Análisis
Multivariante

Colección manuales uex - 59
(E.E.E.S.)

Jesús
Montanero Fernández

59

ANÁLISIS
MULTIVARIANTE

MANUALES UEX

59

(E.E.E.S.)
Espacio
Europeo
Educación
Superior

JESÚS MONTANERO FERNÁNDEZ

ANÁLISIS MULTIVARIANTE

UNIVERSIDAD DE EXTREMADURA

2008

La publicación del presente manual forma parte de las "Acciones para el Desarrollo del Espacio Europeo de Educación Superior en la Universidad de Extremadura Curso 2007/08" en el marco de la VI Convocatoria de Acciones para la Adaptación de la UEX al Espacio Europeo de Educación Superior (Proyectos Pilotos: modalidad A1) del Vicerrectorado de Calidad y Formación Continua y financiada por la Junta de Extremadura, el Ministerio de Educación y Ciencia y la Universidad de Extremadura.

UNIVERSIDAD DE EXTREMADURA

UNIÓN EUROPEA
Fondo Social Europeo

JUNTA DE EXTREMADURA

Edita

Universidad de Extremadura. Servicio de Publicaciones
C./ Caldereros, 2 - Planta 2ª - 10071 Cáceres (España)
Telf. 927 257 041 - Fax 927 257 046
publicac@unex.es
www.unex.es/publicaciones

ISSN 1135-870-X
ISBN 978-84-691-6343-6
Depósito Legal M-45.208-2008

Edición electrónica: Pedro Cid, S.A.
Teléf.: 914 786 125

A Lucía

Prólogo

El presente manual esta concebido como un apoyo a la docencia en una asignatura de segundo ciclo que puede cursarse tanto en la Licenciatura de Matemáticas como en la de Ciencias y Técnicas Estadísticas. El objetivo es que pueda ser entendido por alumnos con conocimientos básicos de Matemáticas en general y Estadística en particular.

Los aspectos formales de la materia han sido desarrollados con cierto detalle. En lo que respecta a las competencias cuya adquisición debe posibilitar esta asignatura, no es estrictamente necesaria la compresión exhaustiva de los mismos, aunque se antoje conveniente que el lector interesado tenga al menos un lugar donde acudir si quiere llevar a cabo un estudio más profundo de la materia, al margen de la bibliografía especializada. Por contra, el alumno debe tener en cuenta que el conocimiento teórico de estos contenidos debe complementarse con su aplicación mediante un programa estadístico. En la página web `http://kolmogorov.unex.es/jmf~/` se encuentra material al respecto.

También cabe resaltar que este manual se complementa con otro dedicado a los Modelos Lineales. De hecho podría considerarse como una segunda parte o segundo volumen de una serie de dos.

Introducción

El análisis multivariante es una disciplina difícil de definir e incluso de delimitar. Bajo este sobrenombre se agrupan diversas técnicas estadísticas que, si bien muchas de ellas fueron ideadas por autores que podemos denominar clásicos, deben su auge y puesta en práctica a la difusión del software estadístico y a la creciente demanda que de ellas exige el desarrollo de otras disciplinas, como la Sociología, Psicología, Biología o Economía. Es desde luego impensable poder aplicar procedimientos como el manova, el análisis factorial, el análisis cluster o el de correspondencias si no se dispone de un programa estadístico adecuado. Y no es menos cierto, como hemos apuntado, que si nos preguntamos cuál es el denominador común de los procedimientos mencionados, la respuesta no sea del todo convincente.

Para muchos autores, hablar de análisis multivariante es simplemente hablar del estudio simultáneo de más de dos variables (Hair et. al (1999)). Desde luego que esta definición se aproxima a la idea que tenemos la mayoría, pero ello haría de la regresión lineal múltiple una técnica multivariante (dado que, en la práctica, no sólo los valores de la variable dependiente sino también los valores explicativos suelen ser observaciones de variables aleatorias). En definitiva, estaríamos incluyendo el estudio del modelo lineal dentro del análisis multivariante. No queremos decir que sea mala idea, todo lo contrario. Ambas materias se encuentran estrechamente vinculadas desde el punto de vista epistemológico. De hecho, este volumen está concebido como continuación de otro primero dedicado al Modelo Lineal que debemos tener presente en todo momento.

Pero al margen de estas disquisiciones, la misma definición anterior nos impediría aceptar como multivariante una técnica tan de moda en nuestros días como es el análisis de correspondencias. También habría que preguntarse por qué se considera multivariante el análisis cluster, puesto que una clasificación en conglomerados podría hacerse, estrictamente hablando, a partir de la observación de una única variable. Razonando en sentido contrario y si queremos llevar al extremo esta crítica fácil, nos preguntamos por qué son necesarias al menos tres variables para hablar de multivariante. Desde luego, si admitimos dos, estaríamos incluyendo estudios como el

de correlación simple, en contra de lo que su propio nombre indica, y si siguiéramos tirando del hilo, no quedaría probablemente técnica en la estadística que dejara de ser multivariante.

Desde luego, no son muy justos estos comentarios pues, como sabemos, en Estadística, donde nos movemos en los pantanosos terrenos que van de los datos reales a los modelos matemáticos, resulta más fácil destruir que construir, y mantener en pie una definición que permita caracterizar las distintas técnicas consideradas multivariantes se nos antoja, al menos desde nuestra estrecha visión, poco menos que imposible. De ahí que nos decantemos por una aproximación más sutil al concepto de análisis multivariante, entendiendo que lo que caracteriza al mismo son más bien sus procedimientos. Esta es la opinión de muchos autores que consideran como rasgo más característico del análisis multivariante el estudio de los denominados *valores teóricos*, que son ciertas combinaciones de las variables consideradas. La forma de construirlas difiere según el propósito buscado, pero en todo caso subyace como denominador común la búsqueda de una reducción en la dimensión inicial del problema. Y esta es una particularidad que debemos tener muy presente en todo momento: el objetivo primero y fundamental de las técnicas multivariantes no es la resolución de un problema estadístico sino su simplificación. Para ello se pretende por todos los medios representar nuestros datos en un espacio de escasa dimensión con la menor pérdida posible de información. Este tipo de simplificaciones puede revelarnos particularidades de los datos que no podíamos conocer en primera instancia por la complejidad de los mismos.

Las diversas técnicas multivariantes pueden clasificarse en función de distintos criterios. Uno de ellos, el más extendido, se basa en los tipos de relaciones examinadas. Se dice que la relación es de dependencia cuando una o varias de las variables estudiadas son o pretenden ser explicadas por el resto. Las primeras se denominan dependientes o respuestas y las segundas, explicativas. Por contra, la relación de interdependencia se da cuando no hay ningún tipo de discriminación o distinción entre variables, sino que todas ellas juegan papeles simétricos o intercambiables. No obstante, estas clases pueden subdividirse en función de la naturaleza de las variables (distinguiendo entre métricas y no métricas), aunque no todas las técnicas estudiadas se ajustan de forma idónea a esta clasificación. No obstante, diremos que en la primera clase se encuentran el análisis multivariante de la varianza (manova), la regresiones múltiple y multivariante y el análisis discriminante. En la segunda se encuadran los análisis factorial, el de correlación canónica, de componentes principales, cluster y de correspondencias.

Otro criterio a la hora de clasificar procedimientos consiste en distinguir si la técni-

ca en cuestión supone una generalización multivariante de otra análoga univariante (unidimensional) o no. De esta forma, el manova generaliza el anova; la regresión lineal multivariante generaliza la regresión lineal múltiple, que a su vez generaliza la simple; los coeficientes de correlación canónica generalizan el coeficiente de correlación múltiple que a su vez generaliza el de correlación simple; el test M de Box generaliza el de Barlett, etc. Sin embargo, los análisis de componentes principales y factorial no tienen sentido en dimensión uno. Los análisis discriminante y de correspondencias cabría incluirlos en este último grupo, con algunas reservas.

