$ contiene los valores propios

El algoritmo iterativo para resolver estas ecuaciones es:

1. Partir de una estimación inicial. Si tenemos una estimación $\widehat{\Lambda}_i$, ($i = 1$ la primera vez), por ejemplo por el método del factor principal, se calcula la matriz $\widehat{\boldsymbol{\psi}}_i$ mediante $\widehat{\boldsymbol{\psi}}_i = diag\left(\mathbf{S} - \widehat{\Lambda}_i\widehat{\Lambda}_i'\right)$. Alternativamente, podemos estimar la matriz $\widehat{\boldsymbol{\psi}}_i$ directamente por el método del factor principal.

2. Se calcula la matriz cuadrada simétrica $\mathbf{A}_i = \widehat{\boldsymbol{\psi}}_i^{-1/2}\left(\mathbf{S} - \widehat{\boldsymbol{\psi}}_i\right)\widehat{\boldsymbol{\psi}}_i^{-1/2} = \widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2} - \mathbf{I}$. Esta matriz pondera los términos de $\mathbf{S}$ por su importancia en términos de los componentes específicos.

3. Se obtiene la descomposición espectral de $\mathbf{A}_i$ de forma que

$$\mathbf{A}_i = \mathbf{H}_{1i}\mathbf{G}_{1i}\mathbf{H}_{1i}' + \mathbf{H}_{2i}\mathbf{G}_{2i}\mathbf{H}_{2i}'$$

   donde los m mayores valores propios de $\mathbf{A}_i$ están en la matriz diagonal $(m \times m)$, $\mathbf{G}_{1i}$ y los $p - m$ menores de la $\mathbf{G}_{2i}$ y $\mathbf{H}_{1i}$ y $\mathbf{H}_{2i}$ contienen los correspondientes vectores propios.

4. Se toma $\widehat{\Lambda}_{i+1} = \widehat{\boldsymbol{\psi}}_i^{1/2}\mathbf{H}_{1i}\mathbf{G}_{1i}^{1/2}$ y se sustituye en la función de verosimilitud, que se maximiza respecto a $\boldsymbol{\psi}$. Esta parte es fácil de hacer con un algortimo de optimización no lineal. Con el resultado obtenido se vuelve a (2), iterando hasta la convergencia.

Puede ocurrir que este algoritmo converja a un máximo local donde algunos de los términos de la matriz $\boldsymbol{\psi}$ sean negativos. Esta solución impropia se denomina a veces una solución

de Heywood. Los programas existentes cambian entonces esos valores por números positivos e intentan encontrar otro máximo local, aunque no siempre el algortimo converge.

En el Apéndice 12.1 se comprueba que la estimación MV es invariante ante transformaciones lineales de las variables. En consecuencia, el resultado de la estimación no depende –como ocurre en componentes principales– del uso de la matriz de covarianzas o de correlaciones. Una ventaja adicional del método de máxima verosimilitud es que podemos obtener las varianzas asintóticas de los estimadores mediante la matriz de información en el óptimo.

Observemos que cuando la matriz $\widehat{\boldsymbol{\psi}}$ tiene los términos diagonales aproximadamente iguales, la estimación MV conducirá a resultados similares al método del factor principal. En efecto, sustituyendo en las ecuaciones del estimador MV $\widehat{\boldsymbol{\psi}} = k\mathbf{I}$, ambos métodos utilizan la misma normalización y la ecuación (12.22) es análoga a la (12.11), que se resuelve en el método del factor principal.

El algortimo EM

Un procedimiento alternativo para maximizar la verosimilitud es considerar los factores como valores ausentes y aplicar el algoritmo EM. La función de verosimiltud conjunta de los datos y los factores puede escribirse $f(\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{f}_1, ..., \mathbf{f}_n) = f(\mathbf{x}_1, ..., \mathbf{x}_n|\mathbf{f}_1, ..., \mathbf{f}_n)\ f(\mathbf{f}_1, ..., \mathbf{f}_n)$. El soporte para la muestra completa es

$$\log(\boldsymbol{\psi}, \Lambda|\mathbf{X}, \mathbf{F}) = -\frac{n}{2}\log|\boldsymbol{\psi}| - \frac{1}{2}\sum(\mathbf{x}_i - \Lambda\mathbf{f}_i)'\boldsymbol{\psi}^{-1}\ (\mathbf{x}_i - \Lambda\mathbf{f}_i) - \frac{1}{2}\sum \mathbf{f}_i\mathbf{f}_i', \qquad (12.24)$$

donde suponemos que las variables $\mathbf{x}_i$ tienen media cero, lo que equivale a sustituir la media por su estimador la media muestral. Observemos que, dados los factores, la estimación de Λ podría hacerse como una regresión. Por otro lado, dados los parámetros podríamos estimar los factores, como veremos en la sección 12.7. Para aplicar el algoritmo EM necesitamos:

(1) Paso M: maximizar la verosimilitud completa respecto a Λ y $\boldsymbol{\psi}$ supuesto conocidos los valores $\mathbf{f}_i$ de los factores. Esto es fácil de hacer, ya que las filas de Λ se obtienen haciendo regresiones entre cada variable y los factores, y los elementos diagonales de $\boldsymbol{\psi}$ son las varianzas residuales en estas regresiones.

(2) Paso E: hay que calcular la esperanza de la verosimilitud completa respecto a la distribución de los $\mathbf{f}_i$ dados los parámetros. Desarrollando (12.24) se observa que las expresiones que aparecen en la verosimilitud son la matriz de covarianzas entre los factores y la matriz de covarianzas entre los factores y los datos. Los detalles de su estimación pueden consultarse en Bartholomew y Knott (1999, p.49)

12.4.2 Otros métodos de estimación

Como el método de máxima verosimilitud es complicado, se han propuesto otros métodos aproximados para calcular estimadores con similares propiedades asintóticas pero de cálculo más simple. Uno de ellos es el de mínimos cuadrados generalizados que exponemos a continuación. Para justificarlo, observemos que la estimación MV puede reinterpretarse como sigue: si no existiesen restricciones sobre $\mathbf{V}$ el estimador MV de esta matriz es $\mathbf{S}$ y, sustituyendo esta estimación en (12.19) la función soporte en el máximo es:

$$-\frac{n}{2}\log|\mathbf{S}| - \frac{n}{2}p.$$

Maximizar la función soporte es equivalente a minimizar con respecto a $\mathbf{V}$ la función de discrepancia obtenida restando del máximo valor anterior el soporte (12.19). La función obtenida con esta diferencia es:

$$F = \frac{n}{2} tr\left(\mathbf{S}\mathbf{V}^{-1}\right) - \frac{n}{2} p - \log\left|\mathbf{S}\mathbf{V}^{-1}\right|$$

que indica que se desea hacer $\mathbf{V}$ tan próximo a $\mathbf{S}$ como sea posible, midiendo la distancia entre ambas matrices por la traza y el determinante del producto $\mathbf{S}\mathbf{V}^{-1}$. Observemos que como $\mathbf{V}$ se estima con restricciones $\left|\mathbf{S}\mathbf{V}^{-1}\right| \leq 1$ y el logaritmo será negativo o nulo. Centrándonos en los dos primeros términos, y despreciando el determinante, la función a minimizar es:

$$F_1 = tr\left(\mathbf{S}\mathbf{V}^{-1}\right) - p = tr\left(\mathbf{S}\mathbf{V}^{-1} - \mathbf{I}\right) = tr\left[\left(\mathbf{S} - \mathbf{V}\right)\mathbf{V}^{-1}\right]$$

que minimiza las diferencias entre la matriz $\mathbf{S}$ observada y la estimada $\mathbf{V}$, pero dando un peso a cada diferencia que depende del tamaño de $\mathbf{V}^{-1}$. Esto conduce a la idea de mínimos cuadrados generalizados, MCG (GLS en inglés), donde minimizamos

$$tr\left[\left(\mathbf{S} - \mathbf{V}\right)\mathbf{V}^{-1}\right]^2$$

y puede demostrarse que si se itera el procedimiento de MCG se obtienen estimadores asintóticamente eficientes

Ejemplo 12.6 *Vamos a ilustrar la estimación MV para los datos de INVEST. Suponiendo dos factores se obtienen los resultados de las tablas siguientes:*

	Factor1	Factor2
Variabilidad	6.80	0.53
P_h	0.85	0.06
$\sum_{i=1}^{h} P_h$	0.85	0.91

Tabla 12.3: Variabilidad explicada por los dos primeros factores estimados por máxima verosimilitud.

	Factor1	Factor2	ψ_i^2
INTER.A	0.95	0.25	0.02
INTER.B	0.85	0.08	0.26
AGRIC.	0.92	0.26	0.07
BIOLO.	0.88	0.45	0.01
MEDIC.	0.93	0.3	0.02
QUIMI.	0.86	0.29	0.17
INGEN.	0.95	0.05	0.09
FÍSICA	1	0	0

Tabla 12.4: Matriz de cargas de los factores

Si comparamos estos resultados con los obtenidos por el método del factor principal (ejercicio 12.5) vemos que el primer factor es similar, aunque aumenta el peso de la física y hay

más diferencias relativas entre los pesos de las variables. El segundo factor tiene más cambios pero su interpretación es también similar. Las varianzas de los componentes específicos presentan pocos cambios con ambos enfoques. Las figuras 12.2 y 12.3 presentan los pesos y la proyección de los datos sobre el plano de los factores.

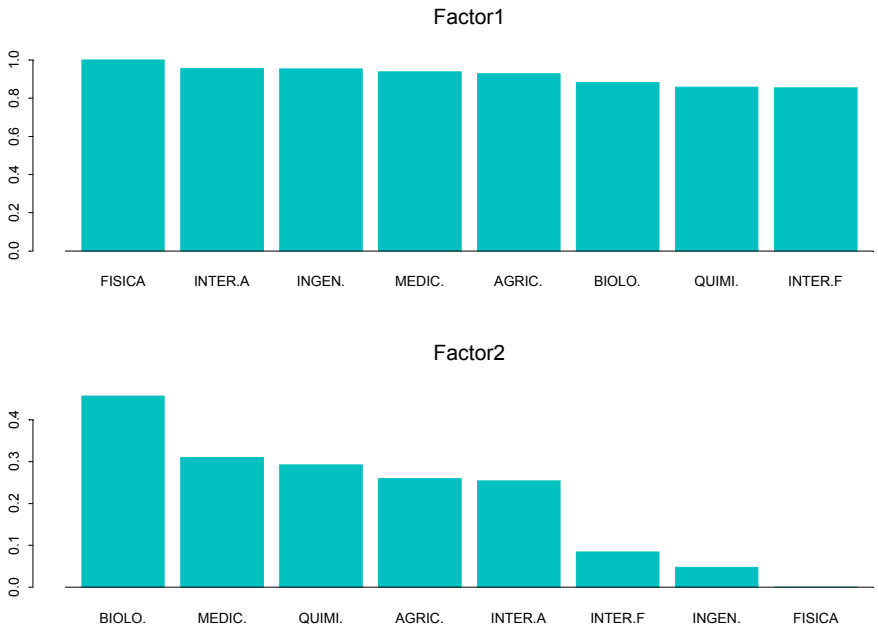


Figura 12.2: Pesos de los variables de INVEST en los dos factores estimados por MV.

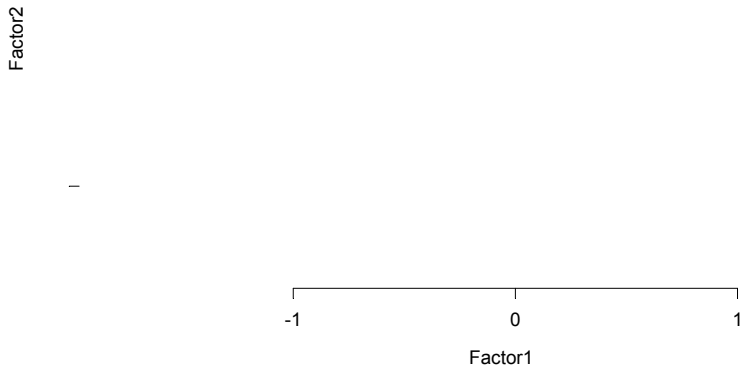
12.5 DETERMINACIÓN DEL NÚMERO DE FACTORES

12.5.1 Contraste de verosimilitud

Supongamos que se ha estimado un modelo con m factores. El contraste de que la descomposición es adecuada puede plantearse como un contraste de razón de verosimilitudes:

$$\begin{aligned} H_0 &: \mathbf{V} = \Lambda\Lambda' + \boldsymbol{\psi} \\ H_1 &: \mathbf{V} \neq \Lambda\Lambda' + \boldsymbol{\psi}. \end{aligned}$$

Este contraste recuerda al de esfericidad parcial que estudiamos en el capítulo 10, aunque existen diferencias porque no exigimos que los componentes específicos tengan igual varianza. El contraste se deduce con los mismos principios que estudiamos en el capítulo 10. Sea $\hat{\mathbf{V}}_0$ el valor de la matriz de varianzas y covarianzas de los datos estimados bajo H_0. Entonces,


Factor2
-1
0
1
Factor1

y, por tanto, mide la distancia entre $\hat{\mathbf{V}}_0$ y e $\mathbf{S}$ en términos del determinante, $-n \log \left|\mathbf{S}\widehat{\mathbf{V}}_0^{-1}\right|$, que es el segundo término de la verosimilitud.

El contraste rechaza H_0 cuando λ sea mayor que el percentil $1-\alpha$ de una distribución χ^2_g con grados de libertad, g, dados por $g = dim(H_1) - dim(H_0)$. La dimensión del espacio paramétrico de H_1 es $p + \begin{pmatrix} p \\ 2 \end{pmatrix} = p(p+1)/2$, igual al número de elementos distintos de $\mathbf{V}$. La dimensión de H_0 es pm – por la matriz Λ – más los p elementos de $\boldsymbol{\psi}$, menos $m(m-1)/2$ restricciones resultantes de la condición que $\Lambda'\boldsymbol{\psi}^{-1}\Lambda$ debe ser diagonal. Por tanto:

$$\begin{aligned} g &= p + p(p-1)/2 - pm - p + m(m-1)/2 = \\ &= (1/2)\left((p-m)^2 - (p+m)\right) \end{aligned} \tag{12.28}$$

Bartlett (1954) ha demostrado que la aproximación asintótica de la distribución χ^2 mejora en muestras finitas introduciendo un factor de corrección. Con esta modificación, el test es rechazar H_0 si

$$(n - 1 - \frac{2p+4m+5}{6}) \ln \frac{|\hat{\Lambda}\hat{\Lambda}' + \hat{\boldsymbol{\psi}}|}{|\mathbf{S}|} > \chi^2_{\left[\left((p-m)^2-(p+m)\right)/2\right]}(1-\alpha) \tag{12.29}$$

Generalmente este contraste se aplica secuencialmente: Se estima el modelo con un valor pequeño, $m = m_1$ (que puede ser $m_1 = 1$) y se contrata H_0. Si se rechaza, se reestima con $m = m_1 + 1$, continuando hasta aceptar H_0.

Un procedimiento alternativo, propuesto por Joreskog (1993), que funciona mejor ante moderadas desviaciones de la normalidad es el siguiente: calcular el estadístico (12.29) para $m = 1, \ldots, m_{\max}$. Sean $X^2_1, \ldots, X^2_{m_{\max}}$ sus valores y $g_1, \ldots, g_{m_{\max}}$ sus grados de libertad. Calcularemos las diferencias $X^2_m - X^2_{m+1}$ y consideramos estas diferencias como valores de una χ^2 con $g_m - g_{m+1}$ grados de libertad. Si el valor obtenido es significativo aumentamos el número de factores y procedemos así hasta que no encontremos una mejora significativa en el ajuste del modelo.

El contraste (12.27) admite una interesante interpretación. El modelo factorial establece que la diferencia entre la matriz de covarianzas, $\mathbf{S}$ $(p \times p)$, y una matriz diagonal de rango p, $\boldsymbol{\psi}$, es aproximadamente una matriz simétrica de rango m, $\Lambda\Lambda'$, es decir:

$$\mathbf{S} - \widehat{\boldsymbol{\psi}} \simeq \widehat{\Lambda}\widehat{\Lambda}'.$$

Premultiplicando y postmultiplicando por $\widehat{\boldsymbol{\psi}}^{-1/2}$ se obtiene que la matriz $\mathbf{A}$, dada por:

$$\mathbf{A} = \widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2} - \mathbf{I}, \tag{12.30}$$

debe ser asintóticamente igual a la matriz:

$$\mathbf{B} = \widehat{\boldsymbol{\psi}}^{-1/2}\widehat{\Lambda}\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1/2}, \tag{12.31}$$

y tener asintóticamente rango m, en lugar de rango p. Se demuestra en el apéndice 12.2 que el contraste (12.27) equivale a comprobar si la matriz $\mathbf{A}$ tiene rango m, lo que debe ser

asintóticamente cierto por (12.31), y que el test (12.27) puede escribirse

$$\lambda = -n \sum_{m+1}^{p} \log(1 + d_i) \tag{12.32}$$

donde d_i son las $p - m$ menores raíces características de la matriz $\mathbf{A}$. La hipótesis nula se rechaza si λ es demasiado grande comparada con la distribución χ^2 con $(1/2)((p-m)^2-p-m)$. En el Apéndice 12.2 se demuestra que este contraste es un caso particular del contraste de verosimilitud sobre la esfericidad parcial de una matriz que presentamos en 10.6.

Cuando el tamaño muestral es grande y m es pequeño con relación a p, si los datos no siguen una distribución normal multivariante el contraste conduce generalmente a rechazar H_0. Este es un problema frecuente en contraste de hipótesis con muestras grandes, donde tendemos a rechazar H_0. Por tanto, es necesario a la hora de decidir el número de factores, diferenciar entre significatividad práctica y significatividad estadística, como ocurre en todo contraste de hipótesis. Este contraste es muy sensible a desviaciones de la normalidad por lo que en la práctica el estadístico (12.27) se utiliza como medida de ajuste del modelo más que como un test formal.

12.5.2 Criterios de selección

Una alternativa a los contrastes es plantear el problema como uno de selección de modelos. Entonces estimaremos el modelo factorial para distinto número de factores, calcularemos la función soporte en el máximo para cada modelo y, aplicando el criterio de Akaike, elegiremos aquel modelo donde

$$AIC(m) = -2L(H_{0,m}) + 2n_p$$

sea mínimo. En esta expresión $2L(H_{0,m})$ es la función soporte para el modelo que establece m factores particularizada en los estimadores MV, que viene dada por (12.25), y n_p es el número de parámetros en el modelo. Observemos que esta expresión tiene en cuenta que al aumentar m la verosimilitud de $L(H_{0,m})$ aumenta, o la desviación $-2L(H_{0,m})$ disminuye, pero este efecto se contrapesa con el número de parámetros que aperece penalizando la relación anterior. Este mismo criterio de selección puede escribirse como minimizar las diferencias $AIC(m) - AIC(H_1)$, donde en todos los modelos restamos la misma cantidad, $AIC(H_1)$, que es el valor del AIC para el modelo que supone que no existe estructura factorial y que estima la matriz de covarianzas sin restricciones. Entonces la función a minimizar es

$$AIC^*(m) = 2(L(H_1) - L(H_{0,m})) - 2g = \lambda(m) - 2g$$

donde $\lambda(m)$ es la diferencia de soportes (12.27), donde en esta expresión $\widehat{\mathbf{V}}_0$ se estima con m factores, y g es el número de grados de libertad dado por (12.28).

Un criterio alternativo es el BIC presentado en el capítulo 11. Con este criterio en lugar de penalizar el número de parámetros con 2 lo hacemos con logn. Este criterio aplicado a la seleccion del modelo factorial mediante las diferencias de soporte es:

$$BIC(m) = \lambda(m) - g \log n$$

Ejemplo 12.7 *Aplicaremos el método de máxima verosimilitud a los datos de INVEST, para realizar un contraste sobre el número de factores. Si basamos el contraste en la expresión* (12.27) *obtenemos la siguiente tabla,*

m	λ	g_m	$p-valor$	AIC	BIC
1	31.1	20	**0.053**	−8.9	−29.79
2	11.73	13	0.55	−14.27	−27.84
3	6.49	7	0.484	−7.51	−14.82
4	5.27	2	0.072	1.27	$-0,73$

por ejemplo, si $m = 1$ *el número de grados de libertad es* $(1/2)\left((7)^2 - (9)\right) = 20$. *Vemos que para* $\alpha = 0.05$ *no podemos rechazar la hipótesis nula de que un factor es suficiente. Sin embargo, el criterio de Akaike indica que el mínimo se obtiene con dos factores, y el criterio BIC confirma, por poca diferencia, la elección de un factor.*

Como el $p-$*valor del contraste anterior está en el límite vamos a comparar este test con el procedimiento propuesto por Joreskog. El primer paso es utilizar la corrección propuesta por Barlet, que realizamos multiplicando los estadísticos* χ^2_m *por* $(n-1-(2p+4m+5)/6)/n$. *Por ejemplo, el estadístico corregido para* $p = 1$ *es,*

$$X_1^2 = ((20 - 1 - (2 * 8 + 4 * 1 + 5)/6)/20) * 31.1 = 23.06$$

en la siguiente tabla presentamos los resultados.

p	X^2_m	$X^2_m - X^2_{m+1}$	$g_m - g_{m+1}$	$p-valor$
1	23.06	14.76	7	0.039
2	8.30	3.92	6	**0.687**
3	4.38	1	5	0.962
4	3.38			

Este método indica que rechazamos la hipótesis de un factor, pero no podemos rechazar la hipótesis de dos factores, con lo que concluimos que el número de factores, escogido con el método de Joreskog, es igual a dos. Como vemos, en este ejemplo el criterio de Joreskov coindice con el criterio de Akaike

Ejemplo 12.8 *Para los datos EPF de la Encuesta de Presupuestos Familiares del Anexo A.3, aplicaremos la técnica de análisis factorial con la estimación máximo verosímil. Los datos han sido transformados en logaritmos para mejorar la asimetría, al igual que se hizo en el análisis de Componentes Principales presentado en los ejemplos 4.2 y 4.3. Para este análisis también hemos estandarizado las observaciones.*

Aceptamos el contraste de un único factor dado que el p-valor es 0.242. *La estimación de los pesos de este factor es aproximadamente una ponderación con menor peso en los epígrafes de alimentación, vestido y calzado, como se muestra en la tabla 12.5*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Factor 1	0.61	0.64	0.86	0.88	0.82	0.84	0.93	0.89	0.72

Tabla 12.5: Vector de cargas de los factores

12.6 ROTACIÓN DE LOS FACTORES

Como vimos en la sección 12.2.3, la matriz de carga no esta identificada ante multiplicaciones por matrices ortogonales, que equivalen a rotaciones. En análisis factorial está definido el espacio de las columnas de la matriz de carga, pero cualquier base de este espacio puede ser una solución. Para elegir entre las posibles soluciones, se tienen en cuenta la interpretación de los factores. Intuitivamente, será más fácil interpretar un factor cuando se asocia a un bloque de variables observadas. Esto ocurrirá si las columnas de la matriz de carga, que representan el efecto de cada factor sobre las variables observadas, contienen valores altos para ciertas variables y pequeños para otras. Esta idea puede plantearse de distintas formas que dan lugar a distintos criterios para definir la rotación. Los coeficientes de la matriz ortogonal que define la rotación se obtendrán minimizando una función objetivo que expresa la simplicidad deseada en la representación conseguida al rotar. El criterio más utilizado es el Varimax, que exponemos a continuación.

Criterio Varimax

La interpretación de los factores se facilita si los que afectan a algunas variables no lo hacen a otras y al revés. Este objetivo conduce al criterio de maximizar la varianza de los coeficientes que definen los efectos de cada factor sobre las variables observadas. Para precisar este criterio, llamemos δ_{ij} a los coeficientes de la matriz de carga asociados al factor j en las $i = 1, ..., p$ ecuaciones después de la rotación y $\boldsymbol{\delta}_j$ al vector que es la columna j de la matriz de carga después de la rotación. Se desea, que la varianza de los coeficientes al cuadrado de este vector sea máxima. Se toman los coeficientes al cuadrado para prescindir de los signos, ya que interesa su valor absoluto. Llamando $\overline{\delta}_{.j} = \sum \delta_{ij}^2/p$ a la media de los cuadrados de los componentes del vector $\boldsymbol{\delta}_j$, la variabilidad para el factor j es:

$$\frac{1}{p}\sum_{i=1}^{p}(\delta_{ij}^2 - \overline{\delta}_{.j})^2 = \frac{1}{p}\sum_{i=1}^{p}\delta_{ij}^4 - (1/p)^2(\sum_{i=1}^{p}\delta_{ij}^2)^2, \tag{12.33}$$

y el criterio es maximizar la suma de las varianzas para todos los factores, dada por:

$$q = (1/p)\sum_{j=1}^{m}\sum_{i=1}^{p}\delta_{ij}^4 - (1/p)^2\sum_{j=1}^{m}(\sum_{i=1}^{p}\delta_{ij}^2)^2. \tag{12.34}$$

Sea Λ la matriz de carga estimada inicialmente. El problema es encontrar una matriz ortogonal $\mathbf{M}$ tal que la matriz $\boldsymbol{\delta}$ dada por

$$\boldsymbol{\delta} = \Lambda\mathbf{M},$$

y cuyos coeficientes δ_{ij} viene dados por

$$\delta_{ij} = \boldsymbol{\lambda}_i' \, \mathbf{m}_j$$

siendo $\boldsymbol{\lambda}_i'$ la fila i de la matriz Λ y $\mathbf{m}_j$ la columna j de la matriz $\mathbf{M}$ que buscamos, verifique la condición de que estos coeficiente maximicen (12.34). Los términos de la matriz $\mathbf{M}$ se obtendrán derivando (12.34) respecto a cada uno de sus términos m_{ij} teniendo en cuenta las restricciones de ortogonalidad $\mathbf{m}_i'\mathbf{m}_i = 1$; $\mathbf{m}_i'\mathbf{m}_j = 0$ $(i \neq j)$. El resultado obtenido es la rotación varimax.

Ejemplo 12.9 *Si aplicamos una rotación varimax a la estimación MV de los datos de INVEST del ejemplo 7.6 se obtiene el resultado presentado en la figura 12.4. Esta nueva matriz*

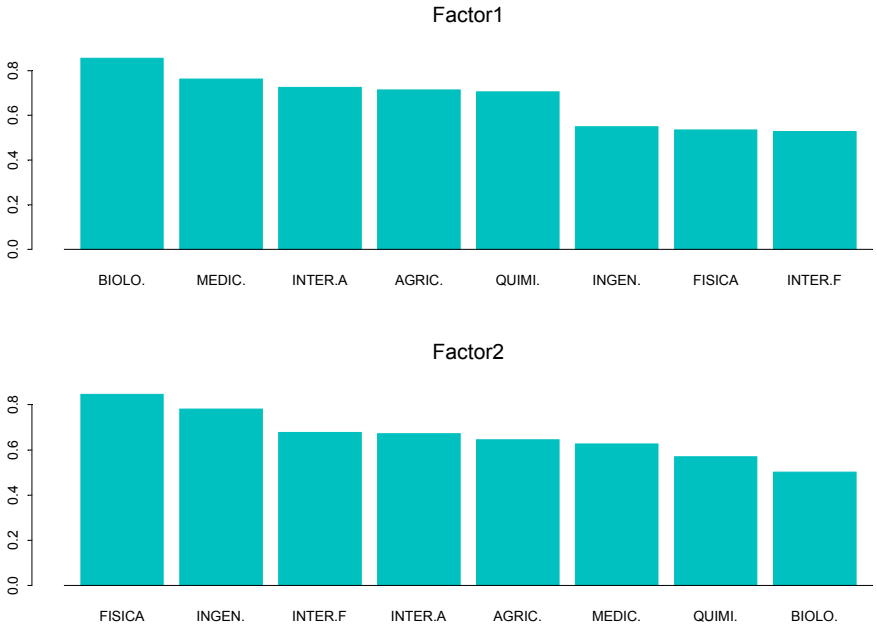


Figura 12.4: Resultado de aplicar una rotación varimax para los factores de INVES.

de cargas resulta al multiplicar los coeficientes de la matriz ortogonal que definen la rotación, M, por la matriz de cargas obtenida en la estimación MV y presentada en el ejemplo 12.6

$$\delta = \overset{\Lambda}{\begin{bmatrix} 0.95 & 0.25 \\ 0.85 & 0.08 \\ 0.92 & 0.26 \\ 0.88 & 0.45 \\ 0.93 & 0.3 \\ 0.86 & 0.29 \\ 0.95 & 0.05 \\ 1 & 0 \end{bmatrix}} \overset{M}{\begin{bmatrix} 0.53 & 0.85 \\ 0.85 & -0.53 \end{bmatrix}} = \begin{bmatrix} 0.71 & 0.67 \\ 0.52 & 0.68 \\ 0.71 & 0.64 \\ 0.85 & 0.51 \\ 0.75 & 0.63 \\ 0.70 & 0.58 \\ 0.55 & 0.78 \\ 0.53 & 0.85 \end{bmatrix}$$

Rotaciones oblicuas

El modelo factorial está indeterminado no sólo ante rotaciones ortogonales sino ante rotaciones oblicuas. En efecto, como vimos en la sección 6.1 el modelo puede establecerse con factores incorrelados o correlados. La solución obtenida de la estimación de Λ corresponde siempre a factores incorrelados, pero podemos preguntarnos si existe una solución con factores correlados que tenga una interpretación más interesante. Matemáticamente esto implica definir nuevos factores $\mathbf{f}^* = \mathbf{H}\mathbf{f}$, donde $\mathbf{H}$ es una matriz no singular que puede interpretarse, en general, como un giro oblicuo. La nueva matriz de varianzas y covarianzas de los factores será $\mathbf{V}_f^* = \mathbf{H}\mathbf{H}'$.

Existen diversos procedimientos para obtener rotaciones oblicuas, como el Quartmin, Oblimax, Promax, etc. que el lector puede consultar en la literatura especializada. El problema de las rotaciones oblicuas es que los factores, al estar correlados, no pueden interpretarse independientemente.

12.7 ESTIMACIÓN DE LOS FACTORES

En muchos problemas el interés del análisis factorial es determinar la matriz de carga, y los valores particulares de los factores en los elementos de la muestra no tienen interés. Sin embargo, en otros casos se desean obtener los valores de las variables factores sobre los elementos observados. Existen dos procedimientos para estimar los factores: el primero, debido a Bartlett, supone que el vector de valores de los factores para cada observación es un parámetro a estimar. El segundo, supone que son variables aleatorias. Vamos a revisar brevemente ambos procedimientos.

12.7.1 Los factores como parámetros

El vector $(p \times 1)$ de valores de las variables en el individuo i, $\mathbf{x}_i$, tiene una distribución normal con media $\Lambda\mathbf{f}_i$, donde $\mathbf{f}_i$ es el vector $(m \times 1)$ de factores para el elemento i en la muestra, y matriz de covarianzas $\boldsymbol{\psi}$, es decir

$$\mathbf{x}_i \sim N_p(\Lambda\mathbf{f}_i, \boldsymbol{\psi})$$

Los parámetros $\mathbf{f}_i$ pueden estimarse por máxima verosimilitud como se indica en el Apéndice 12.3. El estimador resultante es el de mínimos cuadrados generalizados, dado por

$$\widehat{\mathbf{f}}_i = \left(\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}\right)^{-1}\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\mathbf{x}_i. \tag{12.35}$$

que tiene una clara interpretación intuitiva: si conocemos Λ, el modelo factorial

$$\mathbf{x}_i = \Lambda\mathbf{f}_i + \mathbf{u}_i$$

es un modelo de regresión con variable dependiente $\mathbf{x}_i$, variables explicativas las columnas de Λ y parámetros $\mathbf{f}_i$. Como la perturbación, $\mathbf{u}_i$, no se distribuye como $N(\mathbf{0}, \mathbf{I})$ sino $N(\mathbf{0}, \boldsymbol{\psi})$, tendremos que utilizar mínimos cuadrados generalizados, lo que conduce a (12.35).

12.7.2 Los factores como variables aleatorias

El segundo método es suponer que los factores son variables aleatorias, y buscar un predictor lineal que minimice el error cuadrático medio de predicción. Llamando $\mathbf{f}_i$ como antes a los valores de los factores en el individuo i y $\mathbf{x}_i$ al vector de variables observadas, el vector $(\mathbf{f}_i, \mathbf{x}_i)$ tendrá una distribución normal multivariante y el objetivo es encontrar $E[\mathbf{f}_i|\mathbf{x}_i]$. Por los resultados de la sección 8.5.1 tenemos que:

$$E[\mathbf{f}_i|\mathbf{x}_i] = E[\mathbf{f}_i] + Cov(f_i, x_i)\, Var(x_i)^{-1}(x_i - E(\mathbf{x}_i))$$

Como $E[\mathbf{f}_i] = 0$ y las covarianzas entre los factores y las variables son los términos de la matriz de carga, podemos escribir, suponiendo variables de media cero y los parámetros conocidos:

$$\widehat{\mathbf{f}}_i = E[\mathbf{f}_i|\mathbf{x}_i] = \Lambda'\mathbf{V}^{-1}\mathbf{x}_i \tag{12.36}$$

que es el predictor regresión lineal de los factores sobre los datos. En efecto Λ' representa las covarianzas entre factores y variables y $\mathbf{V}$ las covarianzas entre variables. Esta ecuación puede también escribirse (vease el Apéndice 12.3) como

$$\widehat{\mathbf{f}}_i = (\mathbf{I} + \Lambda'\boldsymbol{\psi}^{-1}\Lambda)^{-1}\Lambda'\boldsymbol{\psi}^{-1}\mathbf{x}_i. \tag{12.37}$$

Comparando (12.35) y (12.36) vemos que éste último método puede interpretarse como una regresión cresta (ridge regression). Supone sumar la unidad a los elementos diagonales de la matriz $\Lambda'\boldsymbol{\psi}^{-1}\Lambda$. Este estimador tiene también una interpretación bayesiana que se presenta en el Apéndice 12.4.

Si en las ecuaciones (12.35) y (12.37) sustituimos los valores teóricos por los estimados, obtenemos un vector $\widehat{\mathbf{f}}_i$ que representa la estimación del valor de los m factores en el individuo i. Aplicando sucesivamente estas ecuaciones a los n datos muestrales, $\mathbf{x}_1, ..., \mathbf{x}_n$, obtendremos los valores de los factores para los n individuos, $\mathbf{f}_1, ..., \mathbf{f}_n$, donde cada $\mathbf{f}_i$ es un vector $(m \times 1)$.

Ejemplo 12.10 *Con los datos de ACCIONES estimaremos los valores del factor supuesta la matriz de carga estimada en ejemplo 12.3 para las variables en logaritmos. Vamos a detallar su obtención para las primeras 5 acciones de dicho ejemplo. La matriz de datos X contendrá estas 5 observaciones.*

$$X = \begin{bmatrix} 1.22 & 4.5 & 3.41 \\ 1.63 & 4.02 & 2.29 \\ 1.5 & 3.96 & 2.44 \\ 1.25 & 3.85 & 2.42 \\ 1.77 & 3.75 & 1.95 \end{bmatrix}$$

Comencemos por el primer método, mínimos cuadrados generalizados. Los estimadores de $\widehat{\Lambda}$, $\widehat{\boldsymbol{\psi}}^{-1}$, *obtenidos en el ejemplo 12.3 son,*

$$\widehat{\Lambda} = \begin{bmatrix} -0.269 \\ -0.229 \\ 0.407 \end{bmatrix}; \quad \widehat{\boldsymbol{\psi}}^{-1} = \begin{bmatrix} 1.984 & 0 & 0 \\ 0 & 3.834 & 0 \\ 0 & 0 & 9.534 \end{bmatrix}$$

y aplicando las fórmulas obtenemos,

$$\left(\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}\right)^{-1} = 0.5; \qquad \left(\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}\right)^{-1}\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1} = \begin{bmatrix} -0.55 \\ -1.75 \\ 19.24 \end{bmatrix}'$$

Los 5 primeros valores del primer factor se calculan con $\widehat{\mathbf{f}}_i = X\left(\left(\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}\right)^{-1}\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\right)$,

$$\widehat{\mathbf{f}}_i = \begin{bmatrix} 1.22 & 4.5 & 3.41 \\ 1.63 & 4.02 & 2.29 \\ 1.5 & 3.96 & 2.44 \\ 1.25 & 3.85 & 2.42 \\ 1.77 & 3.75 & 1.95 \end{bmatrix} \begin{bmatrix} -0.55 \\ -1.75 \\ 19.24 \end{bmatrix} = \begin{bmatrix} 57.062 \\ 36.128 \\ 39.191 \\ 39.136 \\ 29.982 \end{bmatrix}$$

Para estimar los valores por el segundo método calcularemos :

$$\left(\mathbf{I}+\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}\right)^{-1} = 0.342; \qquad \left(\mathbf{I}+\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}\right)^{-1}\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1} = \begin{bmatrix} -0.36 \\ -1.15 \\ 12.65 \end{bmatrix}'$$

y los 5 primeros valores del primer factor se calculan con $\widehat{\mathbf{f}}_i = X\left(\mathbf{I} + \widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}\right)^{-1}\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1'}$,

$$\widehat{\mathbf{f}}_i = \begin{bmatrix} 1.22 & 4.5 & 3.41 \\ 1.63 & 4.02 & 2.29 \\ 1.5 & 3.96 & 2.44 \\ 1.25 & 3.85 & 2.42 \\ 1.77 & 3.75 & 1.95 \end{bmatrix} \begin{bmatrix} -0.36 \\ -1.15 \\ 12.65 \end{bmatrix} = \begin{bmatrix} 37.52 \\ 23.75 \\ 25.77 \\ 25.73 \\ 19.72 \end{bmatrix}$$

Observemos que ambas estimaciones presentan la misma estructura, pero el efecto de contracción del segundo método hace que los valores obtenidos sean menores.

12.8 DIAGNOSIS DEL MODELO

Residuos de los factores

Para contrastar si el modelo es adecuado conviene calcular los factores $\widehat{\mathbf{f}}$ y los residuos $\mathbf{e}$ y estudiar sus propiedades. De acuerdo con las hipótesis:

$$\mathbf{u} \sim \mathbf{N}_p(\mathbf{0}, \boldsymbol{\psi})$$

Por tanto, si la matriz de covarianzas de los residuos no es diagonal debemos aumentar el número de factores hasta que los residuos estimados:

$$\widehat{\mathbf{u}}_i = \mathbf{e}_i = \mathbf{x}_i - \Lambda\widehat{\mathbf{f}}_i$$

verifiquen las hipótesis. En concreto, contrastaremos si los residuos tienen una distribución normal. Los residuos pueden indicarnos también la presencia de observaciones atípicas o de grupos de observaciones que no se ajustan bien al modelo construido.

Ejemplo 12.11 *Para los datos de EPF calculamos la matriz de varianzas covarianzas de los residuos del modelo estimado en el ejercicio 12.8, donde se estimó un único factor.*

$$\hat{\boldsymbol{\psi}} = \begin{bmatrix} \mathbf{0.61} & 0.13 & -0.04 & -0.03 & -0.06 & -0.03 & 0.02 & -0.11 & 0.05 \\ 0.13 & \mathbf{0.57} & -0.01 & 0.04 & -0.01 & 0.01 & -0.05 & -0.13 & 0.04 \\ -0.04 & -0.01 & \mathbf{0.22} & -0.07 & -0.07 & -0.05 & -0.01 & 0.01 & -0.06 \\ -0.03 & 0.04 & -0.07 & \mathbf{0.19} & -0.03 & -0.01 & -0.05 & -0.03 & 0.02 \\ -0.06 & -0.01 & -0.07 & -0.03 & \mathbf{0.3} & 0 & -0.02 & -0.04 & -0.01 \\ -0.03 & 0.01 & -0.05 & -0.01 & 0 & \mathbf{0.26} & -0.06 & -0.05 & 0.1 \\ 0.02 & -0.05 & -0.01 & -0.05 & -0.02 & -0.06 & \mathbf{0.1} & -0.01 & -0.1 \\ -0.11 & -0.13 & 0.01 & -0.03 & -0.04 & -0.05 & -0.01 & \mathbf{0.18} & -0.05 \\ 0.05 & 0.04 & -0.06 & 0.02 & -0.01 & 0.1 & -0.1 & -0.05 & \mathbf{0.45} \end{bmatrix}$$

En la diagonal figura la varianza específica, se aprecia que los términos de fuera de la diagonal son relativamente pequeños. Compararemos esta matriz con la resultante de estimar dos factores, que en el contraste tiene un p-valor de 0.64*, la nueva matriz de varianzas de los residuos es:*

$$\hat{\boldsymbol{\psi}} = \begin{bmatrix} \mathbf{0.6} & 0.12 & -0.04 & -0.04 & -0.07 & -0.06 & 0.02 & -0.1 & 0.01 \\ 0.12 & \mathbf{0.53} & 0.03 & 0.02 & -0.01 & -0.04 & -0.01 & -0.09 & -0.08 \\ -0.04 & 0.03 & \mathbf{0.22} & -0.04 & -0.05 & -0.01 & -0.03 & 0.01 & 0 \\ -0.04 & 0.02 & -0.04 & \mathbf{0.19} & -0.02 & -0.04 & -0.02 & 0 & -0.05 \\ -0.07 & -0.01 & -0.05 & -0.02 & \mathbf{0.3} & -0.01 & -0.01 & -0.02 & -0.03 \\ -0.06 & -0.04 & -0.01 & -0.04 & -0.01 & \mathbf{0.18} & -0.01 & -0.01 & -0.05 \\ 0.02 & -0.01 & -0.03 & -0.02 & -0.01 & -0.01 & \mathbf{0.03} & -0.04 & 0 \\ -0.1 & -0.09 & 0.01 & 0 & -0.02 & -0.01 & -0.04 & \mathbf{0.19} & 0.01 \\ 0.01 & -0.08 & 0 & -0.05 & -0.03 & -0.05 & 0 & 0.01 & \mathbf{0.17} \end{bmatrix}$$

y la variabilidad especifica ha disminuido en las variables X_7 y X_9 y fuera de la diagonal, en general, los valores son más pequeños. Se podía incrementar el número de factores en uno más, pero se corre el peligro de sobreajustar el modelo y que la interpretación de los pesos esté demasiado sujeta a los datos en concreto usados.

En las figuras 12.5 y 12.6 se presentan las cargas de los factores y la representación de las distintas provincias en el espacio formado por estos dos factores.Las cargas del segundo factor permiten una interpretación análoga a la descrita para el segundo componente principal. En la figura 12.7 presentamos los histogramas de las distribuciones marginales de los residuos. Algunos de estos histogramas no parecen seguir una distribución normal.

Residuos del ajuste

Se definen los residuos del ajuste como los términos de $\mathbf{S}-\widehat{\mathbf{V}}$. Frecuentemente es más cómodo utilizar los residuos estandarizados, donde cada residuo se divide por su desviación típica asintótica.

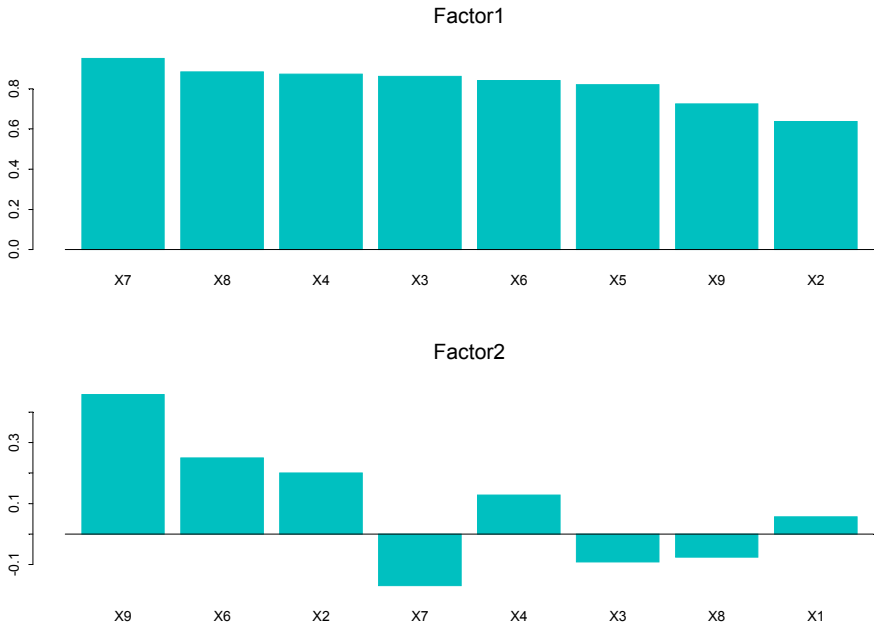


Figura 12.5: Representación de la matriz de cargas de los dos primeros factores estimados por máxima verosimilitud.

Medidas de ajuste del modelo

Podemos construir una medida del ajuste del modelo factorial para cada variable mediante

$$\gamma_i^2 = \frac{h_i^2}{s_i^2} = 1 - \frac{\psi_i^2}{s_i^2}$$

que suele denominarse coeficiente de correlación al cuadrado entre la variable y los factores. El coeficiente de determinación para todo el sistema puede construirse con

$$R^2 = 1 - \frac{\left|\widehat{\boldsymbol{\psi}}\right|^{1/p}}{\left|\widehat{\mathbf{V}}\right|^{1/p}},$$

donde $\left|\widehat{\boldsymbol{\psi}}\right|$es el determinante de la matriz de varianzas residuales y $\left|\widehat{\mathbf{V}}\right|$ el estimado por el modelo.

El estadístico χ^2 dado por (12.27) proporciona otra medida global de ajuste. Para calibrar su valor lo compararemos con sus grados de libertad. Cuando los datos no son normales la distribución de (12.27) puede desviarse mucho de la χ^2 pero, en cualquier caso, su valor puede utilizarse como un criterio de ajuste.

Ejemplo 12.12 *Calculamos el coeficiente de correlación al cuadrado entre la variable y los factores para los datos el ejemplo 12.8 en el modelo con dos factores. Como las variables*

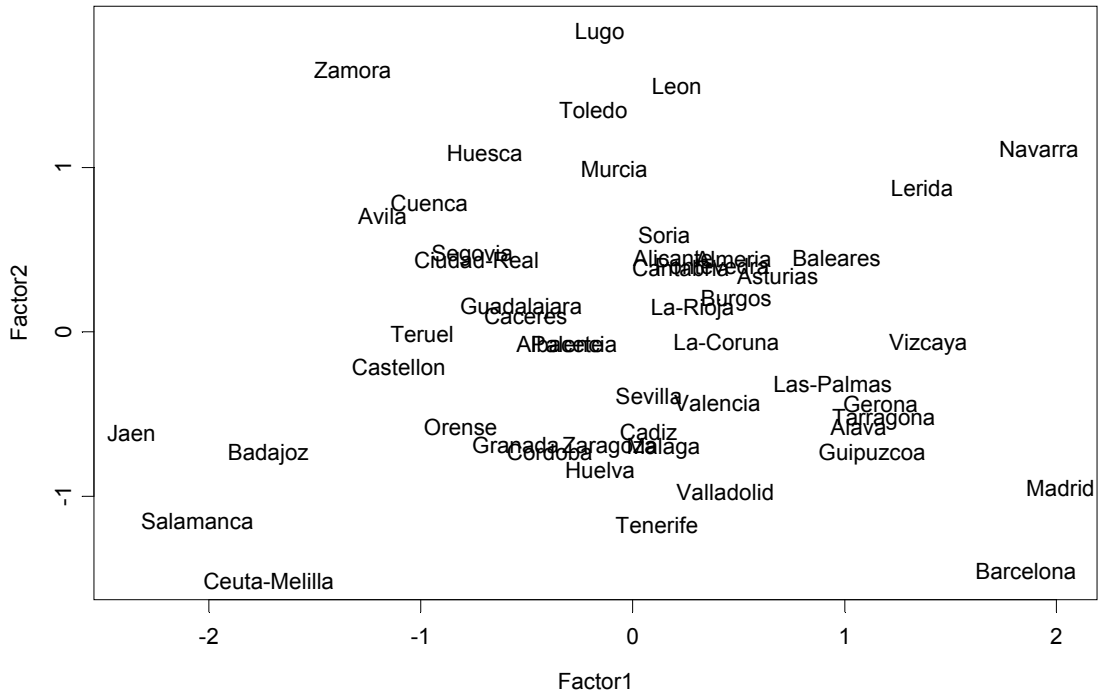


Figura 12.6: Representación de las provincias en los dos primeros factores.

originales estaban estandarizadas, $s_i^2 = 1$ *para* $i = 1, \ldots, 9$, *los coeficientes se calculan como 1 menos la varianza especifica.*

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
γ_i^2	0.4	0.47	0.78	0.81	0.7	0.82	0.97	0.81	0.83

El coeficiente de determinación es

$$R^2 = 1 - (\frac{1.781499 \times 10^{-8}}{0.0002415762})^{1/9} = .652$$

y vemos que proporciona un valor promedio de las relaciones de dependencia en el sistema.

12.9 Análisis Factorial Confirmatorio

El análisis factorial puede aplicarse como una herramienta exploratoria o como un modelo para contrastar teorías. En este segundo caso, el número de factores se supone conocido a priori y se establecen restricciones sobre los elementos de la matriz de carga. Por ejemplo, algunos pueden ser cero o iguales entre sí. Dada la existencia de información adicional,

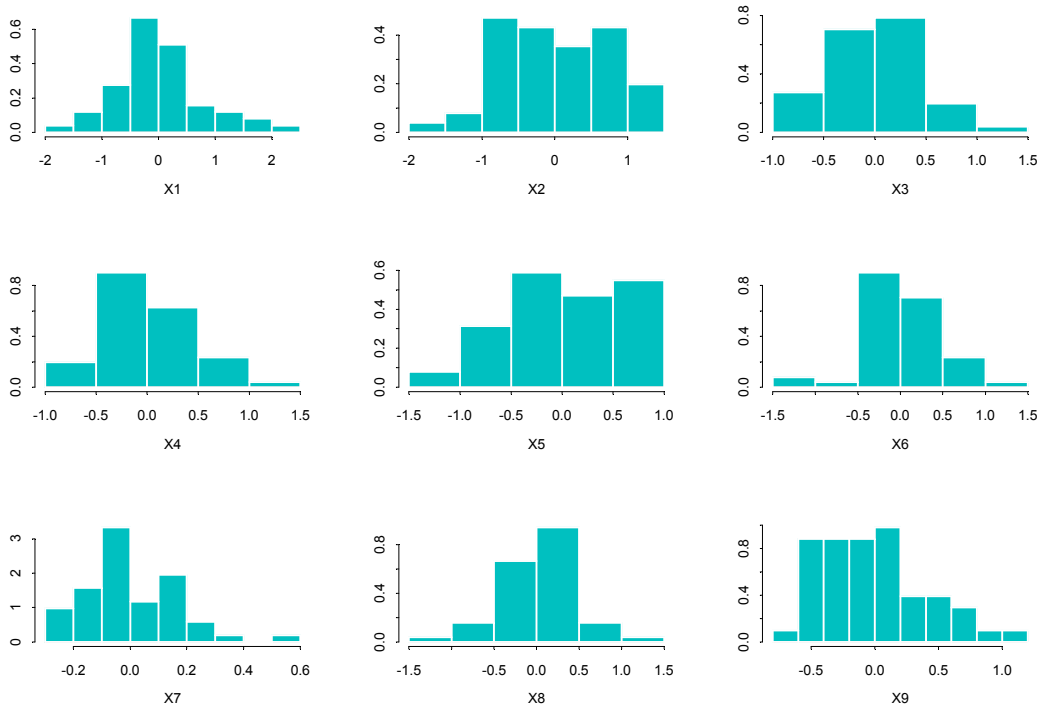


Figura 12.7: Histograma de las distribuciones marginales de los residuos.

se supone habitualmente que los factores tienen matriz de varianza y covarianzas $\mathbf{V}_f$ no necesariamente identidad, aunque con restricciones en sus términos.

La ecuación fundamental se convierte en

$$\mathbf{V}_x = \Lambda \mathbf{V}_f \Lambda' + \boldsymbol{\psi}$$

pero ahora las tres matrices desconocidas del sistema, $\Lambda, \mathbf{V}_f$ y $\boldsymbol{\psi}$ contienen numerosas restricciones de manera que el número total de parámetros libres, t, verifica

$$t \leq \frac{p\,(p+1)}{2}$$

para que el modelo esté identificado.