El proceso a seguir en cualquier estudio estadístico, y muy especialmente en análisis multivariante, es complejo y requiere, además de los conocimientos teóricos que puedan aportar el estudio académico, de cierta experiencia en el manejo de datos en general y, por supuesto, de una buena dosis de humildad. Por desgracia, el procedimiento dista mucho de ser un algoritmo que nos conduce del planteamiento del problema a la solución del mismo. No obstante, en Anderson , Hair, Tatham, Black (2000), libro de carácter eminentemente aplicado, se propone un proceso de modelización en seis pasos para resolver, o mejor dicho afrontar, un problema multivariante, que puede resultar orientativo: en primer lugar, se plantean la definición del problema, objetivos y técnica conveniente; a continuación, el proyecto de análisis; le sigue la evaluación de los supuestos básicos requeridos; posteriormente, se efectúa la estimación del modelo y valoración del ajuste del mismo; seguidamente, se lleva a cabo la interpretación del valor teórico; para finalizar, se procede a la validación del modelo.

La validación es necesaria en aras de garantizar cierto grado de generalidad al modelo obtenido. Esto puede conseguirse mediante diversos procedimientos, como la validación cruzada, el *jackknife* o las técnicas bootstrap.

La evaluación de los supuestos básicos es uno de los asuntos más delicados de la Estadística en general y del análisis multivariante en especial. Decimos en especial porque, si bien podemos aceptar, aunque a regañadientes, que una variable aleatoria se ajuste satisfactoriamente a un modelo de distribución normal, resulta más difícil aceptar la normalidad multivariante de un vector aleatorio de dimensión 10. Además, esta dependencia del análisis multivariante respecto a la distribución normal queda patente ante la dificultad de traducir al lenguaje multivariante los procedimientos clásicos de la estadística no paramétrica basados en los rangos (tests de Mann-Whitney, Kruskall-Wallis). No obstante, debemos anticipar que en algunos casos, no estaremos en condiciones de garantizar al lector una solución plenamente satisfactoria del problema.

Para acabar esta introducción mencionaremos algunos aspectos particulares del programa que desarrollamos a continuación. El lector podrá tal vez reconocer en

el mismo la influencia de *The Theory of Lineal Models and Multivariate Analysis,* S.F. Arnold (1981), ed. Wiley. Efectivamente, la visión que aporta este libro sobre el análisis multivariante queda bien patente en la estructura de este volumen en la del volumen dedicado a los Modelos Lineales; muy especialmente en todo lo referente al modelo lineal normal multivariante. También han resultado de gran utilidad referencias como Rencher (1995), Bilodeau (1999), Flury (1997), Dillon, Goldstein (1984), sin olvidar otros clásicos como Anderson (1958) o Mardia et al. (1979).

Cada uno de los capítulos consta de una introducción donde se comentan los aspectos generales del mismo, la exposición de la materia correspondiente y una serie de cuestiones que se proponen como trabajo personal para el lector. La distribución y secuenciación de los mismos se ha realizado teniendo en cuenta las ideas aportadas por los autores anteriormente citados. El lector podrá apreciar sin duda una evolución en el estilo en función del tema a tratar. Así, los primeros capítulos, dedicados a los distintos modelos de distribución y al modelo lineal normal multivariante, pueden resultar más teóricos que los que se dedican a técnicas concretas del análisis multivariante.

Por último, contamos con un apéndice dedicado, en primer lugar, al Álgebra de matrices. La demostración de los resultados que aquí se exponen puede encontrarse en el Apéndice del volumen dedicado a la Modelos Lineales, mencionado anteriormente que, insistimos, debemos tener muy presente dado que éste es una continuación de aquél. En el apéndice de dicho volumen puede encontrarse, además, un breve repaso de nociones fundamentales de la Probabilidad y Estadística que pueden ser de utilidad para el lector. No los hemos incluido en éste por no resultar redundantes. Por último, en la segunda sección de nuestro apéndice podemos encontrar también la demostración muy técnica y extensa de un resultado correspondiente al capítulo 6.

Índice general

1. Distribuciones del análisis multivariante **19**
1.1. Distribución normal multivariante 20
1.2. Distribución normal matricial. 30
1.3. Distribuciones de Wishart y Hotelling 43

2. Modelo lineal normal multivariante **55**
2.1. Estimación 56
2.2. Contrastes lineales sobre la media 60
2.3. Test de Wilks 69
2.4. Tests de Lawley-Hotelling y Pillay 72
2.5. Test de Roy 75
2.6. Estudio Asintótico del Modelo 78
2.7. Contraste de hipótesis generalizado. 92

3. Contrastes para la matriz de covarianzas. **95**
3.1. Test de correlación. 96
3.2. Test M de Box. 99
3.3. Contraste para una matriz de covarianza. 100
3.4. Test de esfericidad de Barlett. 101
3.5. Ejemplos 102

4. Análisis Multivariante de la Varianza **105**
4.1. Contraste de una media 106
4.2. Contraste de dos medias 108
4.3. Manova con un factor 112
4.4. Análisis de perfiles 116

5. Regresión Lineal Multivariante **121**
5.1. El modelo de Regresión 122

5.2. Regresión y correlación 123
5.3. Estimación de los parámetros 125
5.4. Tests de hipótesis 127
5.5. Estudio asintótico. 133
5.6. Regresión con variables ficticias. Mancova 134

6. Análisis de correlación canónica 137
6.1. Definición 137
6.2. Inferencias 141
6.3. Relación con el test de correlación 143
6.4. Relación con regresión y manova 144
6.5. Reducción de dimensión en correlación lineal 144

7. Análisis de componentes principales 151
7.1. Punto de vista probabilístico 152
7.2. Punto de vista muestral 158
7.3. Relación con las variables originales 169

8. Aplicaciones de componentes principales 175
8.1. Multicolinealidad 175
8.1.1. Ejemplo 182
8.2. Análisis de correspondencias 184
8.2.1. Ejemplo 193

9. Análisis discriminante I 197
9.1. Ejes discriminantes 198
9.2. Análisis discriminate y correlación canónica 203
9.3. Caso de dos grupos. 204
9.4. Variables observadas y discriminación. 205

10. Análisis discriminante II 211
10.1. Dos grupos: planteamiento general. 212
10.2. Dos normales con covarianzas común 218
10.3. Caso general: r distribuciones p-normales 224
10.4. Relación con los ejes discriminantes. 226
10.5. Caso de matriz de covarianzas distintas 229
10.6. Validez de la estrategia. 230
10.7. Estimación de densidades 232
10.8. Regresión logística 235

10.9. k-proximidad 236

11. Análisis factorial **239**
11.1. Planteamiento del problema 240
11.2. Método de componentes principales 245
11.3. Modelo basado en el concepto de factor 250
11.4. Ejemplo 256

12. Análisis cluster **259**
12.1. Medidas de afinidad 260
12.2. Formación de conglomerados 261
12.3. Interpretación de los conglomerados 263

13. Apéndice **265**

Capítulo 1

Distribuciones del análisis multivariante

En este capítulo se estudiarán cuatro distribuciones multidimensionales que desempeñarán un papel fundamental en el modelo lineal normal multivariante: las distribuciones normal multivariante y matricial, la distribución de Wishart y la distribución T^2 de Hotelling. De la segunda y tercera podemos decir que son distribuciones matriciales, pues son generadas por matrices aleatorias. Este concepto de matriz aleatoria, recogido de Arnold (1981) y que trataremos a continuación, no es ni mucho menos común a todos los textos consultados. No obstante, consideramos que facilita una elegante exposición del modelo lineal normal multivariante, teniendo en cuenta que n observaciones de datos p-dimensionales configuran una matriz de dimensión $n \times p$. Veremos que, si las observaciones son independientes y generadas según distribuciones normales p-variantes con matriz de covarianzas común, la matriz de datos sigue un modelo normal matricial[1]. Igualmente, la distribución de Wishart, que generaliza la χ^2 de Pearson, es inducida por matrices aleatorias definidas positivas, como puede ser un estimador de la matriz de covarianzas. El teorema 1.28 establece la importancia de esta distribución en el modelo lineal normal multivariante.

No obstante, dado que podemos establecer una identificación natural entre las matrices de orden $m \times q$ y los vectores en $\mathbb{R}^{mq}$, los conceptos de matriz aleatoria y distribución matricial no son en esencia nuevos. Tampoco lo es la distribución T^2 de Hotelling. El teorema 1.32 demuestra que esta distribución, asociada siempre a la distancia de Mahalanobis y que es, por lo tanto, univariante, difiere de la distribución F Snedecor en una constante, siendo equivalente a una t de Student al cuadrado cuando consideramos una única variable. De hecho, en el análisis multivariante aparece en

[1]Realmente, el modelo lineal matricial es más general y no se restringe a este caso.

las mismas situaciones donde en análisis univariante aparece la t de Student.

En la primera parte del capítulo se aborda el estudio del modelo normal multivariante (junto con las distribuciones relacionadas). Esta sección, aunque no es realmente específica del Análisis Multivariante, es fundamental pues el supuesto de normalidad multivariante de las observaciones es el pilar sobre el que se construyen la mayoría de los modelos a estudiar. A continuación se extenderá su estudio estudio al caso matricial, para definir a generalizaciones multivariantes de las distribuciones asociadas.

1.1. Distribución normal multivariante

En esta sección se aborda el estudio de la distribuciones normal multivariante, haciendo especial hincapié en el caso esférico, junto con otras distribuciones relacionadas con esta última, como son la χ^2, F-Snedecor y t-Student. Realmente, damos por hecho que todas ellas son de sobras conocidas, por lo que nos limitaremos a repasar las definiciones y propiedades fundamentales. Las demostraciones que se echen en falta pueden encontrarse en cualquier referencia clásica, o bien en el primer capítulo del volumen dedicado a los Modelos Lineales.

Dados un vector $\mu \in \mathbb{R}^n$ y una matriz $\Sigma \in \mathcal{M}_{n\times n}$ simétrica y semidefinida positiva, se dice que un vector aleatorio $Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$ sigue un modelo de distribución normal multivariante en dimensión n (o n-normal) con media μ y matriz de varianzas-covarianzas Σ, cuando su correspondiente función característica es la siguiente

$$\varphi_Y(t) = \exp\left\{ \mathbf{i} t'\mu - \frac{1}{2} t'\Sigma t \right\}, \quad t \in \mathbb{R}^n.$$

En ese caso, se denota $Y \sim N_n(\mu, \Sigma)$. Un vector de este tipo puede construirse explícitamente como sigue: si Σ diagonaliza según el teorema 13.4 mediante

$$\Sigma = \Gamma \Delta \Gamma',$$

consideramos Z_i, $i = 1, \ldots, n$, independientes y con distribuciones normales de media 0 y varianza el elemento i-ésimo de la diagonal de Δ, δ_i^2, respectivamente. Si Z denota el vector aleatorio $(Z_1, \ldots, Z_n)'$, se tiene entonces que

$$Y = \mu + \Gamma Z \tag{1.1}$$

sigue la distribución deseada. Dado que $\mathtt{E}[Z] = 0$ y $\mathtt{Cov}[Z] = \Delta$, y teniendo en cuenta que, en general,

$$\mathtt{E}[AZ + b] = A\mathtt{E}[Z] + b, \qquad \mathtt{Cov}[AZ + b] = A\mathtt{Cov}[Z]A'. \tag{1.2}$$

se deduce que una distribución $N_n(\mu, \Sigma)$ tiene por media μ y por matriz de varianzas-covarianzas Σ. También es inmediato comprobar que presenta la siguiente función generatriz, bien definida en todo $\mathbb{R}^n$:

$$g_Y(t) = \exp\left\{t'\mu - \frac{1}{2}t'\Sigma t\right\}, \quad t \in \mathbb{R}^n.$$

En consecuencia, existen los momentos de cualquier orden de la distribución, que pueden calcularse mediante las sucesivas derivadas parciales de g en 0.

Es bien conocido que la normalidad en dimensión 1 se conserva ante transformaciones afines, es decir, que si a una distribución normal se le aplica una homotecia y una traslación, la distribución resultante sigue siendo normal. Operando con las funciones características podemos obtener de manera trivial el siguiente resultado que generaliza al anterior en el caso multivariante.

Proposición 1.1.
Dados $Y : (\Omega, \mathcal{A}, P) \to \mathbb{R}^n$, tal que $Y \in N_n(\mu, \Sigma)$, $A \in \mathcal{M}_{n\times m}$ y $b \in \mathbb{R}^m$, se verifica

$$AY + b \sim N_m(A\mu + b, A\Sigma A').$$

De la proposición 1.1 se deduce que las n componentes de una normal n-variante son todas normales. Sin embargo, no podemos garantizar, en general, que n componentes normales configuren conjuntamente un vector n-normal, cosa que si sucede si las componentes son independientes. En el volumen dedicado a los Modelos Lineales podemos encontrar un ejemplo que ilustra esa situación. El siguiente resultado supone una interesante caracterización de la distribución normal multivariante.

Proposición 1.2.
Un vector aleatorio n-dimensional Y de media μ y matriz de varianzas-covarianzas Σ sigue una distribución n-normal si y sólo si la variable aleatoria real $\mathrm{u}'X$ sigue una distribución $N(\mathrm{u}'\mu, \mathrm{u}'\Sigma\mathrm{u})$, para cada $\mathrm{u} \in \mathbb{R}^n\backslash\{0\}$.

Queremos decir, por lo tanto, que la distribución es n-normal cuando al proyectar sobre cualquier dirección de $\mathbb{R}^n$ obtenemos una normal en dimensión 1. Por otra parte, el siguiente resultado garantiza la equivalencia entre incorrelación e independencia bajo la hipótesis de normalidad multivariante.

Proposición 1.3.
Si $Y = (Y_1'Y_2')'$ sigue un modelo de distribución normal en dimensión $n_1 + n_2$ y $\Sigma_{12} = 0$, entonces Y_1 e Y_2 son independientes.

Nótese que esta propiedad puede extenderse trivialmente a cualquier colección (no necesariamente dos) de subvectores de un vector aleatorio normal multivariante,

en particular, a cualquier subconjunto de componentes del mismo. Queremos decir lo siguiente: si $Y_{n(1)}, \ldots, Y_{n(k)}$ son componentes incorreladas de un vector n-normal, entonces son también independientes.

Con frecuencia suele suponerse que la matriz de covarianzas Σ de la normal es estrictamente definida positiva, es decir, no singular. En caso contrario se dice que la normal es degenerada, es decir, que está *sobredimensionada*[2] . En ese caso, estará contenida en una subvariedad afín de dimensión $n-1$, por lo que no estará dominada por la medida de Lebesgue en $\mathbb{R}^n$. En el caso no degenerado, tendrá sentido hablar de su densidad respecto a dicha medida.

Proposición 1.4.

Si $Y \sim N_n(\mu, \Sigma)$ con $\Sigma > 0$, entonces admite la siguiente densidad respecto a la medida de Lebesgue:

$$f(\mathtt{y}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left\{-\frac{1}{2}(\mathtt{y}-\mu)'\Sigma^{-1}(\mathtt{y}-\mu)\right\}, \quad \mathtt{y} \in \mathbb{R}^n. \tag{1.3}$$

La siguiente propiedad establece una clara conexión entre los supuestos de normalidad y linealidad, arrojando luz sobre los modelos de Regresión y Correlación. Consideremos dos vectores aleatorios Y_1 e Y_2, de dimensiones n_1 y n_2, respectivamente. Construiremos una versión de la probabilidad condicional regular de Y_1 dado Y_2. Bajo la hipótesis de (n_1+n_2)-normalidad no degenerada de $Y = (Y_1', Y_2')'$. Descompongamos la media y matriz de varianzas-covarianzas de Y de forma obvia y consideremos los parámetros siguientes

$$\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \tag{1.4}$$

$$\beta = \Sigma_{12}\Sigma_{22}^{-1}, \qquad \alpha = \mu_1 - \beta\mu_2. \tag{1.5}$$

Nótese que, en virtud del lema 13.6 y al ser $\Sigma > 0$, tiene sentido hablar de e $\Sigma_{11\cdot 2}$ y es definida positiva.

Proposición 1.5.

En las condiciones anteriores, se verifica

$$P^{Y_1|Y_2=\mathtt{y}_2} = N_{n_1}(\alpha + \beta\mathtt{y}_2, \Sigma_{11\cdot 2}), \qquad \forall \mathtt{y}_2 \in \mathbb{R}^{n_2}.$$

Podemos ir incluso algo más lejos. Para poder seguir la siguiente demostración se necesita tener presentes las siguiente propiedad general de la Esperanza Condicional, que serán de gran utilidad en todas nuestra teoría.