La estimación se realiza por máxima verosimilitud, pero la restricción $\Lambda' \boldsymbol{\psi}^{-1} \Lambda = diagonal$ no suele imponerse si no es necesaria para identificar el modelo.

Los contrastes de bondad del modelo son análogos a los estudiados, pero ahora el número de grados de libertad será $\frac{p(p+1)}{2} - t$, siendo t el número de parámetro libres estimados. Sin embargo, los efectos de no normalidad son aquí más graves que cuando estimamos todos los parámetros, como ocurre en análisis factorial exploratorio.

Recomendamos que el análisis factorial confirmatorio, se compare siempre con un análisis exploratorio para confirmar que el modelo impuesto no está en contradicción con los datos observados.

12.10 Relación con componentes principales

En componentes principales descomponemos la matriz de varianzas y covarianzas de las $\mathbf{X}$ como:

$$\begin{aligned}\mathbf{S} &= \mathbf{A}\Gamma\mathbf{A}' = [\mathbf{a}_1 \dots \mathbf{a}_n] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n' \end{bmatrix} \\ &= [\mathbf{a}_1 \dots \mathbf{a}_n] \begin{bmatrix} \lambda_1 \mathbf{a}_1' \\ \vdots \\ \lambda_n \mathbf{a}_n' \end{bmatrix} \\ &= \lambda_1 \mathbf{a}_1 \mathbf{a}_1' + \ldots + \lambda_n \mathbf{a}_n \mathbf{a}_n'\end{aligned}$$

Si $\lambda_j = 0$ para $j < h$, podemos reconstruir $\mathbf{S}$ con los primeros j componentes. Llamando $\mathbf{H} = \mathbf{A}\Gamma^{1/2}$, tenemos que:

$$\mathbf{S} = \mathbf{H}\mathbf{H}'.$$

En Análisis factorial descomponemos $\mathbf{S}$ como:

$$\mathbf{S} = \Lambda\Lambda' + \boldsymbol{\psi},$$

y como la matriz $\boldsymbol{\psi}$ es diagonal puede recoger las varianzas de las variables, mientras que la matriz de carga recoge las covarianzas de las variables. Esta es una diferencia importante entre ambos métodos. El primero trata de explicar las varianzas, mientras que el segundo explica las covarianzas o correlaciones. Observemos que si $\widehat{\boldsymbol{\psi}} \simeq 0$, es lo mismo tomar m componentes principales que estimar m factores. La diferencia es tanto menor cuanto menor sea $\widehat{\boldsymbol{\psi}}$.

Otra forma de estudiar la relación entre ambos métodos es la siguiente: sea $\mathbf{X}$ la matriz de datos original y $\mathbf{Z}$ la matriz de valores de los componentes. Entonces:

$$\mathbf{Z} = \mathbf{X}\mathbf{A},$$

o también, como $\mathbf{A}$ es ortogonal

$$\mathbf{X} = \mathbf{Z}\mathbf{A}',$$

que permite reconstruir las variables originales a partir de los componentes. Escribiendo

$$\begin{aligned} x_1 &= \alpha_{11} z_1 + \ldots + \alpha_{m1} z_m + \ldots + \alpha_{pm} z_p \\ &\vdots \\ x_p &= \alpha_{p1} z_1 + \ldots + \alpha_{pm} z_m + \ldots + \alpha_{pp} z_p \end{aligned}$$

con m componentes tenemos:

$$\begin{aligned} x_1 &= \alpha_{11} z_1 + \ldots + \alpha_{1m} z_m + v_1 \\ &\vdots \\ x_p &= \alpha_{p1} z_1 + \ldots + \alpha_{mp} z_m + v_p \end{aligned}$$

Esta representación es aparentemente análoga al modelo factorial, ya que v_1 estará incorrelada con los factores $(z_1, ..., z_m)$ al incluir únicamente las variables $(z_{m+1}, ..., z_p)$ que son ortogonales a las anteriores. Sin embargo, la diferencia básica es que en el modelo factorial los errores de las distintas ecuaciones están incorrelados, mientras que en ésta representación no lo estarán. En efecto, como $(v_1, ...v_p)$ contienen todas las variables comunes $(z_{m+1}, ..., z_p)$ estarán correladas. Por esta razón, en general los resultados de ambos métodos serán distintos.

Estos resultados indican que si existen m componentes principales que explican una proporción muy alta de la variabilidad, de manera que la variabilidad específica dada por los términos diagonales de $\boldsymbol{\psi}$ sea pequeña, el análisis factorial y el análisis de componentes principales sobre la matriz de correlaciones serán resultados similares. También en este caso la estimación mediante el método de factores principales conducirá a resultados similares al de máxima verosimilitud.

Aparte de estas diferencias, ambas técnicas tienen una interpretación diferente: en componentes principales tratamos de representar graficamente los datos, mientras que en análisis factorial suponemos que los factores generan las variables observadas.

12.11 Lecturas recomendadas

La mayoría de los textos de análisis multivariante incluyen un capítulo de análisis factorial. Buenos textos recomendables para complementar lo aqui expuesto son Cuadras (1991), que presenta una exposición muy detallada y clara, Jobson (1992), Johnson and Wichern (1998), Mardia et al (1979), Rechner (1998) y Seber (1984). Presentaciones más extensas se encuentran en Bartholomew y Knott (1999) y Harman (1980). La estimación máximo verosimil se estudia con detalle en Joreskov (1963) y Lawley y Maxwell (1971), aunque no se incluyen métodos más modernos de estimación basados en el algoritmo EM o en métodos bayesianos. Bartholomew y Knott (1999) presentan de forma clara la estimación mediante el algoritmo EM. Otra buena referencia sobre el tema es Schafer (1997). Para el tratamiento Bayesiano véase O' Hagan (1994)

EJERCICIOS

Ejercicio 12.1 *Dado el modelo factorial* $\mathbf{x} = \Lambda\ f + \mathbf{u}$ *, donde* $\mathbf{x} = (x_1, x_2, x_3, x_4)$ *tiene media cero y varianzas* $(\ 1, 2, 1, 7)$*, y donde* $\Lambda = (.8, 1, 0, 2)'$ *, y Var(f)* $= 1$*, se pide: (1) Calcular las covarianzas entre las variables y el factor; (2) Calcular las correlaciones entre las variables y el factor; (3) Escribir el modelo como un modelo unifactorial con un factor de varianza igual a 5.*

Ejercicio 12.2 *Indicar si es posible el siguiente modelo factorial* $\mathbf{x} = \Lambda\ f + \mathbf{u}$ *, donde* $\mathbf{x} = (x_1, x_2, x_3)$ *tiene media cero y varianzas* $(3, 1, 2)$*, y donde* $\Lambda = (3, 0, 3)'$ *, y Var(f)* $= 1$*.*

Ejercicio 12.3 *Dado un modelo factorial con* $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$ *de media cero y* $\Lambda = (\boldsymbol{\lambda}_1 \boldsymbol{\lambda}_2)$ *con* $\boldsymbol{\lambda}_1 = (1, 1, 1, 0, 0, 0)'$ *, y* $\boldsymbol{\lambda}_2 = (0, 1, 0, 1, 0, 1)$ *y Var*$(\mathbf{f}) = \begin{bmatrix} 1 & .5 \\ .5 & 2 \end{bmatrix}$*, escribirlo en forma estándar con la normalización* $\Lambda'\Lambda =$ *Diagonal*

Ejercicio 12.4 *Demostrar utilizando la relación fundamental que*

a. *Si la varianza específica es igual a la varianza de una variable, la fila de la matriz de carga correspondiente a dicha variable debe tener todos los elementos nulos.*

b. *Si la covarianza entre dos variables es cero, las filas correspondientes a estas variables en la matriz de carga son ortogonales.*

c. *Si las variables están estandarizadas la correlación entre las variables i y j es el producto escalar de las filas de la matriz de carga correspondientes a estas variables.*

d. *Si las variables están estandarizadas, la correlación entre la variable i y el factor j es el término (ij) de la matriz de carga.*

Ejercicio 12.5 *Demostrar utilizando la relación fundamental que si las varianzas de los componentes específicos son idénticos, las columnas de la matriz de carga son vectores propios de la matriz de covarianzas, y obtener los valores propios.*

Ejercicio 12.6 *Indicar cuál será el número máximo de factores incorrelados que podemos estimar con diez variables. ¿Y si los factores están correlados?*

Ejercicio 12.7 *Demostrar mediante un ejemplo que con el método del factor principal no se obtiene la misma matriz de carga con variables estandarizadas y sin estandarizar.*

Ejercicio 12.8 *Se dispone de 20 variables y antes de realizar un análisis factorial se realizan componentes principales y se eligen los 5 primeros componentes. A continuación se realiza un análisis factorial sobre estos componentes. ¿Cuántos factores esperaríamos encontrar?*

Ejercicio 12.9 *Demostrar que si cada columna de la matriz de carga tiene un único elemento no nulo, el modelo factorial no está identificado.*

Ejercicio 12.10 *Demostrar que si rotamos los factores la comunalidad total de cada variable no varía.*

Ejercicio 12.11 *Si $\Lambda = (1, 1, ..., 1)'$ indicar la ecuación para estimar el valor del factor en un individuo si*

a. *diag($\boldsymbol{\psi}$) $= p(1, 1, .., 1)$,*

b. *diag($\boldsymbol{\psi}$) $= (1, 2, .., p)$*

Ejercicio 12.12 *Demostrar que en el modelo unifactorial con $\boldsymbol{\psi} = \sigma^2\mathbb{I}$, el determinante de $\widehat{\mathbf{V}}_0$ es $((\Lambda'\Lambda) + \sigma^2)(\sigma^2)^{p-1}$.*

Ejercicio 12.13 *Demostrar que si todas las variables tienen perturbaciones con la misma varianza y $\boldsymbol{\psi} = \psi_0\mathbf{I}$, y suponiendo $\Lambda'\Lambda = diagonal = \mathbf{D}$, las columnas de Λ son directamente vectores propios de la matriz $\mathbf{V}$, con valores propios $d_i + \psi_0$, donde d_i es el término diagonal de $\mathbf{D}$*

APÉNDICE 12.1:ESTIMACIÓN MÁXIMO-VEROSÍMIL DEL MODELO FACTORIAL

La estimación MV de Λ y $\boldsymbol{\psi}$ requiere escribir la ecuación de verosimilitud con la restricción $\Lambda\boldsymbol{\psi}^{-1}\Lambda = \mathbf{D}$ = diagonal, derivarla y obtener las condiciones de primer orden. Este proceso conduce al mismo resultado que resolver la ecuación de estimación por momentos

$$\mathbf{S} = \widehat{\Lambda}\widehat{\Lambda}' + \widehat{\boldsymbol{\psi}} \tag{12.38}$$

con las restricciones de que $\widehat{\Lambda}$ sea $(p \times m)$ y ψ diagonal. Esta segunda condición se satisface tomando:

$$\widehat{\boldsymbol{\psi}} = diag\left(\mathbf{S} - \widehat{\Lambda}\widehat{\Lambda}'\right). \tag{12.39}$$

Supongamos que a partir de un valor inicial de $\widehat{\Lambda}$ obtenemos la matriz $\widehat{\boldsymbol{\psi}}$ mediante (12.39). El nuevo estimador de Λ debe satisfacer aproximadamente la ecuación $\mathbf{S} - \widehat{\boldsymbol{\psi}} = \widehat{\Lambda}\widehat{\Lambda}'$. Este sistema tiene $p(p+1)/2$ ecuaciones y $p \times m$ incógnitas y en general no tiene solución única. Para reducirlo a un sistema de $p \times m$ ecuaciones post-multipliquemos la ecuación (12.38) por $\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}$. Reordenando términos, se obtiene el sistema de ecuaciones:

$$\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda} = \widehat{\Lambda}(\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda} + \mathbf{I}) = \widehat{\Lambda}(\mathbf{D} + \mathbf{I}_m) \tag{12.40}$$

donde $\mathbf{I}_m$ es la matriz identidad de orden m. Esta ecuación cuando $\widehat{\boldsymbol{\psi}}^{-1}$ es conocido proporciona un sistema no lineal de $(p \times m)$ ecuaciones con $(p \times m)$ incógnitas para obtener $\widehat{\Lambda}$, y sugiere que $\widehat{\Lambda}$ puede obtenerse a partir de los vectores propios $\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1}$, pero esta matriz no es simétrica. Para resolver este problema, premultiplicando por $\widehat{\boldsymbol{\psi}}^{-1/2}$, podemos escribrir

$$\widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2}(\widehat{\boldsymbol{\psi}}^{-1/2}\widehat{\Lambda}) = (\widehat{\boldsymbol{\psi}}^{-1/2}\widehat{\Lambda})(\mathbf{D} + \mathbf{I}_m) \tag{12.41}$$

que muestra que podemos obtener $\widehat{\boldsymbol{\psi}}^{-1/2}\widehat{\Lambda}$ como vectores propios de la matriz simétrica $\widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2}$, o tambien, de la matriz simétrica $\widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2} - \mathbf{I}_p$.

Los estimadores MV satisfacen dos propiedades importantes. La primera es

$$tr\left(\mathbf{S}\widehat{\mathbf{V}}_0^{-1}\right) = p$$

que indica que con la distancia de la traza la matriz estimada está tan cerca como es posible de la matriz de covarianzas observada $\mathbf{S}$. En efecto, $\widehat{\mathbf{V}}_0 = \mathbf{S}$, entonces $tr\left(\mathbf{S}\mathbf{S}^{-1}\right) = tr\left(\mathbf{I}_p\right) = p$. La segunda es que se obtiene el mismo resultado trabajando con variables estandarizadas o sin estandarizar, es decir si estimamos por MV la matriz de carga con las variables originales se obtiene el mismo resultado que si (1) estandarizamos las variables restando las medias y dividiendo por las desviaciones típicas , (2) estimamos por MV la matriz de carga, que es entonces la matriz de correlación entre las variables originales y los factores, (3) pasamos de esa matriz de correlación a la matriz de covarianzas multiplicando por las desviaciones típicas de las variables.

Demostremos la primera propiedad. Si $\widehat{\mathbf{V}}_0 = \widehat{\boldsymbol{\psi}} + \widehat{\Lambda}\widehat{\Lambda}'$ es la estimación MV de la matriz de covarianzas, entonces (véase la sección 2.3.4):

$$\widehat{\mathbf{V}}_0^{-1} = \left(\widehat{\boldsymbol{\psi}} + \widehat{\Lambda}\widehat{\Lambda}'\right)^{-1} = \widehat{\boldsymbol{\psi}}^{-1} - \widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}\left(\mathbf{I}_m + \widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}\right)^{-1}\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}$$

y multiplicando por $\mathbf{S}$ y tomando trazas :

$$tr\left(\mathbf{S}\widehat{\mathbf{V}}_0^{-1}\right) = tr\left(\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1}\right) - tr\left(\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}\left(\mathbf{I}_m + \mathbf{D}\right)^{-1}\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\right)$$

Utilizando la condición (12.40)

$$tr\left(\mathbf{S}\widehat{\mathbf{V}}_0^{-1}\right) = tr\left(\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1}\right) - tr\left(\widehat{\Lambda}\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\right)$$

y, por las propiedades lineales de la traza,

$$tr\left(\mathbf{S}\widehat{\mathbf{V}}_0^{-1}\right) = tr\left[\left(\mathbf{S} - \widehat{\Lambda}\widehat{\Lambda}'\right)\widehat{\boldsymbol{\psi}}^{-1}\right] = tr\left(diag\left(\mathbf{S} - \widehat{\Lambda}\widehat{\Lambda}'\right)\widehat{\boldsymbol{\psi}}^{-1}\right).$$

donde el último paso proviene de que el producto de dos matrices diagonales es diagonal y la traza es simplemente la suma de sus elementos diagonales. Por otro lado, la ecuación (12.39) implica

$$diag\left(\mathbf{S} - \widehat{\Lambda}\widehat{\Lambda}'\right)\widehat{\boldsymbol{\psi}}^{-1} = \mathbf{I}_p$$

y tomando trazas en esta ecuación

$$tr\left(\mathbf{S}\widehat{\mathbf{V}}_0^{-1}\right) = tr\left(\left(\mathbf{S} - \widehat{\Lambda}\widehat{\Lambda}'\right)\widehat{\boldsymbol{\psi}}^{-1}\right) = tr\left(\mathbf{I}_p\right) = p$$

y hemos demostrado que el estimador MV verifica (**??**).

Para demostrar la segunda propiedad, supongamos que hacemos la transformación $\mathbf{y} = \mathbf{DX}$, donde $\mathbf{D}$ es cualquier matriz diagonal (por ejemplo, estandarizamos las variables lo que equivale a trabajar con $\mathbf{R}$, matriz de correlación, en lugar de con $\mathbf{S}$, matriz de covarianzas). Entonces $\mathbf{S}_y = \mathbf{D}\mathbf{S}_x\mathbf{D}$ y $\boldsymbol{\psi}_y = \mathbf{D}\boldsymbol{\psi}_x\mathbf{D}$. Al calcular la nueva matriz para obtener los valores y vectores propios, tendremos

$$\begin{aligned}\boldsymbol{\psi}_y^{-1/2}\mathbf{S}_y{}^{-1}\boldsymbol{\psi}_y^{-1/2} &= (\mathbf{D}\boldsymbol{\psi}_x\mathbf{D})^{-1/2}\mathbf{D}\mathbf{S}_x\mathbf{D}(\mathbf{D}\boldsymbol{\psi}_x\mathbf{D})^{-1/2} \\ &= \boldsymbol{\psi}_x^{-1/2}\mathbf{S}_x{}^{-1}\boldsymbol{\psi}_x^{-1/2}\end{aligned}$$

y es idéntica a la anterior.

APÉNDICE: 12.2 CONTRASTES SOBRE EL RANGO DE UNA MATRIZ

En este apéndice vamos a demostrar que el contraste sobre el número de factores es un caso particular del contraste general de esfericidad parcial, estudiado en 10.6. Necesitamos para ello el Lema siguiente:

Lemma 3 $|\mathbf{I}_m - \mathbf{U}\mathbf{U}'| = |\mathbf{I}_p - \mathbf{U}'\mathbf{U}|$. *Esta igualdad se demuestra aplicando la fórmula del determinante de una matriz particionada a los determinantes de la igualdad*

$$\begin{bmatrix} \mathbf{I}_m & \mathbf{U} \\ \mathbf{U}' & \mathbf{I}_p \end{bmatrix} = \begin{bmatrix} \mathbf{I}_p & \mathbf{U}' \\ \mathbf{U} & \mathbf{I}_m \end{bmatrix}.$$

Vamos a utilizar este lema para demostrar que el contraste de la razón de verosimilitudes (12.27) tiene en cuenta únicamente los $p-m$ menores valores propios de la matriz $\widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2}$. Partiendo del estimador MV de $\widehat{\mathbf{V}}_0$, tenemos:

$$\begin{aligned} \left|\widehat{\mathbf{V}}_0\right| &= \left|\widehat{\Lambda}\widehat{\Lambda}' + \widehat{\boldsymbol{\psi}}\right| = \left|\widehat{\boldsymbol{\psi}}^{1/2}\right| \left|\widehat{\boldsymbol{\psi}}^{-1/2}\widehat{\Lambda}\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1/2} + \mathbf{I}_p\right| \left|\widehat{\boldsymbol{\psi}}^{1/2}\right| \\ &= \left|\widehat{\boldsymbol{\psi}}\right| \left|\widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1/2}\widehat{\boldsymbol{\psi}}^{-1/2}\widehat{\Lambda} + \mathbf{I}_m\right| \end{aligned}$$

Llamando $\mathbf{D} = \widehat{\Lambda}'\widehat{\boldsymbol{\psi}}^{-1}\widehat{\Lambda}$, se obtiene que,

$$\left|\widehat{\mathbf{V}}_0\right| = \left|\widehat{\boldsymbol{\psi}}\right| |\mathbf{D} + \mathbf{I}_m| = \left|\widehat{\boldsymbol{\psi}}\right| \prod_{i=1}^{m}(1 + d_j) \tag{12.42}$$

donde d_j es el elemento diagonal j de $\mathbf{D}$.

Por otro lado, de la estimación de los parámetros por el método de MV del apéndice 12.1, hemos visto en la relación (12.41), que la matriz $\widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2} - \mathbf{I}_p$ tiene m valores propios iguales a los términos diagonales de $\mathbf{D}$. En general esta matriz tendrá rango p, y llamemos también d_i a sus restantes valores propios para $i = m+1, ..., p$. En consecuencia, los valores propios de la matriz $\widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2}$ serán $1 + d_i$ y podemos escribir:

$$\left|\widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2}\right| = |\mathbf{S}| \,/ \left|\widehat{\boldsymbol{\psi}}\right| = \prod_{i=1}^{p}(1 + d_i) \tag{12.43}$$

y utilizando (12.42) y (12.43)

$$\frac{\left|\widehat{\mathbf{V}}_0\right|}{|\mathbf{S}|} = \frac{\left|\widehat{\boldsymbol{\psi}}\right| \prod_{i=1}^{m}(1 + d_i)}{\left|\widehat{\boldsymbol{\psi}}\right| \prod_{i=1}^{p}(1 + d_i)} = \frac{1}{\prod_{i=m+1}^{p}(1 + d_i)}$$

con lo que se obtiene finalmente:

$$\lambda_F = n \log \left|\widehat{\mathbf{V}}_0\right| / |\mathbf{S}| = -n \log \sum_{i=m+1}^{p} (1 + d_i) \tag{12.44}$$

y el contraste de verosimilitud depende únicamente de los valores propios más pequeños de la matriz $\widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2}$.

Vamos a demostrar ahora que este contrate es un caso particular del contraste de esfericidad parcial, presentado en 10.6, cuyo estadístico es:

$$\lambda_{EP} = n(p-m) \log \frac{\sum_{i=m+1}^{p} \lambda_i}{p-m} - n \log \prod_{i=m+1}^{p} \lambda_i \tag{12.45}$$

donde λ_i son los valores propios de $\mathbf{S}$, y que el estadístico (12.44) resulta de aplicar este contrate a la matriz $\widehat{\boldsymbol{\psi}}^{-1/2}\mathbf{S}\widehat{\boldsymbol{\psi}}^{-1/2}$. Si el modelo factorial es correcto, asintóticamente $\mathbf{S} = \boldsymbol{\psi} + \Lambda\Lambda'$ y pre y postmultiplicando por $\boldsymbol{\psi}^{-1/2}$:

$$\boldsymbol{\psi}^{-1/2}\mathbf{S}\boldsymbol{\psi}^{-1/2} = \mathbf{I} + \boldsymbol{\psi}^{-1/2}\Lambda\Lambda'\boldsymbol{\psi}^{-1/2} \tag{12.46}$$

que descompone la matriz $\boldsymbol{\psi}^{-1/2}\mathbf{S}\boldsymbol{\psi}^{-1/2}$ en una de rango m más la matriz identidad. Como los valores propios de esta matriz son $1 + d_i$, tenemos que

$$\lambda_{EP} = n(p-m)\log\frac{\sum_{i=m+1}^{p}(1+d_i)}{p-m} - n\log\sum_{i=m+1}^{p}(1+d_i) \tag{12.47}$$

queda ahora por demostrar que el primer término es cero. Tomando trazas en (12.46).

$$tr(\boldsymbol{\psi}^{-1/2}\mathbf{S}\boldsymbol{\psi}^{-1/2}) = p + tr(\boldsymbol{\psi}^{-1/2}\Lambda\Lambda'\boldsymbol{\psi}^{-1/2}) = p + tr(\Lambda'\boldsymbol{\psi}^{-1}\Lambda) = p + \sum_{i=1}^{m} d_j$$

pero también, por ser $1+d_i$ los valores propios de $\boldsymbol{\psi}^{-1/2}\mathbf{S}\boldsymbol{\psi}^{-1/2}$, sabemos que $tr(\boldsymbol{\psi}^{-1/2}\mathbf{S}\boldsymbol{\psi}^{-1/2}) = \sum_{i=1}^{p}(1+d_i)$, e igulando ambos resultados para la traza:

$$\sum_{i=1}^{p}(1+d_i) = p + \sum_{i=1}^{m} d_j$$

de donde resulta

$$\sum_{i=m+1}^{p} d_i = 0$$

o, lo que es equivalente

$$\sum_{i=m+1}^{p}(1+d_i) = p - m.$$

y sustituyendo en (12.47) el primer término se anula y queda únicamente el segundo, con lo que obtenemos el contraste de la razón de verosimilitud.

APÉNDICE 12.3 :ESTIMACIÓN DE LOS FACTORES

Sea $\mathbf{x}_i$ el vector $(p \times 1)$ en el individuo i. Su función de densidad será:

$$f(\mathbf{x}_i) = |\psi|^{-1/2}(2\pi)^{-p/2}\exp\left\{-1/2(\mathbf{x}_i - \Lambda\mathbf{f}_i)'\boldsymbol{\psi}^{-1}(\mathbf{x}_i - \Lambda\mathbf{f}_i)\right\}$$

supongamos que $\boldsymbol{\psi}$ y Λ son conocidas y se trata de estimar $\mathbf{f}_i$. Entonces, la función de verosimilitud en logaritmos será:

$$L = \log f(\mathbf{x}_i) = K - 1/2(\mathbf{x}_i - \Lambda\mathbf{f}_i)'\boldsymbol{\psi}^{-1}(\mathbf{x}_i - \Lambda\mathbf{f}_i)$$

donde K es una constante. Maximizar L equivale a minimizar:

$$M = (\mathbf{x}_i - \Lambda\mathbf{f}_i)'\boldsymbol{\psi}^{-1}(\mathbf{x}_i - \Lambda\mathbf{f}_i)$$

que es el criterio de mínimos cuadrados. Entonces:

$$M = \mathbf{x}_i'\boldsymbol{\psi}^{-1}\mathbf{x}_i - 2\mathbf{f}_i'\Lambda'\boldsymbol{\psi}^{-1}\mathbf{x}_i + \mathbf{f}_i'\Lambda'\boldsymbol{\psi}^{-1}\Lambda\mathbf{f}_i.$$

Derivando respecto a $\mathbf{f}_i$ e igualando a cero:

$$\frac{dM}{d\mathbf{f}_i} = 0 = -2\Lambda'\boldsymbol{\psi}^{-1}\mathbf{x}_i + 2\Lambda'\boldsymbol{\psi}^{-1}\Lambda\mathbf{f}_i$$

por tanto

$$\mathbf{f}_i = (\Lambda'\boldsymbol{\psi}^{-1}\Lambda)^{-1}\Lambda'\boldsymbol{\psi}^{-1}\mathbf{x}_i$$

sustituyendo en esta expresión Λ y $\boldsymbol{\psi}$ por sus estimadores MV se obtiene el vector $\hat{\mathbf{f}}_i$ para cada observación.

Si el parámetro se considera como variable aleatoria, por las propiedades de la esperanza condicional

$$\widehat{\mathbf{f}}_i = E\left[\mathbf{f}_i|\mathbf{x}_i\right] = \Lambda'\mathbf{V}^{-1}\mathbf{x}_i$$

Utilizando que:

$$\mathbf{V}^{-1} = (\boldsymbol{\psi} + \Lambda\Lambda')^{-1} = \boldsymbol{\psi}^{-1} - \boldsymbol{\psi}^{-1}\Lambda(\mathbf{I} + \Lambda'\boldsymbol{\psi}^{-1}\Lambda)^{-1}\Lambda'\boldsymbol{\psi}^{-1},$$

entonces,

$$\begin{aligned}
\Lambda'\mathbf{V}^{-1} &= \Lambda'\boldsymbol{\psi}^{-1} - \Lambda'\boldsymbol{\psi}^{-1}\Lambda(\mathbf{I} + \Lambda'\boldsymbol{\psi}^{-1}\Lambda)^{-1}\Lambda'\boldsymbol{\psi}^{-1} \\
\Lambda'\mathbf{V}^{-1} &= [\mathbf{I} - \Lambda'\boldsymbol{\psi}^{-1}\Lambda(\mathbf{I} + \Lambda'\boldsymbol{\psi}^{-1}\Lambda)^{-1}]\Lambda'\boldsymbol{\psi}^{-1} \\
\Lambda'\mathbf{V}^{-1} &= (\mathbf{I} + \Lambda'\boldsymbol{\psi}^{-1}\Lambda)^{-1}\Lambda'\boldsymbol{\psi}^{-1}.
\end{aligned}$$

y sustituyendo en la expresión de $\widehat{\mathbf{f}}_i$, se obtiene:

$$\widehat{\mathbf{f}}_i = (\mathbf{I} + \Lambda'\boldsymbol{\psi}^{-1}\Lambda)^{-1}\Lambda'\boldsymbol{\psi}^{-1}\mathbf{x}_i.$$

APÉNDICE 12.4: INTERPRETACIÓN BAYESIANA DEL ESTIMADOR DE LOS FACTORES

El estimador (12.37) tiene una clara interpretación bayesiana. Como a priori, la distribución del factor, $\pi\left(\mathbf{f}_i\right)$, es $N(\mathbf{0}, \mathbf{I})$, y la verosimilitud $f\left(\mathbf{x}|\mathbf{f}, \Lambda, \boldsymbol{\psi}\right)$ es $N\left(\Lambda\mathbf{f}, \boldsymbol{\psi}\right)$, la posterior condicionada a los parámetros es:

$$f\left(\mathbf{f}_i|\mathbf{x}, \Lambda, \boldsymbol{\psi}\right) = k\, f\left(\mathbf{x}|\mathbf{f}_i, \Lambda, \boldsymbol{\psi}\right)\pi\left(\mathbf{f}_i\right) \tag{12.48}$$

donde k es una constante. El exponente de la distribución posterior será:

$$\begin{aligned}
\left(\mathbf{x} - \Lambda\mathbf{f}\right)'\boldsymbol{\psi}^{-1}\left(\mathbf{x} - \Lambda\mathbf{f}\right) + \mathbf{f}'\mathbf{f} &= \mathbf{f}'\left(\mathbf{I} + \Lambda'\boldsymbol{\psi}^{-1}\Lambda\right)\mathbf{f} - \\
&\quad 2\mathbf{f}'\Lambda'\boldsymbol{\psi}^{-1}\mathbf{x} + \mathbf{x}'\boldsymbol{\psi}^{-1}\mathbf{x},
\end{aligned}$$

y completando el cuadrado, el exponente puede escribirse

$$\mathbf{f}'\left(\mathbf{I}+\Lambda'\boldsymbol{\psi}^{-1}\Lambda\right)\mathbf{f}-\mathbf{2f}'\left(\mathbf{I}+\Lambda'\boldsymbol{\psi}^{-1}\Lambda\right)\left(\mathbf{I}+\Lambda'\boldsymbol{\psi}^{-1}\Lambda\right)^{-1}\Lambda'\boldsymbol{\psi}^{-1}\mathbf{x}+\text{Resto}$$

es decir

$$\left(\mathbf{f}-\widehat{\mathbf{f}}\right)'\left(\mathbf{I}+\Lambda'\boldsymbol{\psi}^{-1}\Lambda\right)\left(\mathbf{f}-\widehat{\mathbf{f}}\right)$$

donde

$$\widehat{\mathbf{f}}=E\left[\mathbf{f}|\Lambda,\boldsymbol{\psi}\right]=\left(\mathbf{I}+\Lambda'\boldsymbol{\psi}^{-1}\Lambda\right)^{-1}\Lambda'\boldsymbol{\psi}^{-1}\mathbf{x} \tag{12.49}$$

y

$$Var\left(\widehat{\mathbf{f}}\right)=Var\left[\mathbf{f}|\Lambda,\boldsymbol{\psi}\right]=\left(\mathbf{I}+\Lambda'\boldsymbol{\psi}^{-1}\Lambda\right)^{-1} \tag{12.50}$$

Por tanto, el estimador (12.37) puede interpretarse como la media de la distribución a posteriori de los factores. Observemos que la condición $\Lambda'\boldsymbol{\psi}^{-1}\Lambda$ = diagonal hace que los factores sean, a posteriori, condicionalmente independientes.

Capítulo 13

ANÁLISIS DISCRIMINANTE

13.1 INTRODUCCIÓN

El problema de discriminación o clasificación, que abordaremos en este capítulo, puede plantearse de varias formas y aparece en muchas áreas de la actividad humana: desde la diagnosis médica a los sistemas de concesión de créditos o de reconocimiento de falsas obras de arte. El planteamiento estadístico del problema es el siguiente. Se dispone de un conjunto amplio de elementos que pueden venir de dos o más poblaciones distintas. En cada elemento se ha observado una variable aleatoria $p-$ dimensional $\mathbf{x}$, cuya distribución se conoce en las poblaciones consideradas. Se desea clasificar un nuevo elemento, con valores de las variables conocidas, en una de las poblaciones. Por ejemplo, la primera aplicación del análisis discriminante consitió en clasificar los restos de un cráneo descubierto en una excavación como humano, utilizando la distribución de medidas físicas para los cráneos humanos y los de antropoides.

El problema de discriminación aparece en muchas situaciones en que necesitamos clasificar elementos con información incompleta. Por ejemplo, los sistemas automáticos de concesión de créditos (credit scoring) implantados en muchas instituciones financieras tienen que utilizar variables medibles hoy (ingresos, antigüedad en el trabajo, patrimonio, etc) para prever el comportamiento futuro. En otros casos la información podría estar disponible, pero puede requerir destruir el elemento, como en el control de calidad de la resistencia a la tensión de unos componentes. Finalmente, en otros casos la información puede ser muy costosa de adquirir. En ingeniería este problema se ha estudiado con el nombre de *reconocimiento de patrones (pattern recognition)*, para diseñar máquinas capaces de clasificar de manera automática. Por ejemplo, reconocer voces y sonidos, clasificar billetes o menedas, reconocer caracteres escritos en una pantalla de ordenador o clasificar cartas según el distrito postal. Otros ejemplos de aplicaciones del análisis discriminante son: asignar un texto escrito de procedencia desconocida a uno de varios autores por las frecuencias de utilización de palabras, asignar una partitura musical o un cuadro a un artista, una declaración de impuestos como potencialmente defraudadora o no, una empresa como en riesgo de quiebra o no, las enseñanzas de un centro como teóricas y aplicadas, un paciente como enfermo de cáncer o no, un nuevo método de fabricación como eficaz o no.

Las técnicas que vamos a estudiar reciben también el nombre de *clasificación supervisada,*

para indicar que conocemos una muestra de elementos bien clasificados que sirve de pauta o modelo para la clasificación de las siguientes observaciones.

Existen varios enfoques posibles para este problema. El primero, que se presenta en este capítulo, es el análisis discriminante clásico debido a Fisher, basado en la normalidad multivariante de las variables consideradas y que es óptimo bajo dicho supuesto. Si todas las variables son continuas, es frecuente que aunque los datos originales no sean normales es posible transformar las variables para que lo sean, y los métodos de este capítulo pueden aplicarse a las variables transformadas. Sin embargo, cuando tengamos variables discretas y continuas para clasificar, la hipótesis de normalidad multivariante es poco realista, y en el capítulo siguiente se presentan otros enfoques al problema que pueden funcionar mejor en estos casos.

13.2 CLASIFICACIÓN ENTRE DOS POBLACIONES

13.2.1 Planteamiento del Problema

Sean P_1 y P_2 dos poblaciones donde tenemos definida una variable aleatoria vectorial, $\mathbf{x}$, p-variante. Supondremos que $\mathbf{x}$ es absolutamente continua y que las funciones de densidad de ambas poblaciones, f_1 y f_2, son conocidas. Vamos a estudiar el problema de clasificar un nuevo elemento, $\mathbf{x}_0$, con valores conocidos de las p variables en una de estas poblaciones. Si conocemos las probabilidades a priori π_1, π_2, con $\pi_1 + \pi_2 = 1$, de que el elemento venga de cada una de las dos poblaciones, su distribución de probabilidad será una distribución mezclada

$$f(\mathbf{x}) = \pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})$$

y una vez observado $\mathbf{x}_0$ podemos calcular las probabilidades a posteriori de que el elemento haya sido generado por cada una de las dos poblaciones, $P(i/\mathbf{x}_0)$, con $i = 1, 2$. Estas probabilidades se calculan por el teorema de Bayes

$$P(1|\mathbf{x}_0) = \frac{P(\mathbf{x}_0|1)\pi_1}{\pi_1 P(\mathbf{x}_0|1) + \pi_2 P(\mathbf{x}_0|2)}$$

y como $P(\mathbf{x}_0|1) = f_1(\mathbf{x}_0)\Delta\mathbf{x}_0$, tenemos que:

$$P(1/\mathbf{x}_0) = \frac{f_1(\mathbf{x}_0)\pi_1}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2}, \tag{13.1}$$

y para la segunda población

$$P(2|\mathbf{x}_0) = \frac{f_2(\mathbf{x}_0)\pi_2}{f_1(\mathbf{x}_0)\pi_1 + f_2(\mathbf{x}_0)\pi_2}. \tag{13.2}$$

Clasificaremos $\mathbf{x}_0$ en la población más probable a posteriori. Como los denominadores son iguales, clasificaremos $\mathbf{x}_0$ en P_2 si:

$$\pi_2 f_2(\mathbf{x}_0) > \pi_1 f_1(\mathbf{x}_0)$$

Si las probabilidades a priori son iguales, la condición de clasificar en P_2 se reduce a:

$$f_2(\mathbf{x}_0) > f_1(\mathbf{x}_0)$$

es decir, clasificamos a $\mathbf{x}_0$ en la población más probable, o donde su verosimilitud es más alta.

Consideración de las consecuencias

En muchos problemas de clasificación los errores que podemos cometer tienen distintas consecuencias que podemos cuantificar. Por ejemplo, si una máquina automática clasifica equivocadamente un billete de 10 euros como de 20, y devuelve el cambio equivocado, el coste de clasificación es de 10 euros. En otros casos estimar el coste puede ser más complejo: si no concedemos un crédito que sería devuelto podemos perder un cliente y los ingresos futuros que este podría generar, mientras que si el crédito no se devuelve el coste es la cantidad impagada. Como tercer ejemplo, si clasificamos un proceso productivo como en estado de control, el coste de equivocarnos será una producción defectuosa, y si, por error, paramos un proceso que funciona adecuadamente, el coste será el de la parada y revisión.

En general supondremos que las posibles decisiones en el problema son únicamente dos: asignar en P_1 o en P_2. Una regla de decisión es una partición del espacio muestral E_x (que en general será R^p) en dos regiones A_1 y $A_2 = E_x - A_1$, tales que:

$$\begin{array}{l} \text{si } \mathbf{x}_0 \in A_1 \Longrightarrow d_1 \text{ (clasificar en } P_1). \\ \text{si } \mathbf{x}_0 \in A_2 \Longrightarrow d_2 \text{ (clasificar en } P_2). \end{array}$$

Si las consecuencias de un error de clasificación pueden cuantificarse, podemos incluirlas en la solución del problema formulándolo como un problema bayesiano de decisión. Supongamos que:

1. las consecuencias asociadas a los errores de clasificación son, $c(2|1)$ y $c(1|2)$, donde $c(i|j)$ es el coste de clasificación en P_i de una unidad que pertenece a P_j. Estos costes se suponen conocidos;

2. el decisor quiere maximizar su función de utilidad y esto equivale a minimizar el coste esperado.

Con estas dos hipótesis la mejor decisión es la que minimiza los costes esperados, o funciones de pérdida de oportunidad, en la terminología de Wald. Los resultados de cada decisión que se presenta esquemáticamente en la figura 13.1. Si clasificamos al elemento en el grupo 2 las posibles consecuencias son:

(a) acertar, con probabilidad $P(2|\mathbf{x}_0)$, en cuyo caso no hay ningún coste de penalización;

(b) equivocarnos, con probabilidad $P(1|\mathbf{x}_0)$, en cuyo caso incurrimos en el coste asociado $c(2|1)$.

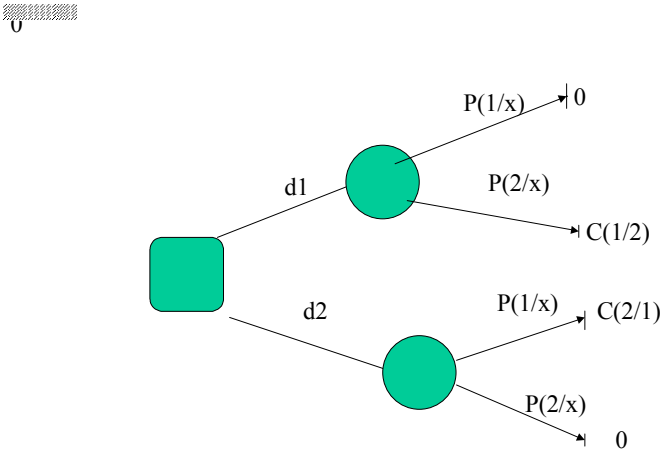


Figura 13.1: Representación de un problema de clasificación entre dos grupos como un problema de decisión.

El coste promedio, o valor esperado, de la decisión "d_2: clasificar $\mathbf{x}_0$ en P_2" será:

$$E(d_2) = c(2|1)P(1|\mathbf{x}_0) + 0P(2|\mathbf{x}_0) = c(2|1)P(1|\mathbf{x}_0). \tag{13.3}$$

Análogamente, el coste esperado de la decisión "d_1: clasificar en el grupo 1" es:

$$E(d_1) = 0P(1|\mathbf{x}_0) + c(1|2)P(2|\mathbf{x}_0) = c(1|2)P(2|\mathbf{x}_0). \tag{13.4}$$

Asignaremos al elemento al grupo 2 si su coste esperado es menor, es decir, utilizando (13.1) y (13.2), si:

$$\frac{f_2(\mathbf{x}_0)\pi_2}{c\,(2|1)} > \frac{f_1(\mathbf{x}_0)\pi_1}{c\,(1|2)}. \tag{13.5}$$

Esta condición indica que, a igualdad de los otros términos, clasificaremos en la población P_2 si

(a) su probabilidad a priori es más alta;

(b) la verosimilitud de que $\mathbf{x}_0$ provenga de P_2 es más alta;

(c) el coste de equivocarnos al clasificarlo en P_2 es más bajo.

En el Apéndice 13.1 se demuestra que este criterio es equivalente a minimizar la probabilidad total de error en la clasificación.

13.2.2 Poblaciones Normales: Función lineal discriminante

Vamos a aplicar el análisis anterior al caso en que f_1 y f_2 son distribuciones normales con distintos vectores de medias pero idéntica matriz de varianzas. Para establecer la regla con carácter general supondremos que se desea clasificar un elemento genérico $\mathbf{x}$, que si pertenece a la población $i = 1, 2$ tiene función de densidad:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}.$$

La partición óptima, es, de acuerdo con la sección anterior, *clasificar* en la población P_2 si:

$$\frac{f_2(\mathbf{x})\pi_2}{c(2|1)} > \frac{f_1(\mathbf{x})\pi_1}{c(1|2)}. \tag{13.6}$$

Como ambos términos son siempre positivos, tomando logaritmos y sustituyendo $f_i(\mathbf{x})$ por su expresión, la ecuación anterior se convierte en:

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \log\frac{\pi_2}{c(2|1)} > -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \log\frac{\pi_1}{c\,(1|2)},$$

Llamando D_i^2 a la distancia de Mahalanobis entre el punto observado, $\mathbf{x}$, y la media de la población i:

$$D_i^2 = (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$

podemos escribir:

$$D_1^2 - \log\frac{\pi_1}{c\,(1|2)} > D_2^2 - \log\frac{\pi_2}{c(2|1)} \tag{13.7}$$

y suponiendo iguales los costes y las probabilidades a priori, $c(1/2) = c(2/1)$; $\pi_1 = \pi_2$, la regla anterior se reduce a:

$$\boxed{\text{Clasificar en 2 si } D_1^2 > D_2^2}$$

es decir, clasificar la observación en la población de cuya media esté más próxima, midiendo la distancia con la medida de Mahalanobis. Observemos que si las variables $\mathbf{x}$ tuvieran $\mathbf{V} = \mathbf{I}\sigma^2$, la regla equivale a utilizar la distancia euclídea. La figura 13.2 muestra las curvas de equidistancia con la distancia de Mahalanobis para dos poblaciones normales con centros en el origen y el punto (5,10).

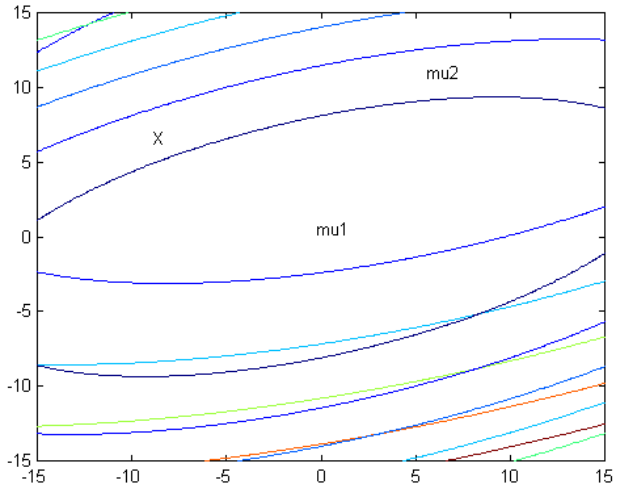


Figura 13.2: Curvas de equidistancia con la distancia de Mahalanobis para clasificar

13.2.3 Interpretación Geométrica

La regla general anterior puede escribirse de una forma equivalente que permite interpretar geométricamente el método de clasificación utilizado. La ecuación (13.7) indica que debemos calcular la distancia de Mahalanobis, corregirla por el término correspondiente a las probabilidades a priori y los costes, y clasificar en el población donde esta distancia modificada sea mínima. Como las distancias tiene siempre el término común $\mathbf{x}'\mathbf{V}^{-1}\mathbf{x}$, que no depende de la población, podemos eliminarlo de las comparaciones y calcular el indicador

$$-\boldsymbol{\mu}_i'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_i'\mathbf{V}^{-1}\boldsymbol{\mu}_i - \log\frac{\pi_i}{c(i|j)},$$

que será una función lineal en $\mathbf{x}$ y clasificar el individuo en la población donde esta función sea mínima. Esta regla divide el conjunto de valores posibles de $\mathbf{x}$ en dos regiones cuya frontera viene dada por:

$$-\boldsymbol{\mu}_1'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_1'\mathbf{V}^{-1}\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2'\mathbf{V}^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_2'\mathbf{V}^{-1}\boldsymbol{\mu}_2 - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1},$$

que, como función de $\mathbf{x}$, equivale a:

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}\mathbf{x} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}\left(\frac{\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1}{2}\right) - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1}. \tag{13.8}$$

Llamando:

$$\boxed{\mathbf{w} = \mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)} \tag{13.9}$$

la frontera puede escribirse como:

$$\mathbf{w}'\mathbf{x} = \mathbf{w}'\frac{\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1}{2} - \log\frac{c(1|2)\pi_2}{c(2|1)\pi_1} \tag{13.10}$$

que es la ecuación de un hiperplano. En el caso particular en que $c(1|2)\pi_2 = c(211)\pi_1$, clasificaremos en P_2 si

$$\mathbf{w}'\mathbf{x} > \mathbf{w}'\left(\frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right). \tag{13.11}$$

o lo que es equivalente, si

$$\mathbf{w}'\mathbf{x} - \mathbf{w}'\boldsymbol{\mu}_1 > \mathbf{w}'\boldsymbol{\mu}_2 - \mathbf{w}'\mathbf{x} \tag{13.12}$$

Esta ecuación indica que el procedimiento para clasificar un elemento $\mathbf{x}_0$ puede resumirse como sigue:

(1) calcular el vector $\mathbf{w}$ con (13.9);

(2) construir la variable indicadora discriminante:

$$z = \mathbf{w}'\mathbf{x} = w_1x_1 + ... + w_px_p$$

que transforma la variable multivariante $\mathbf{x}$ en la variable escalar z , que es una combinación lineal de los valores de la variable multivariante con coeficientes dados por el vector $\mathbf{w}$;

(3) calcular el valor de la variable indicadora para el individuo a clasificar, $\mathbf{x}_0 = (x_{10}, ..., x_{p0})$, con $z_0 = \mathbf{w}'\mathbf{x}_0$ y el valor de la variable indicadora para las medias de las poblaciones, $m_i = \mathbf{w}'\boldsymbol{\mu}_i$. Clasificar en aquella población donde la distancia $|z_0 - m_i|$ sea mínima.

En términos de la variable escalar z, como el valor promedio de z en P_i es :

$$\mathbf{E}(z|P_i) = m_i = \mathbf{w}'\boldsymbol{\mu}_i, \quad i = 1, 2$$

La regla de decisión (13.12) equivale a clasificar en P_2 si:

$$|z - m_1| > |z - m_2| \tag{13.13}$$

Esta variable indicadora, z, tiene varianza:

$$Var(z) = \mathbf{w}'Var(\mathbf{x})\mathbf{w} = \mathbf{w}'\mathbf{V}\mathbf{w} = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = D^2. \tag{13.14}$$

y el cuadrado de la distancia escalar entre las medias proyectadas es la distancia de Mahalanobis entre los vectores de medias originales:

$$(m_2 - m_1)^2 = (\mathbf{w}'(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1))^2 = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = D^2. \tag{13.15}$$

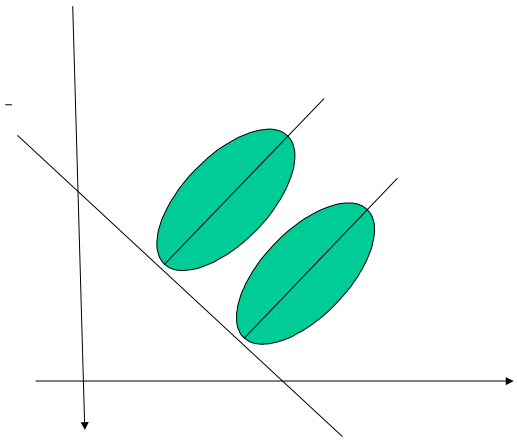


Figura 13.3: Representación de la dirección óptima de proyección para discriminar entre las dos poblaciones.

La variable indicadora z puede interpretarse como una proyección si estandarizamos el vector $\mathbf{w}$. Dividiendo los dos miembros de (13.11) por la norma de $\mathbf{w}$ y llamando $\mathbf{u}$ al vector unitario $\mathbf{w}/\left\|\mathbf{w}\right\|$, la regla de clasificación se convierte en clasificar en P_2 si

$$\mathbf{u}'\mathbf{x} - \mathbf{u}'\boldsymbol{\mu}_1 > \mathbf{u}'\boldsymbol{\mu}_2 - \mathbf{u}'\mathbf{x}, \tag{13.16}$$

donde, al ser $\mathbf{u}$ un vector unitario, $\mathbf{u}'\mathbf{x}$ es simplemente la proyección de $\mathbf{x}$ en la dirección de $\mathbf{u}$ y $\mathbf{u}'\boldsymbol{\mu}_1$ y $\mathbf{u}'\boldsymbol{\mu}_2$ las proyecciones de las medias poblacionales en esa dirección.

En la figura 13.3 se observa que el hiperplano perpendicular a $\mathbf{u}$ por el punto medio $\mathbf{u}'(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2$ divide el espacio muestral en dos regiones A_1 y A_2 que constituyen la partición óptima buscada. Si $c(1|2)\pi_2 \neq c(2|1)\pi_1$ la interpretación es la misma, pero el hiperplano frontera se desplaza paralelamente a sí mismo, aumentando o disminuyendo la región A_2.

La dirección de proyección, $\mathbf{w} = \mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ tiene una clara interpretación geométrica. Consideremos en primer lugar el caso en que las variables están incorreladas y estandarizadas de manera que $\mathbf{V} = \mathbf{I}$. Entonces, la dirección óptima de proyección es la definida por $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. En el caso general, la dirección de proyección puede calcularse en dos etapas: primero, se estandarizan las variables de forma multivariante, para pasar a variables incorreladas con varianza unidad; segundo, se proyectan los datos transformados sobre la dirección que une las medias de las variables estandarizadas.

En efecto, el cálculo de $\mathbf{w}'\mathbf{x}$ puede escribirse como:

$$\mathbf{w}'\mathbf{x} = \left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1/2}\right](\mathbf{V}^{-1/2}\mathbf{x})$$

donde $\mathbf{V}^{-1/2}$ existe si $\mathbf{V}$ es definida positiva. Esta expresión indica que esta operación equivale a: (1) estandarizar las variables $\mathbf{x}$ pasando a otras $\mathbf{y} = \mathbf{V}^{-1/2}\mathbf{x}$ que tienen como matriz

de covarianzas la identidad y como vector de medias $\mathbf{V}^{-1/2}\boldsymbol{\mu}$; (2) proyectar las variables estandarizadas $\mathbf{y}$ sobre la dirección $\boldsymbol{\mu}_2(\mathbf{y}) - \boldsymbol{\mu}_1(\mathbf{y}) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)'\mathbf{V}^{-1/2}$.

La figura 13.4 ilustra algunas direcciones de proyección. En (a) y (b) la dirección de las linea que une las medias coincide con alguno de los ejes principales de la elipse y por tanto la dirección $\mathbf{w} = \mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ coincide con $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, ya que este es un vector propio de $\mathbf{V}$, y por tanto también de $\mathbf{V}^{-1}$. En (c) la dirección óptima es un compromiso entre $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ y las direcciones definidas por los vectores propios de $\mathbf{V}^{-1}$.

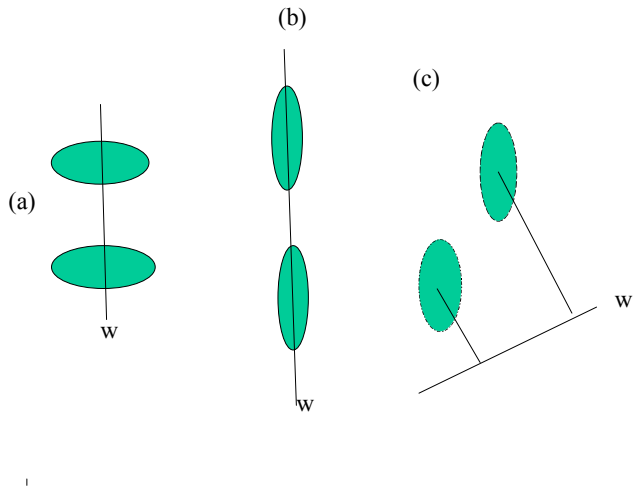


Figura 13.4: En los casos (a) y (b) la dirección óptima coindice con la linea de medias y con los ejes de la elipse. En el caso (c) es un compromiso entre ambos

13.2.4 Cálculo de Probabilidades de error

La utilidad de la regla de clasificación depende de los errores esperados. Como la distribución de la variable $z = \mathbf{w}'\mathbf{x}$ es normal, con media $m_i = \mathbf{w}'\boldsymbol{\mu}_i$ y varianza $D^2 = (m_2 - m_1)^2$, podemos calcular las probabilidades de clasificar erróneamente una observación en cada una de las dos poblaciones. En concreto, la probabilidad de una decisión errónea cuando $\mathbf{x} \in P_1$ es:

$$P(2|1) = P\left\{ z \geq \frac{m_1 + m_2}{2} \mid z \text{ es } N(m_1; D) \right\}$$

y llamando $y = (z - m_1)/D$ a una variable aleatoria $N(0,1)$, y Φ a su función de distribución:

$$P(2|1) = P\left\{ y \geq \frac{\frac{m_1+m_2}{2} - m_1}{D} \right\} = 1 - \Phi\left(\frac{D}{2}\right)$$

Análogamente, la probabilidad de una decisión errónea cuando $\mathbf{x} \in P_2$ es:

$$P(1|2) = P\left\{ z \leq \frac{m_1 + m_2}{2} \mid z \text{ es } N(m_2; D) \right\} =$$

$$= P\left\{y \leq \frac{\frac{m_1+m_2}{2} - m_2}{D}\right\} = \Phi\left(-\frac{D}{2}\right)$$

y ambas probabilidades de error son idénticas, por la simetría de la distribución normal . Podemos concluir que la regal obtenida hace iguales y mínimas (veáse Apéndice 13.1) las probabilidades de error y que los errores de clasificación sólo dependen de las distancias de Mahalanobis entre las medias.

13.2.5 Probabilidades a posteriori

El grado de confianza al clasificar una observación depende de la probabilidad acertar. La probabilidad a posteriori de que la observación pertenezca a la primera población se calcula con :

$$\begin{aligned} P(1|\mathbf{x}) &= \frac{\pi_1 f_1(\mathbf{x})}{\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})} = \\ &= \frac{\pi_1 \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}}{(\pi_1 \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\} + \pi_2 \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\mathbf{V}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\}} \end{aligned}$$

y llamando D_1^2 y D_2^2 a las distancias de Mahalanobis entre el punto y cada una de las dos medias, esta expresión puede escribirse:

$$P(1|\mathbf{x}) = \frac{1}{1 + \frac{\pi_2}{\pi_1}\exp\left\{-\frac{1}{2}(D_2^2 - D_1^2)\right\}}$$

y sólo depende de las probabilidades a priori y de las distancias entre el punto y las medias de ambas poblaciones. Observemos que si $\pi_2/\pi_1 = 1$, cuanto más alejado esta el punto de la primera población, es decir, cuanto mayor sea D_1^2 respecto a D_2^2, mayor serà el denominador y menor será la probabilidad de que pertenezca a ella, $P(1|\mathbf{x})$, y al contrario.

Ejemplo 13.1 *Se desea clasificar un retrato entre dos posibles pintores. Para ello se miden dos variables: la profundidad del trazo y la proporción que ocupa el retrato sobre la superficie del lienzo. Las medias de estas variables para el primer pintor, A, son (2 y .8) y para el segundo, B, (2.3 y .7) y las desviaciones típicas de estas variables son .5 y .1 y la correlación entre estas medidas es .5. La obra a clasificar tiene medidas de estas variables (2.1 y .75). Calcular las probabilidades de error.*

Las distancias de Mahalanobis serán, calculando la covarianza como el producto de la correlación por las desviaciones típicas:

$$D_A^2 = (2.1-2, .75-.8)\begin{bmatrix} .25 & .025 \\ .025 & .01 \end{bmatrix}^{-1}\begin{pmatrix} 2.1-2 \\ 75-.8 \end{pmatrix} = 0,52$$

y para la segunda

$$D_B^2 = (2.1 - 2.3, .75 - .7) \begin{bmatrix} .25 & .025 \\ .025 & .01 \end{bmatrix}^{-1} \begin{pmatrix} 2.1 - 2.3 \\ .75 - .7 \end{pmatrix} = 0,8133$$

Por tanto, asignaremos la obra al primer pintor. El error esperado de clasificación con esta regla depende de la distancia de Mahalanobis entre las medias que es

$$D^2 = (2. - 2.3, .8 - .7) \begin{bmatrix} .25 & .025 \\ .025 & .01 \end{bmatrix}^{-1} \begin{pmatrix} 2. - 2.3 \\ .8 - .7 \end{pmatrix} = 2,6133$$

y $D = 1.6166$. *La probabilidad de equivocarnos es*

$$P(A/B) = 1 - \Phi(\frac{1.6166}{2}) = 1 - \Phi(.808) = 1 - 0,8106 = 0,1894.$$

De manera que la clasificación mediante estas variables no es muy precisa, ya que podemos tener un 18,94% de probabilidad de error. Calculemos la probabilidad a posteriori de que el cuadro pertenezca al pintor A suponiendo que, a priori, ambos pintores son igualmente probables.

$$P(A/\mathbf{x}) = \frac{1}{1 + \exp(-0.5(0,8133 - 0,52)} = \frac{1}{1.86} = 0,5376$$

Esta probabilidad indica que al clasificar la obra como perteneciente al pintor A existe mucha incertidumbre en la decisión, ya que las probabilidades de que pertenezca a cada pintor son semejantes (0,5376 y 0,4624).

13.3 GENERALIZACIÓN PARA VARIAS POBLACIONES NORMALES

13.3.1 Planteamiento General

La generalización de estas ideas para G poblaciones es simple: el objetivo es ahora dividir el espacio E_x en G regiones $A_1, \ldots, A_g, \ldots, A_G$ tales que si $\mathbf{x}$ pertenece a A_i el punto se clasifica en la población P_i. Supondremos que los costes de clasificación son constantes y no dependen de la población en que se haya clasificado. Entonces, la región A_g vendrá definida por aquellos puntos con máxima probabilidad de ser generados por P_g, es decir donde el producto de la probabilidad a priori y la verosimilitud sean máximas:

$$A_g = \{\mathbf{x} \in E_x | \pi_g f_g(\mathbf{x}) > \pi_i f_i(\mathbf{x}); \forall i \neq g\} \tag{13.17}$$

Si las probabilidades a priori son iguales, $\pi_i = G^{-1}, \forall i$, y las distribuciones $f_i(\mathbf{x})$ son normales con la misma matriz de varianzas, la condición (13.17) equivale a calcular la distancia de Mahalanobis del punto observado al centro de cada población y clasificarle en

la población que haga esta distancia mínima. Minimizar las distancias de Mahalanobis $\left(\mathbf{x}-\boldsymbol{\mu}_g\right)' \mathbf{V}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_g\right)$ equivale, eliminando el término $\mathbf{x}'\mathbf{V}^{-1}\mathbf{x}$ que aparece en todas las ecuaciones, a minimizar el indicador lineal

$$L_g\left(\mathbf{x}\right) = -2\boldsymbol{\mu}_g'\mathbf{V}^{-1}\mathbf{x} + \boldsymbol{\mu}_g'\mathbf{V}^{-1}\boldsymbol{\mu}_g. \tag{13.18}$$

y llamando

$$\mathbf{w}_g = \mathbf{V}^{-1}\boldsymbol{\mu}_g$$

la regla es

$$\min_g(\mathbf{w}_g^{'}\boldsymbol{\mu}_g - 2\mathbf{w}_g^{'}\mathbf{x})$$

Para interpretar esta regla, observemos que la frontera de separación entre dos poblaciones, (ij), vendrá definida por:

$$A_{ij}(\mathbf{x}) = L_i(\mathbf{x}) - L_j(\mathbf{x}) = 0 \tag{13.19}$$

sustituyendo con (13.18) y reordenando los términos se obtiene:

$$A_{ij}(\mathbf{x}) = 2(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\mathbf{V}^{-1}\mathbf{x} + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\mathbf{V}^{-1}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) = 0$$

y llamando

$$\mathbf{w}_{ij} = \mathbf{V}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) = \mathbf{w}_i - \mathbf{w}_j$$

la frontera puede escribirse como:

$$\mathbf{w}_{ij}'\mathbf{x} = \mathbf{w}_{ij}'\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j).$$

Esta ecuación admite la misma interpretación como proyección que en el caso de dos poblaciones. Se construye una dirección $\mathbf{w}_{ij}$ y se proyectan las medias y el punto $\mathbf{x}$ que tratamos de clasificar sobre esta dirección. La región de indiferencia es cuando el punto proyectado está equidistante de las medias proyectadas. En otro caso, asignaremos el punto a la población de cuya media proyectada esté más próxima.

Vamos a comprobar que si tenemos G poblaciones sólo necesitamos encontrar

$$r = \min(G-1, p)$$

direcciones de proyección. En primer lugar observemos que, aunque podemos construir $\binom{G}{2} = G\left(G-1\right)/2$ vectores $\mathbf{w}_{ij}$ a partir de las G medias, una vez que tenemos $G-1$ vectores los demás quedan determinados por éstos. Podemos determinar los $G-1$ vectores $\mathbf{w}_{i,i+1}$, para $i = 1, ..., G-1$, y obtener cualquier otro a partir de estas $G-1$ direcciones. Por ejemplo:

$$\mathbf{w}_{i,i+2} = \mathbf{V}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i+2}) = \mathbf{V}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i+1}) - \mathbf{V}^{-1}(\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}_{i+2}) = \mathbf{w}_{i,i+1} - \mathbf{w}_{i+1,i+2}.$$

En conclusión, si $p > G-1$, el número máximo de vectores $\mathbf{w}$ que podemos tener es $G-1$, ya que los demás se deducen de ellos. Cuando $p \leq G-1$, como estos vectores pertenecen a R^p el número máximo de vectores linealmente independientes es p.

Es importante resaltar que, como es natural, la regla de decisión obtenida cumple la propiedad transitiva. Por ejemplo, si $G = 3$, y obtenemos que para un punto $(\mathbf{x})$

$$\begin{aligned} D_1^2(\mathbf{x}) &> D_2^2(\mathbf{x}) \\ D_2^2(\mathbf{x}) &> D_3^2(\mathbf{x}) \end{aligned}$$

entonces forzosamente debemos concluir que $D_1^2(\mathbf{x}) > D_3^2(\mathbf{x})$ y esta será el resultado que obtendremos si calculamos estas distancias, por lo que el análisis es coherente. Además, si $p = 2$, cada una de las tres ecuaciones $A_{ij}(\mathbf{x}) = 0$ será una recta y las tres se cortarán en el mismo punto. En efecto, cualquier recta que pase por el punto de corte de las rectas $A_{12}(\mathbf{x}) = 0$ y $A_{23}(\mathbf{x}) = 0$ tiene la expresión

$$a_1 A_{12}(\mathbf{x}) + a_2 A_{23}(\mathbf{x}) = 0$$

ya que si $\mathbf{x}_0^*$ es el punto de corte como $A_{12}(\mathbf{x}^*) = 0$, por pertenecer a la primera recta, y $A_{23}(\mathbf{x}^*) = 0$, por pertenecer a la segunda, pertenecerá a la combinación lineal. Como, según (13.19), $A_{13}(\mathbf{x}) = L_1(\mathbf{x}) - L_3(\mathbf{x}) = L_1(\mathbf{x}) - L_2(\mathbf{x}) + L_2(\mathbf{x}) - L_3(\mathbf{x})$, tenemos que:

$$A_{13}(\mathbf{x}) = A_{12}(\mathbf{x}) + A_{23}(\mathbf{x})$$

y la recta $A_{13}(\mathbf{x})$ debe siempre pasar por el punto de corte de las otras dos.

13.3.2 Procedimiento operativo

Para ilustrar el procedimiento operativo, supongamos cinco poblaciones con $p > 4$, con lo que existirán cuatro reglas de clasificación independientes y las demás se deducen de ellas. Tenemos dos formas de realizar el análisis. La primera es calcular para las G poblaciones las distancias de Mahalanobis (o lo que es equivalente, las proyecciones (13.18)) y clasificar el elemento en la más próxima. La segunda es hacer el análisis comparando las poblaciones dos a dos. Supongamos que hemos obtenido de las comparaciones 2 a 2 los siguientes resultados: ($i > j$ indica que la población i es preferida a la j, es decir, el punto se encuentra más próximo a la media de la población i que a la de j):

$$\begin{aligned} 1 &> 2 \\ 2 &> 3 \\ 4 &> 3 \\ 5 &> 4 \end{aligned}$$

Las poblaciones 2, 3 y 4 quedan descartadas (ya que $1 > 2 > 3$ y $5 > 4$). La duda no resuelta se refiere a las poblaciones 1 y 5. Construyendo (a partir de las reglas anteriores) la regla para discriminar entre estas dos últimas poblaciones, supongamos que

$$5 > 1$$

y clasificaremos en la población 5.

Cuando $p < G - 1$ el máximo número de proyecciones linealmente independientes que podemos construir es p, y éste será el máximo número de variables a definir. Por ejemplo, supongamos que $p = 2$ y $G = 5$. Podemos definir una dirección de proyección cualquiera, por ejemplo

$$\mathbf{w}_{12} = \mathbf{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

y proyectar todas las medias $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_5)$ y el punto $\mathbf{x}$ sobre dicha dirección. Entonces, clasificaremos el punto en la población de cuya media proyectada está más próxima. Ahora bien, es posible que sobre esta dirección coincidan las medias proyectadas de varias poblaciones. Si esto ocurre con, por ejemplo, las $\boldsymbol{\mu}_4$ y $\boldsymbol{\mu}_5$, resolveremos el problema proyectando sobre la dirección definida por otra pareja de poblaciones.

Ejemplo 13.2 *Una máquina que admite monedas realiza tres mediciones de cada moneda para determinar su valor: peso (x_1), espesor (x_2) y la densidad de estrías en su canto (x_3). Los instrumentos de medición de estas variables no son muy precisos y se ha comprobado en una amplia experimentación con tres tipos de monedas usadas, M_1, M_2, M_3 , que las medidas se distribuyen normalmente con medias para cada tipo de moneda dadas por:*

$$\begin{array}{rccc} \boldsymbol{\mu}_1 = & 20 & 8 & 8 \\ \boldsymbol{\mu}_2 = & 19.5 & 7.8 & 10 \\ \boldsymbol{\mu}_3 = & 20.5 & 8.3 & 5 \end{array}$$

y matriz de covarianzas

$$V = \begin{bmatrix} 4 & .8 & -5 \\ .8 & .25 & -.9 \\ -5 & -.9 & 9 \end{bmatrix}$$

Indicar cómo se clasificaría una moneda con medidas (22, 8.5 ,7) y analizar la regla de clasificación. Calcular las probabilidades de error.

Aparentemente la moneda a clasificar está más próxima a M_3 en las dos primeras coordenadas, pero más próxima a M_1 por x_3, la densidad de estrías. La variable indicador para clasificar entre M_1 y M_3 es

$$z = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)\mathbf{V}^{-1}\mathbf{x} = 1.77x_1 - 3.31x_2 + .98x_3$$

la media de esta variable para la primera moneda, M_1,es $1.77 \times 20 - 3.31 \times 8 + .98 \times 8 = 16.71$ y para la tercera, M_3, $1.77 \times 20.5 - 3.31 \times 8.3 + .98 \times 5 = 13.65$. El punto de corte es la media, 15.17. Como para la moneda a clasificar es

$$z = 1.77 \times 22 - 3.31 \times 8.5 + .98 \times 7 = 17.61$$

la clasificaremos como M_1. Este análisis es equivalente a calcular las distancias de Mahalanobis a cada población que resultan ser $D_1^2 = 1.84$, $D_2^2 = 2.01$ y $D_3^2 = 6.69$. Por tanto clasificamos primero en M_1, luego en M_2 y finalmente como M_3. La regla para clasificar entre la primera y la segunda es

$$z = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\mathbf{V}^{-1}\mathbf{x} = -.93x_1 + 1.74x_2 - .56x_3$$

de estas dos reglas deducimos inmediatamente la regla para clasificar entre la segunda y la tercera, ya que

$$(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)\mathbf{V}^{-1}\mathbf{x} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)\mathbf{V}^{-1}\mathbf{x} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\mathbf{V}^{-1}\mathbf{x}$$

Analicemos ahora las reglas de clasificación obtenidas. Vamos a expresar la reglas inicial para clasificar entre M_1 y M_3 para las variables estandarizadas, con lo que se evita el problema de las unidades. Llamando $\widetilde{x}_i$ a las variables divididas por sus desviaciones típicas $\widetilde{x}_1 = x_1/2;\widetilde{x}_2 = x_2/.5$, y $\widetilde{x}_3 = x_3/3$, la regla en variables estandarizadas es

$$z = 3.54\widetilde{x}_1 - 1.65\widetilde{x}_2 + 2.94\widetilde{x}_3$$

que indica que las variables con más peso para decidir la clasificación son la primera y la tercera, que son la que tienen mayores coeficientes. Observemos que con variables estandarizadas la matriz de covarianzas es la de correlación

$$R = \begin{bmatrix} 1 & .8 & -.83 \\ .8 & 1 & -.6 \\ -.83 & -.6 & 1 \end{bmatrix}$$

El origen de estas correlaciones entre los errores de medida es que si la moneda adquiere suciedad y aumenta ligeramente su peso, también aumenta su espesor y hace más difícil determinar su densidad de estrías. Por eso hay correlaciones positivas entre peso y espesor, al aumentar el peso aumenta el espesor, pero negativas con las estrías. Aunque la moneda que queremos clasificar tiene mucho peso y espesor, lo que indicaría que pertenece a la clase 3, entonces la densidad de estrías debería medirse como baja, ya que hay correlaciones negativas entre ambas medidas, y sin embargo se mide relativamente alta en la moneda. Las tres medidas son coherentes con una moneda sucia del tipo 1, y por eso se clasifica con facilidad en ese grupo.

Vamos a calcular la probabilidad a posteriori de que la observación sea de la clase M_1. Suponiendo que las probabilidades a priori son iguales esta probabilidad será

$$P(1/x_0) = \frac{\exp(-D_1^2/2)}{\exp(-D_1^2/2) + \exp(-D_2^2/2) + \exp(-D_3^2/2)}$$

y sustituyendo las distancias de Mahalanobis

$$P(1/x_0) = \frac{\exp(-1.84/2)}{\exp(-1.84/2) + \exp(-2.01/2) + \exp(-6.69/2)} = .50$$

y análogamente $P(2/x_0) = .46$, y $P(3/x_0) = .04$.

Podemos calcular las probabilidades de error de clasificar una moneda de cualquier tipo en otra clase. Por ejemplo, la probabilidad de clasificar una moneda M_3 con esta regla como tipo M_1 es

$$P(z > 15.17/N(13.64, \sqrt{3.07})) = P(y > \frac{15.17 - 13.64}{1.75}) = P(y > .87) = .192$$

como vemos esta probabilidad es bastante alta. Si queremos reducirla hay que aumentar la distancia de Mahalanobis entre las medias de los grupos, lo que supone "aumentar" la matriz $\mathbf{V}^{-1}$ o "reducir" la matriz $\mathbf{V}$. Por ejemplo, si reducimos a la mitad el error en la medida de las estrías introduciendo medidores más precisos, pero se mantiene las correlaciones con las otras medidas, pasamos a la matriz de covarianzas

$$V_2 = \begin{bmatrix} 4 & .8 & -2.5 \\ .8 & .25 & -.45 \\ -1 & -.2 & 2.25 \end{bmatrix}$$

la regla de clasificación entre la primera y la tercera es ahora

$$z = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)\mathbf{V}^{-1}\mathbf{x} = 3.44x_1 - 4.57x_2 + 4.24x_3$$

y la distancia de Mahalanobis entre las poblaciones 1 y 3 (monedas M_1 y M_3) ha pasado de 3.01 a 12.38, lo que implica que la probabilidad de error entre estas dos poblaciones ha disminuido a $1 - \Phi(\sqrt{12.38}/2) = 1 - \Phi(1.76) = .04$ y vemos que la probabilidad de error ha disminuido considerablemente. Podemos así calcular la precisión en las medidas que necesitaríamos para conseguir unas probabilidades de error determinadas.

13.4 POBLACIONES DESCONOCIDAS. CASO GENERAL

13.4.1 Regla estimada de clasificación

Vamos a estudiar cómo aplicar la teoría anterior cuando en lugar de trabajar con poblaciones disponemos de muestras. Abordaremos directamente el caso de G poblaciones posibles. Como caso particular, la discriminación clásica es para $G = 2$. La matriz general de datos $\mathbf{X}$ de dimensiones $n \times p$, (n individuos y p variables), puede considerarse particionada ahora en G matrices correspondientes a las subpoblaciones. Vamos a llamar x_{ijg} a los elementos de estas submatrices, donde i representa el individuo, j la variable y g el grupo o submatriz. Llamaremos n_g al número de elementos en el grupo g y el número total de observaciones es:

$$n = \sum_{g=1}^{G} n_g$$

Vamos a llamar $\mathbf{x}'_{ig}$ al vector fila $(1 \times p)$ que contiene los p valores de las variables para el individuo i en el grupo g, es decir, $\mathbf{x}'_{ig} = (x_{i1g}, ..., x_{ipg})$. El vector de medias dentro de cada clase o subpoblación será:

$$\overline{\mathbf{x}}_g = \frac{1}{n_g}\sum_{i=1}^{n_g}\mathbf{x}_{ig} \tag{13.20}$$

y es un vector columna de dimensión p que contiene las p medias para las observaciones de la clase g. La matriz de varianzas y covarianzas para los elementos de la clase g será:

$$\widehat{\mathbf{S}}_g = \frac{1}{n_g - 1}\sum_{i=1}^{n_g}(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)(\mathbf{x}_{ig} - \overline{\mathbf{x}}_g)' \tag{13.21}$$

donde hemos dividido por $n_g - 1$ para tener estimaciones centradas de las varianzas y covarianzas. Si suponemos que las G subpoblaciones tienen la misma matriz de varianzas y covarianzas, su mejor estimación centrada con todos los datos será una combinación lineal de las estimaciones centradas de cada población con peso proporcional a su precisión. Por tanto:

$$\widehat{\mathbf{S}}_w = \sum_{g=1}^{G}\frac{n_g - 1}{n - G}\widehat{\mathbf{S}}_g$$

y llamaremos $\mathbf{W}$ a la matriz de sumas de cuadrados *dentro* de las clases que viene dada por:

$$\mathbf{W} = (n - G)\widehat{\mathbf{S}}_w \tag{13.22}$$

Para obtener las funciones discriminantes utilizaremos $\overline{\mathbf{x}}_g$ como estimación de μ_g, y $\widehat{\mathbf{S}}_w$ como estimación de $\mathbf{V}$. En concreto, suponiendo iguales las probabilidades a priori y los costes de clasificación, clasificaremos al elemento en el grupo que conduzca a un valor mínimo de la distancia de Mahalanobis entre el punto $\mathbf{x}$ y la media del grupo. Es decir, llamando $\widehat{\mathbf{w}}_g = \widehat{\mathbf{S}}_w^{-1}\overline{\mathbf{x}}_g$ clasificaremos un nuevo elemento $\mathbf{x}_0$ en aquella población g donde

$$\min_g(\mathbf{x}_0 - \overline{\mathbf{x}}_g)'\widehat{\mathbf{S}}_w^{-1}(\mathbf{x}_0 - \overline{\mathbf{x}}_g) = \min_g \widehat{\mathbf{w}}_g'(\overline{\mathbf{x}}_g - \mathbf{x}_0)$$

que equivale a construir las variables indicadoras escalares

$$z_{g,g+1} = \widehat{\mathbf{w}}_{g,g+1}'\mathbf{x}_0 \quad g = 1, ..., G$$

donde

$$\widehat{\mathbf{w}}_{g,g+1} = \widehat{\mathbf{S}}_w^{-1}(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_{g+1}) = \widehat{\mathbf{w}}_g - \widehat{\mathbf{w}}_{g+1}$$

y clasificar en g frente a $g + 1$ si

$$|z_{g,g+1} - \widehat{m}_g| < |z_{g,g+1} - \widehat{m}_{g+1}|$$

donde $\widehat{m}_g = \widehat{\mathbf{w}}'_{g,g+1}\overline{\mathbf{x}}_g$.

Conviene antes de construir la regla de clasificación realizar un test de que los grupos son realmente distintos, es decir, que no todas las medias $\boldsymbol{\mu}_g$ son iguales. Este contraste puede realizarse siguiendo lo expuesto en la sección 10.7. En la apéndice 13.2 se demuestra que en el caso de dos grupos la función de discriminación lineal $\widehat{w} = \widehat{\mathbf{S}}_{\omega}^{-1}(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)$ puede obtenerse por regresión, definiendo una variable ficticia que tome los valores cero o uno según que el dato pertenezca a una u otra población.

13.4.2 Cálculo de Probabilidades de error

El cálculo de probabilidades de error podría hacerse sustituyendo los parámetros desconocidos por los estimados y aplicando las fórmulas de la sección 13.2, pero este método no es recomendable ya que va a subestimar mucho las probabilidades de error al no tener en cuenta la incertidumbre de estimación de los parámetros. Un mejor procedimiento, que además no depende de la hipótesis de normalidad, es aplicar la función discriminante a las n observaciones y clasificarlas. En el caso de 2 grupos, obtendríamos la tabla:

		Clasificado	
		P_1	P_2
Realidad	P_1	n_{11}	n_{12}
	P_2	n_{21}	n_{22}

donde n_{ij} es el número de datos que viniendo de la población i se clasifica en j. El error aparente de la regla es:

$$\text{Error} = \frac{n_{12} + n_{21}}{n_{11} + n_{22}} = \frac{\text{Total mal clasificados}}{\text{Total bien clasificados}}.$$

Este método tiende a subestimar las probabilidades de error ya que los mismos datos se utilizan para estimar los parámetros y para evaluar el procedimiento resultante. Un procedimiento mejor es clasificar cada elemento con una regla que no se ha construido usándolo. Para ello, podemos construir n funciones discriminantes con las n muestras de tamaño $n-1$ que resultan al eliminar uno a uno cada elemento de la población y clasificar después cada dato con la regla construida sin él. Este método se conoce como *validación cruzada* y conduce a una mejor estimación del error de clasificación. Si el número de observaciones es muy alto, el coste computational de la validación cruzada es alto y una solución más rápida es subdividir la muestra en k grupos iguales y realizar la validación cruzada eliminado en lugar de una observación uno de estos grupos.

Ejemplo 13.3 *Vamos a utilizar los datos de MEDIFIS para clasificar personas por su género conocidas las medidas físicas de las variables de la tabla A.5. Como los datos para toda la población de hombre y mujeres son desconocidos, vamos a trabajar con los datos muestrales. En la muestra hay 15 mujeres (variable sexo=0) y 12 hombres (sexo=1).*

En el ejemplo 10.2 comprobamos que las medias de las poblaciones de las medidas físicas de hombres y mujeres son diferentes. Las funciones discriminantes $\widehat{\mathbf{w}}_g = \widehat{\mathbf{S}}_w^{-1}\overline{\mathbf{x}}_g$ *se indican en la tabla adjunta*

	est	*pes*	*pie*	*lbr*	*aes*	*dcr*	*lrt*
hombres	−1.30	−4.4	20.0	10.0	−2.1	24.4	−4.4
mujeres	−1.0	−4.4	17.7	9.5	−2.5	25.1	−4.7
diferencia	−.3	0	2.3	.5	.4	−.7	.3

La diferencia entre estas dos funciones proporciona la función lineal discriminante. Se observa que la variable con mayor peso en la discriminación es la longitud del pie. Para interpretar este resultado, la tabla siguiente indica las diferencias estandarizadas entre los medias de cada variable en ambas poblaciones. Por ejemplo, la diferencia estandarizada entre las estaturas es ($177.58 - 161.73)/6.4 = 2.477$

	est	*pes*	*pie*	*lbr*	*aes*	*dcr*	*lrt*
dif medias	15.8	18.65	4.83	7.72	5.67	1.36	4.56
desv. típicas	6.4	8.8	1.5	3.1	2.9	1.7	2.2
dif. estand.	2.47	2.11	3.18	2.48	1.97	.78	2.07

La variable que separa más ambas poblaciones es la longitud del pie. Como, además, la longitud del pie esta muy correlada con la estatura y la longitud del brazo, conocida la longitud del pie estas variables no son tan informativas, lo que explica su bajo peso en la función discriminante.

Si aplicamos la función discriminante para clasificar los datos muestrales obtenemos un porcentaje de éxitos del 100%. Todas las observaciones se clasifican bien. Aplicando validación cruzada se obtiene

		Clasificado	
		M	H
Realidad	M	13	2
	H	2	10

Ejemplo 13.4 *que supone una proporción de aciertos de 23/27=0.852. Las observaciones mal clasificadas son las 2, 7, 9, y 18. Vemos que el método de validación cruzada da una idea más realista de la eficacia del procedimiento de clasificación.*

13.5 VARIABLES CANÓNICAS DISCRIMINANTES

13.5.1 El caso de dos grupos

La función lineal discriminante para dos grupos fue deducida por primera vez por Fisher por un razonamiento intuitivo que vamos a resumir brevemente. El criterio propuesto por Fisher es encontrar una variable escalar:

$$z = \boldsymbol{\alpha}'\mathbf{x} \tag{13.23}$$

tal que maximice la distancia entre las medias proyectadas con relación a la variabilidad resultante en la proyección. Intuitivamente, la escala z permitirá separar lo más posible ambos grupos.

La media de la variable z en el grupo 1, que es la proyección del vector de medias sobre la dirección de $\boldsymbol{\alpha}$, es $\widehat{m}_1 = \boldsymbol{\alpha}'\overline{\mathbf{x}}_1$, y la media en el grupo 2 es $\widehat{m}_2 = \boldsymbol{\alpha}'\overline{\mathbf{x}}_2$. La varianza de

la variable z será la misma en ambos grupos, $\boldsymbol{\alpha}'\mathbf{V}\boldsymbol{\alpha}$, y la estimaremos con $s_z^2 = \boldsymbol{\alpha}' S_w \boldsymbol{\alpha}$. Se desea escoger $\boldsymbol{\alpha}$ de manera que la separación entre las medias m_1 y m_2 sea máxima. Una medida adimensional de esta separación es:

$$\phi = \left(\frac{\widehat{m}_2 - \widehat{m}_1}{s_z} \right)^2,$$

y esta expresión es equivalente a:

$$\phi = \frac{(\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1))^2}{\boldsymbol{\alpha}' S_w \boldsymbol{\alpha}}. \tag{13.24}$$

En esta relación $\boldsymbol{\alpha}$ representa una dirección, ya que ϕ es invariante ante multiplicaciones de $\boldsymbol{\alpha}$ por una constante: si $\boldsymbol{\beta} = p\boldsymbol{\alpha}$, $\phi(\boldsymbol{\beta}) = \phi(\boldsymbol{\alpha})$. Para encontrar la dirección $\boldsymbol{\alpha}$ que maximice ϕ, derivando (13.24) e igualando a cero:

$$\frac{d\phi}{d\boldsymbol{\alpha}} = \mathbf{0} = \frac{2\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)'\boldsymbol{\alpha}' S_w \boldsymbol{\alpha} - 2S_w \boldsymbol{\alpha} \left(\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)\right)^2}{(\boldsymbol{\alpha}' S_w \boldsymbol{\alpha})^2}$$

que escribiremos:

$$(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)\boldsymbol{\alpha}'\mathbf{S_w}\boldsymbol{\alpha} = \mathbf{S_w}\boldsymbol{\alpha}\left(\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)\right)$$

o también

$$(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) = S_w \boldsymbol{\alpha} \frac{(\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1))}{\boldsymbol{\alpha}' S_w \boldsymbol{\alpha}}$$

que resulta en

$$\boldsymbol{\alpha} = \lambda {S_w}^{-1}(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)$$

donde $\lambda = (\boldsymbol{\alpha}'\mathbf{S}_w\boldsymbol{\alpha})/\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)$. Como, dada $\boldsymbol{\alpha}$, λ es una constante y la función a optimizar es invariante ante constantes, podemos tomar $\boldsymbol{\alpha}$ normalizado para que $\lambda = 1$, con lo que resulta:

$$\boldsymbol{\alpha} = \mathbf{S_w}^{-1}(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) \tag{13.25}$$

que es la dirección $\mathbf{w}$ de proyección que hemos encontrado en la sección anterior. Además:

$$\boldsymbol{\alpha}'\mathbf{S}_w\boldsymbol{\alpha} = (\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)'\mathbf{S_w}^{-1}(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) = D^2(\overline{\mathbf{x}}_2, \overline{\mathbf{x}}_1) = (\widehat{m}_2 - \widehat{m}_1)^2$$

y la varianza de la variable resultante de la proyección es la distancia de Mahalanobis entre las medias. También:

$$\boldsymbol{\alpha}'(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) = (\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1)'\mathbf{S_w}^{-1}(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_1) = D^2(\overline{\mathbf{x}}_2, \overline{\mathbf{x}}_1)$$

y comparando con (13.24) vemos que ϕ es la distancia de Mahalanobis entre las medias. El procedimiento obtenido conduce a buscar una dirección de proyección que maximice la distancia de Mahalanobis entre los centros de ambas poblaciones. Observemos que si $\mathbf{S}_w = \mathbf{I}$ la distancia de Mahalanobis se reduce a la euclídea y la dirección de proyección es paralela al vector que une ambas medias. Finalmente, observemos que esta regla se ha obtenido sin imponer ninguna hipótesis sobre la distribución de la variable $\mathbf{x}$ en las poblaciones.

13.5.2 Varios Grupos

El enfoque de Fisher puede generalizarse para encontrar variables canónicas que tengan máximo poder discriminante para clasificar nuevos elementos entre G poblaciones. El objetivo es, en lugar de trabajar con las p variables originales $\mathbf{x}$, definir un vector $\mathbf{z} = (z_1, ..., z_r)'$ de r variables canónicas, donde $r = min(G-1, p)$, que se obtengan como combinación lineal de las originales, $z_i = \mathbf{u}_i'\mathbf{x}$, y que permitan resolver el problema de clasificación de la forma siguiente:

(1) Proyectamos las medias de las variables en los grupos, $\overline{\mathbf{x}}_g$, sobre el espacio determinado por las r variables canónicas. Sean $\overline{\mathbf{z}}_1, ..., \overline{\mathbf{z}}_g$ las variables $r \times 1$ cuyas coordenadas son estas proyecciones.

(2) Proyectamos el punto $\mathbf{x}_0$ a clasificar y sea $\mathbf{z}_0$ su proyección sobre dicho espacio.

(3) Clasificamos el punto en aquella población de cuya media se encuentre más próxima. Las distancias se miden con la distancia euclídea en el espacio de las variables canónicas z. Es decir, clasificaremos en la población i si:

$$(\mathbf{z}_0 - \overline{\mathbf{z}}_i)'(\mathbf{z}_0 - \overline{\mathbf{z}}_i) = \min_g (\mathbf{z}_0 - \overline{\mathbf{z}}_g)'(\mathbf{z}_0 - \overline{\mathbf{z}}_g)$$

Con varios grupos la separación entre las medias la mediremos por el cociente entre la variabilidad entre grupos, o variabilidad explicada por los grupos, y la variabilidad dentro de los grupos, o no explicada o residual. Este es el criterio habitual para comparar varias medias en el análisis de la varianza y conduce al estadístico F de Fisher. Para obtener las variables canónicas discriminantes comenzamos buscando un vector $\mathbf{u}_1'$, de norma uno, tal que los grupos de puntos proyectados sobre él tengan separación relativa máxima. La proyección de la media de las observaciones del grupo g en esta dirección será la variable escalar:

$$\overline{z}_g = \mathbf{u}_1'\overline{\mathbf{x}}_g$$

y la proyección de la media para todos los datos será:

$$z_T = \mathbf{u}_1'\overline{\mathbf{x}}_T$$

donde $\overline{\mathbf{x}}_T$ es el vector $p \times 1$ que contiene las medias de las p variables para las n observaciones de la muestra uniendo todos los grupos. Tomando como medida de la distancia entre las medias de los grupos proyectadas $\overline{z}_1, ..., \overline{z}_g$ su variación total dada por $\sum_{g=1}^{G} n_g(\overline{z}_g - \overline{z}_T)^2$, y comparando esta cantidad con la variabilidad dentro de los grupos, dada por $\sum\sum(z_{ig} - \overline{z}_g)^2$, la separación relativa entre las medias, vendrá dada por el estadístico :

$$\phi = \frac{\sum n_g(\overline{z}_g - \overline{z}_T)^2}{\sum\sum(z_{ig} - \overline{z}_g)^2}.$$

y si todos los datos provienen de la misma población y no existen grupos distintos esta variable se distribuye como una F con $G-1$ y $n-G+1$ grados delibertad. Vamos a expresar este criterio en función de los datos originales. La suma de cuadrados *dentro* de grupos, o variabilidad no explicada, para los puntos proyectados, es:

$$VNE = \sum_{j=1}^{n_g}\sum_{g=1}^{G}(z_{jg} - \overline{z}_g)^2 = \sum_{j=1}^{n_g}\sum_{g=1}^{G}\mathbf{u}'(\mathbf{x}_{jg} - \overline{\mathbf{x}}_g)(\mathbf{x}_{jg} - \overline{\mathbf{x}}_g)'\mathbf{u} = \mathbf{u}'\mathbf{W}\mathbf{u}$$

donde $\mathbf{W}$ está dada por

$$\mathbf{W} = \sum_{j=1}^{n_g} \sum_{g=1}^{G} (\mathbf{x}_{jg} - \overline{\mathbf{x}}_g)(\mathbf{x}_{jg} - \overline{\mathbf{x}}_g)'$$

que coincide con (13.22). Esta matriz tiene dimensiones $p \times p$ y tendrá, en general, rango p, suponiendo $n - G \geq p$. Estima la variabilidad de los datos respecto a sus medias de grupo, que es la misma, por hipótesis, en todos ellos.

La suma de cuadrados *entre* grupos, o variabilidad explicada, para los puntos proyectados es:

$$\begin{aligned} VE &= \sum_{g=1}^{G} n_g(\overline{z}_g - \overline{z}_T)^2 = \\ &= \sum n_g \mathbf{u}'(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)'\mathbf{u} = \\ &= \mathbf{u}'\mathbf{B}\mathbf{u} \end{aligned} \tag{13.26}$$

siendo $\mathbf{B}$ la matriz de suma de cuadrados entre grupos, que puede escribirse:

$$\mathbf{B} = \sum_{g=1}^{G} n_g \mathbf{a}_g \mathbf{a}_g'$$

siendo $\mathbf{a}_g = \overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T$. La matriz $\mathbf{B}$ es cuadrada $p \times p$ y simétrica y se obtiene como suma de G matrices de rango uno formadas por los vectores $\mathbf{a}_g$, que no son independientes, ya que están ligados por la relación $\sum_{g=1}^{G} n_g \mathbf{a}_g = \mathbf{0}$,que implica que el rango de $\mathbf{B}$ será $G - 1$.

En resumen, la matriz $\mathbf{W}$ mide las diferencias dentro de grupos y la $\mathbf{B}$ las diferencias entre grupos. La cantidad a maximizar puede también escribirse:

$$\phi = \frac{\mathbf{u}_1'\mathbf{B}\mathbf{u}_1}{\mathbf{u}_1'\mathbf{W}\mathbf{u}_1}, \tag{13.27}$$

derivando e igualando a cero de la forma habitual:

$$\frac{d\phi}{du_1} = 0 = \frac{2\mathbf{B}\mathbf{u}_1(\mathbf{u}_1'\mathbf{W}\mathbf{u}_1) - 2(\mathbf{u}_1'\mathbf{B}\mathbf{u}_1)\mathbf{W}\mathbf{u}_1}{(\mathbf{u}_1'\mathbf{W}\mathbf{u}_1)^2} = 0$$

entonces:

$$\mathbf{B}\mathbf{u}_1 = \mathbf{W}\mathbf{u}_1 \left(\frac{\mathbf{u}_1'\mathbf{B}\mathbf{u}_1}{\mathbf{u}_1'\mathbf{W}\mathbf{u}_1} \right)$$

es decir, por

$$\mathbf{B}\mathbf{u}_1 = \phi \mathbf{W}\mathbf{u}_1$$

y suponiendo $\mathbf{W}$ no singular:

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{u}_1 = \phi \mathbf{u}_1$$

que implica que $\mathbf{u}_1$ debe de ser un vector propio de $\mathbf{W}^{-1}\mathbf{B}$ y entonces ϕ es su valor propio asociado. Como queremos maximizar ϕ, que es el valor del estadístico F en un contrate escalar sobre las medias proyectadas, $\mathbf{u}$ será el vector propio asociado al mayor valor propio de la matriz $\mathbf{W}^{-1}\mathbf{B}$.

Podemos plantearnos obtener un segundo eje tal que maximice la separación ϕ, pero con la condición de que la nueva variable canónica $z_2 = \mathbf{u}_2'\mathbf{x}$ esté incorrelada con la primera, $z_1 = \mathbf{u}_1'\mathbf{x}$. Puede demostrarse análogamente que esto ocurre si tomamos el segundo vector propio (ligado al segundo valor propio) de la matriz $\mathbf{W}^{-1}\mathbf{B}$. En general, sean $\alpha_1, \ldots, \alpha_r$ los valores propios no nulos de $\mathbf{W}^{-1}\mathbf{B}$ y $\mathbf{u}_1, \ldots, \mathbf{u}_r$ los vectores propios ligados a los valores propios no nulos. Las variables escalares $z_j = \mathbf{u}_j'\mathbf{x}$ ordenadas por los valores propios $\alpha_1 > \alpha_2 > \ldots > \alpha_r$ proporcionan máxima separación en el sentido de que el estadístico F para contrastar si existen diferencias entre los G grupos proyectados tiene un valor igual a α_j. Además, estas variables escalares z_j están incorreladas, tanto dentro de grupos como en toda la muestra. Para comprobarlo sea $\mathbf{z}_j$ el vector $n \times 1$ resultado de proyectar los puntos muestrales en la dirección $\mathbf{u}_j'$, es decir, $\mathbf{z}_j = \mathbf{X}\mathbf{u}_j$. Esta variable tendrá media $\overline{z_j} = \mathbf{1}'\mathbf{z}_j/n = \mathbf{1}'\mathbf{X}\mathbf{u}_j/n = \overline{\mathbf{x}}_T'\mathbf{u}_j$ y la covarianza entre dos variables escalares, z_j y z_h vendrá dada por

$$cov(z_j, z_h) = \frac{1}{n}\sum_{i=1}^{n}(z_{ji} - \overline{z}_j)(z_{hi} - \overline{z}_h) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}_j'(\mathbf{x}_i - \overline{\mathbf{x}}_T)(\mathbf{x}_i - \overline{\mathbf{x}}_T)'\mathbf{u}_h$$

y llamando $\mathbf{T}$ a la matriz de suma de cuadrados totales la covarianzas entre las variables canónicas son $\mathbf{u}_j'\mathbf{T}\mathbf{u}_h$. Si descomponemos estas variables en grupos, de manera que cada variable $\mathbf{z}_j$ da lugar a G variables z_{jg} donde g indica el grupo, puede comprobarse analogamente que las covarianzas entre z_{jg} y z_{hg}, sumadas para todos los grupos vienen dadas por $\mathbf{u}_j'\mathbf{W}\mathbf{u}_h$. Vamos a demostrar que, para dos vectores propios distintos, $h \neq j$:

$$\mathbf{u}_h'\mathbf{W}\mathbf{u}_j = \mathbf{u}_h'\mathbf{T}\mathbf{u}_j = 0,$$

donde $\mathbf{T} = \mathbf{W} + \mathbf{B}$.

Comprobemos esta propiedad. Supongamos que $\alpha_h > \alpha_j$. Los vectores propios de $\mathbf{W}^{-1}\mathbf{B}$ verifican que

$$(\mathbf{W}^{-1}\mathbf{B})\mathbf{u}_h = \alpha_h\mathbf{u}_h$$

es decir

$$\mathbf{B}\mathbf{u}_h = \alpha_h\mathbf{W}\mathbf{u}_h. \tag{13.28}$$

Por tanto, para otro vector propio distinto $\mathbf{u}_j$, donde $\alpha_h \neq \alpha_j$, tenemos:

$$\mathbf{B}\mathbf{u}_j = \alpha_j\mathbf{W}\mathbf{u}_j \tag{13.29}$$

multiplicando (13.28) por $\mathbf{u}_j'$ y (13.29) por $\mathbf{u}_h'$:

$$\begin{aligned}\mathbf{u}_j'\mathbf{B}\mathbf{u}_h &= \alpha_h\mathbf{u}_j'\mathbf{W}\mathbf{u}_h \\ \mathbf{u}_h'\mathbf{B}\mathbf{u}_j &= \alpha_j\mathbf{u}_h'\mathbf{W}\mathbf{u}_j\end{aligned}$$

Como los primeros miembros son iguales, los segundos deben de serlo y al ser $\alpha_h \neq \alpha_j$, forzosamente $\mathbf{u}_j'\mathbf{W}\mathbf{u}_h = 0 = \mathbf{u}_j'\mathbf{B}\mathbf{u}_h = \mathbf{u}_j'\mathbf{T}\mathbf{u}_h$.

Observemos que los vectores propios de la matriz $\mathbf{W}^{-1}\mathbf{B}$ *no* serán, en general, ortogonales ya que aunque las matrices $\mathbf{W}^{-1}$ y $\mathbf{B}$ son simétricas, su producto no necesariamente lo es. Además, el rango de esta matriz, $\mathbf{W}^{-1}\mathbf{B}$, será $r = \min(p, G-1)$, (recordemos que el rango del producto de dos matrices es menor o igual que el de las originales) y éste es el máximo número de factores discriminantes que podemos obtener.

La matriz $\mathbf{W}^{-1}\mathbf{B}$ ha sido llamada por Rao matriz de distancias de Mahalanobis generalizada, ya que su traza es la suma de las distancias de Mahalanobis entre la media de cada grupo y la media total. En efecto, tenemos que

$$tr(\mathbf{W}^{-1}\mathbf{B}) = \mathbf{tr}\sum(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)'(\mathbf{W}/n_g)^{-1}(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)$$

13.5.3 Variables canónicas discriminantes

Este procedimiento proporciona $r = min(p, G-1)$ variables canónicas discriminantes que vienen dadas por

$$\mathbf{z} = \mathbf{U}_r'\mathbf{x} \tag{13.30}$$

donde $\mathbf{U}_r$ es una matriz $p \times r$ que contiene en columnas los vectores propios de $\mathbf{W}^{-1}\mathbf{B}$ y $\mathbf{x}$ un vector $p \times 1$. El vector $r \times 1$, $\mathbf{z}$, recoge los valores de las variables canónicas para el elemento $\mathbf{x}$, que son las coordenadas del punto en el espacio definido por las variables canónicas.

Las variables canónicas así obtenidas resuelven el problema de clasificación. En efecto, para clasificar un nuevo individuo $\mathbf{x}_0$ basta calcular sus coordenadas $\mathbf{z}_0$ con (13.30) y asignarlo al grupo de cuya media transformada esté más próxima con la distancia euclídea.

Un problema importante es investigar cuántas dimensiones necesitamos para la discriminación, ya que es posible que la mayoría de la capacidad de separación de las poblaciones se consiga con las primeras variables canónicas. Para estudiar este problema supongamos que los vectores propios de $\mathbf{W}^{-1}\mathbf{B}$ en vez de tomarlos con norma unidad, $\mathbf{u}_i$,los estandarizamos con $\mathbf{v}_i = \mathbf{u}_i / \left|\mathbf{u}_i'\mathbf{W}\mathbf{u}_i\right|^{1/2}$ de manera que estos vectores $\mathbf{v}_i$ sigan siendo vectores propios de $\mathbf{W}^{-1}\mathbf{B}$ pero ahora verifican $\mathbf{v}_i'\mathbf{W}\mathbf{v}_i = 1$. Entonces, la variabilidad explicada por la variable canónica $\mathbf{v}_i$ es, por (13.26),

$$VE(\mathbf{v}_i) = \mathbf{v}_i'\mathbf{B}\mathbf{v}_i$$

pero al ser $\mathbf{v}_i$ un vector propio de $\mathbf{W}^{-1}\mathbf{B}$ verifica

$$\mathbf{B}\mathbf{v}_i = \alpha_i\mathbf{W}\mathbf{v}_i$$

y multiplicando por $\mathbf{v}_i'$ y teniendo en cuenta que por construcción $\mathbf{v}_i'\mathbf{W}\mathbf{v}_i = 1$:

$$VE(\mathbf{v}_i) = \mathbf{v}_i'\mathbf{B}\mathbf{v}_i = \alpha_i,$$

que indica que la variabilidad explicada por la variable canónica $\mathbf{v}_i$ es igual a su valor propio asociado. Por tanto, los valores propios de $\mathbf{W}^{-1}\mathbf{B}$ estandarizados para que $\mathbf{v}_i'\mathbf{W}\mathbf{v}_i = 1$ nos

indican la variabilidad explicada que cada variable canónica aporta al problema de discriminación. Cuando p y G son grandes es frecuente que la mayor capacidad de discriminación se consiga con unas pocas variables canónicas.

Los resultados de clasificación que se obtienen con las variables canónicas son idénticos a los obtenidos con la distancia de Mahalanobis (véase Hernández y Velilla, 2001, para un estudio completo de este problema). Esto es inmediato de comprobar si $G = 2$, caso de dos poblaciones, o cuando las medias sean colineales. En ambos casos la matriz $\mathbf{B}$ tiene rango uno y el vector propio de $\mathbf{W}^{-1}\mathbf{B}$ unido al valor propio no nulo proporciona automáticamente la función lineal discriminante de Fisher. Para comprobarlo basta notar que si $G = 2$ la matriz $\mathbf{B}$ es:

$$\mathbf{B} = \frac{n_1 n_2}{n_1 + n_2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'$$

y el vector propio asociado al valor propio no nulo de $\mathbf{W}^{-1}\mathbf{B}$ es $\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$, que ya obtuvimos anteriormente. Si las medias de las G poblaciones están en una línea recta, entonces:

$$(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_T) = p_1(\overline{\mathbf{x}}_2 - \overline{\mathbf{x}}_T) = ... = k_j(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T) = c(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$$

y la matriz $\mathbf{B}$ puede escribirse

$$\mathbf{B} = \sum_{g=1}^{G}(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_g - \overline{\mathbf{x}}_T)' = k^*(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'$$

y su vector propio asociado al valor propio no nulo de $\mathbf{W}^{-1}\mathbf{B}$ es proporcional a $\mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$.