[2]El objetivo del análisis de componentes principales es, precisamente, encontrar la manera de dar a la distribución su *verdadera* dimensión.

Proposición 1.6.
Si f es variable aleatoria real definida sobre $\mathbb{R}^{n_1+n_2}$, se verifica que

$$\mathrm{E}[f \circ (Y_1, Y_2)|Y_2 = \mathrm{y}_2] = \int_{\mathbb{R}^{n_2}} f(\cdot, \mathrm{y}_2)\, dP^{Y_1|Y_2=\mathrm{y}_2}, \tag{1.6}$$

donde $f(\cdot, \mathrm{y}_2)$ es la variable aleatoria real que asigna a cada $\mathrm{y}_1 \in \mathbb{R}^{n_1}$ el número $f(\mathrm{y}_1, \mathrm{y}_2)$, y $\left(P^{Y_1|Y_2=\mathrm{y}_2}\right)^{f(\cdot,\mathrm{y}_2)}$ denota la distribución de dicha variable respecto de $P^{Y_1|Y_2=\mathrm{y}_2}$. Como consecuencia, se tiene que

$$P^{f\circ(Y_1,Y_2)|Y_2=\mathrm{y}_2} = \left(P^{Y_1|Y_2=\mathrm{y}_2}\right)^{f(\cdot,\mathrm{y}_2)}, \quad (\mathrm{y}_1, \mathrm{y}_2) \in \mathbb{R}^{n_1+n_2}. \tag{1.7}$$

En consecuencia, si la probabilidad de $f \circ (Y_1, Y_2)$ condicionada a Y_2 resulta no depender de el valor que tome esta última, se deduce que ambas son independientes, coincidiendo la distribución condicional anterior con la propia distribución marginal de $f \circ (Y_1, Y_2)$. Aplicando este resultado a nuestro estudio obtenemos inmediatamente lo siguiente:

Proposición 1.7.
En las condiciones anteriores, se verifica

$$Y_1 = \alpha + \beta Y_2 + \mathcal{E},$$

donde $\mathcal{E} \sim N_{n_1}(0, \Sigma_{11\cdot 2})$ y es independiente de Y_2.

Así pues, entre dos vectores aleatorios que componen una distribución normal multivariante sólo es posible una relación lineal (o, mejor dicho, afín), salvo un error aleatorio independiente de media 0. Realmente, a esta conclusión podríamos haber llegado sólo con tener en cuenta que, si Y sigue una distribución norma multivariante, $Y_1 - (\alpha + \beta Y_2)$ es incorrelada con Y_2 si, y sólo si, son independientes.

En general, $\Sigma_{11\cdot 2}$, que es la matriz de varianzas-covarianzas de la diferencia $Y_1 - (\alpha + \beta Y_2)$ o, lo que es lo mismo, de la distribución condicional de Y_1 dado Y_2 (no depende del valor concreto que tome Y_2), se denomina, matriz de varianzas-covarianzas parciales de las componentes de Y_1 dado Y_2, y se interpreta en este caso como la parte de la matriz de varianzas-covarianzas de Y_1 no explicada por Y_2. En el caso $n_1 = 1$, obtenemos

$$Y_1 = \alpha + \beta Y_2 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2_{11\cdot 2}),$$

donde

$$\sigma^2_{11\cdot 2} = \sigma^2_1 - \Sigma_{12}\Sigma^{-1}_{22}\Sigma_{21} = \sigma^2_1(1 - \rho^2_{12}).$$

Según hemos dicho anteriormente, una varianza parcial $\sigma^2_{11\cdot 2}$ nula, equivale a una dependencia funcional de Y_1 respecto a Y_2, y ρ^2_{12} puede interpretarse como la proporción de varianza de Y_1 explicada por Y_2.

Volviendo a la expresión (1.3), correspondiente a la densidad de una distribución normal multivariante no degenerada podemos apreciar que la densidad en el punto $\mathtt{y}$ depende exclusivamente de la distancia de Mahalanobis a la media de la distribución, es decir,

$$\Delta^2(\mathtt{y},\mu) = (\mathtt{y}-\mu)'\Sigma^{-1}(\mathtt{y}-\mu).$$

En esas condiciones, el lugar geométrico de los puntos con una misma densidad es un elipsoide, cuya centro coincide con la media μ y cuya forma viene determinada por la matriz de varianzas-covarianzas Σ. Concretamente, los ejes del elipsoide quedan determinados por una base de autovectores de Σ y su excentricidad por la relación existente entre los autovalores. De hecho, puede demostrarse que los elipsoides son esferas si y sólo si los autovalores de Σ son idénticos, es decir, si Σ es de la forma $\sigma^2\mathtt{Id}$, para algún $\sigma^2 > 0$, en cuyo caso, la densidad en $\mathtt{y}$ dependerá únicamente del cuadrado de su distancia euclídea a la media $\|\mathtt{y}-\mu\|^2$. Por esa razón, la distribución $N_n(\mu,\sigma^2\mathtt{Id})$ se denomina normal multivariante esférica.

Ésta es, como se puede apreciar en el volumen dedicado a los Modelos Lineales, la distribución de partida en el modelo lineal normal. Su función de densidad es pues la siguiente

$$f_Y(\mathtt{y}) = \frac{1}{\sqrt{2\pi\sigma}}\exp\left\{-\frac{1}{2\sigma^2}\|\mathtt{y}-\mu\|^2\right\}. \tag{1.8}$$

De las proposiciones 1.1 y 1.3 se sigue sin dificultad que, dados un vector aleatorio Y n-normal multivariante esférico y dos matrices $A \in \mathcal{M}_{m\times n}$ y $B \in \mathcal{M}_{k\times n}$, los vectores AY y BY son independientes, si y sólo si, $A'B = 0$. Como consecuencia inmediata se obtiene la siguiente proposición.

Proposición 1.8.

Si $Y \sim N_n(\mu,\sigma^2\mathtt{Id})$ y V_1, V_2 son subespacios lineales de $\mathbb{R}^n$ ortogonales, si y sólo si, entonces $P_{V_1}Y$ y $P_{V_2}Y$ son independientes.

La familia de distribuciones normales esféricas (con restricciones de carácter lineal para la media) poseen excelentes propiedades estadísticas. En primer lugar, son familias exponenciales, por lo que la función de verosimilitud cumple con todas las condiciones de regularidad[3] que puedan exigirse en diversos teoremas; son invariantes ante diversos grupos de transformaciones bimedibles, cosa que permitirá obtener profundas reducciones por invarianza[4], de una de las cuales resulta, por ejemplo, el test $\mathtt{F}$; el Principio de Máxima Verosimilitud será aquí de fácil aplicación, conduciendo a la obtención del Estimador de Máxima Verosimilitud y el Test de la Razón de

[3] Continuidad, derivabilidad...

[4] Ver Apéndice del volumen anterior.

Verosimilitudes, etc.

Es especialmente llamativa la invarianza ante rotaciones que presenta cualquier distribución normal esférica de media 0, hasta el punto de que esta propiedad está cerca de caracterizar dicha distribución. Efectivamente, si $\Gamma \in \mathcal{O}_n$ y $Y \sim N_n(0, \sigma^2)$, con $\sigma^2 > 0$, entonces ΓY sigue exactamente la misma distribución. En Bilodeau (1999) podemos encontrar la demostración de una especie de recíproco, debida a Maxwell-Hershell.

Proposición 1.9.

Todo vector aleatorio n-dimensional con componentes independientes e invariante por rotaciones es n-normal esférico de media 0. Concretamente, si Y_1 denota la primera componente del mismo, el parámetro σ que caracteriza la distribución se obtiene mediante

$$\sigma = -\ln \varphi_{Y_1}(1).$$

Por último, una propiedad de demostración trivial, de utilidad en el estudio de la distribución χ^2. Realmente, la tesis de la proposición es cierta para cualquier distribución de media μ y matriz de varianzas-covarianzas $\sigma^2 \mathtt{Id}$.

Proposición 1.10.

Si $Y \sim N_n(\mu, \sigma^2 \mathtt{Id})$, entonces $\mathtt{E}\big[\|Y\|^2\big] = n\sigma^2 + \|\mu\|^2$.