Ejemplo 13.5 *Vamos a estudiar la discriminación geográfica entre los países del mundo del banco de datos MUNDODES. Los 91 países incluidos se han clasificado a priori como del este de Europa (9 países, clave 1), América central y del sur (12 países, clave 2), Europa Occidental mas Canadá y EEUU (18 países, clave 3), Asia (25 países, clave 4) y Africa (27 países, clave 5). La variable PNB se ha expresado en logaritmos neperianos, de acuerdo con los resultados descriptivos que obtuvimos en el capítulo 3.*

Se presenta la salida del programa SPSS para la discriminación múltiple que proporciona los resultados del análisis discriminante utilizando las variables canónicas

Las medias de los cinco grupos en cada variable son:

G TN TM MI EH EM LPNB
1 15.15 10.52 18.14 67.35 74.94 7.48
2 29.17 9.416 51.32 62.70 68.53 7.25
3 13.01 9.583 8.044 71.25 77.97 9.73
4 30.31 8.072 56.488 63.08 65.86 7.46
5 44.52 14.622 99.792 50.63 54.14 6.19
Total 29.46 10.734 55.281 61.38 66.03 7.51

y la columna total indica las medias para el conjunto de los datos.

Las desviaciones típicas en los grupos son:

G TN TM MI EH EM LPNB

1 3.97 2.16 6.97 2.22 1.50 .48
2 7.38 5.51 31.69 4.92 5.31 .66
3 1.85 1.37 1.734 2.51 2.18 .44
4 10.01 3.77 46.02 7.92 9.73 1.69
5 5.685 4.79 30.58 7.09 7.03 1.04
Total 13.69 4.68 46.30 9.72 11.13 1.64
y la matriz W con 86 grados de libertad es
TN TM MI EH EM LPNB
TN 46.90
TM 11.89 15.63
MI 139.41 87.49 1007.64
EH -27.42 -18.76 -169.71 37.55
EM -31.88 -21.08 -194.29 40.52 46.20
LPNB -3.54 -2.18 -22.42 4.60 5.43 1.25
que podemos expresar como matriz de correlaciones:
TN TM MI EH EM LPNB
TN 1.00000
TM .43930 1.00000
MI .64128 .69719 1.00000
EH -.65345 -.77451 -.87247 1.00000
EM -.68487 -.78452 -.90052 .97278 1.00000
LPNB -.46341 -.49350 -.63275 .67245 .71588 1.00000
Las funciones de clasificación lineales para cada grupo son:
G = 1 2 3 4 5
TN 3.4340 3.7363 3.3751 3.6194 3.9314
TM 9.7586 9.1856 9.6773 8.6879 8.9848
MI 1.7345 1.7511 1.7387 1.7107 1.6772
EH -.1319 .28153 .7638 1.7363 .59934
EM 16.962 16.3425 15.780 14.347 15.342
LPNB -9.422 -8.2661 -5.999 -6.703 -7.053
(Constante) -690.658 -683.135 -690.227 -642.071 -647.1495
y los valores propios de $W^{-1}B$ y la proporción de variacion explicada son:
Fcn Val. pr. Var. %
1 3.9309 69.33 69.33*
2 1.1706 20.65 89.97*
3 .4885 8.62 98.59*
4 .0802 1.41 100.00*

Se observa que la primera función discriminante o variable canónica, definida por el primer vector propio de la matriz $\mathbf{W}^{-1}\mathbf{B}$ explica el 69% de la variabilidad y que las dos primeras explican conjuntamente el 89,97%.

Los coeficientes de las variables canónicas indican que las variable más importantes son globalmente la esperanza de vida de la mujer y la tasa de natalidad.

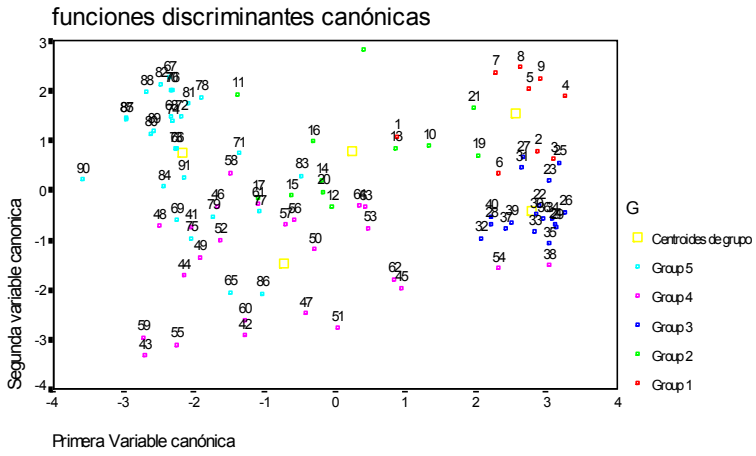


Figura 13.5: Gráfico de proyección de los puntos sobre las dos primeras variables canónicas.

En el gráfico pueden verse las proyecciones de los paises sobre las dos primeras variables canónicas. Los resultados de la clasificación con las 4 variables canónicas se resumen en siguiente tabla, donde r representa la clasificación real y p la previsa por el modelo.

	$p1$	$p2$	$p3$	$p4$	$p5$
$r1$	8	1			
$r2$	1	9	1		1
$r3$			18		
$r4$		2	2	19	2
$r5$		1		4	22

Se observa que se clasifican bien los paises europeos y entre los asiáticos es donde aparece más variabilidad.

En este caso los errores aparentes obtenidos clasificando con la distancia de Mahalanobis (sin validación cruzada) y con las variables canónicas son los mismos. La salida adjunta de MINITAB incluye la información básica. En primer lugar se presenta el resultado de clasificar con las funciones discriminantes:

True Group....
Group 1 2 3 4 5
1 8 1 0 0 0
2 1 9 0 2 1
3 0 1 18 2 0
4 0 0 0 19 4
5 0 1 0 2 22
N Total 9 12 18 25 27
N Correct 8 9 18 19 22
Propor. 0.89 0.75 1.0 0.76 0.82
N = 91 N Correct = 76 Proportion Correct = 0.835
y a continuación el resultado aplicando validación cruzada:
Colocado Verdadero grupo

Group 1 2 3 4 5
1 8 1 1 0 0
2 1 8 0 4 1
3 0 1 17 2 0
4 0 1 0 16 5
5 0 1 0 3 21
N. Total 9 12 18 25 27
N Correct 8 8 17 16 21
Propor. 0.89 0.67 0.94 0.64 0.78
N = 91 N Correct = 70 Propor. Correct = 0.769
funciones lineales discriminantes
1 2 3 4 5
Con. -689.05 -681.53 -688.62 -640.46 -645.54
C2 3.43 3.74 3.38 3.62 3.93
C3 9.76 9.19 9.68 8.69 8.98
C4 1.73 1.75 1.74 1.71 1.68
C5 -0.13 0.28 0.76 1.74 0.60
C6 16.96 16.34 15.78 14.35 15.34
C9 -9.42 -8.27 -6.00 -6.70 -7.05
La tabla siguiente resume los resultados con validación cruzada.

	$p1$	$p2$	$p3$	$p4$	$p5$
$r1$	8	1			
$r2$	1	8	1	1	1
$r3$	1		17		
$r4$		4	2	16	3
$r5$		1		5	21

Finalmente la matriz de distancias entre las medias de los grupos con la distancia de Mahalanobis es

	$EO(1)$	$AL(2)$	$E(3)$	$AS(4)$	$AF(5)$
$EO(1)$		7.2	7.8	20.3	25.2
$AL(2)$			10.9	6.5	7.6
$E(3)$				15.4	30.0
$AS(4)$					1.9
$AF(5)$					

Se observa que la mayor distancia aparece entre el grupo E (que incluye los paises de europa occidental más Canadá y EEUU) y Africa. La segunda es entre EO y Africa. La distancia menor es entre Asia y Africa.

13.6 DISCRIMINACIÓN CUADRÁTICA. DISCRIMINACIÓN DE POBLACIONES NO NORMALES

Si admitiendo la normalidad de las observaciones la hipótesis de igualdad de varianzas no fuese admisible, el procedimiento de resolver el problema es clasificar la observación en el

grupo con máxima probabilidades a posteriori. Esto equivale a clasificar la observación $\mathbf{x}_0$ en la grupo donde se minimice la función :

$$\min_{j\in(1,...,G)}\left[\frac{1}{2}\log|\mathbf{V}_j|+\frac{1}{2}(\mathbf{x}_0-\boldsymbol{\mu}_j)'\mathbf{V}_j^{-1}(\mathbf{x}_0-\boldsymbol{\mu}_j)-\ln(C_j\pi_j)\right]$$

Cuando $\mathbf{V}_j$ y $\boldsymbol{\mu}_j$ son desconocidos se estiman por $\mathbf{S}_j$ y $\overline{\mathbf{x}}_j$ de la forma habitual. Ahora el término $\mathbf{x}_0{}'\mathbf{V}_j^{-1}\mathbf{x}_0$ no puede anularse, al depender del grupo, y las funciones discriminantes no son lineales y tendrán un término de segundo grado. Suponiendo que los costes de clasificación son iguales en todos los grupos, clasificaremos nuevas observaciones con la regla :

$$\min_{j\in(1,...,G)}\left[\frac{1}{2}\log|\widehat{\mathbf{V}}_j|+\frac{1}{2}(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_j)'\widehat{\mathbf{V}}_j^{-1}(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_j)-\ln\pi_j\right]$$

En el caso particular de dos poblaciones y suponiendo las mismas probabilidades a priori clasificaremos una nueva observación en la población 2 si

$$\log|\widehat{\mathbf{V}}_1|+(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_1)'\widehat{\mathbf{V}}_1^{-1}(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_1)>\log|\widehat{\mathbf{V}}_2|+(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_2)'\widehat{\mathbf{V}}_2^{-1}(\mathbf{x}_0-\widehat{\boldsymbol{\mu}}_2)$$

que equivale a

$$\mathbf{x}_0'(\widehat{\mathbf{V}}_1^{-1}-\widehat{\mathbf{V}}_2^{-1})\mathbf{x}_0-2\mathbf{x}_0'(\widehat{\mathbf{V}}_1^{-1}\widehat{\boldsymbol{\mu}}_1-\widehat{\mathbf{V}}_2^{-1}\widehat{\boldsymbol{\mu}}_2)>c \tag{13.31}$$

donde $c=\log(|\widehat{\mathbf{V}}_2|/|\widehat{\mathbf{V}}_1|)+\widehat{\boldsymbol{\mu}}_2{}'\widehat{\mathbf{V}}_2^{-1}\widehat{\boldsymbol{\mu}}_2-\widehat{\boldsymbol{\mu}}_1{}'\widehat{\mathbf{V}}_1^{-1}\widehat{\boldsymbol{\mu}}_1$. Llamando

$$\widehat{\mathbf{V}}_d^{-1}=(\widehat{\mathbf{V}}_1^{-1}-\widehat{\mathbf{V}}_2^{-1})$$

y

$$\widehat{\boldsymbol{\mu}}_d=\widehat{\mathbf{V}}_d(\widehat{\mathbf{V}}_1^{-1}\widehat{\boldsymbol{\mu}}_1-\widehat{\mathbf{V}}_2^{-1}\widehat{\boldsymbol{\mu}}_2)$$

y definiendo las nuevas variables

$$\mathbf{z}_0=\widehat{\mathbf{V}}_d^{-1/2}\mathbf{x}_0$$

y llamando $\mathbf{z}_0=(z_{01},...,z_{0p)'}$ y definiendo el vector $\mathbf{m}=(m_1,...,m_p)'=\widehat{\mathbf{V}}_d^{1/2}(\widehat{\mathbf{V}}_1^{-1}\widehat{\boldsymbol{\mu}}_1-\widehat{\mathbf{V}}_2^{-1}\widehat{\boldsymbol{\mu}}_2)$, la ecuación (13.31) puede escribirse

$$\sum_{i=1}^{p}z_{0i}^2-2\sum_{i=1}^{p}z_{0i}m_i>c$$

Esta es una ecuación de segundo grado en las nuevas variables z_{0i}. Las regiones resultantes con estas funciones de segundo grado son típicamente disjuntas y a veces difíciles de interpretar en varias dimensiones. Por ejemplo, la figura (13.6) muestra un ejemplo unidimensional del tipo de regiones que se obtienen con la discriminación cuadrática.

Figura 13.6: Ejemplo de discriminación cuadrática. La zona de clasificación de P1 es la zona central y la de P2 la de las colas

El número de parámetros a estimar en el caso cuadrático es mucho mayor que en el caso lineal. En el caso lineal hay que estimar $Gp + p(p+1)/2$ y en el caso cuadrático $G(p+p(p+1)/2)$. Por ejemplo con 10 variables y 4 grupos pasamos de estimar 95 parámetros en el caso lineal a 260 en el caso cuadrático. Este gran número de parámetros hace que, salvo en el caso en que tenemos muestras muy grandes, la discriminación cuadráticas sea bastante inestable y, aunque las matrices de covarianzas sean muy diferentes, se obtengan con frecuencia mejores resultados con la función lineal que con la cuadrática. Un problema adicional con la función discriminante cuadrática es que es muy sensible a desviaciones de la normalidad de los datos. La evidencia disponible indica que la clasificación lineal es en estos casos más robusta. Recomendamos siempre calcular los errores de clasificación con ambas reglas utilizando validación cruzada y en caso de que las diferencias sean muy pequeñas quedarse con la lineal.

Aparece también un problema de discriminación cuadrática en el análisis de determinadas poblaciones no normales. (Véase Lachenbruch (1975)). En el caso general de poblaciones arbitrarias tenemos dos alternativas: (a) aplicar la teoría general expuesta en 13.2 y obtener la función discriminante que puede ser complicada, b) aplicar la teoría de poblaciones normales, tomar como medida de distancia la distancia de Mahalanobis y clasificar $\mathbf{x}$ en la población P_j para la cual la $\mathbf{D}^2$:

$$\mathbf{D}^2 = (\mathbf{x} - \overline{\mathbf{x}}_j)' \widehat{\mathbf{V}}_j^{-1} (\mathbf{x} - \overline{\mathbf{x}}_j)$$

es mínima.

Para poblaciones discretas estas aproximaciones no son buenas. Se han propuesto métodos alternativos basados en la distribución multinomial o en la distancia χ^2 cuya eficacia está aun por determinarse.

Ejemplo 13.6 *Si aplicamos la discriminación cuadrática a los datos de las medidas físicas*

se obtiene la tabla de errores de clasificación por validación cruzada (sin aplicar validación cruzada se acierta el 100% como en el caso lineal)

		Clasificado	
		M	H
Realidad	M	11	4
	H	5	7

que supone un porcentaje de aciertos del 67%, menor que en el caso lineal. No hay evidencia de que la discriminación cuadrática suponga ninguna ventaja en este caso.

13.7 DISCRIMINACIÓN BAYESIANA

Hemos visto en la sección 13.2 que el enfoque Bayesiano permite dar una solución general del problema de clasificación cuando los parámetros son conocidos. Cuando los parámetros deben estimarse a partir de los datos, el enfoque Bayesiano aporta también una solución directa del problema que tiene en cuenta la incertidumbre en la estimación de los parámetros, a diferencia del enfoque clásico que ignora esta incertidumbre. La solución es válida sean o no iguales las matrices de covarianza. El procedimimiento para clasificar una observación, $\mathbf{x}_0$, dada la muestra de entrenamiento $\mathbf{X}$, es asignarla a la población más probable. Para ello se obtiene el máximo de las probabilidades a posteriori de que la observación a clasificar, $\mathbf{x}_0$, venga de cada una de las poblaciones dada la muestra $\mathbf{X}$. Estas probabilidades se calculan por

$$P(i/\mathbf{x}_0, \mathbf{X}) = \frac{f_i(\mathbf{x}_0|\mathbf{X})\boldsymbol{\pi}_i}{\sum_{g=1}^{G} f_g(\mathbf{x}|\mathbf{X})\boldsymbol{\pi}_j}$$

donde las densidades $f_g(\mathbf{x}_0|\mathbf{X})$, que se denominan predictivas a posteriori o simplemente predictivas, son proporcional a las probabilidades de que la observación $\mathbf{x}_0$ se genere por la población g. Estas densidades se obtienen a partir de la verosimilitud promediando sobre los posibles valores de los parámetros en cada población con su distribución a posteriori:

$$f_g(\mathbf{x_0}|\mathbf{X}) = \int f(\mathbf{x_0}|\boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g|\mathbf{X})\mathbf{d}\boldsymbol{\theta}_g \tag{13.32}$$

donde $\boldsymbol{\theta}_g$ son los parámetros de la población g.

Vamos a estudiar como obtener estas probabilidades. En primer lugar, la distribución a posteriori de los parámetros se calcula de la forma habitual mediante

$$p(\boldsymbol{\theta}_g|\mathbf{X}) = kf(\mathbf{X}|\boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g).$$

Como vimos en la sección 9.2.2 la verosimilitud para la población g con n_g elementos muestrales y media muestra $\overline{\mathbf{x}}_g$ y varianza S_g es

$$f(\mathbf{X}|\boldsymbol{\theta}_g) = k\left|\mathbf{V}_g^{-1}\right|^{n_g/2}\exp(-\frac{n_g}{2}tr(\mathbf{V}_g^{-1}\left\{S_g + (\overline{\mathbf{x}}_g - \boldsymbol{\mu}_g)(\overline{\mathbf{x}}_g - \boldsymbol{\mu}_g)'\right\})$$

y con la prior de referencia

$$p(\boldsymbol{\mu}_g, \mathbf{V}_g^{-1}) = k \left|\mathbf{V}_g^{-1}\right|^{-(p+1)/2}$$

se obtiene la posterior

$$p(\boldsymbol{\mu}_g, \mathbf{V}_g^{-1}/\mathbf{X}) = k \left|\mathbf{V}_g^{-1}\right|^{(n_g-p-1)/2} \exp(-\frac{n_g}{2} tr(\mathbf{V}_g^{-1} \left\{S_g + (\overline{\mathbf{x}}_g - \boldsymbol{\mu}_g)(\overline{\mathbf{x}}_g - \boldsymbol{\mu}_g)'\right\}$$

La distribución predictiva se obtendrá con (13.32), donde ahora $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_g, \mathbf{V}_g^{-1})$. Integrando respecto a estos parámetros puede obtenerse, (véase Press, 1989, para los detalles de la integración) que la distribución predictiva es t multivariante

$$p(\mathbf{x}_0/\mathbf{X}, \mathbf{g}) = \left[\frac{n\pi(n_g+1)}{n_g}\right]^{-p/2} \frac{\Gamma\left(\frac{n_g}{2}\right)}{\Gamma\left(\frac{n_g-p}{2}\right)} |\mathbf{S}_g|^{-1/2} \left[1 + \frac{1}{n_g+1}(\mathbf{x}_0 - \overline{\mathbf{x}}_g)'\mathbf{S}_g^{-1}(\mathbf{x}_0 - \overline{\mathbf{x}}_g)\right]^{-n_g/2}$$

Con esta distribución podemos calcular las probabilidades a posteriori para cada población. Alternativamente, para decidir entre la población i y la j podemos calcular el ratio de las probabilidades a posteriori, dado por :

$$\frac{P(i|\mathbf{x})}{P(j|\mathbf{x})} = c_{ij}\frac{\pi_i}{\pi_j} \cdot \frac{|\mathbf{S}_j|^{1/2}}{|\mathbf{S}_i|^{1/2}} \frac{\left(1 + \frac{1}{n_j+1}(\mathbf{x}_0 - \overline{\mathbf{x}}_j)'\mathbf{S}_j^{-1}(\mathbf{x}_0 - \overline{\mathbf{x}}_j)\right)^{n_j/2}}{\left(1 + \frac{1}{n_i+1}(\mathbf{x}_0 - \overline{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}_0 - \overline{\mathbf{x}}_i)\right)^{n_i/2}}$$

donde π_i son las probabilidades a priori, $\mathbf{S}_j$ las matrices de varianza y covarianzas estimadas, y

$$c_{ij} = \left[\frac{n_i(n_j+1)}{n_j(n_i+1)}\right]^{p/2} \frac{\Gamma\left(\frac{n_i}{2}\right)\Gamma\left(\frac{n_j-p}{2}\right)}{\Gamma\left(\frac{n_j}{2}\right)\Gamma\left(\frac{n_i-p}{2}\right)}$$

Si los tamaños muestrales son aproximadamente iguales, $n_i \simeq n_j$, entonces $c_{ij} \simeq 1$. El clasificador óptimo es cuadrático. Si suponemos que las matrices de covarianza de los grupos son iguales de nuevo obtenemos la función lineal discriminante (véase Aitchinson y Dunsmore, 1975).

13.8 Lecturas complementarias

El análisis discriminante clásico aquí presentado se estudia en todos los libros de análisis multivariante. Presentaciones a un nivel similar al aquí expuesto se encuentran en Cuadras (1991), Flury (1997), Johnson and Wichern (1998), Mardia et al (1979), Rechner (1998) y Seber (1984). Un texto básico muy detallado y con muchas extensiones y regerencias se encuentra en McLachlan (1992). Lachenbruch, (1975) contiene muchas referencias históricas. Enfoques más aplicados centrados en al análisis de ejemplos y salidas de ordenador se persenta en Huberty (1994), Hair el al (1999) y Tabachnick y Fidell (1996). Un enfoque bayesiano al problema de la clasificación puede consultarse en Press (1989).

Ejercicios

13.1 Suponga que se desea discriminar entre dos poblaciones normales con vectores de medias (0,0) y (1,1) , varianzas (2,4) y coeficiente de correlación lineal r=.8. Construir la función lineal discriminante e interpretarla.

13.2 Discutir cómo varían las probabilidades de error en el problema anterior como función del coeficiente de correlación. ¿Ayuda la correlación a la discriminación?

13.3 Las probabilidades a priori en el problema 13.1 son 0.7 para la primera población y 0,3 para la segunda. Calcular la función lineal discriminante en este caso.

13.4 Se desea discriminar entre tres poblaciones normales con vectores de medias (0,0), (1,1) y (0,1) con varianzas (2,4) y coeficiente de correlación lineal r =.5. Calcular y dibujar las funciones discriminantes y hallar su punto de corte.

13.5 Si los costes de equivocarnos en el problema anterior no son los mismos, de manera que el coste de clasificar en la tercera población cuando viene de la primera es el doble de los demás, calcular las funciones discriminantes.

13.6 Justifique que los valores propios de $W^{-1}B$ son positivos, demostrando que esta matriz tiene los mismos valores propios que la matriz $W^{-1/2}BW^{-1/2}$.

13.7 Justificar que se obtienen las mismas variables canónicas discriminantes utilizando las matrices W y B, que las matrices asociadas de varianzas corregidas por grados de libertad.

13.8 Demostrar que es lo mismo obtener el mayor vector propio de $W^{-1}B$ y el menor de $T^{-1}W$.

13.9 Demostrar que el primer componente principal cuando hay dos grupos viene dado por $\mathbf{v} = c(\mathbf{W} - \boldsymbol{\lambda}\mathbf{I})^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ (sugerencia: Si $\mathbf{T} = \mathbf{W} + \mathbf{B}$, el primer componente es el mayor autovector (ligado al mayor autovalor) de $\mathbf{T}$ y verifica $\mathbf{T}\,\mathbf{v} = \mathbf{W}\mathbf{v} + \mathbf{B}\mathbf{v} = \lambda\mathbf{v}$. Como $\mathbf{B} = \mathbf{k}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'$, tenemos que $\mathbf{W}\mathbf{v} + c(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) = \lambda\mathbf{v}$).

13.10 Demostrar que si $(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$ es un vector propio de $\mathbf{W}^{-1}$ la dirección discriminante es el eje natural de distancia entre las medias y coincide con el primer componente principal.

13.11 Demostrar que la distancia de Mahalanobis es invariante a transformaciones lineales comprobando que si $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, con $\mathbf{A}$ cuadrada y no singular, se verifica que $D^2(y_i, y_j) = D^2(x_i, x_j)$. (Sugerencia: utilizar que $\mathbf{V}_y = \mathbf{A}\mathbf{V}_\mathbf{x}\mathbf{A}'$ y $\mathbf{V}_\mathbf{y}^{-\mathbf{1}} = (\mathbf{A}')^{-1}\mathbf{V}_\mathbf{x}^{-\mathbf{1}}\mathbf{A}'$)

APÉNDICE 13.1:EL CRITERIO MINIMIZAR LA PROBABILIDAD DEL ERROR

El criterio de minimizar la probabilidad de error puede escribirse como minimizar P_T, donde:

$$P_T(error) = P(1|\mathbf{x} \in 2) + P(2|\mathbf{x} \in 1)$$

siendo $P(i|\mathbf{x} \in j)$ la probabilidad de clasificar en la población i una observación que proviene de la j. Esta probabilidad viene dada por el área encerrada por la distribución j en la zona de clasificación de i, es decir:

$$P(i|\mathbf{x} \in j) = \int_{A_i} f_j(\mathbf{x})d\mathbf{x}$$

por tanto:

$$P_T = \int_{A_1} f_2(\mathbf{x})d\mathbf{x} + \int_{A_2} f_1(\mathbf{x})d\mathbf{x}$$

y como A_1 y A_2 son complementarios:

$$\int_{A_1} f_2(\mathbf{x})d\mathbf{x} = \mathbf{1} - \int_{A_2} f_2(\mathbf{x})d\mathbf{x}$$

que conduce a:

$$P_T = 1 - \int_{A_2} (f_2(\mathbf{x}) - f_1(\mathbf{x}))d\mathbf{x}$$

y para minimizar la probabilidad de error debemos maximizar la integral. Esto se consigue definiendo A_2 como el conjunto de puntos donde el integrando es positivo, es decir:

$$A_2 = \{\mathbf{x} | f_2(\mathbf{x}) > f_1(\mathbf{x})\}$$

y obtenemos de nuevo el criterio antes establecido.

APÉNDICE 13.2: DISCRIMINACIÓN Y REGRESIÓN

Un resultado interesante es que la construcción de una función discriminante en el caso de dos poblaciones, puede abordarse como un problema de regresión.

Consideremos las n observaciones como datos en un modelo lineal y definamos unas variable respuesta y que toma el valor $+k_1$, cuando $\mathbf{x} \in P_1$ y $-k_2$, cuando $\mathbf{x} \in P_2$. Podemos asignar a k_1 y k_2 valores cualesquiera, aunque, como veremos, los cálculos se simplifican si hacemos estas constantes iguales al número de elementos de la muestra en cada clase. El modelo será:

$$y_i = \beta_1(x_{1i} - \overline{x}_1) + \beta_2(x_{2i} - \overline{x}_2) + ... + \beta_p(x_{pi} - \overline{x}_p) + u_i \qquad i = 1, 2 \tag{13.33}$$

donde hemos expresado las x en desviaciones. El estimador de mínimos cuadrados es:

$$\widehat{\beta} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{Y} \tag{13.34}$$

donde $\widetilde{\mathbf{X}}$ es la matriz de los datos en desviaciones.

Sea $\overline{\mathbf{x}}_1$ el vector de medias en el primer grupo, $\overline{\mathbf{x}}_2$ en el segundo y $\overline{\mathbf{x}}_T$ el correspondiente a todas las observaciones. Supongamos que en la muestra hay n_1 datos del primer grupo y n_2 del segundo. Entonces,

$$\overline{\mathbf{x}}_T = \frac{n_1\overline{\mathbf{x}}_1 + n_2\overline{\mathbf{x}}_2}{n_1 + n_2}. \tag{13.35}$$

Sustituyendo (13.35) en el primer término de (13.34):

$$\begin{aligned} \widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} &= \sum_{i=1}^{n_1+n_2} (\mathbf{x}_i - \overline{\mathbf{x}}_T)(\mathbf{x}_i - \overline{\mathbf{x}}_T)' = \\ &= \sum_{i=1}^{n_1} (\mathbf{x}_i - \overline{\mathbf{x}}_T)(\mathbf{x}_i - \overline{\mathbf{x}}_T)' + \sum_{i=1+n_1}^{n_1+n_2} (\mathbf{x}_i - \overline{\mathbf{x}}_T)(\mathbf{x}_i - \overline{\mathbf{x}}_T)'. \end{aligned}$$

Como

$$\begin{aligned}\sum_{i=1}^{n_1}(\mathbf{x}_i-\overline{\mathbf{x}}_T)(\mathbf{x}_i-\overline{\mathbf{x}}_T)' &= \sum_{i=1}^{n_1}(\mathbf{x}_i-\overline{\mathbf{x}}_1+\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_T)(\mathbf{x}_i-\overline{\mathbf{x}}_1+\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_T)'= \\ &= \sum_{i=1}^{n_1}(\mathbf{x}_i-\overline{\mathbf{x}}_1)(\mathbf{x}_i-\overline{\mathbf{x}}_1)'+n_1(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_T)'\end{aligned}$$

ya que los términos cruzados se anulan al ser $\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}}_1)=0$. Procediendo análogamente para el otro grupo, podemos escribir:

$$\begin{aligned}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} &= \sum_{i=1}^{n_1}(\mathbf{x}_i-\overline{\mathbf{x}}_1)(\mathbf{x}_i-\overline{\mathbf{x}}_1)'+\sum_{i=n_1+1}^{n_1+n_2}(\mathbf{x}_i-\overline{\mathbf{x}}_2)(\mathbf{x}_i-\overline{\mathbf{x}}_2)' \\ &\quad +n_1(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_T)'+n_2(\overline{\mathbf{x}}_2-\overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_2-\overline{\mathbf{x}}_T)'\end{aligned} \tag{13.36}$$

Los primeros dos términos conducen a la matriz $\mathbf{W}$ de sumas de cuadrados dentro de grupos que, como hemos visto, estima $\mathbf{V}$, mediante:

$$\widehat{\mathbf{V}}=\mathbf{S}=\frac{1}{n_1+n_2-2}\mathbf{W}. \tag{13.37}$$

Los segundos dos términos son las sumas de cuadrados entre grupos. Sustituyendo $\overline{\mathbf{x}}_T$ por (13.35):

$$\overline{\mathbf{x}}_1-\frac{n_1\overline{\mathbf{x}}_1+n_2\overline{\mathbf{x}}_2}{n_1+n_2}=\frac{1}{n_1+n_2}n_2(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2), \tag{13.38}$$

con lo que resulta:

$$(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_T)'=\left(\frac{n_2}{n_1+n_2}\right)^2(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2)' \tag{13.39}$$

$$(\overline{\mathbf{x}}_2-\overline{\mathbf{x}}_T)(\overline{\mathbf{x}}_2-\overline{\mathbf{x}}_T)'=\left(\frac{n_1}{n_1+n_2}\right)^2(\overline{\mathbf{x}}_2-\overline{\mathbf{x}}_1)(\overline{\mathbf{x}}_2-\overline{\mathbf{x}}_1)' \tag{13.40}$$

Sustituyendo (13.39) y (13.40) en (13.36) obtenemos que:

$$\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}=(n_1+n_2-2)\mathbf{S}+\frac{n_1n_2}{n_1+n_2}(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2)'$$

que implica

$$\left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1}=(n_1+n_2-2)^{-1}\mathbf{S}^{-1}+a\mathbf{S}^{-1}(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2)(\overline{\mathbf{x}}_1-\overline{\mathbf{x}}_2)'\mathbf{S}^{-1} \tag{13.41}$$

donde a es una constante. Por otro lado:

$$\widetilde{\mathbf{X}}'\mathbf{Y}=\sum_{i=1}^{n_1+n_2}y_i(\mathbf{x}_i-\overline{\mathbf{x}}_T)=k_1\sum_{i=1}^{n_1}(\mathbf{x}_i-\overline{\mathbf{x}}_T)-\ k_2\sum_{i=1}^{n_2}(\mathbf{x}_i-\overline{\mathbf{x}}_T)$$

sustituyendo $\overline{\mathbf{x}}_T$ por su expresión (13.35), resulta, por (13.38):

$$\widetilde{\mathbf{X}}'\mathbf{Y} = \frac{k_1 n_1 n_2}{n_1+n_2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) + \frac{k_2 n_1 n_2}{n_1+n_2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)k_T \tag{13.42}$$

siendo $k_T = n_1 n_2(k_1+k_2)/(n_1+n_2)$. Sustituyendo (13.41) y (13.42) en la fórmula (13.34) se obtiene que:

$$\widehat{\beta} = k.\mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$$

que es la expresión de la función discriminante clásica.

Capítulo 14

DISCRIMINACIÓN LOGÍSTICA Y OTROS MÉTODOS DE CLASIFICACIÓN

14.1 INTRODUCCIÓN

El problema de discriminación o clasificación cuando conocemos los parámetros de las distribuciones admite una solución general que hemos visto en el capítulo anterior. Sin embargo, en la mayoría de las aplicaciones los parámetros son desconocidos y deben estimarse a partir de los datos. Si la distribución conjunta de las observaciones es normal multivariante, utilizar las distancias de Mahalanobis estimadas suele dar buenos resultados y será óptimo con muestras grandes. Sin embargo, es frecuente que los datos disponibles para la clasificación no sean normales. Por ejemplo, en muchos problemas de clasificación se utilizan variables discretas. En estos casos no tenemos garantías de que los métodos estudiados en el capítulo 13 sean óptimos.

En este capítulo presentamos otros métodos de clasificación. Una posibilidad es intentar construir un modelo que explique los valores de la variable de clasificación. Por ejemplo, si se desea discriminar entre créditos que se devuelven o que presentan problemas para su cobro, puede añadirse a la base de datos una nueva variable, y, que tome el valor cero cuando el crédito se devuelve sin problemas, y el valor uno en otro caso. El problema de discriminación se convierte en prever el valor de la variable ficticia, y, en un nuevo elemento del que conocemos el vector de variables $\mathbf{x}$. Si el valor previsto está más próximo a cero que a uno, clasificaremos al elemento en la primera población. En otro caso, lo haremos en la segunda. Para modelar este tipo de relaciones se utilizan los modelos de respuesta cualitativa, que se revisan en la sección siguiente. Dentro de esta clase el modelo más utilizado es el modelo logístico, que se estudia con cierto detalle en las secciones siguientes.

Además del modelo logístico, presentamos brevemente en este capítulo otros métodos de discriminación, que pueden verse como procedimientos generales de aproximar la función de clasificación en casos complejos no lineales y que requiren el uso intensivo del ordenador. El primero de estos métodos es el de los árboles de clasificación, CART, que es un algoritmo para llevar a la práctica una idea simple pero efectiva, especialmente cuando muchas de

las variables de clasificación son binarias. El segundo es el de las redes neuronales, que son aproximaciones universales de funciones y, convenientemente construidas, pueden dar buenos resultados en casos no lineales. El tercero es los métodos no paramétricos, que utilizan aproximaciones locales. El cuarto ha sido propuesto recientemente por Vapnik (2000) y constituye una filosofía alternativa a las redes neuronales para aproximar funciones generales. La eficacia de estos procedimientos es todavía objeto de investigación.

14.2 EL MODELO LOGIT

14.2.1 Modelos con respuesta cualitativa

Consideremos el problema de la discriminación entre dos poblaciones. Una forma de abordar el problema es definir una variable de clasificación, y, que tome el valor cero cuando el elemento pertenece a la primera población, P_1, y uno cuando pertenece a la segunda, P_2. Entonces, la muestra consistirá en n elementos del tipo $(y_i, \mathbf{x}_i)$, donde y_i es el valor en ese elemento de la variable binaria de clasificación y $\mathbf{x}_i$ un vector de variables explicativas. A continuación, construiremos un modelo para prever el valor de la variable ficticia binaria en un nuevo elemento cuando se conocen las variables $\mathbf{x}$. El primer enfoque simple es formular el modelo de regresión:

$$y = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x} + \mathbf{u} \tag{14.1}$$

y, hemos visto en el capítulo anterior, que si estimamos los parámetros por mínimos cuadrados este procedimiento es equivalente a la función lineal discriminante de Fisher y es óptimo para clasificar si la distribución conjunta de las variables explicativas es normal multivariante, con la misma matriz de covarianzas. Sin embargo, este modelo presenta problemas de interpretación. Tomando esperanzas en (14.1) para $\mathbf{x} = \mathbf{x}_i$:

$$E[y|\mathbf{x}_i] = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i \tag{14.2}$$

Llamemos p_i a la probabilidad de que y tome el valor 1 (pertenezca a la población P_2) cuando $\mathbf{x} = \mathbf{x}_i$:

$$p_i = P(y = 1|\mathbf{x}_i) \tag{14.3}$$

la variable y es binomial y toma los valores posibles uno y cero con probabilidades p_i y $1 - p_i$. Su esperanza será :

$$E[y|\mathbf{x}_i] = p_i \times 1 + (1 - p_i) \times 0 = p_i \tag{14.4}$$

y de (14.2) y (14.4), concluimos que:

$$\boxed{p_i = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i} \tag{14.5}$$

Esta formulación tiene dos problemas principales:

1. Si estimamos el modelo lineal (14.1), la predicción $\widehat{y}_i = \widehat{p}_i$ estima, por (14.5), la probabilidad de que un individuo con características definidas por $\mathbf{x} = \mathbf{x}_i$ pertenezca a la segunda población. Sin embargo p_i debe estar entre cero y uno, y no hay ninguna garantía de que la predicción $\widehat{y}_i$ verifique esta restricción: podemos obtener probabilidades mayores que la unidad o negativas. Esto no es un problema para clasificar la observación, pero sí lo es para interpretar el resultado de la regla de clasificación.

2. Como los únicos valores posibles de y son cero y uno la perturbación u_i sólo puede tomar los valores $1 - \beta_0 + \boldsymbol{\beta}_1 \mathbf{x}_i = 1 - p_i$ y $-\beta_0 - \boldsymbol{\beta}_1' \mathbf{x}_i = -p_i$ con probabilidades p_i y $(1 - p_i)$. La esperanza de la perturbación es cero ya que:

$$E[u_i] = p_i(1 - p_i) + (1 - p_i)(-p_i) = 0$$

pero la perturbación no sigue una distribución normal. En consecuencia, los estimadores minimocuadráticos de los coeficientes del modelo (14.1) no serán eficientes. La varianza de u_i es:

$$Var(u_i) = (1 - p_i)^2 p_i + (1 - p_i) p_i^2 = (1 - p_i) p_i,$$

y las perturbaciones son heterocedásticas. Para estimar los parámetros del modelo se debería utilizar mínimos cuadrados ponderados.

A pesar de estos dos inconvenientes, este modelo simple estimado por mínimos cuadrados conduce a una buena regla de clasificación, ya que, según la interpretación de Fisher, maximiza la separación entre los grupos, sea cual sea la distribución de los datos. Sin embargo, cuando los datos no son normales, o no tienen la misma matriz de covarianzas, la clasificación mediante una ecuación de relación lineal no es necesariamente óptima.

Si queremos que el modelo construido para discriminar nos proporcione directamente la probabilidad de pertenecer a cada población, debemos transformar la variable respuesta para garantizar que la respuesta prevista esté entre cero y uno. Escribiendo:

$$p_i = F(\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i),$$

p_i estará entre cero y uno si escogemos F para que tenga esa propiedad. La clase de funciones no decrecientes acotadas entre cero y uno es la clase de las funciones de distribución, por lo que el problema se resuelve tomando como F cualquier función de distribución. Algunas posibilidades consideradas son:

(1) Tomar como F la función de distribución de una uniforme. Esto equivale a truncar el modelo de regresión, ya que entonces:

$$\begin{array}{lll} p_i = 1 & \text{si} & \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i \geq 1 \\ p_i = \beta_0 + \boldsymbol{\beta}' \mathbf{x}_i & & 0 < \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i < 1 \\ p_i = 0 & & \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i \leq 0. \end{array}$$

Esta solución no es sin embargo satisfactoria ni teóricamente (un pequeño incremento de $\mathbf{x}$ produce en los extremos un salto muy grande, cuando sería más lógico una evolución gradual), ni prácticamente: la estimación del modelo es difícil e inestable debido a la discontinuidad.

(2) Tomar como F la función de distribución logística, dada por:

$$p_i = \frac{1}{1 + e^{-\beta_0 - \boldsymbol{\beta}_1' \mathbf{x}_i}}. \tag{14.6}$$

Esta función tiene la ventaja de la continuidad. Además como:

$$1 - p_i = \frac{e^{-\beta_0 - \boldsymbol{\beta}_1' \mathbf{x}_i}}{1 + e^{-\beta_0 - \boldsymbol{\beta}_1' \mathbf{x}_i}} = \frac{1}{1 + e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}$$

resulta que:

$$g_i = \log \frac{p_i}{1 - p_i} = \beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i \tag{14.7}$$

que es un modelo lineal en esta transformación que se denomina *logit*. La variable *Logit*, g, representa en una escala logarítmica la diferencia entre las probabilidades de pertenecer a ambas poblaciones, y al ser una función lineal de las variables explicativas nos facilita la estimación y la interpretación del modelo.

(3) Tomar otra distribución, como por ejemplo escoger F igual a la distribución normal estándar. Se obtiene entonces el modelo *probit*, que es muy similar al logit, sin tener las ventajas de interpretación del modelo logístico, como veremos a continuación.

14.2.2 El modelo logit con datos normales

El modelo logit se aplica a una amplia gama de situaciones donde las variables explicativas no tienen una distribución conjunta normal multivariante. Por ejemplo, si algunas son categóricas, podemos introducirlas en el modelo logit mediante variables ficticias como se hace en el modelo de regresión estándar. Una ventaja adicional de este modelo es que si las variables son normales verifican el modelo logit. En efecto, supongamos que las variables $\mathbf{x}$ provienen de una de dos poblaciones normales multivariantes con distinta media pero la misma matriz de varianzas covarianzas. Hemos visto en el capítulo anterior (sección 12.2) que, suponiendo las probabilidades a priori de ambas poblaciones iguales:

$$p_i = P(y = 1 | \mathbf{x}_i) = \frac{f_1(\mathbf{x}_i)}{f_1(\mathbf{x}_i) + f_2(\mathbf{x}_i)} \tag{14.8}$$

y, utilizando la transformación logit, (14.7):

$$g_i = \log \frac{f_1(\mathbf{x}_i)}{f_2(\mathbf{x}_i)} = -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_1)' \mathbf{V}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_2)' \mathbf{V}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_2)$$

y simplificando

$$g_i = \frac{1}{2}\left(\boldsymbol{\mu}_2 \mathbf{V}^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \mathbf{V}^{-1} \boldsymbol{\mu}_1\right) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \mathbf{V}^{-1} \mathbf{x}_i.$$

Por tanto, g_i es una función lineal de las variables $\mathbf{x}$, que es la característica que define el modelo logit. Comparando con (14.7) la ordenada en el origen, β_0, es igual

$$\beta_0 = \frac{1}{2}\left(\boldsymbol{\mu}_2 \mathbf{V}^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \mathbf{V}^{-1} \boldsymbol{\mu}_1\right) = -\frac{1}{2} \mathbf{w}' (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

donde $\mathbf{w} = \mathbf{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, y el vector de pendientes

$$\boldsymbol{\beta}_1 = \mathbf{w}$$

Observemos que la estimación de $\widehat{\mathbf{w}}$ mediante el modelo logístico no es eficiente en el caso normal. En efecto, en lugar de estimar los $p(p+1)/2$ términos de la matriz $\widehat{\mathbf{V}}$ y los $2p$ de las medias $\overline{\mathbf{x}}_1$ y $\overline{\mathbf{x}}_2$, con el modelo logístico estimamos únicamente $p+1$ parámetros $\beta_0, \beta_1, \ldots \beta_p$. En el caso de normalidad se obtiene un mejor procedimiento con la regla de Fisher, que estima $\widehat{\mathbf{V}}, \overline{\mathbf{x}}_1$ y $\overline{\mathbf{x}}_2$, la distribución completa de las $\mathbf{x}$, mientras que el modelo logístico estima sólo los $p+1$ parámetros de la distribución de y condicionada a $\mathbf{x}$. Como:

$$f(\mathbf{x}, y) = f(y|\mathbf{x}) f(\mathbf{x})$$

perdemos información al considerar sólo la condicionada $f(y|\mathbf{x})$ —como hace el modelo logístico— en lugar de la conjunta $f(\mathbf{x}, y)$, que se utiliza en el enfoque del capítulo anterior. Efron (1975) demostró que cuando los datos son normales multivariantes y estimamos los parámetros en la muestra, la función de discriminación lineal de Fisher funciona mejor que regresión logística

En resumen, en el caso de normalidad la regla discriminante es mejor que el modelo logístico. Sin embargo, la función logística puede ser más eficaz cuando los poblaciones tengan distinta matriz de covarianzas o sean marcadamente no normales. En el campo de la concesión automática de créditos (Credit Socoring) existen numerosos estudios comparando ambos métodos. La conclusión general es que ninguno de los dos métodos supera al otro de manera uniforme y que depende de la base de datos utilizada. Rosenberg y Gleit (1994) y Hand y Henley (1997) han presentado estudios sobre este problema.

14.2.3 Interpretación del Modelo Logístico

Los parámetros del modelo son β_0, la ordenada en el origen, y $\boldsymbol{\beta}_1 = (\beta_1, ..., \beta_p)$, las pendientes. A veces se utilizan también como parámetros $\exp(\beta_0)$ y $\exp(\beta_i)$, que se denominan los odds ratios o ratios de probabilidades, e indican cuanto se modifican las probabilidades por unidad de cambio en las variables x. En efecto, de (14.7) deducimos que

$$O_i = \frac{p_i}{1-p_i} = \exp(\beta_0).\prod_{j=1}^{p} \exp(\beta_j)^{x_j}.$$

Supongamos dos elementos, i, k, con todos los valores de las variables iguales excepto la variable h y $x_{ih} = x_{jh} + 1$. El cociente de los ratios de probabilidades (odds ratio) para estas dos observaciones es:

$$\frac{O_i}{O_k} = e^{\beta_h}$$

e indica cuanto se modifica el ratio de probabilidades cuando la variable x_h aumenta una unidad. Sustituyendo $\widehat{p}_i = .5$ en el modelo logit, entonces,

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = 0,$$

es decir,

$$x_{i1} = -\frac{\beta_0}{\beta_1} - \sum_{j=2}^{p} \frac{\beta_j x_{ij}}{\beta_1},$$

y x_{i1} representa el valor de x_1 que hace igualmente probable que un elemento, cuyas restantes variables son $x_{i2}, ..., x_{ip}$, pertenezca a la primera o la segunda población.

14.3 LA ESTIMACIÓN DEL MODELO LOGIT

14.3.1 Estimación MV

Supondremos una muestra aleatoria de datos $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$. La función de probabilidades para una respuesta y_i cualquiera es:

$$P(y_i) = p_i^{y_i}(1-p_i)^{1-y_i} \quad y_i = 0, 1$$

y para la muestra :

$$P\left(y_1, \ldots y_n\right) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}.$$

Tomando logaritmos:

$$\log P(\mathbf{y}) = \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum \log\left(1-p_i\right) \tag{14.9}$$

la función soporte (de verosimilitud en logaritmos) puede escribirse como

$$\log P(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i \log p_i + (1-y_i)\log(1-p_i)). \tag{14.10}$$

donde $\boldsymbol{\beta}' = \left(\beta_0, \beta_1, \ldots, \beta_p\right)$ es un vector de $p+1$ componentes, incluyendo la constante β_0 que determina las probabilidades p_i. Maximizar la verosimilitud puede expresarse como minimizar una función que mide la desviación entre los datos y el modelo. En el capítulo 10 se definió la desviación de un modelo mediante $D(\boldsymbol{\theta}) = -\mathbf{2L}(\boldsymbol{\theta})$ y por tanto la desviación del modelo será:

$$D(\boldsymbol{\beta}) = -2\sum_{i=1}^{n} (y_i \log p_i + (1-y_i)\log(1-p_i)). \tag{14.11}$$

y hablaremos indistintamente de de maximizar el soporte o minimizar la desviación del modelo. Se define la desviación de cada dato (deviance) por:

$$d_i = -2(y_i \log p_i + (1 - y_i) \log(1 - p_i)). \tag{14.12}$$

y miden el ajuste del modelo al dato $(y_i, \mathbf{x}_i)$. En efecto, observemos en primer lugar que como los $\widehat{p}_i$ son menores que uno, sus logaritmos son negativos, por lo que la *desviación es siempre positiva*. Además, en el calculo de la desviación sólo interviene uno de sus dos términos, ya que y_i solo puede valer cero o uno. Entonces:

- Si $y_i = 1$, y la observación pertenece a la segunda población, el segundo término de la desviación es nulo y $d_i = -2 \log p_i$. La observación tendrá una desviación grande si la probabilidad estimada de pertenecer a la segunda población, p_i, es pequeña, lo que indica que esta observación está mal explicada por el modelo.
- Si $y_i = 0$, y la observación pertenece a la primera población, sólo interviene el segundo término de la desviación $d_i = -2 \log (1 - p_i)$. La desviación será grande si p_i es grande, lo que indica que la probabilidad de pertenecer a la verdadera población es pequeña y el modelo ajusta mal dicho dato.

Para maximizar la verosimilitud, expresando p_i en función de los parámetros de interés, $\boldsymbol{\beta}$,en (14.9) obtenemos la función soporte:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \mathbf{x}_i' \boldsymbol{\beta} - \sum_{i=1}^{n} \log \left(1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}\right) \tag{14.13}$$

que derivaremos para obtener los estimadores MV. Escribiendo el resultado como vector columna:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} y_i \mathbf{x}_i - \sum_{i=1}^{n} \mathbf{x}_i \left(\frac{e^{\mathbf{x}_i' \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}}\right) \tag{14.14}$$

e igualando este vector a cero y llamando $\widehat{\boldsymbol{\beta}}$ a los parámetros que satisfacen el sistema de ecuaciones:

$$\sum_{i=1}^{n} y_i \mathbf{x}_i = \sum \mathbf{x}_i \left(\frac{1}{1 + e^{-\mathbf{x}_i' \widehat{\boldsymbol{\beta}}}}\right) = \sum \widehat{y}_i \mathbf{x}_i \tag{14.15}$$

Estas ecuaciones establecen que el producto de los valores observados por las variables explicativas debe ser igual al de los valores previstos. También, que los residuos del modelo, $e_i = y_i - \widehat{y}_i$, deben ser ortogonales a las variables $\mathbf{x}$. Esta condición es análoga a la obtenida en el modelo de regresión estándar, pero ahora el sistema (14.15) resultante no es lineal en los parámetros $\widehat{\boldsymbol{\beta}}$. Para obtener el valor $\widehat{\boldsymbol{\beta}}_{MV}$ que maximiza la verosimilitud acudiremos a

un algoritmo tipo *Newton-Raphson*. Desarrollando el vector $(\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta})$ alrededor de un punto $\boldsymbol{\beta}_a$, se tiene

$$\frac{\partial L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \frac{\partial L(\boldsymbol{\beta}_a)}{\partial\boldsymbol{\beta}} + \frac{\partial^2 L(\boldsymbol{\beta}_a)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}(\boldsymbol{\beta}-\boldsymbol{\beta}_a);$$

para que el punto $\boldsymbol{\beta}_a$ corresponda al máximo de verosimilitud su primera derivada debe anularse. Imponiendo la condición $\partial L(\boldsymbol{\beta}_a)/\partial\boldsymbol{\beta}=0$, se obtiene:

$$\boldsymbol{\beta}_a = \widehat{\boldsymbol{\beta}} + \left(-\frac{\partial^2 L(\boldsymbol{\beta}_a)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)^{-1}\left(\frac{\partial L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right) \tag{14.16}$$

que expresa cómo obtener el punto máximo $\boldsymbol{\beta}_a$, a partir de un punto próximo cualquiera $\boldsymbol{\beta}$. La ecuación depende de la matriz de segundas derivadas, que, en el óptimo, es la inversa de la matriz de varianzas y covarianzas asintótica de los estimadores MV. Para obtener su expresión, derivando por segunda vez en (14.14), se obtiene:

$$\widehat{\mathbf{M}}^{-1} = \left(-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right) = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\omega_i \tag{14.17}$$

donde los coeficientes ω_i están dados por:

$$\omega_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{\left(1+e^{\mathbf{x}_i'\beta}\right)^2} = p_i(1-p_i) \tag{14.18}$$

Sustituyendo en (14.16) las expresiones (14.17) y (14.14) y evaluando las derivadas en un estimador inicial $\widehat{\boldsymbol{\beta}}$, se obtiene el siguiente método para obtener un nuevo valor del estimador, $\boldsymbol{\beta}_a$, a partir del $\widehat{\boldsymbol{\beta}}$

$$\boldsymbol{\beta}_a = \widehat{\boldsymbol{\beta}} + \left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\widehat{\omega}_i\right)^{-1}\left(\sum\mathbf{x}_i(y_i-\widehat{p}_i)\right) \tag{14.19}$$

donde $\widehat{p}_i$ y $\widehat{\omega}_i$ se calculan con el valor $\widehat{\boldsymbol{\beta}}$. El algortimo puede escribirse como:

$$\boxed{\boldsymbol{\beta}_a = \widehat{\boldsymbol{\beta}} + \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{Y}-\widehat{\mathbf{Y}}\right)} \tag{14.20}$$

donde $\widehat{\mathbf{W}}$ es una matriz diagonal con términos $\widehat{p}_i(1-\widehat{p}_i)$ y $\widehat{\mathbf{Y}}$ el vector de valores esperados de $\mathbf{Y}$. La matriz de varianzas y covarianzas de los estimadores así obtenidos es aproximadamente, según (14.20), $\left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}$. Observemos que la ecuación (14.20) indica que debemos

modificar el estimador si los residuos no son ortogonales a las variables explicativas, es decir si $\mathbf{X}'\left(\mathbf{Y}-\widehat{\mathbf{Y}}\right) \neq 0$. La modificación del estimador depende de esta diferencia y se reparte entre los componentes de $\widehat{\boldsymbol{\beta}}$ en función de su matriz de varianzas y covarianzas estimada.