A continuación abordaremos un breve estudio de cuatro distribuciones directamente derivadas de la normal esférica: χ^2, F-Snedecor, Beta y t-Student. Un estudio más detallado de las mismas con todas las demostraciones que quedarán pendientes puede encontrarse, por ejemplo, en Nogales (1998). En primer lugar, la distribución χ^2 central con n grados de libertad (se denota χ^2_n) está definida sobre $\mathbb{R}^+$ mediante la siguiente función de densidad[5]

$$g_n(\mathtt{y}) = [\Gamma(n/2)2^{n/2}]^{-1} e^{-\mathtt{y}/2} \mathtt{y}^{\frac{n}{2}-1} I_{(0,+\infty)}(\mathtt{y}). \tag{1.9}$$

Puede probarse que tiene por media n y por varianza $2n$. La distribución χ^2 no central con m grados de libertad y parámetro de no centralidad $\lambda > 0$ (se denota $\chi^2_m(\lambda)$) se define mediante la función de densidad

$$\sum_{n=0}^{\infty} P_n(\lambda) g_{2n+1}(\mathtt{y}),$$

donde

$$P_n(\lambda) = \lambda^n \frac{\mathtt{e}^{-\lambda}}{n!}, \quad \mathtt{n} \in \mathbb{N}.$$

[5]Recordemos previamente que la función $\Gamma(\cdot)$ se define mediante $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\mathtt{e}^{-x}dx$, donde $\alpha > 0$.

Se obtiene, por lo tanto, a partir de una composición entre una distribución de Poisson en $\mathbb{N}$ y la familia de las distribuciones χ^2_n, cuando n recorre $\mathbb{N}$. La distribución χ^2 central se corresponde con el caso $\lambda = 0$. En general, dado $\gamma > 0$, la expresión $Y \sim \gamma\chi^2_m(\lambda)$ debe entenderse como $\gamma^{-1}Y \sim \chi^2_n(\lambda)$.

Puede demostrarse que, si $Y_1, \ldots, Y_n$ son variables aleatorias reales independientes tales que

$$Y_i \sim N(\mu_i, \sigma^2),\ i = 1, \ldots, n,\ \sigma^2 > 0,$$

entonces

$$\sigma^{-2} \sum_{i=1}^{n} Y_i^2 \sim \chi^2_n \left(\sigma^{-2} \sum_{i=1}^{n} \mu_i^2 \right).$$

En otras palabras, considerar una colección de variables en esas condiciones equivale a considerar un vector aleatorio $Y \sim N_n(\mu, \sigma^2 \mathtt{Id})$, para algún $\mu \in \mathbb{R}^n$ y $\sigma^2 > 0$, y estamos afirmando que

$$\|Y\|^2 \sim \sigma^2 \chi^2_n \left(\frac{\|\mu\|^2}{\sigma^2} \right).$$

En consecuencia, debemos entender el modelo χ^2 no central como la distribución del cuadrado de la distancia euclídea al origen de un vector aleatorio normal esférico. La norma euclídea al cuadrado es una función positiva de gran importancia en nuestra teoría, debida fundamentalmente a su presencia en la función de densidad (1.8). De hecho, ya comentamos que la densidad depende de $\mathtt{y}$ a través del cuadrado de su distancia euclídea a la media. Ello se traducirá en el uso de esta función y, en consecuencia, del modelo χ^2, a la hora de estimar el parámetro σ^2, de reducir por suficiencia y, también, cuando se efectúe una reducción por invarianza respecto al grupo de las rotaciones, según se sigue del teorema 13.9.

Hemos afirmado que el modelo χ^2 no central surge de la necesidad de considerar la norma euclídea de un vector normal esférico. No obstante, podemos generalizar un poco más. Si E es un subespacio vectorial de $\mathbb{R}^n$ y Γ es una base ortonormal del mismo, se verifica trivialmente que $\|P_E Y\|^2 = \|\Gamma' Y\|^2$ y que $\|P_E \mu\|^2 = \|\Gamma' \mu\|^2$. Por lo tanto, se tiene

$$\|P_E Y\|^2 \sim \sigma^2 \chi^2_{\mathtt{dim}E} \left(\frac{\|P_E \mu\|^2}{\sigma^2} \right). \tag{1.10}$$

Así pues, el grado de libertad de la distribución coincide con la dimensión del subespacio. Obtendremos una χ^2 central cuando $\mathtt{E}[Y]$ sea ortogonal al subespacio sobre el cual se proyecta Y. Por lo tanto y en general, se sigue de lo anterior junto con la proposición 1.10, que la media de una distribución χ^2 no central se obtiene mediante

$$\mathtt{E}\left[\sigma^2 \chi^2_m\big(\lambda/\sigma^2\big)\right] = m\sigma^2 + \lambda. \tag{1.11}$$

Dadas dos variables aleatorias reales X_1 y X_2, positivas e independientes, con distribuciones $\chi^2_n(\lambda)$, siendo $\lambda \geq 0$, y χ^2_m, respectivamente, se define la distribución F-Snedecor no central con (n, m) grados de libertad y parámetro de no centralidad λ (de denota por $F_{n,m}(\lambda)$), como la que corresponde a la variable $(n^{-1}X_1)/(m^{-1}X_2)$. Puede demostrarse que su función de densidad es la siguiente:

$$f_{n,m,\lambda}(\mathtt{y}) = \frac{n}{m}\mathtt{e}^{-\lambda}\sum_{k=0}^{\infty} c_k \frac{\lambda^k}{k!}\frac{\left(\frac{n}{m}\mathtt{y}\right)^{\frac{n}{2}-1+k}}{\left(1+\frac{n}{m}\mathtt{y}\right)^{\frac{n+m}{2}+k}} I_{(0,+\infty)}(\mathtt{y}), \tag{1.12}$$

donde 0^0 se entiende como 1 y

$$c_k = \frac{\Gamma\left(\frac{1}{2}(n+m)+k\right)}{\Gamma\left(\frac{1}{2}n+k\right)\Gamma\left(\frac{1}{2}m\right)}, \quad k \in \mathbb{N}.$$

La distribución $F_{n,m}(0)$ se denomina F-Snedecor central con (n, m) grados de libertad, y se denota por $F_{n,m}$. Su función de densidad es pues la siguiente:

$$f_{n,m}(\mathtt{y}) = \frac{n^{\frac{n}{2}}m^{\frac{m}{2}}\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n}{2}\right)}\frac{\mathtt{y}^{\frac{n}{2}-1}}{(n\mathtt{y}+m)^{\frac{n+m}{2}}} I_{(0,+\infty)}(\mathtt{y}).$$

En nuestro caso, si $Y \sim N_n(\mu, \sigma^2\mathtt{Id})$ y dados dos subespacios ortogonales $V_1, V_2 \subset \mathbb{R}^n$ tales que $\mu \in V_2^{\perp}$, se verifica que

$$\frac{\mathtt{dim}V_2}{\mathtt{dim}V_1}\frac{\|P_{V_1}Y\|^2}{\|P_{V_2}Y\|^2} \sim F_{\mathtt{dim}V_1,\mathtt{dim}V_2}\left(\frac{\|P_{V_1}\mu\|^2}{\sigma^2}\right). \tag{1.13}$$

Así pues, la distribución F de Snedecor resulta de relacionar las distancias al origen de dos proyecciones sobre sendos subespacio ortogonales. Si $\mu \in V_1^{\perp} \cap V_2^{\perp}$ tendremos una distribución F central. Una operación de este tipo surge al reducir por invarianza en el proceso de obtención del test `F` (ver volumen anterior). Otras distribuciones íntimamente relacionadas con la F-Snedecor central son la Beta y la t-Student.

La distribución Beta de parámetros $\alpha, \beta > 0$, que se denotará por $B(\alpha, \beta)$, se define mediante la función de densidad[6]

$$f_{\alpha,\beta}(\mathtt{y}) = \mathtt{B}(\alpha,\beta)^{-1}\mathtt{y}^{\alpha-1}(1-\mathtt{y})^{\beta-1}I_{(0,1)}(\mathtt{y}).$$

Se trata pues de una distribución sobre el intervalo $(0, 1)$. Presenta un estrecha relación con la distribución F-Snedecor central. Concretamente, se verifica

$$X \sim F(n,m) \;\Leftrightarrow\; \left(1+\frac{n}{m}X\right)^{-1} \sim B\left(\frac{m}{2},\frac{n}{2}\right). \tag{1.14}$$

[6]Recordar que la función $\mathtt{B}$ se define mediante $\mathtt{B}(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$, donde $\alpha, \beta > 0$.