La forma habitual de implementar este método es el siguiente algoritmo iterativo que proporciona en convergencia el estimador MV de $\boldsymbol{\beta}$.

1. Fijar un valor arbitrario inicial, $\widehat{\boldsymbol{\beta}}_1$, para los parámetros y obtener el vector $\widehat{\mathbf{Y}}_1$ para dicho valor en el modelo logit. Por ejemplo, si $\widehat{\boldsymbol{\beta}}_1 = 0$,

$$\widehat{y}_i = \widehat{p}_i = \frac{1}{1+e^{-0}} = \frac{1}{2}$$

y el vector $\widehat{\mathbf{Y}}$ tiene todas sus componentes iguales a $1/2$.

2. Definir una variable auxiliar z_i de residuos estandarizados por:

$$z_i = \frac{y_i - \widehat{y}_i}{\sqrt{\widehat{y}_i\left(1-\widehat{y}_i\right)}} = \frac{y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i\left(1-\widehat{p}_i\right)}}$$

o vectorialmente:

$$\mathbf{Z} = \widehat{\mathbf{W}}^{-1/2}(\mathbf{Y}-\widehat{\mathbf{Y}})$$

donde $\widehat{\mathbf{W}}$ es una matriz diagonal con términos $\widehat{y}_i\left(1-\widehat{y}_i\right)$.

3. Estimar por mínimos cuadrados una regresión con variable dependiente $\mathbf{Z}$ y matriz de regresores $\mathbf{T} = \widehat{\mathbf{W}}^{1/2}\mathbf{X}$. Los parámetros estimados con esta regresión, $\widehat{\mathbf{b}}_1$, vendrán dados por:

$$\begin{aligned} \widehat{\mathbf{b}}_1 &= \left(\mathbf{T}'\mathbf{T}\right)^{-1}\mathbf{T}'\mathbf{Z} \\ &= \left(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{Y}-\widehat{\mathbf{Y}}\right) \end{aligned}$$

y, comparando con (14.20), vemos que $\mathbf{b}_1$ estima el incremento $\boldsymbol{\beta}_a - \widehat{\boldsymbol{\beta}}_1$ de los parámetros que nos acerca al máximo.

4. Obtener un nuevo estimador de los parámetros $\widehat{\boldsymbol{\beta}}_2$ del modelo logístico mediante

$$\widehat{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\beta}}_1 + \widehat{\mathbf{b}}_1$$

5. Tomar el valor estimado resultante de la etapa anterior, que en general llamaremos $\widehat{\boldsymbol{\beta}}_h$, y sustituirlo en la ecuación del modelo logístico para obtener el vector de estimadores $\widehat{\mathbf{Y}}\left(\widehat{\boldsymbol{\beta}}_h\right) = \widehat{\mathbf{Y}}_{\mathbf{h}}$. Utilizando este vector $\widehat{\mathbf{Y}}_h$ construir la matriz $\widehat{\mathbf{W}}_h$ y las nuevas variables $\mathbf{Z}_h$ y $\mathbf{T}_h$

$$\mathbf{Z}_h = \widehat{\mathbf{W}}_h^{-1/2}\left(\mathbf{Y}-\widehat{\mathbf{Y}}_h\right),$$

$$\mathbf{T}_h = \widehat{\mathbf{W}}_h^{1/2}\mathbf{X},$$

y volver a la etapa 2. El proceso se repite hasta obtener la convergencia $\left(\widehat{\boldsymbol{\beta}}_{\mathbf{h+1}} \simeq \widehat{\boldsymbol{\beta}}_{\mathbf{h}}\right)$.

14.3.2 Contrastes

Si queremos contrastar si una variable o grupo de variables incluidas dentro de la ecuación es significativo, podemos construir un contraste de la razón de verosimilitudes comparando el máximo de la función de verosimilitud para el modelo con y sin estas variables. Supongamos que $\boldsymbol{\beta} = (\boldsymbol{\beta}_1\boldsymbol{\beta}_2)$, donde $\boldsymbol{\beta}_1$ tiene dimensión $p-s$, y $\boldsymbol{\beta}_2$ tiene dimensión s. Se desea contrastar si el vector de parámetros:

$$H_0 : \boldsymbol{\beta}_2 = 0.$$

frente a la alternativa

$$H_1 : \boldsymbol{\beta}_2 \neq 0$$

El contrate de razon de verosimilitudes utiliza que $\lambda = 2L(H_1) - 2L(H_0)$, donde $L(H_1)$ es el máximo del soporte cuando estimamos los parámetros bajo H_1 y $L(H_0)$ es el máximo cuando estimamos los parámetros bajo H_0 es, si H_0 es cierta, una χ^2_s. Una manera equivalente de definir el contraste es llamar $D(H_0) = -2L(\widehat{\boldsymbol{\beta}}_1)$ a la desviación cuando el modelo se estima bajo H_0, es decir, suponiendo que $\boldsymbol{\beta}_2 = 0$, y $D(H_1) = -2L(\widehat{\boldsymbol{\beta}}_1\widehat{\boldsymbol{\beta}}_2)$ a la desviación bajo H_1. La desviación será menor con el modelo con más parámetros (la verosimilitud será siempre mayor bajo H_1 y, si H_0 es cierta, la diferencia de desviaciones, que es el contraste de verosimilitudes

$$\chi^2_s = D(H_0) - D(H_1) = 2L(\widehat{\boldsymbol{\beta}}_1\widehat{\boldsymbol{\beta}}_2) - 2L(\widehat{\boldsymbol{\beta}}_1)$$

se distribuye como una χ^2_s con s grados de libertad.

En particular este test puede aplicarse para comprobar si un parámetro es significativo y debe dejarse en el modelo. Sin embargo, es más habitual en estos casos comparar el parámetro estimado con su desviación típica. Los cocientes

$$w_j = \frac{\widehat{\beta}_j}{s(\widehat{\beta}_j)}$$

se denominan *estadísticos de Wald* y en muestras grandes se distribuyen, si el verdadero valor del parámetro es cero, como una normal estándar.

Una medida global del ajuste es

$$R^2 = 1 - \frac{D(\widehat{\boldsymbol{\beta}})}{D_0} = 1 - \frac{L(\widehat{\boldsymbol{\beta}})}{L(\beta_0)}$$

donde el numerador es la desviación (verosimilitud en el máximo) para el modelo con parámetros estimados $\widehat{\boldsymbol{\beta}}$ y el denominador la desviación (verosimilitud) para el modelo que sólo incluye la constante β_0. Observemos que, en este último caso, la estimación de la probabilidad p_i es constante para todos los datos e igual a m/n siendo m el número de elementos en la muestra con la variable $y = 1$. Entonces, sustituyendo en (14.11) la desviación máxima

que corresponde al modelo más simple posible con sólo β_0 que asigna la misma probabilidad a todos los datos, es

$$D_0 = -2L(\beta_0) = -2m\log m - 2(n-m)\log(n-m) + 2n\log n.$$

Por otro lado, si el ajuste es perfecto, es decir todas las observaciones con $y = 1$ tienen $p_i = 1$ y las de $y = 0$ tienen $p_i = 0$, entonces, según (14.9) la desviación es cero y $L(\widehat{\boldsymbol{\beta}}) = 0$ y $R^2 = 1$. Por el contrario, si las variables explicativas no influyen nada la desviación con las variables explicativas será igual que sin ellas, $L(\widehat{\boldsymbol{\beta}}) = L(\beta_0)$ y $R^2 = 0$.

Ejemplo 14.1 *Vamos a utilizar los datos de MEDIFIS para construir un modelo logit que clasifique una persona como hombre o mujer en función de sus medidas físicas.*

Si intentamos explicar la variables binaria, género, en función de todas las variables observadas obtenemos que el modelo es

$$\begin{aligned}\log\frac{p_i}{1-p_i} &= -488.84 + 11.84pie + \\ &\quad 2.93aes - 5.10dcr - 10.5lrt - 3.73est + 4.59lbr + .14pes\end{aligned}$$

el modelo ajusta perfectamente los datos y no es posible calcular desviaciones típicas para los coeficientes. El modelo no es único. El problema es que tenemos sólo 27 observaciones y con las siete variables clasificamos fácilmente todas las observaciones. Sin embargo, el modelo obtenido puede ser muy malo para clasificar otras observaciones.

Vamos a contruir el modelo paso a paso. La variable con mayor coeficiente en la ecuación anterior es el pie con lo que comenzaremos con esta variable. Estimando el modelo estimado con el programa SPSS (con MINITAB no se produce convergencia del algoritmo), el modelo es:

$$\log\frac{p_i}{1-p_i} = -433 + 11.08pie$$

Los dos parámetros están muy correlados y las desviaciones típicas son muy grandes. El valor inicial de la desviación es D_0=37.1. Después de estimar el modelo la desviación es D= 3.8. La diferencia entre desviaciones nos proporciona el contraste para ver si la variable pie es significativa. Esta diferencia es 33.27 que bajo la hipótesis de que el parámetro es cero será aproximadamente una distribución ji-cuadrado con 1 grado de libertad. El valor es tan grande que rechazamos la hipótesis a cualquier nivel de significación y concluimos que el pie es muy útil para discriminar.

Es interesante que, en este caso, el estadístico de Wald lleva a un resultado distinto. Como los parámetros están muy correlados, la desviación típica del coeficiente del pie es 108.9, y el estadístico de Wald es 11.08/ 108.= .01 que no es significativo.

Si aplicamos esta ecuación para clasificar los datos muestrales obtenemos un porcentaje de éxitos del 96%. Sólo una observación se clasifica mal como indica la tabla

		Clasificado	
		M	H
Realidad	M	15	0
	H	1	11

Ejemplo 14.2 *Vamos a intentar introducir una variable adicional en el modelo logístico anterior que contiene sólo el pie. Introducimos la estatura, y el modelo estimado es*

$$\log \frac{p_i}{1-p_i} = -440.85 + 12.0pie - -.169est$$

*con desviaciones típicas (4092), (*104.9) *y* $(0, 5243)$ *. El coeficiente de estatura es -.1693 con error estándar .5243 dando lugar a un estadístico de Wald de .3, con lo que concluimos que este coeficiente no es significativo. La desviación de este modelo es 3.709. Por tanto la reducción en desviación debida a la variable estatura con respecto al modelo que sólo incluye el pié es sólo de 3.80-3.71=.09, que no es significativa comparada con una ji-cuadrado con 1 grado de libertad. Este resultado es previsible ya que el modelo con las siete variables tiene una desviación de cero y el que contiene sólo el pie una desviación de D= 3.8: el contraste de que las seis variables adicionales no influyen lleva, en consecuencia, a un valor del estadístico de 3.8, y este valor en la hipótesis de que las variable no influyen debe provenir de una* χ^2_6 *con seis grados de libertad, lo que concuerda con lo observado, y debemos concluir que ninguna de las variables adicionales influye. La conclusión es pues que únicamente con el pie podemos clasificar estos datos con poco error.*

Si comparemos estos resultados con los del capítulo anterior (ejemplo 13.3) son bastante consistentes porque allí ya observamos que la variable más importante era el pié. El porcentaje de clasificación de la función lineal discriminante era del 100% que disminuía al 85% con validación cruzada. En el modelo logístico con sólo una variable hemos obtenido el 96% de éxito. Con validación cruzada este valor disminuye algo, pero mucho menos que en el ejemplo 13.3 debido a la economía de parámetros que hace que se produzca menos sobreajuste.

Ejemplo 14.3 *Vamos a utilizar los datos de MUNDODES para ver cuáles son las variables que clasifican mejor a un país como perteneciente al continente africano. La función logit estimada es*

$$\begin{aligned}\log \frac{p_i}{1-p_i} &= 15.58 + .18tn - .14tm \\ &\quad -.033mi + .05lpnb + .13em - .47eh\end{aligned}$$

Las variables tn, mi y eh son significativas, con un cociente entre la estimación del parámetro y la desviación típica de 6.8, 2.09 y 2.5 respectivamente. La desviación inicial es $D(\beta_0)$*=-2(27*log*27+64*log*64-91*log*91)=* 110.66*. Por otro lado, la desviación del modelo estimado es -2* $L(\widehat{\boldsymbol{\beta}}) = 41.41$ *y la diferencia entre estas dos cantidades proporciona el contraste de que las variables no influyen. Si esto es cierto la diferencia, 69.25 será una ji-cuadrado con 6 grados de libertad. Como este valor es muy grande rechazamos esta hipótesis y admitimos que las variables influyen. El pseudo coeficiente de determinación es*

$$R^2 = 1 - \frac{41.41}{110.66} = .63$$

La tabla de clasificación con este modelo es

		Clasificado	
		nA	A
Realidad	nA	61	3
	A	5	22

que supone una proporción de éxitos de 83/91, o de 91%.

14.3.3 Diagnosis

Los residuos del modelo logit (que a veces se denominan residuos de Pearson) se definen por:

$$e_i = \frac{y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}}$$

y, si el modelo es correcto, serán variables de media cero y varianza unidad que pueden servirnos para hacer la diagnosis del modelo. El estadístico $\chi^2 = \sum_{i=1}^{n} e_i^2$ permite realizar un contraste global de la bondad del ajuste. Si el modelo es adecuado se distribuye asintóticamente como una χ^2 con $gl = n - p - 1$, donde $p + 1$ es el número de parámetros en el modelo.

En lugar de los residuos de Pearson se utiliza mucho las *desviaciones* de las observaciones o pseudoresiduos, definidas en (14.12) por $d_i = -2(y_i \log \widehat{p}_i + (1-y_i)\log(1-\widehat{p}_i))$, que aparece, de manera natural, en la maximización de la función de verosimilitud.

Podemos hacer un contraste de razón de verosimilitudes de la bondad del modelo como sigue: la hipótesis nula será que el modelo es adecuado, es decir, las probabilidades pueden calcularse con el modelo logístico con $p + 1$ parámetros. La hipótesis alternativa será que el modelo no es adecuado y las n probabilidades son libres (supuesto que las x son distintas). Entonces, la desviación bajo H_0 es $D(H_0)$, mientras que bajo H_1 cada observación queda perfectamente clasificada dando $p_i = 0$ si pertenece a la primera población y $p_i = 1$ si pertenece a la segunda, y la desviación es cero porque todas las observaciones se clasifican sin error. El contraste de la razón de verosimilitudes se reduce al estadístico desviación global:

$$D(H_0) = -2\sum(y_i \log \widehat{p}_i + (1-y_i)\log(1-\widehat{p}_i))$$

que, si el modelo es correcto, será también asintóticamente una χ^2 con $n - p - 1$ grados de libertad

14.4 EL MODELO MULTILOGIT

El modelo logit puede generalizarse para más de dos poblaciones, es decir, para variables respuesta cualitativas con más de dos niveles posibles. Supongamos G poblaciones, entonces, llamando p_{ig} a la probabilidad de que la observación i pertenezca a la clase g, podemos escribir:

$$p_{ig} = \frac{e^{\beta_{0g} + \boldsymbol{\beta}'_{1g}\mathbf{x}_i}}{1 + \sum_{j=1}^{G-1} e^{-\beta_{0j} - \boldsymbol{\beta}'_{1j}\mathbf{x}_i}} \quad j = 1, ..., G-1 \tag{14.21}$$

y

$$p_{iG} = \frac{1}{1 + \sum_{j=1}^{G-1} e^{-\beta_{0j} - \boldsymbol{\beta}'_{1j}\mathbf{x}_i}}$$

con lo que automáticamente garantizamos que $\sum_{g=1}^{G} p_{ig} = 1$. Diremos que las probabilidades p_{ig} satisfacen una distribución logística multivariante. La comparación entre dos categorías se hace de la forma habitual

$$\frac{p_{ig}}{p_{ij}} = \frac{e^{\beta_{0g} + \boldsymbol{\beta}'_{1g}\mathbf{x}_i}}{e^{\beta_{0j} + \boldsymbol{\beta}'_{1j}\mathbf{x}_i}} = e^{(\beta_{0g} - \beta_{0j})} e^{(\boldsymbol{\beta}'_{1g} - \boldsymbol{\beta}'_{1j})\mathbf{x}_i}$$

Esta ecuación indica que las probabilidades relativas entre dos alternativas no dependen del resto. Esa hipótesis puede generalizarse (véase Maddala, 1983).

La estimación y contrastes de esos modelos son extensiones directas de los logit ya estudiados y no entraremos en los detalles que el lector interesado puede encontrar en Fox (1984)

14.5 OTROS MÉTODOS DE CLASIFICACIÓN

14.5.1 Árboles de Clasificación

Un procedimiento alternativo de clasificación debido a Breiman y Friedman (véase Breiman et al., 1984) son los árboles de clasificación (Classification and Regression Trees, CART). Este procedimiento no utiliza un modelo estadístico formal y es más bien un algoritmo para clasificar utilizando particiones binarias sucesivas utilizando los valores de una variable cada vez. La idea del procedimiento se resume en la figura 14.1. Suponemos que se dispone de una muestra de entrenamiento, que incluye la información del grupo al que pertenece cada dato y que servirá para construir el criterio de clasificación. Posteriormente se aplicará el criterio para clasificar nuevos datos. Comenzamos con un nudo inicial y nos preguntamos como dividir el conjunto de datos disponibles en dos partes más homogéneas utilizando el valor de una de las variables. La variable se escoge de manera que una partición de los datos en función de que su valor para esta variable sea mayor o menor que una constante proporcione una división de los datos en dos conjuntos lo más homogeneos posibles.

El algortimo comienza seleccionando una variable, supongamos que la x_i, y obteniendo un punto de corte, c, de manera que separemos los datos que tienen $x_1 \leq c$ de aquellos con $x_1 > c$. De este nudo incial saldrá ahora dos, uno al que llegarán las observaciones con $x_1 \leq c$ y otro al que llegarán las que tienen $x_1 > c$. En cada uno de estos nudos se vuelve a repetir el proceso de seleccionar una variable y un punto de corte dividir la muestra en dos partes más homogéneas. El proceso termina cuando hayamos clasificado todas las observaciones (o casi todas) correctamente en su grupo. La construcción del árbol requiere las decisiones siguientes:

1. La selección de las variables y de sus puntos de corte para hacer las divisiones.
2. Cuando un nudo se considera terminal y cuando se continua dividiendo.

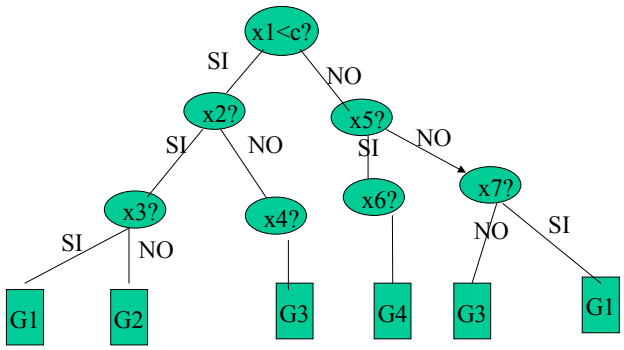


Figura 14.1: Ejemplo de árbol de clasificación

3. La asignación de las clases a los nudos terminales.

Supongamos que se desea clasificar las observaciones en G grupos. Por el primer nudo, llamado *nudo raiz* y marcado en la figura por $x1$, pasarán todas las observaciones, mientras que por cualquier otro nudo sólo pasarán las observaciones que verifican las condiciones de llegada a ese nudo. Por ejemplo, en el nudo $x2?$. sólo se encuentran aquellas observaciones para las que se responde SI a la pregunta : ¿es $x_1 < c$?. Podemos entonces asociar a cada nudo el subconjunto de observaciones que pasarán por él. Para decidir la variable que va a utilizarse para hacer la partición en un nudo se calcula primero la proporción de observaciones que pasan por el nudo para cada uno de los grupos. Llamando a los nudos $t = 1, ..., T$, y $p(g|t)$ a las probabilidades de que las observaciones que llegan al nudo t pertenezcan a cada una de las clases, se define la *impureza* del nudo t por

$$I(t) = -\sum_{g=1}^{G} p(g|t) \log p(g|t)$$

esta medida se llama *entropía*, es no negativa y mide la diversidad. Se utilizó en la sección 4.2.3 para obtener criterios de proyección. Por ejemplo, con dos grupos la impureza es

$$I(t) = -p \log p - (1-p) \log(1-p)$$

Esta función está representada la figura 14.2. Se observa que la heterogeneidad o diversidad es máxima es cuando p=.5, y tiende a cero cuando p se aproxima a cero o a uno.

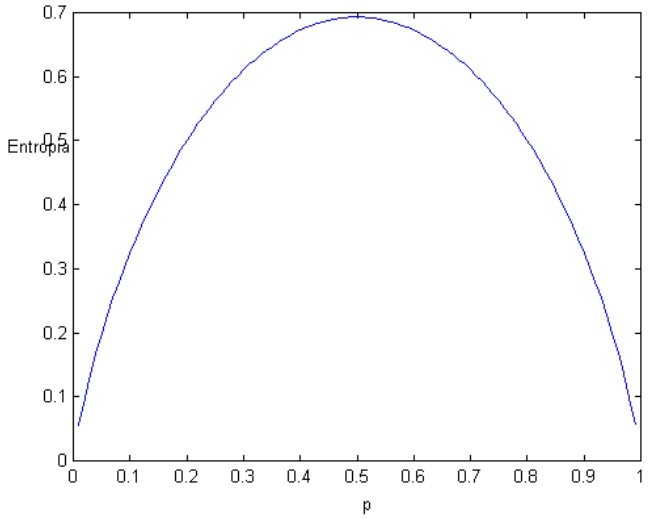


Figura 14.2: Representacion de la entropia como funcion de la probabilidad de un grupo

Con G grupos, si en el nudo t todas las observaciones pertenecen al grupo g, de manera que $p(g|t) = 1$, y $p(i|t) = 0, i \neq g$, la entropía o impureza del nudo es $I(t) = 0$ (tomamos $0log0 = 0$, que es su valor límite). En otro caso, la impureza será positiva y será máxima cuando $p(g|t) = G^{-1}$. La variable utilizada para realizar la división de los datos en un nudo se selecciona minimizando la heterogeneidad o impureza resultante de la división. Consideremos preguntas q posibles del tipo: ¿es $x_i < a$? y sean p_S y p_N las proporciones de las observaciones del nudo t que irán a los nudos resultantes de responder "Si" a la pregunta q (nudo t_S) y al responder "No" (nudo t_N). Llamemos $I(t_S)$ y $I(t_N)$ a las impurezas resultantes de estos nudos que surgirán como consecuencia de la pregunta q. El cambio en entropía o heterogeneidad después de la pregunta q, será la diferencia entre la entropía del nudo, $I(t)$, y la entropía después del nudo, que vendrá dada por $p_S I(t_S) + p_N I(t_N)$. El cambio en entropía producido por la pregunta q es:

$$\Delta I(t, q) = I(t) - p_S I(t_S) - p_N I(t_N) \tag{14.22}$$

y se desea escoger q para maximizar el cambio de entropía en el nudo. El procedimiento es el siguiente: se define un conjunto de preguntas q del tipo $x_i < a$, para $i = 1, ..., p$ y $a \in (-\infty, \infty)$. Para cada pregunta se calcula la disminución de impureza o entropía que implica y se escoge aquella pregunta que maximice la impureza resultante medida por (14.22).

La clasificación en los nudos terminales se hace asignado todas las observaciones del nudo al grupo más probable en ese nudo, es decir, aquel grupo con máxima $p(g|t)$. Si la impureza del nudo es cero, todas las observaciones pertenecen al mismo grupo, y la clasificación de las observaciones podría hacerse sin error, en otro caso, si la impureza del nudo no es cero, la clasificación tendrá un cierto error.

Este proceso de construcción del árbol puede generar muchos nudos cuando el número de variables es grande y se plantea el problema de cómo simplificar o podar el árbol para hacerlo más manejable con poca pérdida de información. Además, pueden utilizarse otras medidas

de diversidad para realizar las divisiones. No entraremos en los detalles de los algoritmos existentes que el lector interesado puede encontrar en Breiman et al. (1984)

Los árboles de clasificación suelen dar buenos resultados cuando muchas de las variables de clasificación son cualitativas y cuando las relaciones entre las variables son muy lineales. Sin embargo, son más ineficaces que los procedimientos clásicos cuando las variables son aproximadamente normales. La evidencia disponible sobre su eficacia es distintos tipos de problemas es todavía pequeña.

Ejemplo 14.4 *Ilustraremos la idea de los árboles de clasificación con los datos de MEDIFIS. La impureza total de los datos inicialmente es*

$$I_0 = -(12/27)\log(12/27) - (15/27)\log(15/27) = .687$$

consideremos preguntas del tipo pie< a?. Si tomamos a = 41, y llamamos q_1 a la pregunta pie<41? , la muestra se divide en 18 casos con respuesta SI y 9 con respuesta NO. La impureza en los nudos resultantes será: Para el nudo NO, la impureza es cero, porque todos los elementos del nudo son hombres. En el nudo SI de los 18 elementos 15 son mujeres y 3 hombres con lo que la impureza será

$$I(S, q_1) = -(15/18)\log(15/18) - (3/18)\log(3/18) = .451$$

con lo que la impureza resultante con esta pregunta es

$$\Delta I(q_1) = .687 - (9/27).0 - (18/27)(.451) = .387$$

Comparemos este resultado con el obtenido con la pregunta q_2 : pie< 40?. Ahora la muestra se divide en 16 y 11. La impureza del grupo SI es

$$I(S, q_2) = -(15/16)\log(15/16) - (1/16)\log(1/16) = .234$$

y la impureza del grupo No es cero. La reducción de impureza de esta pregunta es

$$\Delta I(q_2) = .687 - (11/27).0 - (16/27)(.234) = .548$$

por tanto, es mejor dividir con la pregunta q_2 que con la q_1, ya que obtenemos un grupo más homogeneo como resultado de la división. Otras posibles preguntas se analizan de la misma forma.

14.5.2 Redes Neuronales

Las redes neuronales son algoritmos generales de análisis de datos basados en un uso intensivo del ordenador. Su justificación proviene de que, en condiciones generales, pueden aproximar cualquier función del tipo

$$y = f(x_1, ..., x_p)$$

Supongamos para simplificar la discriminación entre dos grupos. Entonces, y es una variable binaria, cero- uno, y el problema es encontrar la función f que mejor se ajusta

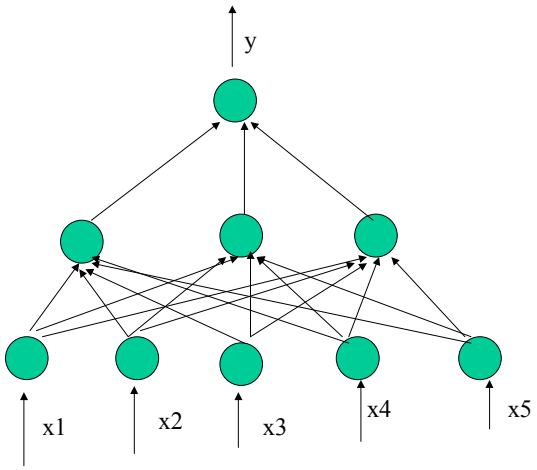


Figura 14.3: Representación de una red neuronal con una capa intermedia.

a los datos observados. Las redes neuronales se construyen a partir de elementos llamados nodos, unidades, o neuronas. Estas unidades reciben un conjunto de entradas, x, que representaremos por la variable vectorial $\mathbf{x}$, y calculan una variable escalar de salida aplicando una ponderación a los componentes de la entrada, añadiendo una constante de sesgo y transformando el resultado de forma no lineal como:

$$z = g(\mathbf{w}'\mathbf{x})$$

donde $\mathbf{w}$ es un vector y g una función generalmente no lineal. El vector $\mathbf{x}$ incorpora una variable adicional, $x_0 = 1$, de manera existe un término de sesgo w_0 incluido en el vector de pesos $\mathbf{w}' = (w_0, ..., w_p)$. Se han propuesto distintos tipos de funciones g pero la función más utilizada en problemas de clasificación es la función logística

$$g(t) = 1/(1 + e^{-t})$$

que proporcionará una salida entre cero y uno. Otra función utilizada es $g(t) = sign(t)$ que produce una respuesta binaria con valores posibles 1 y -1. Por ejemplo, el modelo logístico puede verse como una red neuronal de un sólo nodo y función g logística. Aunque existen muchas estructuras de redes neuronales posibles, la más utilizada para clasificación es el perceptrón, que consiste en un conjunto de neuronas clasificadas en capas y cuya representación gráfica se ilustra en la figura (14.3). Esta estructura contiene cinco neuronas de entrada, tres en la capa intermedia y una de salida.

Las variables de entrada, $(x_1, ..., x_p)$, que caracterizan al elemento a clasificar se introducen por las neuronas iniciales. Por ejemplo, la red de la figura 14.3 es adecuada para clasificar elementos donde se han medido cinco variables. En la neuronas iniciales la función de respuesta es la unidad, es decir $x = f(x)$, y estas neuronas sirven para distribuir las

variables x en la segunda capa de neuronas. Supongamos que la capa intermedia contiene M neuronas (en la figura $M = 3$), entonces cada neurona intermedia $i = 1, ..., M$ recibe un vector de entrada $\mathbf{x}_i$ y genera una variable escalar de salida

$$z_i = g_i(\mathbf{w}_i'\mathbf{x}_i) \tag{14.23}$$

Por ejemplo, en la figura 14.3 la neurona de la capa terminal recibe el vector $\mathbf{z} = (z_1, z_2, z_3)'$ y genera una salida y. Pueden existir varias neuronas terminales y las respuestas de la capa intermedia pasan a las neuronas de salida, y cada una de ellas producirá una variable de salida de acuerdo con una función f. Supongamos que existen $j = 1, ..., J$ neuronas terminales. La salida que producirá la neurona j será

$$y_j = f_j(\mathbf{w}_j'\mathbf{z}_j) = f_j(\mathbf{w}_j'G_j(W, \mathbf{x}))$$

donde $\mathbf{w}_j'$ es la función de ponderación en esta neurona y $\mathbf{z}_j$ es el vector de entrada a la neurona que esta compuesto por las salidas de las neuronas de la capa intermedia. Este vector de entrada tiene M componentes, tantos como salidas intermedias, y cada uno viene dado por (14.23), por lo que puede expresarse como $\mathbf{z}_j = (z_{1j}, ..., z_{Mj}) = (g_1(\mathbf{w}_1'\mathbf{x}_1), ..., g_M(\mathbf{w}_M'\mathbf{x}_M)) = G_j(W, \mathbf{x})$. Por ejemplo, si suponemos que tanto las tres funciones de la capa intermedia como la terminal de la figura 14.3 proporciona una respuesta logística, la respuesta final es:

$$y = g(w_{40} + w_{41}z_1 + w_{42}z_2 + w_{43}z_3)$$

donde g es la función logística. Sustituyendo z por su expresión, que es una función logística de las variables de entrada, tenemos finalmente que

$$y = g(w_{40} + w_{4i}\sum_{i=1}^{3} w_{4i}g(\mathbf{w}_i'\mathbf{x}))$$

La clasificación se realiza en función del valor de la variable y. Por ejemplo, si $y > .5$ el dato se clasifica en la primera población y si $y < .5$ en la segunda.

Para llevar a la práctica este método, hay que estimar los parámetros que definen cada función g_i. Por ejemplo, en la red de la figura 14.3 para cada neurona hay que estimar seis pesos, el coeficiente constante del sesgo más los pesos para cada una de las cinco variables, lo que supone $6\times3 = 18$ parámetros para la capa intermedia, más los 6 de la capa final, lo que supone un total de 24 parámetros. Si llamamos c_i a la variable binaria que incluye las etiquetas de los datos, por ejemplo, $c_i = 1$ cuando pertenece a la primera clase y $c_i = 0$ cuando pertenece a la segunda, la estimación de los pesos se obtiene minimizando :

$$E = \sum_{i=1}^{n}(c_i - y_i)^2 = \sum_{i=1}^{n}(c_i - f(x, w))^2$$

con respecto a los parámetros w o pesos que caracterizan la red. Esta función, que es no lineal en los parámetros, se minimiza frecuentemente con un algoritmo del gradiente, donde los pesos se modifican en cada iteración proporcionalmente al gradiente de esta función

$$w_{ij} \rightarrow w_{ij} + \eta\frac{\partial E}{w_{ij}}$$

siendo η la longitud de paso. El lector interesado en los detalles puede acudir a Ripley (1996). Las redes neuronales necesitan muestras grandes para estimar eficientemente los muchos parámetros que se introducen. La experiencia disponible parece indicar que si la estructura de la red neuronal se diseña bien, en situaciones estándar pueden dar resultados similares a los métodos clásicos, y pueden comportarse mejor en situaciones donde las relaciones entre las variables de clasificación sean muy no lineales. Sin embargo la experiencia disponible es muy limitada por la falta de reglas precisas sobre el número de neuronas a colocar en la capa intemedia y los problemas de convergencia de los algoritmos de estimación, dado el alto número de parámetros a estimar. Por ejemplo, supongamos que queremos clasificar en 4 clases posibles datos de dimensión 10. Si utilizamos cinco neuronas en la capa intermedia y 3 neuronas de salida (necesitamos $G-1$ neuronas de salida) esto supone 88 parámetros a estimar. Un inconveniente de las redes neuronales es la falta de una teoría que oriente sobre las situaciones en que darán buenos resultados y aquellas en las que pueden ser muy ineficientes.

14.5.3 Métodos no Paramétricos

Vecinos más próximos

Un procedimiento de clasificación simple y que ha dado buenos resultados con poblaciones no normales es el siguiente:

(1) Definir una medida de distancia entre puntos, habitualmente la distancia de Mahalanobis.

(2) Calcular las distancias del punto a clasificar, $\mathbf{x}_0$, a todos los puntos de la muestra.

(3) Seleccionar los m puntos muestrales más próximos al que pretendemos clasificar. Calcular la proporción de estos m puntos que pertenece a cada una de las poblaciones. Clasificar el punto $\mathbf{x}_0$ en la población con mayor frecuencia de puntos entre los m.

Este método se conoce como $m-$ vecinos próximos. En el caso particular de $m=1$ el método consiste en asignarle a la población al que pertenece el elemento más próximo. Un problema clave de este método es claramente la selección de m. Una práctica habitual es tomar $m=\sqrt{n_g}$ donde n_g es un tamaño de grupo promedio. Otra posibilidad es probar con distintos valores de m, aplicarselo a los puntos de la muestra cuya clasificación es conocida y obtener el error de clasificación en función de m. Escoger aquel valor de m que conduzca al menor error observado.

Estimación de densidades

Si la densidad de las observaciones no es normal y tenemos una muestra grande podemos intentar estimar directamente la distribución de los datos. Para clasificar un punto $\mathbf{x}_0$, no necesitamos estimar toda la densidad sino sólo la densidad en ese punto, ya que la clasificación se realiza maximizando la probabilidad a posteriori, es decir, maximizando

$$\max_g \pi_g f_g(\mathbf{x}_0)$$

Un estimador ingenuo de la densidad de la población g en el punto $\mathbf{x}_0$ es construir un hipercubo con centro en $\mathbf{x}_0$ y lado h, contar los puntos provenientes de la densidad g incluidos

en el, $n_g(\mathbf{x}_0)$ y estimar la densidad por el cociente entre la frecuencia relativa de puntos en el cubo y su volumen. Es decir

$$f_g(\mathbf{x}_0) = \frac{n_g(\mathbf{x}_0)}{n_g h^p}$$

Este es un procedimiento similar al de los vecinos más próximos, pero en lugar de fijar el número m de puntos más próximos fijamos un entorno h y contamos cuantos puntos de cada distribución están en dicho intervalo. La regla anterior puede escribirse como

$$f_g(\mathbf{x}_0) = \frac{1}{n_g h^p} \sum_{i=1}^{n} \prod_{j=1}^{p} K\left(\frac{x_{ij} - x_{0j}}{h}\right)$$

donde la función K se denomina el núcleo y tiene la propiedad

$$K\left(\frac{x_{ij} - x_{0j}}{h}\right) = \begin{array}{ll} 1, & \text{si} \quad |x_{ij} - x_{0j}| \leq h \\ 0, & \text{en otro caso} \end{array}$$

Este estimador de la densidad es muy irregular, ya que los puntos o bien entran en el hipercubo y contribuyen con valor uno a la densidad en el punto, o no cuentan en absoluto para determinar la densidad. Un estimador mejor es permitir que los puntos contribuyan a la estimación de la densidad en función de su distancia, lo que puede hacerse sustituyendo el núcleo rectangular por una función suave que promedie la información de los puntos en función de su distancia. Un núcleo muy utilizado es el normal, dado por

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}).$$

El procedimiento puede mejorarse teniendo en cuenta la dependencia de las variables y utilizando un núcleo multivariante que tenga en cuanta esta dependencia. Por ejemplo, un núcleo normal multivariante conduciría a :

$$f_g(\mathbf{x}_0) = \frac{1}{nh^p \left|\mathbf{S}_g\right|^{1/2}} \sum_{i=1}^{n} \exp\left\{-\frac{1}{2h^2} \left(\mathbf{x}_i - \mathbf{x}_0\right)' \mathbf{S}_g^{-1} \left(\mathbf{x}_i - \mathbf{x}_0\right)\right\}$$

donde $\mathbf{S}_g$ es una estimación de la matriz de covarianzas en el grupo g. En general es frecuente tomar la misma matriz $\mathbf{S}_g$ en todos los grupos y estimarla como una media ponderada de las matrices de cada grupo.

El problema de este método es que la estimación de la densidad depende críticamente de la elección del parámetro h, que es desconocido. Puede utilizarse el error de clasificación como comentamos antes para estimar h. Remitimos al lector interesado a McLachan (1992) para un estudio detallado de este tema. Más recientemente algunos autores han tratado de mejorar la idea de vecinos más próximos introduciendo criterios de invarianza ante rotaciones para mejorar las propiedades del método, véase Hastie y Simard (1998). Otros autores han utilizado estimación no paramétrica para estimar directamente las relaciones no lineales de clasificación, vease por ejemplo Hastie, Tibshirani y Buja (1994).

14.5.4 Otros Métodos

Un método reciente que está adquiriendo cierta popularidad es el denominado SVM (support vector machines), véase Vapnik (2000) y Cherkassky y Mulier (1998). Este método adopta un punto de vista distinto del habitual: en lugar de buscar una reducción del espacio de los datos y resolver el problema en ese espacio de dimensión menor busca un espacio de dimensión mayor donde los puntos pueden separarse de forma lineal. Para entender la filosofía del método definamos problemas separables linealmente. Supongamos que tenemos una muestra de n elementos del tipo $(y_i, \mathbf{x}_i)$ donde y_i es la variable binaria de clasificación, que, por conveniencia ahora, tomamos con valores posibles -1 y +1 en lugar de cero y uno, y $\mathbf{x}_i \in \mathbf{R}^p$ es el vector de variables para la observación i. Este conjunto de n datos es linealmente separable si es posible encontrar un vector $\mathbf{w} \in \mathbf{R}^p$ que defina un plano que nos separe perfectamente las observaciones. Es decir, todas las observaciones de un grupo, por ejemplo las de $y_i = -1$ se encuentran a un lado del plano y verifican que $\mathbf{w}'\mathbf{x}_i + b < -1$, para un cierto escalar b, mientras que los puntos del otro grupo con $y_i = 1$ están al otro lado y verifican $\mathbf{w}'\mathbf{x}_i + b > 1$. Estas dos desigualdades pueden también escribirse conjuntamente como

$$y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 \quad \text{para } i = 1, ..., n$$

Sea

$$f(\mathbf{x_i}) = \mathbf{w}'\mathbf{x}_i + b$$

el valor del hiperplano de separación óptima entre los dos conjuntos para un punto $\mathbf{x}_i$. La distancia entre un punto cualquiera, $\mathbf{x}_i$, y el hiperplano viene dada por la proyección del punto en la dirección $\mathbf{w}$, que es el vector ortogonal al plano. Esta proyección se calcula mediante $\mathbf{w}'\mathbf{x}_i / \|\mathbf{w}\|$. Como los puntos verifican $y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1$, maximizaremos las distancias de los puntos al plano maximizando

$$\frac{y_i(\mathbf{w}'\mathbf{x}_i + b)}{\|\mathbf{w}\|}$$

para todos los puntos muestrales. Esto ocurrirá si el numerador es positivo y el denominador tan pequeño como sea posible, lo que conduce al problema de programación cuadrática

$$\begin{aligned} & \min \|\mathbf{w}\|^2 \\ y_i(\mathbf{w}'\mathbf{x}_i + b) \quad \geq \quad & 1, \quad i = 1, ..., n \end{aligned}$$

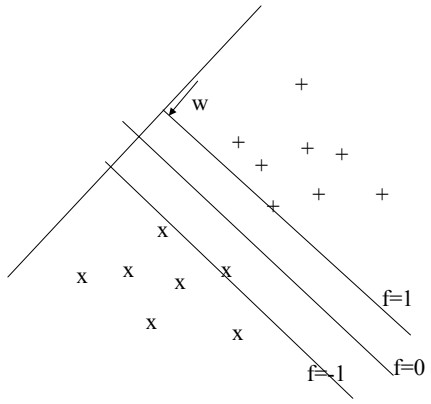


Figura 14.4: Representación de dos clases de puntos separables linealmente en cierto espacio, el plano separador f y el vector w ortogonal al plano.

Cuando los datos no son linealmente separables el procedimiento habitual es proyectar los datos sobre un espacio de dimensión menor y utilizar funciones no lineales para separarlos grupos. Por ejemplo, en discriminación con dos poblaciones normales multivariantes con distinta matriz de covarianzas es frecuente que los datos no sean linealmente separables y entonces como hemos visto proyectamos los datos sobre un espacio de dimensión menor y utilizamos una función cuadrática para discriminar sobre las proyecciones en dicho espacio. En enfoque del vector soporte (VSM) es aplicar una transformación a los datos que los lleve a un espacio de dimensión mucho mayor que p y entonces aplicar una discriminación lineal como la que se ha presentado. La clave del procedimiento es darse cuenta que para resolver el problema en un espacio de dimensión alta solo necesitamos conocer los productos escalares entre las observaciones y si definimos adecuadamente un producto escalar en el espacio ampliado tenemos resuelto el problema. Los detalles técnicos de implantación del método dejan todavía muchos interrogantes por resolver y el lector interesado puede acudir a la referencias indicadas al comienzo de la sección.

14.6 Lecturas complementarias

El modelo logistíco se estudia en Hosmer y Lemeshow (1989), Cramer (1991) y Fox (1984). Estos modelos son casos particulares de los modelos lineales generalizado estudiados por Mc Cullagh y Nelder (1989). Su aplicación para discriminación se presentan claramente en McLachlan (1992), que contiene numerosas referencias a otros métodos no presentados en este libro. Las redes neuronales para clasificación se explican en Hand (1997), Ripley (1996), Hertz et al. (1991) y Bishop (1995). El libro de McLachlan (1992) es una buena referencia general para métodos alternativos de discriminación. Veáse también Fukunaga (1990). La teoria de aprendizaje estadístico y su aplicación para la construcción de máquinas

de clasificación se presenta en Cherkassky y Mulier (1998) y Vapnik (2000).

Ejercicios

14.1 Escribir la transformación logit como la diferencia entre las funciones soporte del dato a clasificar bajo los dos modelos posibles.

14.2 Demostrar que el estimador de β_0 en el modelo logístico si $\beta_1 = ... = \beta_p = 0$ viene dado por $\widehat{\beta}_0 = \log \widehat{p}/(1-\widehat{p})$ donde $\widehat{p} = \sum y_i/n$. Interpretar este resultado.

14.3 Supongamos que los datos tienen observaciones repetidas de manera que para cada $\mathbf{x}_i$ se han observado n_i datos y r_i de ellos pertenecen a la población con $y = 1$ y $n_i - r_i$ a la de $y = 0$. Demostrar que la función soporte puede escribirse entonces como $L(\boldsymbol{\beta}) = \sum r_i \log p_i + \sum (n_i - r_i) \log(1 - p_i)$.

14.4 Demostrar que si las dos poblaciones bajo consideración son normales multivariantes con distiantas medias y matriz de covarianzas, la función logit para clasificar la observación $\mathbf{x}$ es $g(\mathbf{x}) = a + \mathbf{b}'\mathbf{x} + \mathbf{x}'\mathbf{A}\mathbf{x}$, determinar la expresión de $a, \mathbf{b}$ y $\mathbf{A}$ en función de los parámetros del modelo.

14.5 Demostrar que para clasificar una observación en las condiciones del ejemplo 14.2 la función logística es lineal en los parámetros y por lo tanto el modelo logístico puede aplicarse analogamente en este caso, aunque el número de parámetros a estimar es mayor.

14.6 Demostrar que en el caso de una única variable explicativa, si las observaciones de los dos grupos estan separadas, en el sentido de que todas las observaciones con $y = 0$ están en una zona $x_i < c$ y todas las de $y = 1$ en la zona $x_i > c$, la función soporte es $L(\boldsymbol{\beta}) = \sum_{x_i > c} \log p_i + \sum_{x_i < c} \log(1 - p_i)$ y por tanto puede tomar el valor máximo cero si hacemos $p_i = 1$ para las observaciones con $x_i > c$ y $p_i = 0$ para las observaciones con $x_i < c$.

14.7 Demostrar que el ejercicio 14.6 podemos aproximarnos arbitrariamente al valor $L(\boldsymbol{\beta}) = \mathbf{0}$ tomando $\widehat{\beta}_0 = -c\widehat{\beta}_1$ y haciendo que $\widehat{\beta}_1$ sea arbitrariamente grande.

14.8 Explicar, a la vista de los ejercicios 14.6 y 14. 7 porque el método de máxima verosimilitud va a fallar si todas las observaciones están perfectamente separadas, en el sentido definido en 14.6

Capítulo 15

CLASIFICACIÓN MEDIANTE MEZCLAS DE DISTRIBUCIONES

15.1 FUNDAMENTOS

En este capítulo volveremos al problema del análisis de conglomerados para analizar la homogeneidad de una muestra y encontrar grupos si existen, problema que ya estudiamos, desde un punto de vista descriptivo, en el capítulo 8. En este capítulo supondremos que los datos se han generado por una mezcla de G distribuciones desconocidas y preentaremos métodos para identificar los grupor. Hay tres métodos principales para partir una muestra heterogénea en grupos más homogéneos.

El primero, y más antiguo, es el método de k-medias, (o G medias en nuestra notación) que se presentó en el capítulo 8 como un algortimo heurístico para maximizar una medida de homogeneidad al partir la muestra en G grupos. En este capítulo veremos que las hipótesis que hagamos respecto a los componentes de la mezcla implican distintos criterios a maximizar con el algoritmo de k-medias.

El segundo, es tratar de estimar los parámetros de los componentes de la mezcla y clasificar después las observaciones a los grupos por sus probabilidades de pertenencia a las distintas poblaciones. En este capítulo estudiaremos con detalle el caso en el que los datos provienen de una mezcla de G poblaciones normales.

El tercero, es proyectar los puntos sobre distintas direcciones que separen los grupos lo más posible y clasificar las observaciones en grupos mediante estas proyecciones univariantes.

Los procedimientos de partición de la muestra en grupos o conglomerados a partir de la hipótesis de que los datos provienen de mezclas de distribuciones, están relacionados con el análisis discriminante que vimos en el capítulo 13. En ambos casos suponemos mezclas, y queremos encontrar criterios para asignar nuevas observaciones a las distintas poblaciones. Sin embargo, en análisis discriminante suponemos que las poblaciones son conocidas, o tenemos una muestra de cada población (a veces llamada muestra de entrenamiento), donde las observaciones están clasificadas sin error, de manera que podemos estimar los parámetros de cada distribución. En análisis de conglomerados ni conocemos el número de poblaciones ni disponemos de datos previos de clasificación, y toda la información sobre el número de grupos y su estructura debe obtenerse de la muestra disponible.

15.2 EL METODO de K-MEDIAS para mezclas

Para obtener criterios que podamos aplicar al caso de conglomerados, volvamos a revisar el problema de discriminar entre G poblaciones normales multivariantes $N\left(\boldsymbol{\mu}_g, \mathbf{V}_g\right)$, cuando se dispone de una muestra de entrenamiento donde se conoce la procedencia de las observaciones, y sea n_g al número de elementos de la muestra que provienen de la población g, donde $g = 1, \ldots, G$, y $\sum n_g = n$. Aplicando los resultados de la sección 10.2.2, la verosimilitud de la muestra será, sumando los soportes :

$$\log f\left(\mathbf{x}_1, \ldots \mathbf{x}_n\right) = -\sum_{g=1}^{G} \frac{n_g}{2} \log\left|\mathbf{V}_g\right| - \sum_{g=1}^{G} \frac{n_g}{2} tr\left(\mathbf{V}_g^{-1} \mathbf{S}\left(\boldsymbol{\mu}_g\right)\right)$$

donde $\mathbf{S}\left(\boldsymbol{\mu}_g\right) = \frac{1}{n_g} \sum_{i=1}^{n_g} \left(\mathbf{x}_i - \boldsymbol{\mu}_g\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_g\right)'$. Según esta ecuación la estimación de cada vector de medias, $\boldsymbol{\mu}_g$, será $\overline{\mathbf{x}}_g$, la media muestral y la función soporte concentrada en este parámetro será:

$$\log f\left(\mathbf{x}_1, \ldots \mathbf{x}_n\right) = -\sum_{g=1}^{G} \frac{n_g}{2} \log\left|\mathbf{V}_g\right| - \sum_{g=1}^{G} \frac{n_g}{2} tr\left(\mathbf{V}_g^{-1} \mathbf{S}_g\right)$$

donde

$$\mathbf{S}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_g\right)\left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_g\right)'.$$

Supongamos que admitimos la hipótesis $\mathbf{V}_g = \sigma^2 \mathbf{I}$, es decir las variables están incorreladas y tienen la misma varianza entre sí y en todos los grupos. Entonces, la función soporte se reduce a:

$$\log f\left(\mathbf{x}_1, \ldots \mathbf{x}_n\right) = -\frac{np}{2} \log \sigma^2 - \frac{1}{2\sigma^2} tr\left(\sum_{i=1}^{g} n_g \mathbf{S}_g\right)$$

y llamando

$$\mathbf{W} = \sum_{i=1}^{g} n_g \mathbf{S}_g$$

maximizar la verosimilitud supondría

$$\boxed{\min tr(\mathbf{W})}$$

que es el *criterio de la traza,* equivale a minimizar la suma ponderada de las varianzas estimadas en cada grupo. Este criterio se obtuvo, por otros métodos, en el capítulo 8, y es el que se utiliza en el algoritmo de k-medias. Tiene la ventaja de ser simple y fácil de calcular, pero no es invariante ante transformaciones lineales y no tiene en cuenta las correlaciones.

Si admitimos la hipótesis $\mathbf{V}_g = \mathbf{V}$, la verosimilitud es equivalente a la del problema de discriminación clásica estudiado en el capítulo 13, y viene dada por :

$$\log f(\mathbf{x}_1, \ldots \mathbf{x}_n) = -\frac{n}{2}\log|\mathbf{V}| - \frac{1}{2}tr\left(\mathbf{V}^{-1}\left(\sum_{i=1}^{g} n_g \mathbf{S}_g\right)\right)$$

y la estimación MV de $\mathbf{V}$ es entonces

$$\widehat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{g} n_g \mathbf{S}_g = \frac{1}{n}\mathbf{W},$$

e insertando esta estimación en la función de verosimilitud, maximizar la verosimilitud equivale a:

$$\boxed{\min|\mathbf{W}|}$$

que es el *criterio del determinante,* propuesto por Friedman and Rubin (1967). Este criterio si es invariante a transformaciones lineales, y, como veremos, tiende a identificar grupos elípticos.

En el caso general en que las poblaciones tienen distinta matriz de varianzas y covarianzas, la estimación MV de $\mathbf{V}_g$ es $\mathbf{S}_g$ y el máximo de la función de verosimilitud es

$$\log f(\mathbf{x}_1, \ldots \mathbf{x}_n) = -\frac{1}{2}\sum n_g \log|\mathbf{S}_g| - \frac{np}{2}, \tag{15.1}$$

y maximizar esta verosimilitud equivale a:

$$\min \sum n_g \log|\mathbf{S}_g| \tag{15.2}$$

En otros términos, cada grupo debe tener "volumen" mínimo. Suponemos que cada grupo tiene $n_g > p+1$, de manera que $|\mathbf{S}_g|$ sea no singular, lo que exige que $n > G(p+1)$.

Un criterio adicional propuesto por Friedman y Rubin (1967) es partir de la descomposición del análisis de la varianza multivariante y maximizar el tamaño de la distancia de Mahalanobis generaliza entre los grupos dada por $\mathbf{W}^{-1}\mathbf{B}$. De nuevo el tamaño de esta matriz puede medirse por la traza o el determinante, pero este criterio no ha dado buenos resultados en el análisis de conglomerados (vease Seber, 1984).

Cualquiera de estos criterios puede maximizarse con un algoritmo similar al k–medias que vimos en la capítulo 8. El criterio del determinante es fácil de implementar y, como se demuestra en el apéndice 15.1, tiende a producir grupos elípticos, mientras que el de la traza

produce grupos esféricos. El criterio (15.2) tiene el inconveniente de que es necesario imponer restricciones fuertes sobre el número de observaciones en cada grupo para que las matrices no sean singulares, ya que si el número de grupos es grande, el número de parámetros a estimar puede ser muy alto. En la práctica, parece mejor permitir algunos rasgos comunes en las matrices de covarianzas y esto no es fácil de imponer con este criterio. Además, si un grupo tiene pocas observaciones y $|\mathbf{S}_g|$ es casi singular, este grupo tendrá un peso desproporcionado en el criterio y el algoritmo tiende a caer es ese tipo de soluciones. Por esta razón ese criterio, aunque tiene interés teórico, se utiliza poco en la práctica.

15.2.1 Número de grupos

En la práctica el número de grupos, G, es desconocido y el algoritmo se calcula para distinto valores de G, $G = 1, 2, ...$ Comparar las soluciones obtenidas no es simple, porque cualquiera de los criterios disminuirá si aumentamos el número de grupos. En efecto, según el análisis de la varianza multivariante, la variabilidad total puede descomponerse como:

$$\boxed{\mathbf{T} = \mathbf{W} + \mathbf{B}} \tag{15.3}$$

Intuitivamente, el objetivo de la división en grupos es conseguir que $\mathbf{B}$, la variabilidad entre los grupos, sea lo más grande posible, mientras que $\mathbf{W}$, la variabilidad dentro de cada grupo, sea lo más pequeña posible. Dada una división cualquiera en grupos, si elegimos uno cualquiera de ellos podemos aplicarle de nuevo esta descomposición, con lo que reduciremos de nuevo la variabilidad descomponiendo más este grupo. Por tanto, no podemos utilizar ningún criterio basado en el tamaño de $\mathbf{W}$ para comparar soluciones con grupos distintos, ya que simpre podemos disminuir $\mathbf{W}$ haciendo más grupos.

Como vimos en el capítulo 8 podemos realizar un test F aproximado calculando la reducción proporcional de variabilidad que se obtiene aumentando un grupo adicional. El test es:

$$H = \frac{tr(\mathbf{W}_G) - tr(\mathbf{W}_{G+1})}{tr(\mathbf{W}_{G+1})/(n - G - 1)} \tag{15.4}$$

y, en la hipotesis de que G grupos son suficientes, el valor de H puede compararse con una F con $p, p(n - G - 1)$ grados de libertad. La regla de Hartigan (1975), implantada en algunos programas informáticos, es continuar dividiendo el conjunto de datos si este cociente es mayor que 10.

Un criterio adicional para seleccionar los grupos es el propuesto por Calinski y Harabasz (1974). Este criterio parte de la descomposición (15.3) y selecciona el valor de G maximizando

$$CH = \max \frac{tr(\mathbf{B})/(G - 1)}{tr(\mathbf{W})/(n - G)} \tag{15.5}$$

Ambos criterios parecen funcionar bien en las aplicaciones y están relacionados (veáse los ejercicios 15.1 y 15.2)

	$G=2$	$G=3$	$G=4$	$G=5$	$G=6$
eh	30	20	14	15	14
em	35	22	13	16	12
mi	509	230	129	76	83
tm	15	11	9	9	9
tn	64	58	37	35	26
Total	653	341	202	151	144
H		77.4	61.5	30.4	6.2
CH	265.9	296.3	356.6	359.7	302

Tabla 15.1: Tabla con la varianza promedio dentro de los grupos para cada variable con distinto número de grupos con el algoritmo de k-medias.

Ejemplo 15.1 *Vamos a comprar los dos criterios para seleccionar el número de grupos en el algoritmo de k-medias con el criterio de la traza para los datos de los países. Vamos a utilizar únicamente las 5 variables demográficas de MUNDODES. Comenzaremos con los resultados del programa SPSS. Para decidir el número de grupos este programa nos proporciona una tabla con las varianzas de cada variable dentro de los grupos. La tabla 15.1 resume esta información:*

Por filas esta tabla da la suma de cuadrados en todos los grupos para cada variable dividida por el número de grados de libertad de esta suma que es $n-G$, que es la información obtenida directamente del programa, com veremos a continuación. La table total es la suma de estas varianzas que es la traza de W. A continuación tenemos los estadísticos H y CH para determinar el número de grupos. Ambos criterios conducen a cinco grupos en este ejemplo.

Para ilustrar una salida estandar de un programa informatico, la tabla siguiente proprociona la salida del programa SSPS para 5 grupos. El programa proporciona los centros de los cinco grupos y la suma de cuadrados entre los grupos para cada variable dividida por sus grados de libertad, $G-1=4$. Esta es la columna Cluster MS. En la tabla de Análisis de la varianza tenemos también las varianzas de cada variable dentro de los grupos, que están en la columna Error MS. Esta columna es la que se ha copiado redondeada en la tabla 15.1, que contiene la suma de las varianzas de las variables.

```
Final Cluster Centers.

 Cluster    EH      EM        MI        TM       TN

       1  64.475  68.843     37.575    7.762   29.868
       2  43.166  46.033    143.400   20.022   46.577
       3  70.122  76.640     11.288    8.920   15.017
       4  57.342  60.900     74.578   10.194   39.057
       5  51.816  54.458    110.558   13.875   43.008

Analysis of Variance.

Variable       Cluster MS   DF         Error MS        DF            F   Prob
```

EH	1805.3459	4	15.059	86.0	119.8792	.000
EM	2443.0903	4	16.022	86.0	152.4812	.000
MI	46595.0380	4	76.410	86.0	609.7986	.000
TM	289.2008	4	9.507	86.0	30.4191	.000
TN	3473.4156	4	34.840	86.0	99.6950	.000

Tabla Salida del Programa SPSS para 5 grupos con los datos de MUNDODES\

Ejercicio 15.1 *Se observa en la tabla 15.1 que la variable mi tiene mucha más varianza que las demás, y, por tanto, va a tener un peso muy importante en la construcción de los grupos, que van a hacerse principalmente por los valores de esta variable. La tabla de las varianzas muestra que el número de grupos es cinco, ya que al aumentar a seis la disminución de las varianzas es muy pequeña.*

Para ilustrar el cálculo de los estadísticos para determinar el número de grupos, llamando MS(G) a la fila de totales en la tabla 15.1, esta fila será igual a $tr(\mathbf{W}_G)/(n-G)$*, y el estadistico se calcula como*

$$H = \frac{tr(\mathbf{W}_G) - tr(\mathbf{W}_{G+1})}{tr(\mathbf{W}_{G+1})/(n-G-1)} = \frac{(n-G)MS(G) - (n-G-1)MS(G+1)}{MS(G+1)}$$

Ejemplo 15.2 *donde* $n = 91$ *y* G *es el número de grupos indicado por columnas. Así se obtiene la fila de H, y de acuerdo con el criterio de Hartigan escogeríamos cinco grupos. .*

Esta tabla incluye también la información para calcular el criterio (15.5). El numerador de esta expresión es la suma de los términos cluster MS para todas las variables y el denominador la suma de la columna Error MS. Para G=5 el criterio CH es

$$CH = \frac{1805.34 + .. + 3473.42}{15.06 + ... + 34.84} = \frac{54606}{151.8} = 359.7$$

y la aplicación de este criterio lleva también a cinco grupos.

Comparando las medias de la solución para cinco grupos, vemos que el grupo con menor mortalidad infantil es el tres, que incluye los paises de Europa menos Albania, y el de mayor mortalidad el dos con los países más pobres de Africa. La figura 15.1 presenta un histograma de la variable mi. Se observa que esta variable que va a tener un peso dominante en la formación de los grupos indica claramente la heterogeneidad de la muestra.

Para ilustra el funcionamiento de distintos programas la tabla siguiente indica la salida del programa MINITAB para cinco grupos con las variables sin estandarizar, y estandarizadas

A. Resultados de MUNDODES, variables sin estandarizar. MINITAB

Number of	Within cluster	Average distance	Maximum distance
	observations	sum of squares	from centroid	from centroid
Cluster1	21	10060.590	19.308	57.601
Cluster2	14	797.147	7.200	10.897

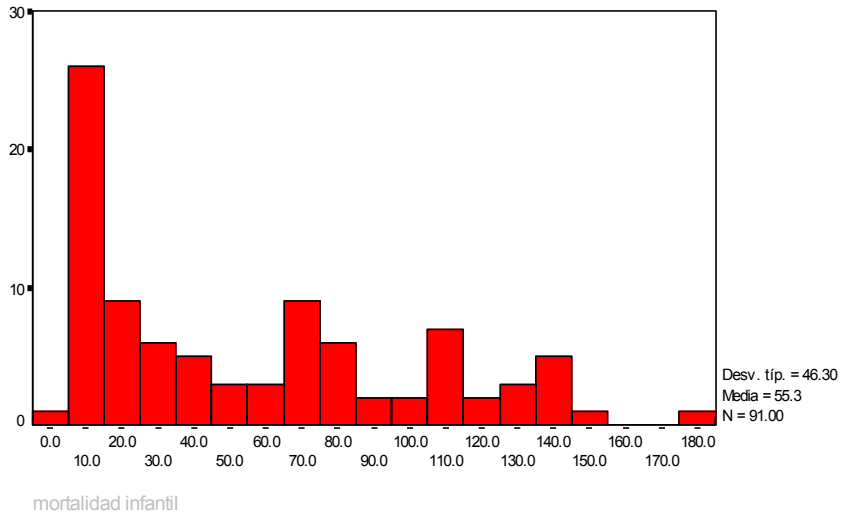


Figura 15.1: Histograma de la variable mortalidad infantil indicando la presencia de entre cuatro y cinco grupos de paises

Cluster3	28	829.039	5.005	10.008
Cluster4	9	826.444	8.724	15.306
Cluster5	19	2713.755	11.338	19.143

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
tn	4.4762	2.7857	2.7143	3.3333	4.3684
tm	44.5381	22.8571	13.4429	34.1222	39.0579
mi	16.5095	6.4714	9.4250	9.1000	10.1947
eh	124.6333	23.5500	9.1143	45.7111	74.5789
em	48.1095	67.3643	70.7464	62.4333	57.3421

B: Resultados de MUNDODES, variables estandarizadas

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	20	14.440	0.817	1.275
Cluster2	10	9.932	0.736	2.703
Cluster3	29	20.771	0.792	1.535
Cluster4	22	32.443	1.134	2.132
Cluster5	10	6.679	0.727	1.621

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
tn	0.6955	-1.7841	-0.1214	0.9070	-1.2501
tm	0.6007	-0.9921	-0.9665	1.2233	-0.0978
mi	-0.3585	0.3087	-0.5867	1.3417	-0.8421
eh	0.3300	-0.7676	-0.9455	1.3758	-0.1771
em	-0.2078	0.5478	0.9537	-1.4424	0.2754

Salida del programa MINITAB para 5 grupos con los datos de MUNDODES
con y sin estandarizar.

Ejemplo 15.3 *Este programa nos da la suma de cuadrados dentro de los grupos por clusters (grupos) en lugar de por variables. Los resultados para datos sin estandarizar son parecido pero no idénticos, como puede verse comparando las medias de las variables en los grupos. Al estandarizar las variables los resultados cambian sustancialmente, al tener un peso mayor el resto de las variables. Los grupos son más homogéneos por continentes y en Europa se separan los paises occidentales y los orientales.*

15.3 ESTIMACIÓN DE MEZCLAS DE NORMALES

Un enfoque natural para realizar la subdivisión de la muestra en grupos o conglomerados es suponer que los datos se han generado como una mezcla de distribuciones normales multivariantes y estimar conjuntamente los parámetros de las distribuciones que forman la mezcla y las probabilidades a posteriori de cada dato de pertenecer a cada una de los componentes de la mezcla. Vamos a presentar este enfoque.

15.3.1 Las ecuaciones de máxima verosimilitud para la mezcla

Supongamos que los datos provienen de una mezcla de distribuciones

$$f(\mathbf{x}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}),$$

la función de verosimilitud será

$$l(\boldsymbol{\theta}|\mathbf{X}) = \prod_{i=1}^{n} (\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i))$$

y puede escribirse como la suma de G^n términos correspondientes a todas las posibles clasificaciones de la n observaciones entre los G grupos. La función soporte de la muestra será

$$L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i) = \sum_{i=1}^{n} \log \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i) \tag{15.6}$$

Supongamos que cada $f_g(\mathbf{x})$ es normal k-dimensional con vector de medias $\boldsymbol{\mu}_g$ y matriz de covarianzas $\mathbf{V}_g$, de manera que $\boldsymbol{\theta} = (\pi_1, ..., \pi_G, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_G, \mathbf{V}_1, ..., \mathbf{V}_G)$. Sustituyendo estas densidades por su expresión, la verosimilitud será

$$L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^{n} \log(\sum_{g=1}^{G} \pi_g \left|\mathbf{V}_g\right|^{-1/2} (2\pi)^{-p/2} \exp(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_g)'\mathbf{V}_g^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g)). \tag{15.7}$$

Observemos que si hacemos en esta función $\widehat{\boldsymbol{\mu}}_g = \mathbf{x}_i$, la estimación de $\mathbf{V}_g$ es cero y si $\pi_g \neq 0$, el cociente $\pi_g \left|\mathbf{V}_g\right|^{-1/2}$ tiende a infinito y también lo hará la función soporte. Por tanto, esta función tiene muchos máximos, ligados a soluciones donde cada densidad viene determinada exactamente por una observación. Para evitar estas singularidades supondremos que, como mínimo, hay p observaciones de cada modelo, y trataremos de encontrar un máximo local de esta función que proporcione un estimador consistente de los parámetros.

Un problema adicional de esta función de verosimilitud es que las distribuciones normales no están identificadas, ya que el orden $1, ..., G$ es arbitrario. Para resolver este problema podemos suponer que las distribuciones $1, ..., G$ corresponden a $\pi_1 \geq \pi_2 \geq ... \geq \pi_G$ o definir el orden de las distribuciones por una medida del tamaño de la media o la matriz de covarianzas.

Para maximizar esta función con relación a las probabilidades π_i hay que tener en cuenta que $\sum_{g=1}^{G} \pi_g = 1$. Introduciendo esta restricción con un multiplicador de Lagrange en (15.6), la función a maximizar es

$$L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^{n} \log \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i) - \lambda(\sum_{g=1}^{G} \pi_g - 1). \tag{15.8}$$

Derivando respecto a las probabilidades:

$$\frac{\partial L(\boldsymbol{\theta}|\mathbf{X})}{\partial \pi_g} = \sum_{i=1}^{n} \frac{f_g(\mathbf{x}_i)}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)} - \lambda = 0$$

y multiplicando por π_g, supuesto $\pi_g \neq 0$ ya que en otro caso el modelo g es redundante, podemos escribir

$$\lambda \pi_g = \sum_{i=1}^{n} \pi_{ig} \tag{15.9}$$

donde hemos llamado π_{ig} a :

$$\boxed{\pi_{ig} = \frac{\pi_g f_g(\mathbf{x}_i)}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i)}} \tag{15.10}$$

Estos coeficientes representan la probabilidad de que una vez observado el dato $\mathbf{x}_i$ haya sido generada por la normal $f_g(\mathbf{x})$. Estas probabilidades se denominan a posteriori y se calculan por el teorema de Bayes. Su interpretación es la siguiente. Antes de observar $\mathbf{x}_i$ la probabilidad de que cualquier observación, y en particular la $\mathbf{x}_i$, venga de la clase g es π_g. Sin

embargo, después de observar $\mathbf{x}_i$, esta probabilidad se modifica en función de lo compatible que sea este valor con el modelo g. Esta compatibilidad se mide por $f_g(\mathbf{x}_i)$: si este valor es relativamente alto, aumentará la probabilidad de que venga del modelo g. Naturalmente para cada dato $\sum_{g=1}^{G} \pi_{ig} = 1$.

Para determinar el valor de λ, sumando (15.9) para todos los grupos

$$\lambda = \sum_{i=1}^{n} \sum_{g=1}^{G} \pi_{ig} = n$$

y sustituyendo en (15.9), las ecuaciones para estimar las probabilidades a priori son

$$\boxed{\widehat{\pi}_g = \frac{1}{n} \sum_{i=1}^{n} \pi_{ig}.} \tag{15.11}$$

que proporcionan las probabilidades a priori como promedio de las probabilidades a posteriori.

Vamos a calcular ahora las estimaciones de los parámetros de las distribuciones normales. Derivando la función soporte respecto a las medias:

$$\frac{\partial L(\boldsymbol{\theta}|\mathbf{X})}{\partial \boldsymbol{\mu}_g} = \sum_{i=1}^{n} \frac{\pi_g f_g(\mathbf{x}) \mathbf{V}_g^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_g)}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x})} = 0 \quad g = 1, ..., G$$

que puede escribirse como

$$\boxed{\widehat{\boldsymbol{\mu}_g} = \sum_{i=1}^{n} \frac{\pi_{ig}}{\sum_{i=1}^{n} \pi_{ig}} \mathbf{x}_i.} \tag{15.12}$$

La media de cada distribución se estima como una media ponderada de todas las observaciones con pesos $\omega_i = \pi_{ig} / \sum_{i=1}^{n} \pi_{ig}$, donde $\omega_{ig} \geq 0$, y $\sum_{i=1}^{n} \omega_{ig} = 1$. Los pesos, ω_{ig}, representan la probabilidad relativa de que la observación i pertenezca a la población g. Análogamente, derivando respecto a $\mathbf{V}_g$ y utilizando los resultados de la sección 10.2 obtenemos que:

$$\boxed{\widehat{\mathbf{V}_g} = \sum_{i=1}^{n} \frac{\pi_{ig}}{\sum_{i=1}^{n} \pi_{ig}} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}_g})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}_g})'} \tag{15.13}$$

que tiene un interpretación similar, como promedio de desviaciones de los datos respecto a sus medias, con pesos proporcionales a las probabilidades a posteriori.

Para resolver estas ecuaciones (15.11), (15.12) y (15.13) y obtener los estimadores necesitamos las probabilidades π_{ig}, y para calcular estas probabilidades con (15.10) necesitamos los parámetros del modelo. Intuitivamente podríamos iterar entre ambas etapas y esta es la solución que se obtiene con el algoritmo EM.

15.3.2 Resolución mediante el algoritmo EM

Para aplicar el algoritmo EM transformemos el problema introduciendo un conjunto de variables vectoriales no observadas $(\mathbf{z}_1, ..., \mathbf{z}_n)$, que tienen como función indicar de qué modelo

proviene cada observación. Con este objetivo, $\mathbf{z}_i$ será una variable binaria vectorial $G \times 1$ que tendrá sólo un componente igual a uno, el correspondiente al grupo del que proviene el dato $\mathbf{x}_i$, y todos los demás igual a cero. Por ejemplo $\mathbf{x}_i$ vendrá de la población 1 si $z_{i1} = 1$ y $z_{i2} = z_{i2} = ... = z_{iG} = 0$. Se verificará que $\sum_{g=1}^{G} z_{ig} = 1$ y $\sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} = n$. Con estas nuevas variables, la función de densidad de $\mathbf{x}_i$ condicionada a $\mathbf{z}_i$ puede escribirse

$$f(\mathbf{x}_i/\mathbf{z}_i) = \prod_{g=1}^{G} f_g(\mathbf{x}_i)^{z_{ig}}. \tag{15.14}$$

En efecto, en $\mathbf{z}_i$ solo un componente z_{ig} es distinto de cero y ese componente definirá cual es la función de densidad de las observaciones. Análogamente, la función de probabilidades de la variable $\mathbf{z}_i$ será

$$p(\mathbf{z}_i) = \prod_{g=1}^{G} {\pi_g}^{z_{ig}}. \tag{15.15}$$

Por otro lado, la función de densidad conjunta es

$$f(\mathbf{x}_i, \mathbf{z}_i) = f(\mathbf{x}_i/\mathbf{z}_i)p(\mathbf{z}_i),$$

que, por (15.14) y (15.15), podemos escribir

$$f(\mathbf{x}_i, \mathbf{z}_i) = \prod_{g=1}^{G} (\pi_g f_g(\mathbf{x}_i))^{z_{ig}}$$

El soporte de la verosimilitud conjunta es

$$L_C(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i, \mathbf{z}_i) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log \pi_g + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log f_g(\mathbf{x}_i) \tag{15.16}$$

Si las variables z_{ig} que definen la población de la que proviene cada dato fueran conocidas, la estimación de los parámetros es inmediata, y la hemos comentado en el problema de análisis discriminante. La media de cada componente se estima como promedio de las observaciones generadas por el componente, que puede escribirse

$$\widehat{\boldsymbol{\mu}}_g = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \mathbf{x}_i,$$

y la matriz de covarianzas de cada grupo se calculará teniendo en cuenta sólo las observaciones de ese grupo mediante

$$\widehat{\mathbf{V}_g} = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} (\mathbf{x}_i - \overline{\mathbf{x}}_g)(\mathbf{x}_i - \overline{\mathbf{x}}_g)'.$$

Sin embargo, el problema es que ahora las variables de clasificación no son conocidas. La solución que proporciona el algoritmo EM es estimar las variables z_{ig} mediante las probabilidades a posteriori, y después utilizar estas fórmulas.

El algoritmo EM comienza con una estimación inicial $\widehat{\boldsymbol{\theta}}^{(0)}$. En el paso E calcularemos el valor esperado de las observaciones ausentes en la verosimilitud completa (15.16) condicionando a los parámetros iniciales y a los datos observados. Como la verosimilitud es lineal en z_{ig}, esto equivale a sustituir las variables ausentes por sus esperanzas. Las variables ausentes, z_{ig}, son variables binomiales con valores 0,1, y

$$E(z_{ig}/\mathbf{X},\widehat{\boldsymbol{\theta}}^{(0)}) = p(z_{ig} = 1/\mathbf{X},\widehat{\boldsymbol{\theta}}^{(0)}) = p(z_{ig} = 1/\mathbf{x}_i,\widehat{\boldsymbol{\theta}}^{(0)}) = \widehat{\pi}_{ig}^{(0)}$$

donde $\widehat{\pi}_{ig}^{(0)}$ es la probabilidad de que la observación $\mathbf{x}_i$ venga del modelo j cuando ya se ha observado $\mathbf{x}_i$ y los parámetros de los modelos son los dados por $\widehat{\boldsymbol{\theta}}^{(0)}$. Estas son las probabilidades a porteriori que se calculan por (15.10) utilizando como valores de los parámetros los especificados en $\widehat{\boldsymbol{\theta}}^{(0)}$. Al sustituir las variables ausentes por sus esperanzas se obtiene

$$L_C^*(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^{n}\sum_{g=1}^{G} \widehat{\pi}_{ig}^{(0)} \log \pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G} \pi_{ig}^{(0)} \log f_g(\mathbf{x}_i)$$

En la etapa M se maximiza esta función respecto a los parámetros $\boldsymbol{\theta}$. Observemos que los parámetros π_g aparecen sólo en el primer término y los de las normales sólo en el segundo. Podemos pues obtenerlos independientemente. Comenzando por los π_g, estos parámetros están sujetos a que su suma debe ser uno, por lo que la función a maximizar es

$$\sum_{i=1}^{n}\sum_{g=1}^{G} \widehat{\pi}_{ig}^{(0)} \log \pi_g - \lambda(\sum_{g=1}^{G} \pi_g - 1)$$

que conduce a (15.11) con los valores π_{ig} ahora fijos a $\widehat{\pi}_{ig}^{(0)}$. Para obtener los estimadores de los parámetros de la normal, derivando el segundo término se obtienen las ecuaciones (15.12) y (15.13), donde ahora las probabilidades π_{ig} son iguales a $\widehat{\pi}_{ig}^{(0)}$. La resolución de estas ecuaciones conduce a un nuevo vector de parámetros, $\widehat{\boldsymbol{\theta}}^{(1)}$, y el algoritmo se itera hasta obtener convergencia. En resumen, el algoritmo es:

1. Partir de un valor $\widehat{\boldsymbol{\theta}}^{(0)}$ y calcular $\widehat{\pi}_{ig}^{(0)}$ con (15.10)
2. Resolver (15.11), (15.12) y (15.13) para obtener $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\mu}},\widehat{\mathbf{V}})$
3. Volver con este valor a 1 e iterar 1 y 2 hasta convergencia.

15.3.3 Aplicación al análisis de conglomerados

Se han propuesta distintas implementaciones de las mezclas de normales para resolver problemas de conglomerados. En nuestra opinión el método más prometedor es debido a Banfield y

Raftery (1993) y Dasgupta y Raftery (1988), que han diseñado un método basado en mezclas de distribuciones normales y un algoritmo, MCLUST, que funciona bien en la práctica. La bases del procedimiento son comenzar el algoritmo EM con una estimación inicial obtenida mediante análisis jerárquico y reparameterizar las matrices de covarianzas para que puedan tener partes comunes y partes específicas. En síntesis el procedimiento consiste en :

1. Seleccionar un valor M para el máximo número de grupos y reparametrizar las distribuciones normales que forman los grupos.

2. Estimar los parámetros de la mezcla con el algoritmo EM para $G = 1, ..., M$. Las condiciones iniciales del algoritmo se establecen con un método jerárquico de los estudiados en el capítulo 9.

3. Seleccionar el número de grupos por el criterio BIC.

Vamos a analizar cada una de estas etapas

Reparametrizaciones

Un inconveniente de la parametrización habitual de las mezclas de normales es que las matrices de covarianzas se suponen o bien iguales, resultando en $p(p+1)/2$ parámetros, o bien desiguales, con $Gp(p+1)/2$ parámetros. Si la dimensión del espacio es grande y podemos tener mucho grupos, suponer que todas las matrices son distintas puede implicar un número gigantesco de parámetros que puede hacer la estimación mediante el algoritmo EM muy lenta e incluso impracticable.

Una solución propuesta por Banfield y Raftery (1993) es parametrizar las matrices de covarianza por su descomposición espectral como

$$\mathbf{V}_g = \lambda_g \mathbf{C}_g \mathbf{A}_g \mathbf{C}_g'$$

donde $\mathbf{C}_g$ es una matriz ortogonal con los vectores propios de $\mathbf{V}_g$ y $\lambda_g \mathbf{A}_g$ es la matriz de valores propios, siendo el escalar λ_g el valor mayor propio de la matriz. Recordemos que los vectores propios de la matriz indican orientación, mientras que los valores propios indican tamaño, en el sentido de volumen ocupado en el espacio por el grupo. De esta forma podemos permitir que las orientaciones de ciertos grupos sean distintas, o que el tamaño de otros sea distinto. Por ejemplo, podemos suponer que el tamaño es el mismo pero las orientaciones son diferentes. Entonces $\mathbf{V}_g = \lambda \mathbf{C}_g \mathbf{A} \mathbf{C}_g'$. No entraremos en los detalles, que el lector puede consultar en Dasgupta y Raftery (1998) y las referencias que allí se indican.

Número de Grupos

El criterio para seleccionar el número de grupos es minimizar el BIC. Vimos en la sección 10.5.3 que el criterio de Schwartz aproxima las probabilidades a posteriori de cada modelo. Sustituyendo la expresión de la verosimilitud en el máximo de la mezcla de normales en la expresión del BIC y eliminando constantes, este criterio en este caso equivale a:

$$BIC = \min \sum n_g \log |\mathbf{S}_g| + n\,(p, G) \ln n$$

donde $n\,(p, G)$ es el número de parámetros en el modelo. Conviene indicar que, aunque este criterio parece funcionar bien en la práctica para escoger el número de grupos, las hipótesis

de regularidad que se efectúan para deducir el BIC como aproximación de la probabilidad a posteriori no se verifican en el caso de las mezclas, por lo que este criterio puede aplicarse como una guía y no como una regla automática. El criterio AIC es

$$AIC = \min \sum n_g \log |\mathbf{S}_g| + n\,(p, G)\, n$$

15.4 MÉTODOS BAYESIANOS

15.4.1 Estimación Bayesiana de Mezclas de Normales

El enfoque bayesiano puede aplicarse a la estimación de mezclas de distribuciones. Hemos visto en la sección 15.3 que la verosimilitud de una mezcla contiene G^n términos, correspondientes a las posibles asignaciones de las n observaciones de la muestra a las G poblaciones posibles. Al multiplicar por la prior, $p(\boldsymbol{\theta})$, la posterior tendrá también G^n términos y salvo en el caso de n muy pequeño es inmanejable.

La introducción de variables faltantes ha permitido resolver la estimación mediante el algoritmo EM. Este mismo enfoque lleva a una solución rápida del problema mediante Muestreo de Gibbs. Introduciendo las variables no observadas, $\mathbf{z}_i$, tenemos que, conocida $\mathbf{z}_i$, la densidad de $\mathbf{x}_i$ es normal multivariante, con parámetros determinados por la componente de $\mathbf{z}_i$ igual a uno. Podemos escribir

$$f(\mathbf{x}_i|\mathbf{z}_i\boldsymbol{\theta}) \sim N_p(\prod_{g=1}^{G} \boldsymbol{\mu}_i^{z_{ig}}, \prod_{g=1}^{G} \mathbf{V}_i^{z_{ig}})$$

por otra parte, la variable $\mathbf{z}_i$ tiene una distribución multinomial con parámetros

$$f(\mathbf{z}_i|\boldsymbol{\theta}) \sim M_G(1; \pi_{i1}, ..., \pi_{iG}).$$

Estas dos funciones determinan la verosimilitud de la muestra

$$\ell(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}) = \prod_{i=1}^{n} f(\mathbf{x}_i|\mathbf{z}_i\boldsymbol{\theta}) f(\mathbf{z}_i|\boldsymbol{\theta})$$

que será el producto de $2n$ términos: n distribuciones normales con parámetros determinados por las componentes de $\mathbf{z}_i$ distintos de cero y n coeficientes π_{ig}, determinados también por los $\mathbf{z}_i$.

Los problemas de singularidad indicados al estudiar la verosimilitud de distribuciones mezcladas se acentúan si tomamos distribuciones impropias, por lo que en la estimación de mezclas conviene tomar distribuciones propias. Eligiendo distribuciones conjugadas, tomaremos una distribución de Dirichlet sobre las proporciones de la mezcla, una normal para la media dada la varianza, y una de Wishart para la precisión. Es decir, a priori

$$p(\boldsymbol{\pi}) \sim \mathbf{D}(\boldsymbol{\alpha})$$

$$p(\boldsymbol{\mu}_i|\mathbf{V}_i^{-1}) \sim N_p(\boldsymbol{\mu}_{i0}, \mathbf{V}_i/n_{i0})$$

$$p(\mathbf{V}_i^{-1}) \sim W_p(m_{i0}, \mathbf{M}_i/m_{i0})$$

La posterior de los parámetros y las observaciones faltantes dados los datos será

$$p(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{X}) = f(\mathbf{X}|\boldsymbol{\theta}, \mathbf{Z}) f(\mathbf{Z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}) p(\boldsymbol{\theta})$$

Podemos aplicar el Muestro de Gibbs para obtener muestras de esta distribución. La idea es muestrear iterando entre las dos distribuciones condicionadas $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z})$ y $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$. En la primera suponemos una asignación de las observaciones entre los grupos y obtenemos la estimación de los parámetros. Esta distribución es fácil de obtener como veremos, ya que cada observación queda identificada dentro de un grupo. En la segunda suponemos un valor para los parámetros y calculamos la probabilidad de que cada observación venga de cada grupo. De nuevo esta distribución es fácil de obtener. Al final, tendremos un conjunto de muestras de Monte Carlo de estas distribuciones, $(\boldsymbol{\theta}^{(1)}, \mathbf{Z}^{(1)}), ..., (\boldsymbol{\theta}^{(N)}, \mathbf{Z}^{(N)})$. Los valores $(\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)})$ permitirán estimar la distribución a posteriori de los parámetros dados los datos, mientras que las secuencias $(\mathbf{Z}^{(1)}, ..., \mathbf{Z}^{(N)})$ proporcionaran las probabilidad a posteriori de que cada observación pertenezca a cada grupo, dados los datos.

Comencemos con el muestreo de $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z})$. Fijado $\mathbf{Z} = \mathbf{Z}^{(h)}$, podemos, para cada grupo, obtener una muestra de valores de los parámetros como sigue:

(1) Para cada media con

$$p(\boldsymbol{\mu}_g|\mathbf{V}_g^{-1}, \mathbf{X}, \mathbf{Z}^{(h)}) \sim N_p(\boldsymbol{\mu}_{gp}, \mathbf{V}_{gp})$$

donde la media a posteriori se calcula de la forma habitual, pero utilizando solo las observaciones que han sido clasificadas en $\mathbf{Z}^{(g)}$ como pertenecientes al grupo g :

$$\boldsymbol{\mu}_{gp} = \frac{n_{g0}\boldsymbol{\mu}_{g0} + n_g(\mathbf{Z}^{(h)})\bar{\mathbf{x}}_g(\mathbf{Z}^{(h)})}{n_{g0} + n(\mathbf{Z}^{(h)})}$$

donde $n_g(\mathbf{Z}^{(h)})$ es el número de observaciones en el grupo g dado por

$$n_g(\mathbf{Z}^{(h)}) = \sum_{i=1}^{n} z_{ig}$$

y $\bar{\mathbf{x}}_g(\mathbf{Z}^{(h)})$ la media de estas observaciones, dada por

$$\bar{\mathbf{x}}_g(\mathbf{Z}^{(h)}) = \frac{\sum_{i=1}^{n} z_{ig}\mathbf{x}_i}{\sum_{i=1}^{n} z_{ig}}.$$

Análogamente, la varianza de la posterior será

$$\mathbf{V}_{gp} = \frac{\mathbf{V}_g}{n_{g0} + n_g(\mathbf{Z}^{(h)})}.$$

(2) Para las matrices de precisión con

$$p(\mathbf{V}_g^{-1}|\mathbf{X}, \mathbf{Z}^{(g)}) \sim W_p(m_{gp}, \mathbf{M}_{gp})$$

donde

$$m_{gp} = n_{g0} + m_{g0}$$

y

$$\mathbf{M}_{gp}^{-1} = m_{g0}\mathbf{M}^{-1} + n_g(\mathbf{Z}^{(h)})\mathbf{S}_g + \frac{n_g(\mathbf{Z}^{(h)})n_{g0}}{n_g(\mathbf{Z}^{(h)}) + n_0}(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_{g0})(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_{g0})',$$

y la varianza muestral en cada grupo se estima con las observaciones de ese grupo por

$$n_g(\mathbf{Z}^{(h)})\mathbf{S}_g = \sum_{i=1}^{n} z_{ig}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'.$$

(3) Para las probabilidades con

$$p(\pi_g|\mathbf{X}, \mathbf{Z}^{(\mathbf{g})}) \sim \mathbf{D}(\alpha_1 + n_1(\mathbf{Z}^{(h)}), ..., \alpha_G + n_G(\mathbf{Z}^{(h)})).$$

Una vez obtenido el vector de parámetros $\boldsymbol{\theta}^{(g)}$ obtendremos un nuevo valor de $\mathbf{Z} = \mathbf{Z}^{(g+1)}$ mediante:

(4) Simular $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_n)$ de

$$f(\mathbf{z}_i|\mathbf{X}, \boldsymbol{\theta}^{(\mathbf{g})}) \sim M_G(1; \pi_{i1}(\boldsymbol{\theta}^{(\mathbf{g})}), ..., \pi_{iG}(\boldsymbol{\theta}^{(\mathbf{g})})),$$

donde las probabilidades a posteriori de las observaciones vienen dadas por:

$$\pi_{ig}(\boldsymbol{\theta}^{(\mathbf{g})}) = \frac{\pi_g f_g(\mathbf{x}_i|\boldsymbol{\theta}^{(\mathbf{g})})}{\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i|\boldsymbol{\theta}^{(\mathbf{g})})}.$$

Un problema adicional en la estimación mediante muestreo de Gibbs es la falta de identificación de los componentes de la mezcla, que señalamos en la sección 15.3.3. Una posibilidad es introducir un orden entre las distribuciones, pero esta solución puede no ser siempre adecuada. Vease Stephens (2000) y Celeux y otros (2000) para discusiones recientes de este problema.

15.5 MÉTODOS DE PROYECCIÓN

Una alternativa a los métodos anteriores es buscar direcciones de proyección de los datos donde puedan aparecer los distintos grupos y después buscar los grupos sobre estas direcciones univariantes. Alternativamente podemos proyectar sobre dos direcciones, un plano, y buscar grupos en el plano. La ventaja de este enfoque es que no necesitamos especificar a priori el número de grupos, ni comparar soluciones con número de grupos muy distintos.

Una intuición sobre las posible buenas direcciones de proyección nos lo proporciona el siguiente resultado: supongamos que tenemos una muestra donde cada dato puede venir de una de dos poblaciones normales que tienen la misma matriz de covarianzas, que es desconocida, y medias distintas, también desconocidas. La procedencia de cada observación

es también desconocida. Se desea encontrar una dirección de proyección de las observaciones que nos proporcione, si es posible, la máxima separación entre las poblaciones. Puede demostrarse que aunque no conocemos la procedencia de cada obervación es posible clasificar con la función lineal discriminante, que sabemos es óptima para clasificar en este problema. Si suponemos que la probabilidad de que cada dato venga de cada una de las dos poblaciones es la misma, Peña y Prieto (2000) han demostrado que entonces la dirección que minimiza la kurtosis de la proyección es la función lineal discriminante de Fisher (véase el apéndice 15.3). Este resultado puede extenderse a varias poblaciones: las reglas óptimas de clasificación de Fisher se obtiene minimizando la kurtosis de las proyecciones.

Este resultado sugiere separar los grupos buscando las direcciones que nimimizan la kurtosis y proyectando los datos sobre estas direcciones. Como es posible que además de los grupos existan datos atípicos aislados, o grupos muy pequeños alejados del resto, y hemos visto en el capítulo 4 que estos atípicos se manifiestan en las direcciones con máxima kurtosis, podemos pensar en un algortimo que incluya en la búsqueda también estas direcciones. Esto conduce al siguiente método propuesto por Peña y Prieto (2001b):

(1) Comenzar proyectando los datos sobre las direcciones que maximizan el coeficiente de kurtosis de los datos proyectados. A continuación proyectar los datos sobre el espacio ortogonal a la direccion encontrada y seleccionar ahora la dirección sobre ese subespacio donde se maximiza el coeficiente de kurtosis. Repetir el proceso de proyección ortogonal a las direcciones ya encontradas y selección de una nueva dirección maximizando la kurtosis. De esta manera se obtienen p direcciones ortogonales de proyección.

(2) Repetir el cálculo de direcciones de (1) pero ahora buscando direcciones que minimizan el coeficiente kurtosis.

(3) Explorar cada una de estas $2p$ direcciones para encontrar grupos y datos atípicos de la forma siguiente. Se obtienen los estadísticos ordenados de los datos proyectados y se consideran los saltos entre estadísticos ordenados. Si los datos proyectados provienen de una distribución unimodal, estos saltos deben tener una pauta conocida, con saltos grandes en los extremos y pequeños saltos en el centro de la distribución.

Para decidir cuando se produce un salto sobre una dirección se utilizan las propiedades de los estadísticos de orden. Puede demostrarse que si tenemos una muestra de datos normales y los transformamos con su función de distribución, los estadísticos ordenados de los datos transformados tienen una distribución uniforme, y entonces puede estudiarse fácilmente la distribución de los espacios o saltos. Por esta razón el procedimiento estandariza los datos proyectados con la función de distribución inversa de la normal univariante antes de comenzar a buscar saltos en los datos proyectados. El algoritmo para explorar las direcciones es:

1. Para cada dirección $\mathbf{d}_k$, $k = 1, \ldots, 2p$, se calcula la proyección de los datos sobre ella mediante $u_{ki} = \mathbf{x}_i'\mathbf{d}_k$.

2. Se estandarizan las observaciones , $z_{ki} = (u_{ki} - m_k)/s_k$, donde $m_k = \sum_i u_{ki}/n$ es la media de las proyecciones y $s_k^2 = \sum_i (u_{ki} - m_k)^2/(n-1)$ su varianza.

3. Se ordenan las proyecciones z_{ki} para cada k, y se obtienen los estadísticos ordenados $z_{k(i)}$. A continuación estos estadísticos se transforman con la función de distribución inversa de la normal estandar $\bar{z}_{ki} = \Phi^{-1}(z_{k(i)})$.

4. Se calculan los espacios, que son las diferencias entre valores consecutivos de los estadísticos de orden transformados $\gamma_{ki} = \bar{z}_{k,i+1} - \bar{z}_{ki}$.

5. Se buscan valores altos de los espacios γ_{ki}, que corresponderán a huecos en la distribución de los datos. En efecto, si el estadístico de orden trece es 10 y el de orden catorce 20 el espacio es 20-10=10 y si este valor es mucho mayor que los otros espacios indica que hay un hueco sin datos entre estos valores, que puede corresponder a la sepación entre dos grupos de datos. Un valor alto del espacio indicará la presencia de más de un grupo de datos. Para determinar los valores altos de γ_{ki} introducimos una constante, $\kappa = 1 - 0.1^{1/n}/p^{10/(3n)}$, donde κ se calcula a partir de la distribución de los espacios (veáse Peña y Prieto para los detalles), y decidimos que comienza un nuevo grupo de datos cuando $\gamma_{kj} > \kappa$. En concreto, definimos $i_{0k} = 0$ y calculamos

$$r = \inf_j \{n > j > i_{0k} : \gamma_{kj} > \kappa\}.$$

   Si $r < \infty$, esto indica la presencia de varios grupos, en otro caso se pasa a analizar la dirección siguiente.

6. Marcar todas las observaciones l que verifican $\bar{z}_{kl} \leq \bar{z}_{kr}$ como pertenecientes a grupos diferentes que las que verifican $\bar{z}_{kl} > \bar{z}_{kr}$. Hacer $i_{0k} = r$ y volver a 5 para repetir la búsqueda de huecos en los datos.

Despues de este análisis se realiza un paso final para asignar las observaciones a los grupos identificados, como sigue.

1. Sea G el número de grupos identificados, se ordenan los grupos por número de observaciones de manera que el grupo 1 es el mayor y el G el menor. Suponemos que las observaciones se han renumerado de manera que ahora las observaciones $i_{g-1} + 1$ to i_g pertenecen al grupo g ($i_0 = 0$ and $i_G = n$).

2. Para cada grupo $g = 1, \ldots, G$:

   (a) Se calcula la media y la matriz de covarianzas de las observaciones del grupo si hay la menos $p + 1$ datos.

   (b) Calcular las distancias de Mahalanobis para todas las observaciones que no están en el grupo g,

$$D_j^2 = (\mathbf{x}_j - \mathbf{m}_g)' \mathbf{S}_g^{-1} (\mathbf{x}_j - \mathbf{m}_g), \quad j \leq i_{g-1},\ j > i_g.$$

   (c) Asignar al grupo g las observaciones que satisfacen $D_j^2 \leq \chi^2_{p,0.99}$.

   (d) Si ninguna observación se reclasifica ir al grupo $g + 1$. En otro caso, renumerar las observacione como en 1 y repetir el proceso para el mismo grupo g.

15.6 Lecturas complementarias

La literatura sobre los métodos aquí presentados es extensa. Anderberg (1973), Everitt (1993), Gordon (1981), Hartigan (1975), Mirkin (1996) , Spath y Bull (1980) y Spath (1985) están dedicados a métodos de agrupamiento y presentan los métodos clásicos. Las ideas de proyección y estimación de normales son más recientes y están descritos en artículos. Banfield and Raftery (1993) and Dasgupta and Raftery (1998) describen el algoritmo Mclust para estimar mezclas de normales. Para métodos de proyección véase Friedman (1987), Jones and Sibson (1987), Posse (1995), Nason (1995) y Peña y Prieto (2001).

EJERCICIOS

Ejercicio 15.2 *Demostrar que el criterio de Hartigan para el algoritmo de k-medias equivale a continuar añadiendo grupos hasta que* $tr(\mathbf{W}_G) < tr(\mathbf{W}_{G+1})(n-G+9)/(n-G-1)$ *(Sugerencia utilizar que* $tr(\mathbf{W}) = SCDG$*, e imponer la condición de que el valor de F sea mayor que 10)*

Ejercicio 15.3 *Demostrar que para n grande el criterio de r Calinski y Harabasz para el algoritmo de k medias equivale aproximadamente a seleccionar el número de grupos G si* $tr(\mathbf{W}_G) < tr(\mathbf{W}_{G+1})G/(G-1)$

Ejercicio 15.4 *Demostrar que en la estimación de mezclas de normales, si las distribuciones tienen* $\mathbf{V}_g = \boldsymbol{\sigma}_g^2 I$*, la estimación MV de* $\boldsymbol{\sigma}_g^2$ *se obtiene con* $\widehat{\boldsymbol{\sigma}_g^2} = \sum_{i=1}^{n} \frac{\pi_{ig}}{\sum_{i=1}^{n} \pi_{ig}} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}_g})'(\mathbf{x}_i - \widehat{\boldsymbol{\mu}_g})$.

Ejercicio 15.5 *Demostrar que si tenemos una mezcla de g poblaciones normales* $N_i(\mu_i, V)$ *con distinta media pero la misma matriz de covarianzas y con probabilidades* π_i *con* $\sum_i \pi_i = 1$*, si proyectamos las observaciones sobre una dirección u y llamamos* $z = u'x$ *a los datos proyectados, su media es* $E(z) = \bar{m} = u'\mu$ *donde* $\mu = \sum_i \pi_i \mu_i$ *y su varianza* $var(z) = u'(V+B)u$*, donde* $B = \sum_i \pi_i (\mu_i - \mu)(\mu_i - \mu)'$.

Ejercicio 15.6 *Demostrar que en el ejercicio anterior el cuarto momento de los datos proyectados es* $3\left(u'(V+B)u\right)^2 - 3\left(u'Bu\right)^2 + \sum_i \pi_i (u'B_i u)^2$.

APÉNDICE 15.1 COMPARACIÓN DEL CRITERIO DE LA TRAZA Y EL DETERMINANTE

Supongamos, para simplificar dos grupos. Estudiemos, en primer lugar, cómo aumenta la variabilidad dentro del primer grupo al incluir un nuevo elemento del segundo. Supongamos que el primer grupo tiene n puntos $(\mathbf{x}_1, ..., \mathbf{x}_n)$ con centro en $\mathbf{m}$, y le añadimos un nuevo punto, $\mathbf{x}^*$. La nueva media al incluir ese elemento será, llamando $\mathbf{d} = (\mathbf{x}^* - \mathbf{m})$

$$\mathbf{m}^* = \mathbf{m} + \frac{1}{n+1}\mathbf{d}$$

y el cambio en la matriz de suma de cuadrados

$$\mathbf{W}^* = \sum (\mathbf{x}_i - \mathbf{m} - \frac{1}{n+1}\mathbf{d})(\mathbf{x}_i - \mathbf{m} - \frac{1}{n+1}\mathbf{d})'$$

descomponiendo esta suma en la de los primeros n elementos originales del grupo más el último, tenemos que

$$\mathbf{W}^* = \mathbf{W} + \frac{n}{(n+1)^2}\mathbf{d}\mathbf{d}' + (\mathbf{x}^* - \mathbf{m}^*)(\mathbf{x}^* - \mathbf{m}^*)'$$

y como

$$\mathbf{x}^* - \mathbf{m}^* = \frac{n}{n+1}\mathbf{d}$$

sustituyendo, tenemos finalmente que:

$$\mathbf{W}^* = \mathbf{W} + \frac{n}{n+1}\mathbf{d}\mathbf{d}' \tag{15.17}$$

Con el criterio de la traza, el cambio en la matriz $\mathbf{W}$ es:

$$tr(\mathbf{W}^* - \mathbf{W}) = \frac{n}{n+1}tr(\mathbf{d}\mathbf{d}') = \frac{n}{n+1}\mathbf{d}'\mathbf{d}$$

y el cambio será mínimo si incluimos el punto de mínima distancia euclídea respecto al centro de grupo. Por un razonamiento análogo, podemos analizar el problema simétrico de la disminución de variabilidad al eliminar un elemento de un grupo y concluir que el criterio de la traza lleva a minimizar las distancias euclídeas entre los puntos y sus medias de grupo.

Analicemos el criterio del determinante. Suponiendo en (15.17), para simplificar el razonamiento, que $\frac{n}{n+1}$ es la unidad, tenemos que:

$$|\mathbf{W}^*| - |\mathbf{W}| = |\mathbf{W} + \mathbf{d}\mathbf{d}'| - |\mathbf{W}|\,, \tag{15.18}$$

y como:

$$|\mathbf{W} + \mathbf{d}\mathbf{d}'| = |\mathbf{W}(\mathbf{I} + \mathbf{W}^{-1}\mathbf{d}\mathbf{d}')| = |\mathbf{W}|\ |\mathbf{I} + \mathbf{W}^{-1}\mathbf{d}\mathbf{d}'|, \tag{15.19}$$

y, por otro lado, al tener $\mathbf{W}^{-1}\mathbf{d}\mathbf{d}'$ rango uno:

$$\begin{aligned} |\mathbf{I} + \mathbf{W}^{-1}\mathbf{d}\mathbf{d}'| &= \Pi(1+\lambda_i) = 1 + \lambda_1 = 1 + tr(\mathbf{W}^{-1}\mathbf{d}\mathbf{d}') \\ &= 1 + tr(\mathbf{d}'\mathbf{W}^{-1}\mathbf{d}) = 1 + \mathbf{d}'\mathbf{W}^{-1}\mathbf{d}, \end{aligned}$$

con lo que concluimos con la relación

$$|\mathbf{W} + \mathbf{d}\mathbf{d}'| = |\mathbf{W}| + |\mathbf{W}|\mathbf{d}'\mathbf{W}^{-1}\mathbf{d}.$$

Sustituyendo ahora este valor en (15.18), resulta que minimizaremos el efecto de añadir un punto sobre el determinante si minimizamos:

$$|\mathbf{W}|\mathbf{d}'\mathbf{W}^{-1}\mathbf{d},$$

que equivale a minimizar la distancia de Mahalanobis entre el punto y el centro de los datos. Podemos concluir que, en general, el criterio de la traza minimiza distancias euclídeas, mientras que el del determinante minimiza distancias de Mahalanobis.