La distribución t de student central con n grados de libertad (se denota por t_n) es la que corresponde al cociente $X_1/\sqrt{X_2/n}$, donde $X_1 \sim N(0,1)$ y $X_2 \sim \chi^2_n$, siendo ambas independientes. Su densidad es la siguiente:

$$f_n(\mathtt{y}) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)\left(1+\frac{\mathtt{y}^2}{n}\right)^{-\frac{n+1}{2}}}.$$

La distribución t_n puede considerarse un caso particular de la distribución F-Snedecor central, concretamente $F_{1,n}$ dado que es la única distribución simétrica cuyo cuadrado es una $F_{1,n}$. En ese sentido decimos que $t_n^2 = F_{1,n}$.

Para acabar con esta sección, hablaremos de las distintas pruebas de bondad de ajuste a la distribución normal multivariante. Primeramente, debemos preguntarnos si alguno de los procedimientos clásicos para contrastar la normalidad en dimensión uno son válidos o pueden extenderse fácilmente al caso multivariante. En ese sentido, hemos de tener en cuenta que el test de Kolmogorov-Smirnov-Lilliefords, que analiza la similitud entre la función de distribución empírica y la que correspondería a una normal, sólo tiene sentido en dimensión uno. Por otra parte, el test χ^2, que analiza la semejanza entre el histograma de frecuencias relativas y la función de densidad, requiere, en el caso multivariante, de una enorme cantidad de datos para ser aproximadamente válido. Sin embargo, Srivastava, Hui (1987) proponen dos tests que son generalizaciones multivariantes del test de Shapiro-Wilks, y que quedan recogidos en Bilodeau (1999).

De todas formas, dado que la normalidad del vector aleatorio implica la normalidad de sus componentes, es costumbre bastante extendida (Hair et al. (1999)) resolver el problema contrastando la normalidad de cada una de sus componentes. Este método, que no es un test propiamente dicho, puede refinarse mediante una análisis del diagrama de dispersión matricial para comprobar visualmente si las distribuciones bidimensionales se asemejan a lo que cabría esperar en el caso normal. De este último libro recogemos otra interesante prueba de bondad de ajuste basada en la distribución de las distancias de mahalanobis de los datos a la media muestral. Consiste en lo siguiente:

Consideremos una muestra aleatoria simple $X_1, \ldots, X_n$ de determinada distribución sobre $\mathbb{R}^p$ con matriz de covarianzas positiva. Nuestro problema consiste en decidir si dicha distribución es p-normal. Consideremos las distancias de Mahalanobis

$$d_i^2 = (X_i - \overline{X})' S^{-1} (X_i - \overline{X}),\ i = 1, \ldots, n$$

Nótese que en caso $p = 1$ estamos hablando de $s^{-2}(x_i - \overline{x})^2$, que son los valores x_i tipificados. En lo que respecta a la distribución de la distancia de Mahalanobis

bajo la hipótesis de normalidad, enunciamos el siguiente resultado, cuya demostración podemos encontrar en Bilodeau (1999), pag. 185, junto con Wilks (1963):

Teorema 1.11.
Si $X_1, \ldots, X_n$ iid $N_p(\mu, \Sigma)$, entonces

$$\frac{n}{(n-1)^2} d_i^2 \sim \beta\left(\frac{1}{2}p;\ \frac{1}{2}(n-p-1)\right), \qquad r_{d_i^2, d_j^2} = -\frac{1}{n-1}.$$

Ello invita a considerar, en el caso p-normal, el conjunto

$$\frac{n}{(n-1)^2}\ d_1^2, \ldots, \frac{n}{(n-1)^2}\ d_n^2$$

como una muestra aleatoria simple de la distribución $\beta\big(\frac{1}{2}p;\ \frac{1}{2}(n-p-1)\big)$, siempre y cuando n sea *suficientemente grande*. En ese caso y en virtud del teorema de Glivenko-Cantelli, la función de distribución empírica debería converger uniformemente a la función de distribución del modelo Beta anterior, lo cual puede contrastarse mediante el test de Kolmogorov-Smirnov-Lilliefords. No obstante y como opción alternativa, podemos realizar la validación mediante procedimientos meramente gráficos. Efectivamente, si ordenamos las distancias $n(n-1)^{-2}d_{(1)}^2 \geq \ldots \geq n(n-1)^{-2}d_{(n)}^2$ y asignamos al elemento i-ésimo de la lista anterior la proporción acumulada i/n, obtendremos la función de distribución empírica. Los n valores obtenidos pueden compararse con los que corresponderían a una distribución Beta con los parámetros anteriores. El gráfico de dispersión bidimensional que los confronta se denomina P-P Plot. Además, dado que la igualdad o convergencia de las funciones de distribución implica a igualdad o convergencia, respectivamente, de sus inversas, podemos confrontar el valor i-ésimo de la lista con con el cuantil i/n correspondiente a la distribución Beta, obteniéndose así el denominado Q-Q Plot. Una buena aproximación al cuantil i-ésimo de la distribución Beta anterior es según Blom (1958)[7] el siguiente:

$$\gamma_i = (i-\alpha)/(n-\alpha-\beta+1), \quad \text{donde } \alpha = \frac{p-2}{2p}, \quad \beta = \frac{n-p-2}{2(n-p-1)}.$$

Tanto en los gráficos tipo P-P como en los Q-Q cabe esperar un buen ajuste de la nube de puntos a la recta $y = x$. En caso contrario, la hipótesis inicial debe ser rechazada.

Otro método similar consiste en calcular $n^{-1}d_i^2$, $i = 1, \ldots, n$ y confrontarlos mediante el test de Kolmogorov- Smirnov-Lilliefords o, en su defecto, mediante gráficos tipo Q-Q o P-P con la distribución χ_p^2. Este procedimiento se basa en el hecho de

[7] Blom (1958), Statistical Estimation Transformed Beta-variables. Wiley.

que $n(X_i-\mu)'\Sigma^{-1}(\mathbf{X}_i-\mu)$ sigue una distribución χ^2_p. Teniendo en cuenta que tanto $\overline{X}$ como S convergen en probabilidad a μ y Σ, respectivamente, la Ley Débil de os Grandes Números garantiza la convergencia en distribución de $n^{-1}d_i^2$ a χ^2_p.

1.2. Distribución normal matricial.

Este modelo probabilístico generaliza la distribución normal multivariante en un contexto matricial. En esta sección demostraremos sus principales propiedades. Antes de definir esta nueva modelo distribución hemos de aclarar algunos conceptos previos relacionados con el mismo: el concepto de matriz aleatoria, que se distingue del de vector aleatorio sólo en un sutil matiz, el de función característica de una matriz aleatoria y el producto de Kronecker de dos matrices, cuyas propiedades se exponen más ampliamente en Bilodeau (1999). La distribución normal matricial se definirá aquí mediante la función característica, aunque algunos autores (por ejemplo Arnold (1981)) prefieren utilizar la función generatriz de momentos, o incluso la función de densidad.

Primeramente, definimos una matriz aleatoria $n\times p$ como una variable X sobre un espacio de probabilidad $(\Omega,\mathcal{A},P)$ con valores en $\mathcal{M}_{n\times p}$. Se denota $X=(X_{i,j})$, donde $1\leq i\leq n$ y $1\leq j\leq p$. Si consideramos un orden de lectura determinado, toda matriz $n\times p$ se identifica con un vector de $\mathbb{R}^{np}$. Es decir, si $\mathtt{Vec}_{n\times p}$ es el conjunto de todas las posibles formas de ordenar una matriz $n\times p$, que se corresponde con el conjunto de todas las posibles permutaciones de np elementos, de cardinal $(np)!$, podemos establecer una aplicación ϕ del producto cartesiano $\mathcal{M}_{n\times p}\times\mathtt{Vec}_{n\times p}$ en $\mathbb{R}^{np}$, tal que, para cada orden de lectura $\mathtt{vec}\in\mathtt{Vec}_{n\times p}$, $\phi(\cdot,\mathtt{vec})$ es una biyección de $\mathcal{M}_{n\times p}$ en $\mathbb{R}^{np}$. De esta forma, determinado previamente el orden de lectura $\mathtt{vec}$, una matriz aleatoria X sobre $\mathcal{M}_{n\times p}$ es un vector aleatorio sobre $\mathbb{R}^{np}$, que se denota $\mathtt{vec}(X)$. El concepto de matriz aleatoria se precisa por el hecho de que n datos correspondientes a una distribución p-dimensional configuran una matriz de orden $n\times p$.