Capítulo 16

CORRELACIÓN CANÓNICA

16.1 INTRODUCCIÓN

El análisis de correlaciones canónicas es debido a Hotelling. Este investigador estudió en 1936 la relación entre resultados de test de capacidad intelectual y medidas físicas de un grupo de personas. Hotelling pretendía investigar las relaciones entre ambos conjuntos de variables y conocer cuantas dimensiones independientes tenía la relación existente entre ellas. En general, correlación canónica se utiliza cuando un conjunto de variables multivariantes puede dividirse en dos grupos homogéneos (por criterios económicos, demográficos, sociales, etc.), y se desea estudiar la relación entre ambos conjuntos de variables. En particular, los dos grupos pueden corresponder a las mismas variables medidas en dos momentos distintos en el tiempo, el espacio,..., etc.

Supongamos que disponemos de un conjunto de datos de n individuos y k variables que pueden subdividirse en dos grupos: el primero incluye p variables y el segundo q, donde $p+q=k$. Por ejemplo, las primeras p variables representan las inversiones realizadas por una empresa y las restantes q variables representan distintas medidas de beneficios. Es posible que $p=q$. Por ejemplo, cuando medimos las p variables en el instante t y en el instante $t+1$. Llamaremos $\mathbf{X}$ en este capítulo a la matriz $n \times p$ que contienen los valores de las p primeras variables en los n elementos (individuos, países, empresas, etc...) e $\mathbf{Y}$ a la matriz $(n \times q)$ que contienen los valores de las q segundas variables medidas sobre esos mismos n elementos. Para medir la relación entre ambos conjuntos vamos a buscar una combinación lineal de las p primeras variables que tenga la máxima correlación con una combinación lineal de las q segundas variables. Es decir, llamando:

$$\mathbf{x}^* = \mathbf{X}\boldsymbol{\alpha} = \sum_{i=1}^{p} \alpha_i \mathbf{x}_i$$

a una combinación lineal de las variables del primer grupo, y:

$$\mathbf{y}^* = \mathbf{Y}\boldsymbol{\beta} = \sum_{j=1}^{q} \beta_j \mathbf{x}_j$$

a una combinación lineal de las variables del segundo, se desea encontrar los vectores $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ tales que las nuevas variables $\mathbf{x}^*$ e $\mathbf{y}^*$ tengan máxima correlación.

Es posible que una vez encontrada la primera relación entre estas dos variables indicadoras no exista más relacion entre ambos conjuntos de variables y entonces decimos que toda la relación entre ambos conjuntos se resume en una dimensión. Para comprobarlo, podemos buscar una segunda variable indicadora del primer conjunto, que este incorrelada con la primera, y que tenga correlación máxima con otra variable indicadora del segundo conjunto. Procediendo de esta manera, podemos obtener $r = min(p, q)$ relaciones entre variables indicadoras de ambos grupos que pueden ordenarse por orden de importancia. Determinar el número de relaciones entre las variables (variables indicadoras incorreladas de un conjunto relacionadas con el otro) permite juzgar cuantas dimensiones distintas tiene la relación. En el análisis de correlación canónica ambos conjuntos de variables se consideran simetricamente, pero es posibles que queremos explicar el conjunto de variables dependientes $\mathbf{Y}$ mediante las independientes $\mathbf{X}$. El procedimiento habitual (regresión múltiple) es construir un indicador de la variables dependientes y relacionarlo con las independientes mediante una regresión múltiple. Este procedimiento es razonable cuando exista sólo una dimensión en la relación, pero puede ser engañoso si existen varias. El análisis de correlación canónica permite determinar cuantas dimensiones relevantes tiene la relación entre ambos conjuntos de variables.

16.2 Construcción de las variables canónicas

16.2.1 La primera variable canónica

Supondremos que $\mathbf{x}$ es un vector $p \times 1$ con distribución $N_p(\mathbf{0}, \mathbf{V}_{11})$ e $\mathbf{y}$ un vector $q \times 1$ con distribución $N_q(\mathbf{0}, \mathbf{V}_{22})$, de manera que las variables están medidas en desviaciones a la media. Entonces, construyendo el vector $(p+q) \times 1$ con todas las variables, $[\mathbf{x}'\mathbf{y}']'$, la matriz de covarianzas para el conjunto de todas las variables es:

$$\mathbf{V}_{x,y} = E\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}[\mathbf{x}'\mathbf{y}']\right) = \begin{bmatrix}\mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{bmatrix}.$$

Queremos encontrar dos vectores, $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$, que definan dos nuevas variables escalares, $x^* = \boldsymbol{\alpha}'\mathbf{x}$, $y^* = \boldsymbol{\beta}'\mathbf{y}$, con máxima correlación. El coeficiente de correlación entre x^* e y^* es:

$$\rho(x^*y^*) = \frac{E\left[\boldsymbol{\alpha}'\mathbf{x}\mathbf{y}'\boldsymbol{\beta}\right]}{E\left[\boldsymbol{\alpha}'\mathbf{x}\mathbf{x}'\boldsymbol{\alpha}\right]^{1/2}\left[\boldsymbol{\beta}'\mathbf{y}\mathbf{y}'\boldsymbol{\beta}\right]^{1/2}}$$

que puede escribirse:

$$\rho = \frac{\boldsymbol{\alpha}'\mathbf{V}_{12}\boldsymbol{\beta}}{(\boldsymbol{\alpha}'\mathbf{V}_{11}\boldsymbol{\alpha})^{1/2}(\boldsymbol{\beta}'\mathbf{V}_{22}\boldsymbol{\beta})^{1/2}}.$$

Como nos interesa la magnitud, y no el signo, de la correlación vamos a maximizar el cuadrado de la correlación entre (x^*, y^*) respecto a $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$. Para ello, impondremos la condición de varianzas unitarias, es decir,

$$Var(x^*) = \boldsymbol{\alpha}'\mathbf{V}_{11}\boldsymbol{\alpha} = 1 \tag{16.1}$$

$$Var(y^*) = \boldsymbol{\beta}'\mathbf{V}_{22}\boldsymbol{\beta} = 1 \tag{16.2}$$

y la función objetivo es:

$$Maximizar\ \rho^2 = \frac{(\boldsymbol{\alpha}'\mathbf{V}_{12}\boldsymbol{\beta})^2}{(\boldsymbol{\alpha}'\mathbf{V}_{11}\boldsymbol{\alpha})(\boldsymbol{\beta}'\mathbf{V}_{22}\boldsymbol{\beta})} \tag{16.3}$$

con las restricciones (16.1) y (16.2). Introduciendo estas restricciones mediante multiplicadores de Lagrange, la función a maximizar es:

$$M = (\boldsymbol{\alpha}'\mathbf{V}_{12}\boldsymbol{\beta})^2 - \lambda(\boldsymbol{\alpha}'\mathbf{V}_{11}\boldsymbol{\alpha} - 1) - \mu(\boldsymbol{\beta}'\mathbf{V}_{22}\boldsymbol{\beta} - 1).$$

Derivando, respecto a los vectores de coeficientes y escribiendo los resultados como vector columna, utilizando que $\mathbf{V}'_{12} = \mathbf{V}_{21}$:

$$\frac{\partial \mathbf{M}}{\partial \boldsymbol{\alpha}} = 2\mathbf{V}_{12}\boldsymbol{\beta} - 2\lambda\mathbf{V}_{11}\boldsymbol{\alpha}, \tag{16.4}$$

$$\frac{\partial \mathbf{M}}{\partial \boldsymbol{\beta}} = 2\mathbf{V}_{21}\boldsymbol{\alpha} - 2\mu\mathbf{V}_{22}\boldsymbol{\beta}. \tag{16.5}$$

Igualando a cero estas ecuaciones, se obtiene:

$$\mathbf{V}_{12}\boldsymbol{\beta} = \lambda\mathbf{V}_{11}\boldsymbol{\alpha} \tag{16.6}$$

$$\mathbf{V}_{21}\boldsymbol{\alpha} = \mu\mathbf{V}_{22}\boldsymbol{\beta}. \tag{16.7}$$

Para resolver este sistema, multipliquemos la primera ecuación por $\boldsymbol{\alpha}'$ y la segunda por $\boldsymbol{\beta}'$ y utilicemos las igualdades (16.1) y (16.2). Entonces:

$$\boldsymbol{\alpha}'\mathbf{V}_{12}\boldsymbol{\beta} = \lambda\boldsymbol{\alpha}'\mathbf{V}_{11}\boldsymbol{\alpha} = \lambda,$$

$$\boldsymbol{\beta}'\mathbf{V}_{21}\boldsymbol{\alpha} = \mu\boldsymbol{\beta}'\mathbf{V}_{22}\boldsymbol{\beta} = \mu,$$

y como $\lambda = (\boldsymbol{\alpha}'\mathbf{V}_{12}\boldsymbol{\beta}) = (\boldsymbol{\beta}'\mathbf{V}_{21}\boldsymbol{\alpha}) = \mu$, concluimos con el sistema:

$$\mathbf{V}_{12}\boldsymbol{\beta} = \lambda\mathbf{V}_{11}\boldsymbol{\alpha}, \tag{16.8}$$

$$\mathbf{V}_{21}\boldsymbol{\alpha} = \lambda\mathbf{V}_{22}\boldsymbol{\beta}. \tag{16.9}$$

Despejando $\boldsymbol{\beta}$ de la segunda ecuación, $\boldsymbol{\beta} = \lambda^{-1}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\boldsymbol{\alpha}$, y sustituyendo en la primera:

$$\mathbf{V}_{12}\left(\lambda^{-1}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\right)\boldsymbol{\alpha} = \lambda\mathbf{V}_{11}\boldsymbol{\alpha}$$

que conduce a:

$$\left(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\right)\boldsymbol{\alpha} = \lambda^2\boldsymbol{\alpha}. \tag{16.10}$$

Por tanto, $\boldsymbol{\alpha}$ es el vector propio ligado al valor propio λ^2 de la matriz cuadrada de dimensión p:

$$\mathbf{A}_{p\times p} = \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21} \tag{16.11}$$

con valor propio λ^2. Análogamente, se obtiene que $\boldsymbol{\beta}$ debe ser el vector propio ligado al valor propio λ^2 de la matriz

$$\mathbf{B}_{q\times q} = \mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}. \tag{16.12}$$

Observemos que, por (16.3) y (16.10), $\lambda^2 = \mu^2 = \rho^2$ es el cuadrado del coeficiente de correlación entre las variables canónicas x^*, y^*, por lo que tendremos que tomar el vector propio ligado al mayor valor propio.

En resumen, la solución buscada requiere:

1. construir las dos matrices cuadradas de dimensiones p y q, $\mathbf{A}$ y $\mathbf{B}$ definidas por (16.11) y (16.12). El vector propio asociado a su máximo valor propio (que es el mismo en ambas) proporciona las variables canónicas.

2. Este mayor valor propio de ambas matrices es el cuadrado del coeficiente de correlación entre las variables canónicas.

Observemos que de las ecuaciones (16.8) y (16.9) resulta:

$$\boldsymbol{\alpha} = \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\beta}\lambda^{-1} \tag{16.13}$$

$$\boldsymbol{\beta} = \mathbf{V}_{22}^{-1}\mathbf{V}_{21}\boldsymbol{\alpha}\lambda^{-1} \tag{16.14}$$

Por lo que sólo necesitamos obtener vectores propios de una de las matrices. Conocido el vector $\boldsymbol{\alpha}$ podemos obtener el vector $\boldsymbol{\beta}$ con (16.14) y análogamente, conocido $\boldsymbol{\beta}$ obtenemos $\boldsymbol{\alpha}$ con (16.13). Además de (16.10) obtenemos $\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\boldsymbol{\alpha} = \boldsymbol{\lambda}^2\mathbf{V}_{11}\boldsymbol{\alpha}$ y multiplicando por $\boldsymbol{\alpha}'$ e imponiendo la condición (16.1) tenemos que

$$\boldsymbol{\lambda}^2 = \boldsymbol{\alpha}'\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\boldsymbol{\alpha} \tag{16.15}$$

que indica que el coeficiente de correlación canónica, $\boldsymbol{\lambda}^2$,es el cuadrado del coeficiente de correlación múltiple entre la variable $x^* = \boldsymbol{\alpha}'\mathbf{x}$ y las variables $\mathbf{y}$. En efecto, las covarianzas entre x^* e $\mathbf{y}$ vienen dadas por el vector $\mathbf{V}_{21}\boldsymbol{\alpha}$ y las correlaciones por $\mathbf{D}_{22}^{-1/2}\mathbf{V}_{21}\boldsymbol{\alpha}$, donde $\mathbf{D}_{22}$ es una matriz diagonal que contiene las varianzas de las variables y. Entonces el coeficiente de correlación múltiple es

$$\varrho^2 = \mathbf{r}_{12}\mathbf{R}_{22}^{-1}\mathbf{r}_{12} = (\boldsymbol{\alpha}'\mathbf{V}_{12}\mathbf{D}_{22}^{-1/2})(\mathbf{D}_{22}\mathbf{V}_{22}^{-1}\mathbf{D}_{22})(\mathbf{D}_{22}^{-1/2}\mathbf{V}_{21}\boldsymbol{\alpha}) = \boldsymbol{\lambda}^2$$

16.3 Las r variables canónicas

El proceso descrito puede continuarse, buscando una segunda variable escalar canónica, x_2^*, construida como combinación lineal de las originales $\mathbf{x}$, ortogonal a la primera, y que tenga máxima correlación con otra variable escalar $\mathbf{y}_2^*$, combinación lineal de las $\mathbf{y}$ y que sea a su vez ortogonal a $\mathbf{y}_1^*$. Así podemos obtener $2r$ combinaciones lineales, $\left(x_1^*, \ldots x_q^*\right)$, $\left(y_1^*, \ldots, y_q^*\right)$ donde r =mínimo (p, q), que llamaremos variables canónicas. Las matrices $\mathbf{A}$ y $\mathbf{B}$ dadas por (16.11) y (16.12) tienen un rango igual al mínimo de p, q, y si extraemos sus r valores propios no nulos y los vectores propios unidos a dichos valores propios, podemos formar r combinaciones lineales de las variables de ambos grupos que:

(a) Tienen correlación máxima cuando provienen del mismo valor propio.

(b) Están incorreladas dentro de cada grupo.

(c) Están incorreladas si corresponden a distintos vectores propios.

Vamos a demostrar este resultado. Para demostrar que las matrices $\mathbf{A}$ y $\mathbf{B}$ tienen valores propios reales no negativos, probaremos que $\mathbf{A}$ y $\mathbf{B}$ tienen los mismos valores propios que una matriz semidefinida positiva.

Lema: Las matrices $\mathbf{R}^{-1}\mathbf{Q}$ y $\mathbf{R}^{-1/2}\mathbf{Q}\mathbf{R}^{-1/2}$ tienen los mismos valores propios. Además si $\mathbf{v}$ es un vector propio de la primera, $\mathbf{R}^{1/2}\mathbf{v}$ lo es de la segunda.

Demostración: Sea λ un valor propio de $\mathbf{R}^{-1}\mathbf{Q}$ y sea $\mathbf{v}$ su vector propio asociado. Entonces

$$\mathbf{R}^{-1}\mathbf{Q}\mathbf{v} = \lambda\mathbf{v}$$

premultiplicando por $\mathbf{R}^{1/2}$ se obtiene

$$\mathbf{R}^{-1/2}\mathbf{Q}\mathbf{v} = \lambda\mathbf{R}^{1/2}\mathbf{v}$$

y escribiendo esta relación como:

$$\mathbf{R}^{1/2}\mathbf{Q}\mathbf{R}^{-1/2}(\mathbf{R}^{1/2}\mathbf{v}) = \lambda(\mathbf{R}^{1/2}\mathbf{v})$$

y llamando $\mathbf{h} = \mathbf{R}^{1/2}\mathbf{v}$, tenemos que

$$\mathbf{R}^{1/2}\mathbf{Q}\mathbf{R}^{-1/2}\mathbf{h} = \lambda\mathbf{h}$$

Por tanto, λ es un valor propio de las matrices $\mathbf{R}^{-1}\mathbf{Q}$ y $\mathbf{R}^{1/2}\mathbf{Q}\mathbf{R}^{-1/2}$, y su vector propio asociado es, respectivamente, $\mathbf{v}$ y $\mathbf{R}^{1/2}\mathbf{v}$**.**

Corolario

Las matrices $\mathbf{A}$ y $\mathbf{H}\mathbf{H}'$ donde

$$\mathbf{A} = \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}$$

$$\mathbf{H} = \mathbf{V}_{11}^{-1/2}\mathbf{V}_{12}\mathbf{V}_{22}^{-1/2}$$

$$\mathbf{H}'\mathbf{H} = \mathbf{V}_{11}^{-1/2}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1/2}$$

tienen los mismos valores propios.

Este corolario es un caso particular del Lema tomando $\mathbf{R} = \mathbf{V}_{11}$, $\mathbf{Q} = \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}$. La matriz $\mathbf{A}$ tiene los valores propios de $\mathbf{HH}'$ y la $\mathbf{B}$ los de $\mathbf{H}'\mathbf{H}$. Como estas matrices son semidefinidas positivas, los valores propios de $\mathbf{A}$ y $\mathbf{B}$ son reales y no negativos. Vamos a ver las propiedades de los vectores propios.

Vectores propios

Sean $\boldsymbol{\alpha}_i$, $\boldsymbol{\alpha}_j$ dos vectores propios de $\mathbf{A}$ correspondientes a raíces distintas. Los vectores propios correspondientes de $\mathbf{HH}'$ para esas mismas raíces son, según el lema, $\mathbf{V}_{11}^{1/2}\boldsymbol{\alpha}_i$ y $\mathbf{V}_{11}^{1/2}\boldsymbol{\alpha}_j$. Como los vectores propios de matrices simétricas son ortogonales, tendrá que verificarse que

$$\left(\mathbf{V}_{11}^{-1/2}\boldsymbol{\alpha}_i\right)'\left(\mathbf{V}_{11}^{-1/2}\boldsymbol{\alpha}_j\right) = \boldsymbol{\alpha}_i'\mathbf{V}_{11}\boldsymbol{\alpha}_j = 0$$

lo que implica que las variables $x_i^* = \mathbf{x}'\boldsymbol{\alpha}_i$ y $x_j^* = \mathbf{x}'\boldsymbol{\alpha}_j$ correspondientes a distintos indicadores del mismo grupo están incorreladas, ya que, al tener varianza unidad, su correlación será:

$$Cov(x_i^*, x_j^*) = \boldsymbol{\alpha}_i' E\left[\mathbf{x}\mathbf{x}'\right]\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_i'\mathbf{V}_{11}\boldsymbol{\alpha}_j = 0.$$

Demostraremos ahora que las variables indicadoras de grupos distintos correspondientes a distintos vectores propios, x_i^* y y_j^*, también están incorreladas. Como

$$Cov\left(x_i^*, y_j^*\right) = E\left[\boldsymbol{\alpha}_i'\mathbf{x}\mathbf{y}'\boldsymbol{\beta}_j\right] = \boldsymbol{\alpha}_i'\mathbf{V}_{12}\boldsymbol{\beta}_j$$

y utilizando (16.6)

$$Cov\left(x_i^*, y_j^*\right) = \boldsymbol{\alpha}_i'\left(\lambda_j\mathbf{V}_{11}\boldsymbol{\alpha}_j\right) = 0.$$

con lo que hemos comprobado que las variables indicadores ligadas a distintos valores propios estan incorreladas dentro de cada grupo y entre grupos.

16.3.1 Propiedades de las variables y correlaciones canónicas

El procedimiento que hemos expuesto proporciona r variables canónicas cuyas propiedades vamos a resumir brevemente.

1. Las variables canónicas son indicadores de los dos conjuntos de variables que se definen por pares, con la condición de máxima correlacion.

2. Los coeficientes de las variables canónicas son los vectores propios ligados al mismo valor propio de las matrices $\mathbf{V}_{ii}^{-1/2}\mathbf{V}_{ij}\mathbf{V}_{jj}^{-1}\mathbf{V}_{ji}$, para $i = 1, 2$ y $i \neq j$.

3. Si $\boldsymbol{\alpha}_i'\mathbf{x}$ es una variable canónica también lo es -$\boldsymbol{\alpha}_i'\mathbf{x}$, y los signos de las variables canónicas suelen tomarse de manera que las correlaciones entre las variables canónicas $\boldsymbol{\alpha}'\mathbf{x}$ y $\boldsymbol{\beta}'\mathbf{y}$ sean positivos.

4. Las correlaciones canónicas, λ_i^2, son el cuadrado del coeficiente de correlación entre las dos variables canónicas correspondientes.

5. Las correlaciones canónicas son invariantes ante transformaciones lineales de las variables, son propiedades del conjunto de variables y no se modifican si sustituimos las r variables de un conjunto por r combinaciones lineales de ellas linealmente independientes. (véase ejercicio 16.1)

6. La primera correlación canónica, λ_1^2, es mayor o igual que el mayor coeficiente de correlación simple al cuadrado entre una variable de cada conjunto.

7. El coeficiente de correlación canónica λ_i^2 es el coeficiente de determinación en una regresión multiple con respuesta la variable $y_i^* = \boldsymbol{\beta}_i'\mathbf{y}$, y variables explicativas las $\mathbf{x}$. También es el coeficiente de determinación entre la regresión múltiple entre $x_i^* = \boldsymbol{\alpha}_i'\mathbf{x}$ y el conjunto de las $\mathbf{y}$. (ver ejercicio 16.3)

Las variables canónicas son los predictores óptimos en el sentido siguiente: supongamos que se desea encontrar un conjunto de $r = \min(p, q)$ variables combinaciones lineales de las variables de cada grupo, $\mathbf{x}^* = \Gamma\mathbf{x}$, y $\mathbf{y}^* = \Theta\mathbf{y}$, que estén incorreladas, $\Gamma\mathbf{V}_{11}\Gamma' = \mathbf{I}$ y $\Theta\mathbf{V_{22}}\Theta' = \mathbf{I}$ y de manera que $\mathbf{x}^*$ e $\mathbf{y}^*$ estén próximos. Si tomamos como criterio minimizar $E(\|\mathbf{x}^* - \mathbf{y}^*\|^2)$ se obtienen las variables canónicas. Este resultado es debido a Izenman(1975) y Yohai y Garcia Ben (1980).

16.4 ANÁLISIS MUESTRAL

En la práctica, los valores poblacionales no son conocidos y tendremos que estimarlos a partir de la muestra. Supondremos que hemos restado las medias muestrales a cada variable para trabajar con variables de media cero. En la hipótesis de normalidad multivariante, como las variables canónicas son funciones de la matriz de covarianzas entre las variables y el estimador máximo verosimil de esta matriz es $\mathbf{S}$, la matriz de covarianzas muestral, concluimos que los estimadores máximo verosímiles de las variables canónicas se obtienen al extraer los $r = \min(p, q)$ mayores valores propios, y sus vectores propios asociados, de las matrices

$$\hat{\mathbf{A}} = \mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21},$$

y

$$\hat{\mathbf{B}} = \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12},$$

donde las $\mathbf{S}_{ij}$ son las estimación MV de las matrices $\mathbf{V}_{ij}$, y se obtiene particionando convenientemente la matriz $\mathbf{S}$ de covarianzas entre las $p + q$ variables (véase Anderson, cap 12 para un análisis más detallado). Estas matrices son los equivalentes muestrales de (16.11) y (16.12).

En la práctica, suponiendo $p \geq q$, basta obtener los valores propios de la matriz de dimensión menor, $\hat{\mathbf{B}}$ en este caso, y sus vectores propios asociados. Los vectores propios ligados a estos valores propios de la otra matriz, $\hat{\mathbf{A}}$, se obtendrán transformando estos vectores

de la forma $\mathbf{S}_{ii}^{-1}\mathbf{S}_{ij}$ para llevarlos a la dimensión adecuada. Por ejemplo, supongamos que $\mathbf{v}$ es un vector propio de $\hat{\mathbf{B}}$, y verifica $\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{v} = \boldsymbol{\lambda}^2\mathbf{v}$, entonces, multiplicando por $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ tenemos que $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}(\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{v}) = \boldsymbol{\lambda}^{\mathbf{2}}(\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{v})$ y comprobamos que $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{v}$ es el vector propio de $\hat{\mathbf{A}}$ ligado al valor propio $\boldsymbol{\lambda}^{\mathbf{2}}$. A continuación vamos a comprobar que las correlaciones conónicas son invariantes a transformaciones lineales de las variables..

Invarianza del Análisis

Si estandarizamos las variables y trabajamos con las matrices de correlación las correlaciones canónicas no varían. Al estandarizar la matriz $\widehat{\mathbf{A}}$ se convertirá en:

$$\mathbf{R}_1 = \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21},$$

donde las matrices $\mathbf{R}_{ij}$ son las matrices de correlación definidas $\mathbf{R}_{ij} = \mathbf{D}_i^{-1/2}\mathbf{S}_{ij}\mathbf{D}_j^{-1/2}$,$(i, j = 1, 2)$, siendo $\mathbf{D}_i$ y $\mathbf{D}_j$ matrices diagonales que contienen las varianzas de las variables. Vamos a comprobar que las matrices $\widehat{\mathbf{A}}$ y $\mathbf{R}_1$ tienen los mismos valores propios. Utilizando la relación $\mathbf{R}_{ij} = \mathbf{D}_i^{-1/2}\mathbf{S}_{ij}\mathbf{D}_j^{-1/2}$, tenemos que:

$$\mathbf{R}_1 = \mathbf{D}_1^{1/2}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{D}_1^{-1/2} = \mathbf{D}_1^{1/2}\widehat{\mathbf{A}}\mathbf{D}_1^{-1/2}, \tag{16.16}$$

y la ecuación para obtener los valores propios de $\mathbf{R}_1$:

$$|\mathbf{R}_1 - \lambda\mathbf{I}| = 0$$

puede escribirse como

$$|\mathbf{D}^{1/2}||\widehat{\mathbf{A}} - \lambda\mathbf{I}||\mathbf{D}^{-1/2}| = 0$$

y las ecuaciones $|\mathbf{R}_1 - \lambda\mathbf{I}| = 0$ y $|\widehat{\mathbf{A}} - \lambda\mathbf{I}| = 0$ tienen las mismas soluciones, y las matrices $\mathbf{R}_1$ y $\widehat{\mathbf{A}}$ tendrán los mismos valores propios. Por tanto, las correlaciones canónicas son idénticas. Los vectores propios de $\mathbf{R}_1$ pueden obtenerse a partir de los de $\widehat{\mathbf{A}}$, ya que si $\mathbf{v}$ es un vector propio de $\mathbf{R}_1$:

$$\mathbf{R}_1\mathbf{v} = \lambda^2\mathbf{v}$$

entonces, por (16.16):

$$\mathbf{D}_1^{1/2}\widehat{\mathbf{A}}\mathbf{D}_1^{-1/2}\mathbf{v} = \lambda^2\mathbf{v},$$

es decir

$$\widehat{\mathbf{A}}(\mathbf{D}_1^{-1/2}\mathbf{v}) = \lambda^2(\mathbf{D}_1^{-1/2}\mathbf{v}).$$

y $(\mathbf{D}_1^{-1/2}\mathbf{v})$ es un vector propio de $\widehat{\mathbf{A}}$ ligado al mismo valor propio. Por tanto, la variable canónica ligada a λ^2 se obtiene con variables estandarizadas multiplicando la matriz de variables estandarizadas $\left(\mathbf{X}\ \mathbf{D}_1^{-1/2}\right)$ por el vector propio de $\mathbf{R}_1$, $\mathbf{v}$, mientras que partiendo de las variables originales multiplicamos la matriz $\mathbf{X}$ por el vector propio de $\widehat{\mathbf{A}}$, $\mathbf{D}_1^{-1/2}\mathbf{v}$, con lo que las dos variables obtenidas son idénticas.

Ejemplo 16.1 *Como ejemplo consideramos los datos de los hogares españoles. Tomaremos como variables X los gastos en alimentación, bebidas y tabaco (x_1), en vestido y calzado (x_2), en menaje (x_3), en transportes y comunicación (x_4) y en esparcimiento y enseñanza (x_5). La matriz de correlaciones entre estas variables es, redondeada a la segunda cifra decimal*

$$R_{11} = \begin{bmatrix} 1 & & & & \\ .29 & 1 & & & \\ .13 & .25 & 1 & & \\ .23 & .23 & .35 & 1 & \\ .33 & .32 & .22 & .36 & 1 \end{bmatrix}$$

Como variables Y incluiremos el número de personas en la unidad de gastos (y_1), el número de personas mayores de 14 años (y_2), el nivel educativo del sustentador principal (y_3) y el número de perceptores con ingresos (y_4). La matriz de correlación es:

$$R_{22} = \begin{bmatrix} 1 & & & \\ .55 & 1 & & \\ .11 & .04 & 1 & \\ .53 & -.11 & .00 & 1 \end{bmatrix}$$

La matriz de correlaciones cruzadas entre ambos grupos de características es:

$$R_{12} = \begin{bmatrix} .46 & .03 & .22 & .40 \\ .34 & .18 & .32 & .14 \\ .05 & -.02 & .51 & -.02 \\ .33 & .13 & .26 & .25 \\ .29 & .17 & .23 & .17 \end{bmatrix}$$

Esta matriz proporciona por filas las correlaciones de cada variable del grupo x con las y. La suma de los valores absolutos de las filas es una medida descriptiva global de la dependencia. La variable x más correlada con las y es la primera, ya que la suma de las correlaciones $(0,46 + 0,03 + 0,22 + 0,40 = 1,11)$, mientras que para las demás estos números son $0,98$, $0,60$, $0,97$ y $0,86$. Por columnas las sumas son $1,47$, $0,53$, $1,54$, $0,98$. Las variables 1, 2 y 4 del primer grupo y 1, 3 del segundo parecen ser las correlacionadas entre los bloques, por lo que esperamos que estas variables tengan mayor peso en la primera variable canónica.

Los valores y vectores propios de la matriz $\mathbf{A} = \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}\mathbf{R}_{21}$ se encuentran en la tabla 16.1

λ^2	.44	.21	.05	.01
$\boldsymbol{\alpha}$	.76	.41	.63	−.11
	.43	−.05	−.45	−.71
	.31	−.87	.34	.08
	.38	.27	−.24	.65
	.04	−.03	−.48	.22
$\boldsymbol{\beta}$	−.78	−.51	−.28	−.65
	.27	.10	.96	.38
	−.56	.80	.01	.11
	−.07	−.31	.08	.65

Tabla 16.1. Valores y vectores propios para las variables de los hogares

Como vemos la máxima correlación entre una combinación de las variables del primer grupo y otra del segundo se da entre la combinacion que incluye las variables (1, 2, 3, 4) del primer grupo y (1, 3) del segundo. Esta máxima correlación entre las variables de gasto y de estructura de la familia, explica un 44/(44+21+5+1)=62% de la variabilidad y corresponde a la relación entre el indicador de gasto:

$$x_1^* = .76x_1 + .43x_2 + .31x_3 + .38x_4 + .04x_5$$

que es un promedio de los gastos dando mayor peso a la alimentación, con la variable

$$y_1^* = .78y_1 - .27y_2 + .56y_3 + .07y_4$$

donde hemos cambiado el signo del vector propio para facilitar la interpretación. Esta variable pondera principalmente el tamaño de la familia (y_1) y el nivel de educación del sustentador principal (y_3). La interpretación de este resultado es que cuanto mayor sea el indicador y^* mayor será el indicador de gasto. Ambas variables son medidas de "tamaño" global de las variables en ambos conjuntos.

Las segundas variables canónicas explican un 21/71=29,58% y son indicativas de la forma de los datos. El indicador de las variables de gastos contrapone los gastos en alimentación y transporte a los de menaje y encuentra que este indicador está especialmente relacionado un indicador de la diferencia entre el nivel de educación y el tamaño familiar. El gráfico muestra la relación entre las variables x_2^*, y_2^*. Se observa que aquellas familiar con bajo gasto en menaje (x_2^* alto) tienen pocos miembros en la unidad familiar (y_1 bajo) y alto nivel de educación (y_3 alto). Los otros dos componentes explican muy poco de la relación.

Si realizamos este mismo análisis para las variables en logaritmos se obtiene la tabla 8.2. Se observa que ahora el primer componente explica el 50% de la variabilidad y tiene una interpretación similar al caso anterior.

λ^2	.50	.06	.01	.001
	$-.80$	.10	.68	$-.24$
	.18	.38	$-.56$	$-.21$
$\boldsymbol{\alpha}_r$	$-.07$	.28	.18	$-.65$
	.55	.39	.14	.68
	.11	$-.78$	.57	$-.05$
	$-.88$	$-.49$	$-.53$	.43
$\boldsymbol{\beta}$	.38	.23	.60	.52
	$-.27$	.84	.04	$-.11$
	.08	$-.00$	.60	$-.73$

Tabla 16.2. Valores y vectores propios para las variables en logaritmos

16.5 INTERPRETACIÓN GEOMÉTRICA

Vamos a comprobar que las correlaciones canónicas representan relaciones de dependencia entre los subespacios generados por los dos conjuntos de variables. Esta propiedad justifica que el análisis de correlaciones canónicas sea invariante ante reparametrizaciones. Supongamos que disponemos en cada conjunto de variables distintas (no relacionadas linealmente) de manera que las matrices $\mathbf{X}$, de dimensiones $n \times p$, e $\mathbf{Y}$, de dimensiones $n \times q$, son de rango completo, es decir $rg(\mathbf{X}) = p$, $rg(\mathbf{Y}) = q$ y $p + q < n$, y las p columnas de $\mathbf{X}$ generan un subespacio de dimensión p y las q columnas de $\mathbf{Y}$ otro de dimensión q. Supongamos que estamos interesados en encontrar dos vectores, uno en cada subespacio, que estén tan próximos como sea posible. El primer vector, $\mathbf{x}^*$, por pertenecer al espacio $\mathbf{S}(\mathbf{X})$ generado por las columnas de $\mathbf{X}$, será de la forma $\mathbf{x}^* = \mathbf{X}\boldsymbol{\alpha}$ y el segundo, $\mathbf{y}^*$, por pertenecer al subespacio $S(\mathbf{Y})$ generado por las columnas de $\mathbf{Y}$, será $\mathbf{y}^* = \mathbf{Y}\boldsymbol{\beta}$. Los vectores $\mathbf{x}^*$ e $\mathbf{y}^*$ estarán lo más cerca posible (ver figura 16.1), si $\mathbf{x}^*$ es colineal con la proyección de $\mathbf{y}^*$ sobre $\mathbf{S}(\mathbf{X})$ y viceversa.

Para formular esta propiedad, llamemos $\mathbf{P}_1$ y $\mathbf{P}_2$ a las matrices proyección sobre los espacios $\mathbf{S}(\mathbf{X})$ y $\mathbf{S}(\mathbf{Y})$ que vendrán dadas por:

$$\mathbf{P}_1 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{16.17}$$

$$\mathbf{P}_2 = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'. \tag{16.18}$$

entonces la condición exigida es:

$$\mathbf{P}_1\mathbf{y}^* = \lambda\mathbf{x}^*$$

$$\mathbf{P}_2\mathbf{x}^* = \mu\mathbf{y}^*$$

que equivale a:

$$\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{Y}\boldsymbol{\beta} = \lambda\mathbf{X}\boldsymbol{\alpha}$$

$$\left[\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\right]\mathbf{X}\boldsymbol{\alpha} = \mu\mathbf{Y}\boldsymbol{\beta}$$

y multiplicando por $1/n\mathbf{X}'$ la primera y $1/n\mathbf{Y}'$ la segunda, resulta $\mathbf{S}_{12}\boldsymbol{\beta} = \lambda\mathbf{S}_{11}\boldsymbol{\alpha}$ y $\mathbf{S}_{21}\boldsymbol{\alpha} = \mu\mathbf{S}_{22}\boldsymbol{\beta}$. De la primera deducimos que $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\boldsymbol{\beta} = \lambda\boldsymbol{\alpha}$ y de la segunda $\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\boldsymbol{\alpha} = \mu\boldsymbol{\beta}$. Sustituyendo el valor de $\boldsymbol{\beta}$ en la primera ecuación obtenemos

$$\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\boldsymbol{\alpha} = \mu\lambda\boldsymbol{\alpha}$$

que muestra que $\boldsymbol{\alpha}$ es un vector propio de la matriz $\mathbf{A}$ encontrada en (16.11). Las primeras variables canónicas representan los vectores de $\mathbf{S}(\mathbf{X})$ y $\mathbf{S}(\mathbf{Y})$ más próximos.

Una medida de la distancia entre los subespacios generados por las columnas de $\mathbf{X}$ y las $\mathbf{Y}$ es el coseno del ángulo entre ambos subespacios, que se calcula por:

$$Cos\ \theta = \frac{(\boldsymbol{\alpha}'\mathbf{X}')(\mathbf{Y}\boldsymbol{\beta})}{(\boldsymbol{\alpha}'\mathbf{X}'\mathbf{X}\boldsymbol{\alpha})^{1/2}(\boldsymbol{\beta}'\mathbf{Y}'\mathbf{Y}\boldsymbol{\beta})^{1/2}},$$

de donde resulta:

$$Cos\ \theta^2 = \delta^2 = \frac{(\boldsymbol{\alpha}'\mathbf{S}_{12}\boldsymbol{\beta})^2}{(\boldsymbol{\alpha}'\mathbf{S}_{11}\boldsymbol{\alpha})(\boldsymbol{\beta}'\mathbf{S}_{22}\boldsymbol{\beta})},$$

que permite concluir que la máxima correlación canónica es el coseno del ángulo que forman los subespacios generados por $\mathbf{X}$ y por $\mathbf{Y}$.

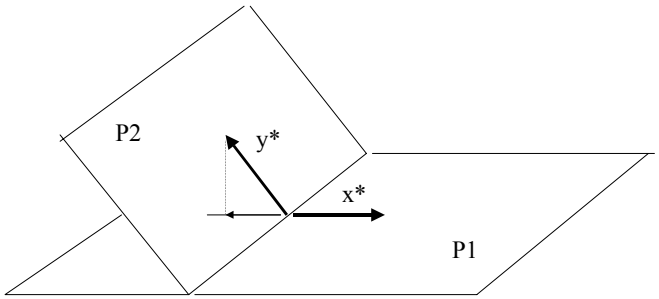


Figura 16.1: Representación de las primeras variables canónicas de los espacios P1 y P2

16.6 CONTRASTES

Podemos construir un contraste de que los dos conjuntos de variables están incorrelados, es decir, $\mathbf{V}_{12} = \mathbf{0}$, bajo la hipótesis de que los vectores $\mathbf{x}$ son $N_p(\mathbf{0}, \mathbf{V}_{11})$ y los $\mathbf{y}$ son $N_q(\mathbf{0}, \mathbf{V}_{22})$.

El contraste de que ambos vectores están incorrelados es equivalente al contraste de que todas las correlaciones canónicas son nulas. Lo estableceremos como:

$$H_0 : \mathbf{V}_{12} = \mathbf{0}$$

$$H_1 : \mathbf{V}_{12} \neq \mathbf{0}$$

Bajo H_0 la verosimilitud conjunta, $f(\mathbf{x}_1, ..., \mathbf{x}_n, \mathbf{y}_1, ..., \mathbf{y}_n)$, se descompone en $f(\mathbf{x}_1, ..., \mathbf{x}_n)$ $f(\mathbf{y}_1, ..., \mathbf{y}_n)$. El ratio de las verosimilitudes máximas entre las dos hipótesis es:

$$\frac{f(H_1)}{f(H_0)} = \frac{(2\pi)^{-n(p+q)} \left|\mathbf{S}\right|^{-n/2} e^{-n(p+q)/2}}{(2\pi)^{-np} \left|\mathbf{S}_{11}\right|^{-n/2} e^{-np/2} (2\pi)^{-nq} \left|\mathbf{S}_{22}\right|^{-n/2} e^{-nq/2}}$$

donde $\mathbf{S}$ es la estimación de la matriz de varianzas y covarianzas conjunta y $\mathbf{S}_{11}$ y $\mathbf{S}_{22}$ las estimaciones correspondientes a cada bloque. El contraste de razón de verosimilitudes será:

$$\lambda = 2(\log(H_1) - \log(H_0)) = -n \log \frac{|\mathbf{S}|}{|\mathbf{S}_{11}| \, |\mathbf{S}_{22}|}. \tag{16.19}$$

Como $|\mathbf{S}| = |\mathbf{S}_{11}||\mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}| = |\mathbf{S}_{11}||\mathbf{S}_{22}||\mathbf{I} - \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}|$, tenemos que

$$\lambda = -n \log(|\mathbf{I} - \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}|) = -n \log \left(\prod_{j=1}^{r} \left(1 - \lambda_j^2\right) \right)$$

donde $r = min(p, q)$ y λ_j^2 son los valores propios de $\mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$, es decir, los coeficientes de correlación canónica al cuadrado. Finalmente el contraste es:

$$\lambda = -n \sum_{j=1}^{r} \log(1 - \lambda_j^2) \tag{16.20}$$

que sigue asintóticamente una distribución χ^2 con grados de libertad igual al número de términos de $\mathbf{V}_{12}$ que son pq. La aproximación mejora si tomamos en lugar de n, la corrección de Bartlett $n - 1/2(p + q + 3)$, con lo que el test resulta:

$$x^2 = -\left(n - \frac{1}{2}(p + q + 3)\right) \sum_{j=1}^{r} log(1 - \lambda_i^2) \tag{16.21}$$

que llevaremos a las tablas de las χ^2 con pq grados de libertad. Rechazaremos H_0 cuando este estadístico sea grande, lo que ocurrirá cuando los coeficientes de correlación canónica λ_j^2 sean grandes.

Este contraste puede extenderse para contrastar que los primeros s coeficientes de correlación canónica son distintos de cero y los restantes $r - s$ son iguales a cero. La hipótesis nula será que la dependencia entre las variables puede expresarse mediante s variables indicadoras, mientras que la alternativa supone que no hay reducción de la dimensión posible y que describir la dependencia require las r dimensiones. El test es entonces:

$$\begin{aligned} H_0 &: \lambda_i > 0 \quad i = 1, \ldots, s; \ \lambda_{S+1} = \ldots = \lambda_r = 0 \\ H_1 &: \lambda_i > 0 \quad i = 1, \ldots, s; \text{ al menos uno } \lambda_j > 0, \ j = s + 1, \ldots, r \end{aligned}$$

y el contraste de la razón de verosimilitudes conduce, como en el caso anterior, a comparar los determinantes de la matriz de covarianzas estimadas bajo H_0 y H_1. Bajo H_1 esta estimación es $\mathbf{S}$ y $|\mathbf{S}| = |\mathbf{S}_{11}|\,|\mathbf{S}_{22}| \prod_{j=1}^{r}\left(1-\lambda_j^2\right)$, mientras que bajo H_0 el determinante estimado debe ser $|\mathbf{S}_{11}|\,|\mathbf{S}_{22}| \prod_{j=1}^{s}\left(1-\lambda_j^2\right)$. Por tanto, el estadístico para el contraste de la razón de verosimilitudes es:

$$\lambda = -\left(n - \frac{1}{2}(p+q+3)\right) \sum_{j=s+1}^{r} \log(1-\lambda_j^2) \tag{16.22}$$

que se distribuye como una χ^2 con $(p-s)\,(q-s)$ grados de libertad. De nuevo rechazamos H_0 cuando el estadístico (16.22) sea grande.

Ejemplo 16.2 *El contraste de que dos conjuntos de variables son independientes para los datos de los hogares del ejemplo 16.1 será, como* $75-(5+4+3)/2 = 69$*,*

$$\begin{aligned} x^2 &= -69\left(\log(1-.44)+\log(1-.21)+\log(1-.05)+\log(1-.01)\right) \\ &= 60,5 \end{aligned}$$

El p valor de 60,5 en una distribución χ^2 con 20 grados de libertad es menor de 0,0001, con lo que rechazamos H_0. Aceptando que la primera es distinta de cero, el contraste es

$$x^2 = -69\left(\log(1-.21)+\log(1-.05)+\log(1-.01)\right) = 20.498$$

y corresponde a una χ^2 con $(5-1)(4-1) = 12$ grados de libertad. El p valor es aproximadamente .05, y rechazaremos H_0. Sin embargo, para las dos raíces restantes:

$$x^2 = -69\left(\log(1-.05)+\log(1-.01)\right) = 4.23$$

y si la hipótesis de que las dos raíces son cero corresponde a una χ^2 con 6 grados de libertad. Como el p valor de 4.23 en esta distribución aceptaremos H_0 y concluiremos que sólo hay dos dimensiones en la relación entre las variables.

16.7 EXTENSIONES A MÁS DE DOS GRUPOS

Supongamos que tenemos 3 conjuntos de variables $\mathbf{X}_1\,(n\times p_1)$, $\mathbf{X}_2\,(n\times p_2)$, $\mathbf{X}_3\,(n\times p_3)$. Podemos buscar tres variables $\mathbf{x}_i^* = \mathbf{X}_i\boldsymbol{\alpha}_i$ $(i=1,2,3)$, una en cada grupo, de manera que la matriz $\mathbf{R}\,(3\times 3)$ de correlaciones entre estas tres variables resultante sea "grande". Según como definamos el tamaño de esta matriz tenemos distintos métodos. Podemos:

(1) Maximizar la suma de las correlaciones al cuadrado, es decir

$$\max\left(\sum r_{ij}^2\right) \tag{16.23}$$

donde $r_{ij} = Corr\left(x_i^*,\, x_j^*\right)$.

(2) Maximizar el mayor valor propio de $\mathbf{R}$.

(3) Minimizar el determinante de $\mathbf{R}$.

En el caso de dos grupos los tres criterios conducen a los coeficientes de correlación canónica. Para varios grupos estos tres criterios no llevan necesariamente a los mismos resultados. Para comprobarlo, en el caso de grupos de variables la matriz de correlación de las dos variables canónicas es:

$$\mathbf{R} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

supuesto $r > 0$ el primer criterio conduce a maximizar r, el segundo también, ya que el mayor valor propio de $\mathbf{R}$ es $1+r$. En el tercer caso el determinante es $|1 - r^2|$ y minimizar el determinante equivale a maximizar r, con lo que en los tres casos maximizamos la correlación entre las combinaciones lineales.

El cálculo de las correlaciones canónicas para varios conjuntos requiere en general cálculos iterativos. El lector interesado puede acudir a Kettenring (1971) y Gnanadesikan (1977).

16.8 RELACIÓN CON OTRAS TÉCNICAS ESTUDIADAS

Además de su interés propio, el análisis de correlaciones canónicas cubre como casos particulares las técnicas de regresión y, por extensión, las de análisis discriminante. En efecto, supongamos primero el caso más simple en que cada uno de los conjuntos tiene únicamente una variable.La correlación canónica entre x e y es el coeficiente de correlación al cuadrado. En efecto, en este caso $p = q = 1$ y $R_{11} = R_{22} = 1$ y $R_{12} = R_{21} = r_{xy}$. Entonces:

$$R_{11}^{-1} R_{12} R_{22}^{-1} R_{21} = r_{xy}^2 = \lambda^2$$

Si el conjunto x tiene varias variables $p = 1$ y $q \geq 1$. La correlación canónica entre la variable endógena o respuesta y el conjunto de las exógenas o regresores es el cuadrado del coeficiente de correlación múltiple. En efecto, ahora $R_{11} = 1$, y llamando $\mathbf{r}_{12}$ al vector de correlaciones entre la endógena y las exógenas y R_{22} a la matriz de correlaciones entre las variables exógenas:

$$\lambda^2 = R_{11}^{-1} R_{12} R_{22}^{-1} R_{21} = \mathbf{r}_{12}' \mathbf{R}_{22}^{-1} \mathbf{r}_{21} = R_{1.2}^2$$

Se obtiene el mismo resultado a partir de las matrices de covarianzas: Entonces $S_{11} = s_y^2$; $\widehat{\beta} = \mathbf{S}_{22}^{-1}\mathbf{S}_{21}$

$$\lambda^2 = \frac{S_{xy}\widehat{\beta}}{s_y^2} = \frac{VE}{VT} = R^2$$

El análisis discriminante puede también abordarse desde correlación canónica. Si definimos $G-1$ variables explicativas y mediante:

$$\begin{array}{rcl} y_i & = & 1 \text{ si la observación pertenece al grupo 1, } \; i = 1, ..., G-1 \\ & & 0 \text{ en otro caso} \end{array}$$

Podemos disponer estas $G-1$ variables binarias en una matriz $\mathbf{Y}$, $n \times (G-1)$. Por otro lado tendremos la matriz $\mathbf{X}$, $n \times p$ de las p variables explicativas. La correlación canónica entre las matrices de datos $\mathbf{Y}$ y $\mathbf{X}$ es análoga al análisis factorial discriminante. Puede demostrarse que, llamando $\mathbf{S}$ a la matriz cuadrada $p+G-1$ de covarianzas entre las variables $\mathbf{Y}$ y $\mathbf{X}$, y utilizando la notación del capítulo 13, $n\mathbf{S}_{11} = \mathbf{T}$ y $n\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} = \mathbf{B}$. Las correlaciones canónicas obtenidas con la matriz $\mathbf{S}_{11}^{-1}\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$ serán las obtenidas con la matriz $\mathbf{T}^{-1}\mathbf{B}$. Para ver la relación entre estas correlaciones canónicas y las obtenidas en análisis dicriminante con la matriz $\mathbf{W}^{-1}\mathbf{B}$, observemos que si llamamos $\mathbf{a}_i$ a los vectores que defienen las variables canónicas y λ a las correlaciones conónicas:

$$\mathbf{T}^{-1}\mathbf{B}\mathbf{a}_i = \lambda_i \mathbf{a}_i$$

entonces

$$(\mathbf{W}+\mathbf{B})^{-1}\mathbf{B}\mathbf{a}_i = \lambda_i \mathbf{a}_i$$

que puede escribirse $(\mathbf{I}+\mathbf{W^{-1}B})^{-1}\mathbf{W}^{-1}\mathbf{B}\mathbf{a}_i = \lambda_i\mathbf{a}_i$, es decir $\mathbf{W}^{-1}\mathbf{B}\mathbf{a}_i = \lambda_i\mathbf{a}_i + \lambda_i\mathbf{W}^{-1}\mathbf{B}\mathbf{a}_i$, que equivale a

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{a}_i = [\lambda_i/(1-\lambda_i)]\mathbf{a}_i$$

con lo que, llamando α_j a los valores propios que definen las variables canónicas discriminantes vemos que estas variables son idénticas a las obtenidas por correlaciones conónicas y los valores propios que defienen estas variables están relacionados por:

$$\alpha_j = \frac{\lambda_i}{1-\lambda_i}$$

16.9 ANÁLISIS CANÓNICO ASIMÉTRICO

El análisis de correlaciones canónicas es simétrico en las variables: si intercambiamos $\mathbf{X}$ por $\mathbf{Y}$, el número de variables canónicas no se modifica, las correlaciones entre las variables canónicas son idénticas y los vectores que definen las variables canónicas se intercambian.

Existen situaciones donde esta simetría no es deseable. Puede ocurrir que las $\mathbf{X}$ sean variables exógenas que queremos utilizar para prever las endógenas $\mathbf{Y}$, y queremos un procedimiento que tenga en cuenta esta asimetría, es decir que maximice la explicación de las variables $\mathbf{Y}$. El análisis de correlaciones canónicas no resuelve el problema. Podemos tener una alta correlación entre $\boldsymbol{\beta}'\mathbf{Y}$ y $\boldsymbol{\alpha}'\mathbf{X}$ y una baja correlación entre cada variable del conjunto $\mathbf{Y}$ y $\boldsymbol{\alpha}'\mathbf{X}$. Por ejemplo, supongamos que en todas las observación se da la relación aproximada :

$$y_1 + \ldots + y_q \simeq \boldsymbol{\alpha}'\mathbf{x}$$

entonces podemos tener un coeficiente próximo a la unidad entre las dos combinaciones lineales $(1, \ldots, 1)\,\mathbf{y}$ y $\boldsymbol{\alpha}'\mathbf{x}$, pudiendo ser, sin embargo, la correlación entre $\boldsymbol{\alpha}'\mathbf{x}$ y cada una de las variables y_i baja.