Analizamos a continuación algunas propiedades de la traza de una matriz cuadrada, que serán de interés en nuestra teoría. Recordemos primeramente que la traza de una matriz cuadrada es la suma de los elementos de su diagonal. Por otra parte, si $a,b\in\mathbb{R}^n$ donde $a=(a_1,\ldots,a_n)'$ y $b=(b_1,\ldots,b_n)'$ se define el producto interior de ambos vectores mediante

$$\langle a,b\rangle := a'b=\sum_{i=1}^{n}a_ib_i \tag{1.15}$$

Pues bien, dadas dos matrices $A,B\in\mathcal{M}_{n\times p}$, con componentes a_{ij} y b_{ij}, respectiva-

mente, donde $i = 1, \ldots, n$ y $j = 1, \ldots, p$,, se verifica

$$\mathtt{tr}(A'B) = \sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}b_{ij}, \tag{1.16}$$

es decir, $\mathtt{tr}(A'B) = \langle \mathtt{vec}(A), \mathtt{vec}(B)\rangle$, para todo $\mathtt{vec} \in \mathtt{Vec}_{n\times p}$. En ese sentido, podemos afirma que la $\mathtt{tr}(A'B)$ generaliza el producto interior de dos vectores, de ahí que definamos

$$\langle A, B\rangle := \mathtt{tr}(A'B), \quad A, B \in \mathcal{M}_{n\times p} \tag{1.17}$$

En particular, se tiene que

$$\langle A, A\rangle = \sum_{ij} a_{ij}^2 = \|\mathtt{vec}(A)\|^2, \quad \forall \mathtt{vec} \in \mathtt{Vec}_{n\times p}.$$

Se deduce fácilmente que $\mathtt{tr}(A'B) = \mathtt{tr}(B'A) = \mathtt{tr}(AB') = \mathtt{tr}(BA')$. Por último, dadas A, B, C matrices cuadradas de orden n, se verifica que

$$\mathtt{tr}(ABC) = \mathtt{tr}(CAB) = \mathtt{tr}(BAC). \tag{1.18}$$

Por otra parte, sabemos que la distribución de un vector aleatorio queda determinada por su función característica. En el caso de una matriz a matriz aleatoria $X \in \mathcal{M}_{n\times p}$, ésta se define mediante

$$\begin{aligned}
\varphi_X(t) &:= \mathtt{E}\left[\exp\left\{\mathbf{i}\langle X, t\rangle\right\}\right] \\
&= \mathtt{E}\left[\exp\left\{\mathbf{i}\mathtt{tr}(X't)\right\}\right] \\
&= \mathtt{E}\left[\exp\left\{\mathbf{i}\sum_{k=1}^{n}\sum_{j=1}^{p} X_{kj}t_{kj}\right\}\right],
\end{aligned}$$

siendo X y t matrices de dimensión $n \times p$. Fijado un orden de lectura $\mathtt{vec}$, se tiene entonces que

$$\begin{aligned}
\varphi_X(t) &= \mathtt{E}\left[\exp\{\mathbf{i}\langle \mathtt{vec}(X), \mathtt{vec}(t)\rangle\right] \\
&= \varphi_{\mathtt{vec}(X)}\big(\mathtt{vec}(t)\big).
\end{aligned}$$

Por lo cual esta nueva definición hereda las propiedades ya conocidas de la función característica de vectores aleatorios, en particular el teorema de inversión de Levy. Se verifica que dos matrices aleatorias X e Y siguen la misma distribución cuando los vectores $\mathtt{vec}(X)$ y $\mathtt{vec}(Y)$ se distribuyen idénticamente, para cualquier orden $\mathtt{vec}$ (si, y sólo si, ocurre para alguno). Luego, la función característica determina unívocamente las distribuciones matriciales. Vamos a destacar, no obstante, tres propiedades, de fácil demostración:

(a) Si $X = (X_1|X_2)$, $\varphi_X\big((t_1|0)\big) = \varphi_{X_1}(t_1)$.
Análogamente, si $X = \left(\dfrac{X_1}{X_2}\right)$, entonces

$$\varphi_X\left(\left(\frac{t_1}{0}\right)\right) = \varphi_{X_1}(t_1).$$

(b) $\varphi_{AXB+C}(t) = \exp\{\mathtt{itr}(C't)\} \cdot \varphi_X(A'tB')$.

(c) Si $X = (X_1 \dots X_p)$, los vectores $X_1, \dots, X_p$ son independientes si, y sólo si, $\varphi_X(t_1 \dots t_p) = \prod_{j=1}^{p} \varphi_{X_j}(t_j)$. Lo mismo sucede si descomponemos X por filas.

Por convenio, consideraremos por defecto el orden `vec` consistente en leer la matriz X por filas, es decir,

$$\mathtt{vec}\begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} = (x_{11}, x_{12}, \dots, x_{1p}, x_{2,1}, \dots, x_{np})' \in \mathbb{R}^{np}. \tag{1.19}$$

Se define la esperanza o media de una matriz aleatoria X como la matriz de las esperanzas de cada una de sus componentes, es decir

$$\mathtt{E}[X] := \big(\mathtt{E}[X_{ij}]\big)_{i,j} \tag{1.20}$$

La matriz de varianzas-covarianzas de la matriz aleatoria X, que se denota por $\mathtt{Covm}[X]$, se define como la matriz $np \times np$ definida de la forma

$$\mathtt{Covm}[X] := \mathtt{Cov}\big[\mathtt{vec}(X)\big]. \tag{1.21}$$

El denominado producto de Kronecker será de gran utilidad a la hora de describir este tipo de matrices, de orden muy elevado. Dadas $A \in \mathcal{M}_{n\times m}$ y $B \in \mathcal{M}_{p\times q}$, se define el producto de Kronecker de ambas mediante

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1m}B \\ \vdots & & \vdots \\ a_{n1}B & \dots & a_{nm}B \end{pmatrix} \in \mathcal{M}_{np\times mq}. \tag{1.22}$$

Podemos encontrar en Bilodeau (1999) las principales propiedades del producto de Kronecker de dos matrices. Destacamos las siguientes, de fácil comprobación:

Proposición 1.12.
Sean $Z \in \mathcal{M}_{n\times p}$, $A \in \mathcal{M}_{m\times n}$, $B \in \mathcal{M}_{p\times k}$, $C \in \mathcal{M}_{n\times r}$, $D \in \mathcal{M}_{k\times s}$ y $H \in \mathcal{M}_{m\times k}$. Entonces

(i) $\mathtt{vec}(AZB + H) = (A \otimes B')\mathtt{vec}(Z) + \mathtt{vec}(H)$.

(ii) $(A \otimes B)' = A' \otimes B'$.

(iii) $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.

Estamos ya en condiciones de definir la distribución normal matricial.

Definición.
Sean $n, p \in \mathbb{N}$, $\mu \in \mathcal{M}_{n\times p}$, $\Gamma \in \mathcal{M}_{n\times n}$, $\Sigma \in \mathcal{M}_{p\times p}$, estas últimas semidefinidas positivas, y X una matriz aleatoria sobre $\mathcal{M}_{n\times p}$. Se dice que X sigue una distribución normal matricial, denotándose $X \sim N_{n,p}(\mu, \Gamma, \Sigma)$, cuando

$$\mathtt{vec}(X) \sim N_{np}\big(\mathtt{vec}(\mu), \Gamma \otimes \Sigma\big).$$

En ese caso, se verificará, por tanto, que

$$\mathtt{E}[X] = \mu, \qquad \mathtt{Covm}[X] = \Gamma \otimes \Sigma.$$

Luego, si $1 \le i, i' \le n$ y $1 \le j, j' \le p$, se verifica que

$$X_{ij} \sim N(\mu_{ij},\ \gamma_{ii} \cdot \sigma_{jj}), \quad \mathtt{cov}\,[X_{ij}, X_{i'j'}] = \gamma_{ii'} \cdot \sigma_{jj'}.$$

Veamos primeramente que esta definición no es vacía, es decir, que existe una matriz aleatoria con valores en $\mathcal{M}_{n\times p}$ siguiendo un modelo de distribución $N_{n,p}(\mu, \Gamma, \Sigma)$:

Teorema 1.13.
Existe una matriz aleatoria en las condiciones de la definición.