Si el objetivo es prever cada uno de los componentes podríamos hacer regresiones entre cada uno de ellos y las variables x. Si las variables están estandarizadas eso supone construir ecuaciones de regresión $\widehat{y}_i = \mathbf{x}'\widehat{\boldsymbol{\beta}}_i$ para cada una de las q variables y donde los coeficientes de cada regresión vienen dados por $\widehat{\boldsymbol{\beta}}_i = \mathbf{R}_x^{-1}\mathbf{r}_{1x}$. Sin embargo, de esta manera obtendríamos q ecuaciones distintas, una para cada variable y. Vamos a ver como obtener una ecuación única, $\boldsymbol{\alpha}'\mathbf{x}$, que tenga buenas propiedades para prever el conjunto de las y.

Para medir la capacidad predictiva de un conjunto de variables respecto al otro se introducen los coeficientes de redundancia, que definimos a continuación.

16.9.1 Coeficientes de redundancia

Supongamos para simplificar que las variables originales están estandarizadas (media cero y varianza unidad) y que las combinaciones lineales $\mathbf{X}\boldsymbol{\alpha}_i$ se obtienen con la condición de varianza unitaria, es decir

$$\frac{1}{n}\boldsymbol{\alpha}_i'\mathbf{X}'\mathbf{X}\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i'\mathbf{R}_{xx}\boldsymbol{\alpha}_i = \mathbf{1}.$$

La correlación entre la variable y_1 y una variable indicadora $\mathbf{x}'\boldsymbol{\alpha}$, construida como combinación lineal de las variables $\mathbf{x}$, será:

$$E\left[y_1\mathbf{x}'\boldsymbol{\alpha}\right] = \mathbf{r}_{1x}'\boldsymbol{\alpha}$$

donde $r_{1x}' = (r_{11}, ..., r_{1p})$ es el vector de correlaciones entre la variable y_1 y las p variables $\mathbf{x}$. La correlación de las q variables y con la variable $\boldsymbol{\alpha}'\mathbf{x}$ será

$$E\left[\mathbf{y}\mathbf{x}'\boldsymbol{\alpha}\right] = \mathbf{R}_{yx}\boldsymbol{\alpha}$$

donde $\mathbf{R}_{yx}\boldsymbol{\alpha}$ es la matriz $(q \times p)$ de correlaciones entre las q variables $\mathbf{y}$ y las variables $\mathbf{x}$. Se define el *coeficiente de redundancia* para explicar el conjunto de las variables $\mathbf{y}$ con la variable $\mathbf{x}'\boldsymbol{\alpha}$, como el valor promedio del cuadrado de las correlaciones entre las $\mathbf{y}$ y la variable indicadora $\boldsymbol{\alpha}'\mathbf{x}$, es decir:

$$CR\left(\mathbf{y}|\mathbf{x}'\boldsymbol{\alpha}\right) = \frac{1}{q}\boldsymbol{\alpha}'\mathbf{R}_{yx}\,\mathbf{R}_{xy}\,\boldsymbol{\alpha} \tag{16.24}$$

Si tenemos $r = \min(p, q)$ combinaciones lineales $\mathbf{x}'\boldsymbol{\alpha}_1, \ldots, \mathbf{x}'\boldsymbol{\alpha}_r$, *la redundancia total* para explicar el conjunto de las variables $\mathbf{y}$ con el conjunto de las variables x a través de estas combinaciones lineales es:

$$R(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{r} CR\left(\mathbf{y}|\mathbf{x}'\boldsymbol{\alpha}_i\right), \tag{16.25}$$

y es una medida de la correlación. Análogamente podríamos construir $R(\mathbf{x}|\mathbf{y})$ que es, en general, distinta de $\mathbf{R}(\mathbf{y}|\mathbf{x})$.

Las expresiones del coeficiente de redundancia y de la redundancia total se definen entre las variables $\mathbf{y}$ y ciertas combinaciones lineales $\mathbf{x}\boldsymbol{\alpha}_i$. En el caso en que los coeficientes $\boldsymbol{\alpha}$ se obtengan por correlación canónica, puede demostrarse que

$$R(\mathbf{y}|\mathbf{x}) = \frac{tr\ (\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy})}{tr\,(\mathbf{R}_{yy})} \tag{16.26}$$

que puede escribirse como:

$$R(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{r} \frac{1}{q} R_j^2 \tag{16.27}$$

donde R_j^2 es el coeficiente de correlación múltiple al cuadrado en una regresión entre la variable y_j y el conjunto de las variables $\mathbf{X}$. En efecto, el numerador de (16.26) es la suma de los términos diagonales de la matriz, que tienen la forma $\mathbf{r}'_{y_jx}\,\mathbf{R}_{xx}^{-1}\,\mathbf{r}_{y_jx}$, que es la expresión del cuadrado del coeficiente de correlación múltiple entre una variable y_j y un vector de variables $\mathbf{x}$, ambas estandarizadas, y el denominador $tr\,(\mathbf{R}_{yy}) = q$, por estar las variables estandarizadas.

16.9.2 Análisis canónico asimétrico

Supongamos que se desea encontrar la combinación $\mathbf{x}'\boldsymbol{\alpha}$ con máxima correlación con cada variable y_i individualmente, de manera que la suma de las correlaciones al cuadrado entre $\mathbf{x}'\boldsymbol{\alpha}$ y las variables $\mathbf{y}$ sea máxima. Esto supone maximizar $\boldsymbol{\alpha}'\mathbf{R}_{xy}\mathbf{R}_{yx}\boldsymbol{\alpha}$ con la restricción $\boldsymbol{\alpha}'\mathbf{R}_{xx}\boldsymbol{\alpha} = 1$, lo que conduce a la ecuación

$$\mathbf{R}_{xy}\mathbf{R}_{yx}\boldsymbol{\alpha} = \gamma\mathbf{R}_{xx}\boldsymbol{\alpha}$$

y vemos que $\boldsymbol{\alpha}$ debe ser un vector propio de la matriz $\mathbf{H} = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yx}$.

Análogamente a correlación canónica podríamos preguntarnos por una segunda variable $\mathbf{x}'\boldsymbol{\alpha}_2$, ortogonal a la primera, que tenga máxima correlación con el vector de variables endógenas. Este problema lleva a tomar el segundo vector propio de la matriz $\mathbf{H}$. De la misma forma, las restantes variables $\mathbf{x}'\boldsymbol{\alpha}_3, \ldots, \mathbf{x}'\boldsymbol{\alpha}_q$ (suponemos $p > q$) se obtendrán como vectores propios de la matriz $\mathbf{H}$.

Este mismo análisis puede aplicarse para explicar las variables $\mathbf{X}$ con las $\mathbf{Y}$, pero ahora el problema no es simétrico, ya que los vectores propios de las matrices $\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yx}$ y $\mathbf{R}_{yy}^{-1}\mathbf{R}_{yx}\mathbf{R}_{xy}$ no serán en general iguales, ni estas matrices tendrán vectores propios ortogonales.

Este tipo de análisis ha sido desarrollado por Stewart y Love (1968) y Gudmundsson (1977).

16.10 Lecturas complementarias

El análisis de correlaciones canónicas se presenta en la mayoría de los textos generales de análisis multivariante. Buenas exposiciones del análisis de correlaciones canónicas se encuentran en Rechner (1998) y Anderson (1984). El lector interesado en un estudio más detallado de las distribuciones de los estadísticos puede acudir a Muirhead (1982). El análisis canónico se ha utilizado mucho en series temporales para determinar la dimensión de una serie temporal. Algunas de estas aplicaciones pueden encontrarse en Peña, Tiao and Tsay (2001).

EJERCICIOS

16.1 Demostrar que si transformamos las variables con $\mathbf{y}_n = \mathbf{F}\mathbf{y} + \mathbf{a}$ y $\mathbf{x}_n = \mathbf{G}\mathbf{x} + \mathbf{a}$, donde $\mathbf{F}$ y $\mathbf{G}$ son matrices no singulares las correlaciones canónicas obtenidas con las matrices de varianzas y covarianzas de estas nuevas variables son idénticas a las de las variables originales (sugerencia, calcular la matriz $\mathbf{A}_n$ (16.11) para estas nuevas variables y comprobar que $\mathbf{A}_n$ se escribe como $\mathbf{A}_n = (\mathbf{G}')^{-1}\mathbf{A}_n\mathbf{G}'$)

16.2 Comprobar que si transformamos las variables como en el ejercicio anterior las variables canónicas si se modifican

16.3 Demostrar que el coeficiente de correlación canónica λ_i^2 es el coeficiente de determinación en una regresión multiple con respuesta la variable $y_i^* = \boldsymbol{\beta}_i'\mathbf{y}$, y variables explicativas las $\mathbf{x}$. (Sugerencia, comprobar que la correlación entre y_i^* y $\mathbf{x}$ es $\boldsymbol{\beta}_i'\mathbf{V}_{21}$ y utilizar la expresión del coeficiente de correlación múltiple como $\mathbf{r}_{yx}\mathbf{R}_x^{-1}\mathbf{r}_{yx}$ y la expresión (16.15)).

B I B L I O G R A F I A

Aitchinson,J. (1986), *The Statistical Analysis of Compositional Data,* Chapman and Hall, London.

Aluja, T. y Morineau, A. (1999), *Aprender de los datos : El análisis de componentes principales,* EUB, Barcelona.

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, (2 ed), New York, Wiley.

Anderberg, G.M.R. (1973), *Cluster Analysis for Applications*, New York, Academic Press.

Arnold, S.F. (1981), *The Theory of Linear Models and Multivariate Observations*, New York, Wiley.

Barnett, V. (1981), *Interpreting Multivariate Data*, New York, Wiley.

Bartholomew, D. J. y Knott, M. (1999), *Latent Variable Models and Factor Analysis,* Arnold.

Basilewsky, A. (1983), *Applied Matrix Algebra, in the Statistical Science*, North-Holland.

Batista, J.M. y Martínez, M.R. (1989), *Análisis Multivariante*, ESADE.

Benzecri, J. (1976), *L'analyse des données*, Dunod.

Bernstein I. H. (1987), *Applied Multivarite Analysis*, Springer - Verlag.

Bertir, P. y Bouroche, J.M. (1975), *Analysis des données multidimensionelles*, Press Univ. France.

Bolche, B. y Huang, C. (1974), *Multivariate Statistical Methods for Business and Economics*, Prentice Hall.

Bollen, K. A. (1989) *Structural Equations with Latent Variables,* Wiley

Breiman, L. et al (1984) *Classification and Regression Trees*, Wadsworth.

Cuadras, C.M. (1991), *Métodos de Análisis Multivariante*, Editorial Universitaria de Barcelona (2 edición).

Coxon A.P.M.(1982),*The User's guide to Multidimensional Scaling*, Exeter,NH: Heinemann Educational Books.

Cherkassky, V. y Mulier, F. (1998), *Learning from Data,* Wiley

Davidson M. L. (1983), *Multidimensional Scaling*, Wiley.

Dempster, A.P., (1969), *Elements of continuous multivariante analysis*, Addison-Wesley.

Dillon, W., Goldstein, M. (1984), *Multivariate Analysis*, New York, Wiley.

Drosbeke, J.J., Fichet, B. et Tassi P. (1992), *Modeles por l'analyse des donnees multidimensionnelle,* Economica.

Eaton, M.L. (1983), *Multivariate Statistics*, New York, Wiley.

Escofier, B. et Pagés, J. (1992), *Análisis factoriales simples y múltiples,* Universidad del País Vasco.

Escudero, L. F. (1977), *Reconocimiento de patrones*, Paraninfo.

Everitt, B.S. (1978), *Graphical Techniques for Multivariate Data*, New York, North-Holland.

Everitt, B.S. (1984), *An Introduction to Latent variable models*, Chapman and Hall.

Everitt, B.S. (1993), *Cluster Analysis* , Oxford University Press.

Finn, J. D. (1974), *A General Model for Multivariate Analysis,* Holt, Rinehart and Winston.

Fox, J. (1984). *Linear Statistical Models and related methods with applications to social research*, Wiley.

Flury, B. (1997), *A First Course in Multivariate Statistics*, Springer.

Flury, B. (1988), *Common Principal Components and Related Multivariate Methods,* Wiley.

Flury, B. y Riedwyl, H. (1988), *Multivariate Statistics*, Chapman and Hall.

Fakunaga, K. (1990), *Introduction to Statistical Pattern Recognition (2ª edición),* Academic Press.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Wiley.

Giri, C.C. (1977), *Multivariate Statistical Inference*, New York, Academic Press.

Gnanadesikan, R. (1997), *Methods for Statistical Data Analysis of Multivariate Observations*, New York, Wiley (2ªedición).

Goldstein, M. y Dillon, W.R. (1978), *Discrete Discriminant Analysis*, New York, Wiley.

Gordon A. D. (1981), *Classification*, Chapman and Hall.

Gower, J and Hand D. (1996), *Biplots*, Chapman and Hall

Green, P. E. (1978) *Mathematical tools for applied multivariate analysis*, Academic Press

Green, P, E, Carmone, F, J, y Smith S, M, (1989), *Multidimensional Scaling: Methods and Applications*, Boston: Allyn y Bacon.

Greenacre, M.J. (1984), *Theory and Applications of correspondence analysis*, Academic Press.

Greenacre, M.J. (1993), *Correspondence Analysis in Practice*, Academic Press.

Greenacre, M.J.and Blasius (1994), *Correspondance Analysis in the Social Science*, Academic Press.

Hair, J. F. et al (1995), *Multivariate Data Analysis with Readings*, Prentice Hall.

Hand, D. J. (1997) *Construction and Assessment of Classification Rules,* Wiley

Harman, H.H. (1980), *Análisis Factorial Moderno*, Saltés, Madrid.

Hartigan, J.A. (1975), *Clustering Algorithms*, New-York, Wiley.

Huberty,C. (1994) *Applied Discriminant Analyisis,* Wiley

Jackson, J.E. (1991), *A User's guide to principal components*, Wiley.

Jambu, M. (1991), *Exploratory and Multivariate Data Analysis*, Academic Press.

Johnson, N.L. y Kotz, (1972), *Distributions in Statistics: Continuos Multivariate Distributions*, New York, Wiley.

Joliffe, I. T. (1986), *Principal Components Analysis*, Springer-Verlag.

Johnson, R.A. y Wichern, D.W. (1998), *Applied Multivariate Statistical Analysis,* (Fourth edition), Prentice Hall

Jöreskog, K,G, (1963) *Statistical estimation in Factor Analysis,* Almqvist and Wicksell.

Jöreskog, K.G. y Sörbon, D. (1979), *Advances in Factor Analysis and Structural Equation Models*, Cambridge, MA:ABT.

Kachigan, S.K. (1982), *Multivariate Statistical Analysis*, New York, Radins Press.

Karson, M.J. (1982), *Multivariate Statistical Methods*, The Iowa State, University Press.

Kendall, M.G. y Stuart, A. (1967), *The Advaced Theory of Statistics*, Vol. 2, New York Harper.

Kendall, M. (1975), *Multivariate Analysis*, Charles Griffin, Londres.

Kruskal, J.B. y Wish, M. (1978), *Multidimensional Scaling*, Murray Hill, N.J. Bell Laboratories.

Krzanowski, W.J. (1988), *Principles of Multivariate Analysis: A. User's Perspective*, Oxford University Press, Oxford.

Krzanowski, W.J.and Marriot, F. H. C. (1994), *Multivariate Analysis: Part I,* Edward Arnold, London.

Krzanowski, W.J.and Marriot, F. H. C. (1995), *Multivariate Analysis: Part II,* Edward Arnold, London.

Lachenbruch, P.A. (1975), *Discriminant Analysis*, New York, Hafner Press.

Lawley, D.N. y Maxwell, A.E. (1971), *Factor Analysis as a Statistical Method*, New York, American Elsevier.

Lebart, L. y Fenelon, J.P. (1973), *Statistique et Informatique Appliquees*, París, Dunod.

Lebart, L., Morineau, A. y Fenelon, J.P. (1985), *Tratamiento estadístico de datos*, Marcombo.

Lebart, L., Morineau, A. y Warwick, K.M. (1984). *Multivariate Descriptive Analysis*, New York, Wiley.

Lebart, L., Morineau, A. y Piron, M. (1997), *Statistique exploratoire multidimensionnelle,* Dunod.

Lebart, L., Salem, A. y Bécue M. (2000), *Análisis estadístico de textos,* Milenio.

Lefebre, J. (1976), *Introduction aux Analyses Statistiques Multidimensionnelles*, Masson, Paris.

McLachan, G. J. y Basford, K. (1988) *Mixture Models*, Marcel Dekker.

McLachan, G. J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*, Wiley

Mardia, K.V., Kent, J.T. y Bibby, J.M. (1979), *Multivariate Analysis*, New York, Academic Press.

Maxwell, A.E. (1977). *Multivariate Analysis in Behavioural Research*, Chapman and Hall

McQuarrie, A. D. R. y Tsai, C. L.(1998): *Regression and time series model selection,* Singapore: World Scientific.

Miller, A. J. (1990): *Subset Selection in Regression,* Chapman and Hall

Mirkin, B. (1996), *Mathematical Classification and Clustering*, Kluwer Academic Publishers

Morrison, B.F. (1976), *Multivariate Statistical Methods*, (2 ed.), New York, Academic Press.

Muirhead, R.J. (1982), *Aspect of Multivariate Analysis*, New York, Wiley.

Press, S.J. (1972), *Applied Multivariate Analysis*, Holt, Rinehart.

Rao, C.R. (1973), *Linear Statistical Inference and its Applications*, New York, Wiley.

Rechner, A. C. (1995) *Methods of Multivariate Analysis,* Wiley

Rechner, A. C. (1998) *Multivariate Inference and its Applications,* Wiley

Sánchez Carrión, J.J. (1984). *Introducción a las técnicas de Análisis Multivariante aplicados a las Ciencias Sociales*, CIS.

Saporta, G. (1990). *Probabilités Analyse de données et statistique*, Tecnip.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data,* Chapman and Hall.

Schiffman, S.S., Reynolds, M.L. y Young, F.W. (1981). *Introduction to Multidimensional Scaling*, New York, Academic Press.

Seber, G.A.F. (1984), *Multivariate Observations*, New York, Wiley.

Spath, H. y Bull, U. (1980), *Cluster Analysis of Algorithms for Data Reduction and Classifications of Objects*, New York, Wiley.

Spath, H. (1985), *Cluster Dissection and Analysis*, Chichester: Ellis Horwood.

Srivastava, M.S. y Carter, E.M., (1983), *An Introduction to Applied Multivariate Statistics*, New York, North Holland.

Srivastava, M.S. y Khatri, C.G. (1979), *An Introduction to Multivariate Statistics*, New York, North Holland.

Stevens, J. (1986). *Applied Multivariate Analysis for the Social Sciences.*

Tabachnik, B. G. y Fidell, L. S. (1996) *Using Multivariate Statistics,* HarperCollins.

Tastsuoka, M (1971). *Multivariate Analysis*, Monterrey CA, Brooks/Cok.

Tatsuoka, M. (1988). *Multivariate thecniques*, Macmillan Publ.

Torrens-Ibern, J. (1972), *Modèles et methodes de l'analyse factorielle*, Dunod.

Vapnik, V. V. (2000) *The Nature of Statistical Learning (2ª edición),* Springer

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Wiley.

Young, F.W. (1987), *Multidimensional Scaling: History, Theory and Applications,* (R.M. Hamer, ed.) HillsdaleNJ: Lawrence Erlbaum Associates.

Otras referencias

Anderson, T.W. (1963). Asymptotic Theory for Principal Componentes. *Annals of Statistics*, 34, 122-148.

Anderson, T.W. (1996). R.A. Fisher and Multivariate Analysis. *Statistical Science*, 11,1 20-34.

Anderson, T.W., and Bahadur, R.R. (1962), "Classification into two multivariate normal distributions with different covariance matrices," *Annals of Mathematical Statistics,* 33, 420-431.

Aitchinson,J. y Dunsmore, I.R. (1975), *Statistical Prediction Analysis,* Cambridge University Press.

Atkinson, A.C. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers", *Journal of the American Statistical Association*, 89, 1329–1339.

Banfield, J.D., and Raftery, A. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics,* 49, 803-821.

Bartlett, M. S. (1954), "A note on multiplying factors for various chi-squared approximations," *J. Royal Statistical Society B*, 16,296-298.

Barnett, V., and Lewis, T. (1978) *Outliers in Statistical Data.* Wiley.

Bartholomew, D. J. (1995). Spearman and the origin and development of factor analysis. *British Journal of Statistical and Mathematical Psycology,* 48, 211-220

Bensmail, H., and Celeux, G. (1997), "Inference in Model- Based cluster analysis," *Statistics and Computing,* 7, 1-10.

Bernardo J. M. y Smith, A. F. M. (1994): *Bayesian Theory.* Wiley.

Bibby ,J. y Toutenberg, H. (1977) *Prediction and improved estimation in linear models.* Wiley.

Binder, D. A. (1978), "Bayesian Cluster Analysis," *Biometrika,* 65, 31-38.

Box, G. E. P. (1949) "A general distribution theory for a class of likelihood criteria", *Biometrika*, 36, 317-346.

Box, G. E. P. y Tiao, G. C. (1968) "A Bayesian Approach to Some Outlier Problems", *Biometrika*, 55, 119–129.

Box, G. E. P. y Tiao, G. C. (1973): *Bayesian Inference in Statistical Analysis.* Addison-Wesley.

Boser, B. Guyon, I. and Vapnik. V. (1992). A training algorithem for optimal margin classifiers. Proceedings of COLT II, Philadelphia.

Breiman, L. et al. (1984). *Classification and Regression Trees.* Wadsworth.

Caballero, B. y Peña, D.(1987) Un estudio estadístico de la Investigación científica en los países de la OCDE". *Estadística Española.* 29, 114, 151-178.

Carlin

Casella, G. y Berger, R. L.(1990). *Statistical Inference.* Thomson.

Cuadras,C. (1993). Interpreting an Inequality in Multiple Regression. *The American Statistician , 47, 256-258.*

Celeux, G., Hurn, M., and Robert, C. P. (2000). "Computational and Inferencial Difficulties with mixture posterior distributions," *Journal of the American Statistical Association,* 95, 957-970.

Cook, D., Buja, A., Cabrera, J., and Hurley, C. (1995), "Grand Tour and Projection Pursuit," *Journal of Computational and Graphical Statistics,* 4, 155-172.

Cherkassky, V. y Mulier, F. (1998), *Learning from Data,* Wiley.

Chow, G. (1981) A Comparison of the Information and Posterior Probability Criteria for model Selection. *Journal of Econometrics,* 16,21-33.

Dasgupta, A., and Raftery, A. E. (1998) " Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering, " *Journal of the American Statistical Association,* 93, 294-302.

David, H.A. (1998)," Statistics in U.S. Universities in 1933 and the establishment of the statistical laboratory at Iowa State," *Statistical Science*, 13, 66-74.

Donoho, D.L. (1982), "Breakdown Properties of Multivariate Location Estimators". Ph.D. qualifying paper, Harvard University, Dept. of Statistics.

Efron, B. (1982). *The Jackknife, the Bootstrap and other resampling plans.* SIAM. Philadelphia.

Efron (1975) demostró que cuando los datos so

Efron, B. y Tibshirani, J. (1993). *An introduction to the Bootstrap.* Chapman and Hall.

Friedman, J.H. (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association,* 82, 249-266.

Friedman, H.P., and Rubin, J. (1967), "On some Invariant Criteria for Grouping Data," *Journal of the American Statistical Association,* 62, 1159-1178.

Friedman, J.H., and Tukey, J.W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers,* C-23, 881-889.

Gill, P.E., Murray, W., and Wright, M.H. (1981), *Practical Optimization.* Academic Press.

Fox, J. (1984). *Linear Statistical Models and related Methods.* Wiley.

Gamerman, D. (1997). *Markov Chain Monte Carlo.* Chapman and Hall.

Gelman, A., Carlin, J., Stern, H. y Rubin, D. B. (1997). *Bayesian Data Analysis.* Chapman and Hall.

Graybill, F. A. (1976): *Theory and Application of the Linear Model.* Wadsworth Publ. Com.

Graybill (1983)

Gudmundsson, G. (1977). "Multivariate Analysis of Economic Variables", *Applied Statistics*, 26, 1, 48-59.

Hample, F. R. y otros (1986): *Robust Statistics*. Wiley.

Hand, D. J. at al (2000) Data mining for fun and profit. *Statistical Science*, 15,111-131.

Hardy, A. (1996), "On the Number of Clusters," *Computational Statistics and Data Analysis,* 23, 83-96.

Hartigan, J.A. (1975), *Clustering Algorithms.* New York: Wiley.

Hartigan, J.A., and Wong, M.A. (1979), "A k-means clustering algorithm," *Applied Statistics,* 28, 100-108.

Huber, P.J. (1985), "Projection Pursuit," *The Annals of Statistics,* 13, 435-475.

Hastie , T. y Simard, P. Y. (1988) Metrics and Models for Handwritten character recognition. *Statistical Science*, 13, 54-65.

Hastie, T., Tibshirani R. y Buja A. (1994) Flexible discriminant analysis for optimal scoring. *Journal of American Statistical Association*, 428, 1255-1270.

Hernández y Velilla, 2001

Izenman A. J. (1975), Reduced Rank Regression for the multivariate linear model. *J. Multivariate Analysis*, 5, 248-264.

Jeffreys, H. (1961), *Theory of Probability.* Oxford Clarendon Press.

Jones, M.C., and Sibson, R. (1987), What Is Projection Pursuit?, *Journal of the Royal Statistical Society, Series A,* 150, 1-18.

Jöreskog, K.G. (1973), A general method for estimating a structural equation model. En A.S. Goldberg and O.D.Ducan, eds, *Structural Equation Models in the Social Sciences.* Academic Press, pp 85-112.

Juan, J. y Prieto, F. J. (2001)," Using Angles to identify concentrated multivariate outliers," *Technometrics*, 3, 311-322.

Justel, A. y Peña, D. (1996)," Gibbs Sampling Will Fail in Outlier Problems with Strong Masking,". *Journal of Computational and Graphical Statistics*, 5,2, 176-189.

Justel A., Peña, D. y Zamar,R. (1997)," A Multivariate Kolmogorov Smirnov goodness of fit test," *Statistics and Probability Letters,* 35, 251-259.

Kettenring, J. R. (1971)," Canonical analysis of several sets of variables," *Biometrika,* 58, 33-451.

Lanterman, A. D. (2001)," Schwarz, Wallace and Rissanen: intertwining themes in theories of model selection," *International Statistical Review*, 69, 185-212.

Little R. J .A. y Rubin, D. (1987), *Statistical analysis con missing Data.* Wiley

Lee, P. M. (1997), *Bayesian Statistics.* Oxford University Press.

MacQueen J. B.(1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkely Simposium in Mathematical Statistics and Probability,* pp 281-297. University of California Press

Maronna, R., and Jacovkis, P.M. (1974), "Multivariate Clustering Procedures with Variable Metrics," *Biometrics,* 30, 499-505.

Maronna, R.A. (1976), "Robust M-Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51–67.

Maronna, R.A. and Yohai, V.J. (1995), "The Behavior of the Stahel-Donoho Robust Multivariate Estimator," *Journal of the American Statistical Association*, 90, 330–341.

Nason, G. (1995), "Three-Dimensional Projection Pursuit," *Applied Statistics,* 44, 411-430.

O'Hagan, A. (1994). *Bayesian Inference*. Edward Arnold.

Peña, D. (2001). *Fundamentos de Estadística*. Alianza Editorial

Peña, D. (2002). *Regresión y Diseño de Experimento*s. Alianza Editorial

Peña, D. y Guttman, I. (1993)

Peña, D. y Prieto,F. J (2000)," The Kurtosis Coefficient and the Linear Discriminant Function " *Statistic and Probability Letters,* 49, 257-261.

Peña, D. y Prieto,F. J. (2001a) "Robust covariance matrix estimation and multivariate outlier detection," (con discusión), *Technometrics*, 3, 286-310,

Peña, D. y Prieto,F. J (2001b) " Cluster Identification using Projections," *The Journal of American Statistical Association*, 96, December 2001

Peña D. y Rodriguez, J. (2002) " A powerful portmanteau test for time series," *The Journal of American Statistical Association,* June 2002, (en prensa)

Peña, D., Tiao, G. C. y Tsay, R. S. (2001), *A Course in Time Series*, Wiley

Posse, C. (1995), "Tools for Two-Dimensional Exploratory Projection Pursuit," *Journal of Computational and Graphical Statistics,* 4, 83-100.

Pollock D. S. G. (1993) *The Algebra of Econometrics.* Wiley

Press, S. J. (1989): *Bayesian Statistics*. Wiley.

Robsert, C. P. (1994). *The Bayesian Choice*. Springer.

Robert, C. P. y Casella, G. (1999). *Montecarlo Statistical Methods*. Springer-Verlag.

Rocke, D.M. and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association*, 91, 1047–1061.

Ruspini, E.H. (1970), "Numerical Methods for Fuzzy Clustering," *Information Science,* 2, 319-350.

Rousseeuw, P.J. and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–639.

Spearman C. (1904) General intelligence, objectively determined and measured, *American Journal of Psycology,* 15, 201-293.

Stahel, W.A. (1981), "Robuste Schatzungen: Infinitesimale Optimalitat und Schatzungen von Kovarianzmatrizen," Ph.D. Thesis, ETH Zurich.

Stewart, D. y Love, W. (1968). "A General Canonical Correlation Index", *Psychological Bulletin*, 70, 160-163.

Titterington, D. M., Smith, A. F. M. y Makov, U.E. (1987). *Statistical Analysis of finite mixture distributions.* Wiley.

Velilla (1993)

Wolfe, J. H. (1970) Pattern Clustering by multivarite mixture analysis. *Multivariate Behavioral Research*, 5, 329-350.

Vapnik, V. (1996) *The nature of Statistical Learning.* Springer-Verlag.

Yohai, V. y Garcia Ben, M. (1980), " Canonical variables as optimal predictors," *Annals of Statistics*, 8, 865-869.

Zellner, A. (editor) (1980): *Bayesian Analysis in Econometrics and Statistics*. North Holland.

Apéndice A

Datos

En este capítulo figuran las tablas de datos y una breve descripción de las variables de todos los conjuntos de datos analizados en el libro.

La siguiente tabla muestra asociado al nombre de cada conjunto las técnicas multivariantes que se le han aplicado.

	EUROALI	EUROSEC	EPF	INVEST	MEDIFIS	MUNDODES
Comp. Principales	★		★			
Anál. Factorial		★			★	
Correlaciones Cano.				★		
Cordenadas Prin.		★				★
Anál. Discriminante		★				
Anál. de Congl.						
Regre. Logist.						
Descriptiva						
Anál. de corres.						

EUROALI

	CR	CB	H	L	P	C	F	N	FV
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Bélgica	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Checoslova.	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Dinamarca	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
Alemania-E.	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finlandia	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
Francia	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Grecia	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungria	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Irlanda	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italia	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Países Bajos	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Noruega	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Polonia	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Rumania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
España	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Suecia	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Suiza	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
Reino Unido	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
URSS	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
Alemania-O.	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

Tabla A.1: Datos EUROALI

EUROALI Este conjunto de datos está constituido por 25 observaciones y 9 variables. Las obsrvaciones corresponden a países Europeos, y las variables al porcentaje de consumo de proteínas que cada tipo de alimento proporciona.

Las variables son: Carnes rojas (CR), Carnes blancas (CB), Huevos (H), Leche (L), Pescado (P), Cereales (C), Fecula (F), Nueces (N), Fruta y verdura (FV).

Datos : Tabla A.1

Fuente: Eurostat, 1999

EUROSEC El número de observaciones es 26 y el de variables 9. Las observaciones corresponden a países Europeos. Las variables miden el porcentaje de empleo en los distintos sectores económicos.

EUROSEC

	Agr	Min	Man	PS	Con	SI	Fin	SPS	TC
Bélgica	3.3	0.9	27.6	0.9	8.2	19.1	6.2	26.6	7.2
Dinamarca	9.2	0.1	21.8	0.6	8.3	14.6	6.5	32.2	7.1
Francia	10.8	0.8	27.5	0.9	8.9	16.8	6	22.6	5.7
Alemania-E	6.7	1.3	35.8	0.9	7.3	14.4	5	22.3	6.1
Irlanda	23.2	1	20.7	1.3	7.5	16.8	2.8	20.8	6.1
Italia	15.9	0.6	27.6	0.5	10	18.1	1.6	20.1	5.7
Luxenburgo	7.7	3.1	30.8	0.8	9.2	18.5	4.6	19.2	6.2
Paises Bajos	6.3	0.1	22.5	1	9.9	18	6.8	28.5	6.8
Reino Unido	2.7	1.4	30.2	1.4	6.9	16.9	5.7	28.3	6.4
Austria	12.7	1.1	30.2	1.4	9	16.8	4.9	16.8	7
Finlandia	13	0.4	25.9	1.3	7.4	14.7	5.5	24.3	7.6
Grecia	41.4	0.6	17.6	0.6	8.1	11.5	2.4	11	6.7
Noruega	9	0.5	22.4	0.8	8.6	16.9	4.7	27.6	9.4
Portugal	27.8	0.3	24.5	0.6	8.4	13.3	2.7	16.7	5.7
España	22.9	0.8	28.5	0.7	11.5	9.7	8.5	11.8	5.5
Suecia	6.1	0.4	25.9	0.8	7.2	14.4	6	32.4	6.8
Suiza	7.7	0.2	37.8	0.8	9.5	17.5	5.3	15.4	5.7
Turkia	66.8	0.7	7.9	0.1	2.8	5.2	1.1	11.9	3.2
Bulgaria	23.6	1.9	32.3	0.6	7.9	8	0.7	18.2	6.7
Checos.	16.5	2.9	35.5	1.2	8.7	9.2	0.9	17.9	7
Alemania-O	4.2	2.9	41.2	1.3	7.6	11.2	1.2	22.1	8.4
Hungria	21.7	3.1	29.6	1.9	8.2	9.4	0.9	17.2	8
Polonia	31.1	2.5	25.7	0.9	8.4	7.5	0.9	16.1	6.9
Rumanía	34.7	2.1	30.1	0.6	8.7	5.9	1.3	11.7	5
URSS	23.7	1.4	25.8	0.6	9.2	6.1	0.5	23.6	9.3
Yugoslavia	48.7	1.5	16.8	1.1	4.9	6.4	11.3	5.3	4

Tabla A.2: Datos EUROSEC

Los sectores son: Agricultura, Minería, Industria, Energía, Construcción, Servicios Industriales, Finanzas, Servicios, Transportes y Comunicaciones.

Datos: Tabla A.2

Fuente: Euromonitor (1979, pp. 76-77).

ENCUESTA DE PRESUPUESTOS FAMILIARES

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Almería	618957	295452	522128	167067	58288	280035	129219	307967	107334
Cádiz	683940	203619	426690	124162	60657	285888	127792	313238	83523
Córdoba	590770	265604	487143	113386	37131	237320	116764	247536	79924
Granada	547353	238822	458338	119540	40340	236694	103901	272308	72813
Huelva	649225	245722	570631	99250	61953	253286	123244	238880	83070
Jaén	556210	183295	332662	86364	37160	136992	57607	189811	57311
Málaga	617778	201348	508252	121010	63518	256973	128336	323632	93971
Sevilla	621570	208156	549399	137408	45101	298000	118269	308524	84514
Huesca	577107	249310	412907	107976	39602	335334	90547	227266	92103
Teruel	545238	199788	343919	122154	42281	224286	90291	237747	77938
Zaragoza	556737	266468	496989	132517	54106	235188	118931	282369	79718
Asturias	624941	280273	530828	132066	57679	340013	149265	315478	120856
Baleares	564220	226816	602397	144005	86803	358290	150551	351555	131802
Las Palmas	632640	201704	522846	153775	84148	327988	173031	305628	114627
Tenerife	523476	171072	467424	118857	65247	303598	142620	283563	80959
Cantabria	604083	287943	654929	119269	63320	302277	116752	276663	105421
Ávila	543595	242609	388063	92808	47035	254563	74522	250853	82061
Burgos	602307	255567	600121	162166	51308	280023	132161	301813	111224
León	623047	245240	500414	136030	41667	333066	119657	267506	146434
Palencia	589710	206048	449113	113714	34787	248620	115825	294068	109264
Salamanca	488784	167814	400049	78217	24476	195065	69846	193056	54442
Segovia	528040	184840	455368	103446	46337	217156	91436	259705	116303
Soria	679722	232673	503695	129768	55000	272249	117587	300014	120803
Valladolid	567361	223201	566433	140573	46111	254216	149041	327774	98430
Zamora	544527	178835	402354	99953	32143	227163	70283	231577	125332
Albacete	535939	199559	425598	137799	55967	232209	104866	291708	91735
Ciudad Real	545912	227255	487651	125740	44001	230820	88650	230213	90886
Cuenca	506814	194156	420488	109533	50961	220678	78673	270038	103288
Guadalajara	546909	179824	477446	115585	40129	299174	94923	287703	87720
Toledo	583325	255527	411896	130747	65345	282127	105872	241749	122189
Barcelona	702920	257429	702315	168696	97595	365255	239187	379319	99929
Gerona	684186	285047	566149	149308	77553	259839	191400	329089	134786
Lérida	696542	283134	508906	146773	90828	402073	180652	353124	152924
Tarragona	586122	283112	557653	150464	62853	331848	185713	381485	114876
Alicante	579355	205685	490235	134254	68141	297939	117710	316675	111756
Castellón	496559	201606	411972	107739	42939	212051	84610	241795	77370
Valencia	539570	228072	464127	138419	62471	285948	134751	384939	96564
Badajoz	430442	204529	332948	91831	41112	187500	77481	203808	61478
Cáceres	569808	222756	403547	119078	47904	248571	100282	285880	89736

EPF (continúa)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
La Coruña	692445	249121	506616	141805	56114	277401	142246	289111	108489
Lugo	719078	286277	414893	142629	62779	301867	101889	216693	119398
Orense	598669	182378	370866	106873	31779	221028	114728	205921	90184
Pontevedra	736441	263479	468984	136204	50815	344289	129685	309349	100255
Madrid	670577	253928	864553	148014	86601	393664	232387	440275	130290
Murcia	610718	210169	470859	128627	46866	318508	102978	311262	114457
Navarra	669082	324877	704572	221954	81180	415313	185493	411027	156493
Alava	664450	234132	631137	189169	58406	313033	164730	355280	98942
Guipuzcua	643141	254653	668435	151454	61985	302491	169527	405259	109995
Vizcaya	635929	283160	677817	156612	67899	337253	176222	423122	132572
La Rioja	634839	209753	542656	127615	54684	269843	126717	322845	121844
Ceuta y Melilla.	678733	192344	362317	81673	27191	138705	81979	226279	65135

Tabla A.3: EPF

EPF Estos datos corresponden a 51 observaciones y 9 variables. Las observaciones son las provincias españolas más Ceuta yMelilla, que aparecen unidas como una única provincia, y las variables los nueve epígrafes en los que se desglosa la Encuesta de Presupuestos Familiares en España.

Las variables son: X_1= alimentación, X_2= vestido y calzado, X_3= vivienda, X_4= moviliario doméstico, X_5= gastos sanitarios, X_6= transporte, X_7= enseñanza y cultura, X_8= turismo y ocio, X_9= otros gastos.

Datos tabla A.3

Fuente: Encuesta de Presupuestos Familiares del año 1990/91

PUBLICACIONES CIENTIFICAS DE LOS PAISES DE LA OCDE

	INTER.A	INTER.F	AGRIC.	BIOLO.	MEDIC.	QUIMI.	INGEN.	FÍSICA
EE.UU	815.319	379.851	88.663	58.104	255.864	440.134	111.015	162.288
UK	162.103	90.332	35.158	29.802	59.63	92.725	6.409	34.349
JP	105.856	78.811	13.978	16.758	55.634	308.926	32.039	40.538
F	118.935	76.186	13.818	11.253	49.938	120.065	9.984	35.792
G	91.099	85.037	11.74	20.337	41.233	66.087	11.304	22.093
C	72.722	49.459	14.041	16.722	23.139	101.9	12.034	14.645
I	42.905	29.734	7.904	13.444	31.078	36.322	5.833	11.351
A	36.121	22.236	12.419	9.863	12.186	19.641	4.898	6.299
H	29.912	18.036	6.563	7.548	13.721	23.029	3.798	6.775
S	28.568	16.19	3.985	9.502	14.852	18.341	2.387	3.549
CH	26.495	14.518	3.378	3.636	11.096	19.304	2.556	5.784
E	16.425	11.818	3.089	3.981	7.196	15.493	1.258	2.692
B	17.311	11.791	3.24	4.011	8.098	11.964	1.772	3.417
D	14.677	555	2.635	5.667	8.368	14.266	1.197	1.999
AU	10.957	13.154	1.433	2.372	5.928	6.713	1.318	2.278
FI	11.012	6.457	2.028	4.756	5.731	6.647	1.001	1.669
N	9.075	5.432	1.803	3.299	4.801	5.326	912	853
Y	4.686	2.957	2.031	1.194	1.806	7.046	801	1.861
GR	3.72	2.749	692	1.293	1.518	2.415	896	1.366
IR	6.786	214	432	1.119	1.355	98	522	941
P	1.221	1.929	388	386	564	12	493	413

Tabla A.4: Datos INVEST

INVEST Este conjunto de datos presenta 21 observaciones de 8 variables. Las observaciones corrsponden a los países de la OCDE y las variables son el número de publicaciones científicas recogidas en el trienio 1982-84 en ocho bases de datos de producción científica. Las variables se han llamado según la orientación de la base de datos: InterA(por interdisciplinaria), Inter F (por interdisciplinaria), Agric., Biolo., Medic., Quimic., Ingen. y Física.

Datos: tabla A.4

Fuente: Caballero y Peña (1987).

MEDIDAS FÍSICAS

Obs.	sexo	est	pes	pie	lbr	aes	dcr	lrt
1	0	159	49	36	68	42	57	40
2	1	164	62	39	73	44	55	44
3	0	172	65	38	75	48	58	44
4	0	167	52	37	73	41.5	58	44
5	0	164	51	36	71	44.5	54	40
6	0	161	67	38	71	44	56	42
7	0	168	48	39	72.5	41	54.5	43
8	1	181	74	43	74	50	60	47
9	1	183	74	41	79	47.5	59.5	47
10	0	158	50	36	68.5	44	57	41
11	0	156	65	36	68	46	58	41
12	1	173	64	40	79	48	56.5	47
13	0	158	43	36	68	43	55	39
14	1	178	74	42	75	50	59	45
15	1	181	76	43	83	51	57	43
16	1	182	91	41	83	53	59	43
17	1	176	73	42	78	48	58	45
18	0	162	68	39	72	44	59	42
19	0	156	52	36	67	36	56	41
20	0	152	45	34	66	40	55	38
21	1	181	80	43	76	49	57	46
22	1	173	69	41	74	48	56	44
23	0	155	53	36	67	43	56	38
24	1	189	87	45	82	53	61	52
25	0	170	70	38	73	45	56	43
26	1	170	67	40	77	46.5	58	44.5
27	0	168	56	37.5	70.5	48	60	40

Tabla A.5: Datos MEDIFIS

MEDIFIS Este conjunto de datos contiene 28 observaciones de 8 variables. Las observaciones cooresponde a estudiantes españoles y las variables a sus características físicas. Las variables son: género (0 mujer, 1 hombre), estatura (en cm), peso (en Kgr), longitud de pie (en cm.), longitud de brazo (en cm.), anchura de la espalda (en cm.), diámetro del cráneo (en cm.), longitud entre la rodilla y el tobillo (en cm.).

Datos: tabla A.5

Fuente: Elaboración propia

DESARROLLO EN EL MUNDO

	Tasa Nat.	Tasa Mort.	Mort.Inf	Esp.Hom	Esp.Muj.	PNB
Albania	24,7	5,7	30,8	69,6	75,5	600
Bulgaria	12,5	11,9	14,4	68,3	74,7	2250
Checos.	13,4	11,7	11,3	71,8	77,7	2980
Hungria	11,6	13,4	14,8	65,4	73,8	2780
Polonia	14,3	10,2	16	67,2	75,7	1690
Rumania	13,6	10,7	26,9	66,5	72,4	1640
URSS	17,7	10	23	64,6	74	2242
Bielorrusia	15,2	9,5	13,1	66,4	75,9	1880
Ucrania	13,4	11,6	13	66,4	74,8	1320
Argentina	20,7	8,4	25,7	65,5	72,7	2370
Bolivia	46,6	18	111	51	55,4	630
Brasil	28,6	7,9	63	62,3	67,6	2680
Chile	23,4	5,8	17,1	68,1	75,1	1940
Colombia	27,4	6,1	40	63,4	69,2	1260
Ecuador	32,9	7,4	63	63,4	67,6	980
Guayana	28,3	7,3	56	60,4	66,1	330
Paraguay	34,8	6,6	42	64,4	68,5	1110
Perú	32,9	8,3	109,9	56,8	66,5	1160
Uruguay	18	9,6	21,9	68,4	74,9	2560
Venezuela	27,5	4,4	23,3	66,7	72,8	2560
Mexico	29	23,2	43	62,1	66	2490
Bélgica	12	10,6	7,9	70	76,8	15540
Finlandia	13,2	10,1	5,8	70,7	78,7	26040
Dinamarca	12,4	11,9	7,5	71,8	77,7	22080
Francia	13,6	9,4	7,4	72,3	80,5	19490
Alemania	11,4	11,2	7,4	71,8	78,4	22320
Grecia	10,1	9,2	11	65,4	74	5990
Irlanda	15,1	9,1	7,5	71	76,7	9550
Italia	9,7	9,1	8,8	72	78,6	16830
Paises Bajos	13,2	8,6	7,1	73,3	79,9	17320
Noruega	14,3	10,7	7,8	67,2	75,7	23120
Portugal	11,9	9,5	13,1	66,5	72,4	7600
España	10,7	8,2	8,1	72,5	78,6	11020
Suecia	14,5	11,1	5,6	74,2	80	23660
Suiza	12,5	9,5	7,1	73,9	80	34064
Reino Unido	13,6	11,5	8,4	72,2	77,9	16100
Austria	14,9	7,4	8	73,3	79,6	17000
Japon	9,9	6,7	4,5	75,9	81,8	25430
Canada	14,5	7,3	7,2	73	79,8	20470

MUNDODES (continúa)

	Tasa Nat	Tasa Mort	Mort.Inf	Esp.Hom	Esp.Muj.	PNB
EEUU	16,7	8,1	9,1	71,5	78,3	21790
Afganistan	40,4	18,7	181,6	41	42	168
Bahrein	28,4	3,8	16	66,8	69,4	6340
Iran	42,5	11,5	108,1	55,8	55	2490
Irak	42,6	7,8	69	63	64,8	3020
Israel	22,3	6,3	9,7	73,9	77,4	10920
Jordania	38,9	6,4	44	64,2	67,8	1240
Kuwait	26,8	2,2	15,6	71,2	75,4	16150
Oman	45,6	7,8	40	62,2	65,8	5220
Arabia Saudi	42,1	7,6	71	61,7	65,2	7050
Turkia	29,2	8,4	76	62,5	65,8	1630
Emiratos Arabes	22,8	3,8	26	68,6	72,9	19860
Bangladesh	42,2	15,5	119	56,9	56	210
China	21,2	6,7	32	68	70,9	380
Hong Kong	11,7	4,9	6,1	74,3	80,1	14210
India	30,5	10,2	91	52,5	52,1	350
Indonesia	28,6	9,4	75	58,5	62	570
Malasia	31,6	5,6	24	67,5	71,6	2320
Mongolia	36,1	8,8	68	60	62,5	110
Nepal	39,6	14,8	128	50,9	48,1	170
Pakistan	30,3	8,1	107,7	59	59,2	380
Filipinas	33,2	7,7	45	62,5	66,1	730
Singapur	17,8	5,2	7,5	68,7	74	11160
Srilanka	21,3	6,2	19,4	67,8	71,7	470
Tailandia	22,3	7,7	28	63,8	68,9	1420
Argelia	35,5	8,3	74	61,6	63,3	2060
Angola	47,2	20,2	137	42,9	46,1	610
Botswana	48,5	11,6	67	52,3	59,7	2040
Congo	46,1	14,6	73	50,1	55,3	1010
Egipto	38,8	9,5	49,4	57,8	60,3	600
Etiopia	48,6	20,7	137	42,4	45,6	120
Gabon	39,4	16,8	103	49,9	53,2	390
Gambia	47,4	21,4	143	41,4	44,6	260
Ghana	44,4	13,1	90	52,2	55,8	390
Kenya	47	11,3	72	56,5	60,5	370
Libia	44	9,4	82	59,1	62,6	5310
Malawi	48,3	25	130	38,1	41,2	200
Marruecos	35,5	9,8	82	59,1	62,5	960

MUNDODES (continúa)

	Tasa Nat.	Tasa Mort	Mort.Inf	Esp.Hom	Esp.Muj.	PNB
Mozambique	45	18,5	141	44,9	48,1	80
Namibia	44	12,1	135	55	57,5	1030
Nigeria	48,5	15,6	105	48,8	52,2	360
Sierra Leona	48,2	23,4	154	39,4	42,6	240
Somalia	50,1	20,2	132	43,4	46,6	120
Surafrica	32,1	9,9	72	57,5	63,5	2530
Sudan	44,6	15,8	108	48,6	51	480
Swaziland	46,8	12,5	118	42,9	49,5	810
Tunez	31,1	7,3	52	64,9	66,4	1440
Uganda	52,2	15,6	103	49,9	52,7	220
Tanzania	50,5	14	106	51,3	54,7	110
Zaire	45,6	14,2	83	50,3	53,7	220
Zambia	51,1	13,7	80	50,4	52,5	420
Zimbabwe	41,7	10,3	66	56,5	60,1	640

Tabla A.6: MUNDODES)

MUNDODES Este conjunto de datos consta de 91observaciones y 6 variables. Las observaciones corresponden a 91 países. Las variables son indicadores de desarrollo. Las seis variables son :

Tasa Nat.:	Ratio de natalidad por 1000 habitantes
Tasa Mort:	Ratio de mortalidad por 1000 habitantes
Mort.Inf:	Mortalidad infantil (por debajo de un año)
Esp.Hom:	Esperanza de vida en hombres
Esp.Muj.:	Esperanza de vida en mujeres
PNB:	Producto Nacional Bruto per cápita

Datos: tabla A.6
Fuente: "UNESCO 1990 Demographic Year Book" y de "The Annual Register 1992".

ACCIONES DE LA BOLSA DE MADRID

Obs.	X_1	X_2	X_3	Obs.	X_1	X_2	X_3
1	3.4	89.7	30.2	18	4.4	58.5	12.1
2	5.1	55.7	9.9	19	7.8	84.3	11.0
3	4.5	52.3	11.5	20	16.0	96.5	6.0
4	3.5	47.0	11.2	21	16.7	100.0	6.8
5	5.9	42.7	7.0	22	15.2	92.3	5.2
6	5.1	30.6	6.9	23	17.5	99.9	6.8
7	4.6	64.4	11.8	24	16.2	93.5	6.1
8	5.0	51.0	9.6	25	14.7	100.0	6.6
9	3.2	54.4	14.7	26	15.3	99.9	5.9
10	3.4	45.7	13.2	27	15.8	100.0	6.9
11	6.5	39.9	5.2	28	18.3	96.3	5.7
12	4.4	40.3	13.7	29	15.9	100.0	6.1
13	5.1	52.4	11.0	30	16.1	92.5	6.1
14	5.8	43.9	8.0	31	9.7	87.6	7.7
15	4.6	52.8	14.4	32	6.9	53.6	6.6
16	7.2	65.8	7.8	33	14.4	87.8	5.2
17	7.2	58.1	7.7	34	14.9	34.5	4.69

Tabla A.7: ACCIONES

ACCIONES Este conjunto de datos presenta 34 observaciones y 3 variables. Las observaciones corresponden a distintas acciones que cotizan en el mercado continuo español y las variables a tres medidas de rentabilidad de estas acciones durante un período de tiempo. Las variables son : X_1 es la rentabilidad efectiva por dividendos, X_2 es la proporción de beneficios que va a dividendos y X_3 el ratio entre precio por acción y beneficios.

Datos : Tabla A.7

Fuente: Elaboración propia