Demostración.
En primer lugar, al ser Γ una matriz $n \times n$ semidefinida positiva, podemos encontrar una matriz $A \in \mathcal{M}_{n\times n}$ tal que $\Gamma = AA'$. Igualmente, podemos encontrar una matriz $B \in \mathcal{M}_{p\times p}$ tal que $\Sigma = B'B$. Sea entonces Z una matriz aleatoria $n \times p$ cuyas componentes son iid $N(0, 1)$. En ese caso, $\mathtt{vec}(Z) \sim N_{np}(0, \mathtt{Id})$. Consideremos entonces la matriz aleatoria $X = AZB + \mu$. En virtud de la proposición 1.12(i), se tiene que

$$\mathtt{vec}(X) = (A \otimes B')\mathtt{vec}(Z) + \mathtt{vec}(\mu),$$

luego, sigue un modelo de distribución normal np-variante de media $\mathtt{vec}(\mu)$ y matriz de covarianzas

$$(A \otimes B')\mathtt{Id}(A \otimes B')',$$

que, en virtud de la proposición 1.12(ii)-(iii) es igual a $AA' \otimes B'B$, es decir, igual a $\Gamma \otimes \Sigma$.

∎

Aplicando las propiedades (b) y (c) obtenemos la función característica de la distribución $N_{n,p}(\mu, \Gamma, \Sigma)$.

Corolario 1.14.
Si $X \sim N_{n,p}(\mu, \Gamma, \Sigma)$, entonces

$$\varphi_X(t) = \exp\left\{\mathbf{i}\mathtt{tr}(\mu' t) - \frac{1}{2}\mathtt{tr}(t'\Gamma t \Sigma)\right\}.$$

Demostración.
Consideremos Z en las condiciones del teorema 1.13. Se verifica entonces, para cada $u \in \mathcal{M}_{n\times p}$,

$$\begin{aligned}\varphi_Z(u) &\overset{(c)}{=} \prod_{i,j}\varphi_{Z_{ij}}(u_{ij}) = \prod_{ij}\exp\left\{-\frac{1}{2}u_{ij}^2\right\} \\ &= \exp\left\{-\frac{1}{2}\sum_{i,j}u_{ij}^2\right\} = \exp\left\{-\frac{1}{2}\mathtt{tr}(u'u)\right\}\end{aligned}$$

En consecuencia, se tiene que, para cada $t \in \mathcal{M}_{n\times p}$,

$$\begin{aligned}\varphi_X(t) &= \varphi_{AZB+\mu}(t) \\ &\overset{(b)}{=} \exp\{\mathbf{i}\mathtt{tr}(\mu' t)\}\cdot\varphi_Z(A'tB') \\ &= \exp\left\{\mathbf{i}\mathtt{tr}(\mu' t) - \frac{1}{2}\mathtt{tr}\big((A'tB')'(A'tB')\big)\right\} \\ &= \exp\left\{\mathbf{i}\mathtt{tr}(\mu' t) - \frac{1}{2}\mathtt{tr}\big(t'AA'tB'B\big)\right\} \\ &= \exp\left\{\mathbf{i}\mathtt{tr}(\mu' t) - \frac{1}{2}\mathtt{tr}\big(t'\Gamma t\Sigma\big)\right\}\end{aligned}$$

∎

Veamos algunas propiedades inmediatas de la distribución normal matricial.

Teorema 1.15.
Supongamos que $X \sim N_{n,p}(\mu, \Gamma, \Sigma)$. Se verifica:

(a) Si $p = 1$ y $\Sigma = \sigma^2 \geq 0$, entonces $N_{n,1}(\mu, \Gamma, \sigma^2) = N_n(\mu, \sigma^2\Gamma)$.

(b) Si $a \in \mathbb{R}$, entonces $aX \sim N_{n,p}(a\mu, a^2\Gamma, \Sigma) = N_{n,p}(a\mu, \Gamma, a^2\Sigma)$.

(c) $X' \sim N_{p,n}(\mu', \Sigma, \Gamma)$. Si $n = 1$, y $\Gamma = \gamma^2$, $X' \sim N_p(\mu', \gamma^2\Sigma)$.

(d) Si $C \in \mathcal{M}_{m\times n}$, $D \in \mathcal{M}_{p\times r}$ y $E \in \mathcal{M}_{m\times r}$, entonces

$$CXD + E \sim N_{m,r}\big(C\mu D + E, C\Gamma C', D'\Sigma D\big). \tag{1.23}$$

(e) Si $p = p_1 + p_2$, $X = (X_1X_2)$, $\mu = (\mu_1\mu_2)$ y $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, entonces

$$X_i \sim N_{n,p_i}(\mu_i, \Gamma, \Sigma_{ii}).$$

(f) Si $n = n_1 + n_2$, $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ y $\Gamma = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix}$, entonces

$$X_i \sim N_{n_i,p}(\mu_i, \Gamma_{ii}, \Sigma).$$

Demostración.
Las propiedades (a), (b) y (c) son triviales. Para demostrar (d) basta considerar la función característica de la transformación. Para probar (e) se considera la propiedad (d) tomando

$$D_1 = \begin{pmatrix} \mathtt{Id} \\ 0 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0 \\ \mathtt{Id} \end{pmatrix}.$$

Igualmente, para demostrar (f), se considera $C_1 = (\mathtt{Id}|0)$ y $C_2 = (0|\mathtt{Id})$.

∎

Sabemos que en la distribución normal multivariante la incorrelación y la independencia sin propiedades equivalentes. Veamos com se traslada este resultado a la distribución normal matricial.

Teorema 1.16.

Supongamos que $X \sim N_{n,p}(\mu, \Gamma, \Sigma)$.

(a) Si $X = (X_1X_2)$ y $\Gamma \neq 0$, X_1 y X_2 son independientes sii $\Sigma_{12} = 0$.

(b) Si $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ y $\Sigma \neq 0$, X_1 y X_2 son independientes sii $\Gamma_{12} = 0$.

Demostración.
(a) Supongamos que X_1 y X_2 son independientes. Al, ser $\Gamma \neq 0$, existen $i, i' \in \{1, \ldots, n\}$ tales que $\gamma_{ii'} \neq 0$. Tenemos que demostrar que, si $j \in \{1, \ldots, p_1\}$ y $j' \in \{p_1 + 1, \ldots, p\}$, entonces $\sigma_{jj'} = 0$. Por hipótesis se tiene que $0 = \mathtt{cov}[X_{ij}, X_{i'j'}] = \gamma_{ii'} \cdot \sigma_{jj'} \Rightarrow \sigma_{jj'} = 0$.
Recíprocamente, si $\Sigma_{12} = 0$, la función característica de X en $(t_1 t_2)$ se expresará como sigue:

$$\begin{aligned} \varphi_{(X_1 X_2)}(t_1 t_2) &= \exp\left\{ \mathtt{itr}(\mu_1' t_1 + \mu_2' t_2) - \frac{1}{2}\mathtt{tr}\left[\Gamma(t_1 t_2) \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \begin{pmatrix} t_1' \\ t_2' \end{pmatrix} \right] \right\} \\ &= \varphi_{X_1}(t_1)\varphi_{X_2}(t_2). \end{aligned}$$

(b) Se demuestra análogamente.

∎

Veamos cuál es la función de densidad de la distribución normal matricial. Previamente, enunciamos un resultado que es consecuencia directa del teorema del cambio de variables:

Lema 1.17.

Sea P una probabilidad sobre $(\mathbb{R}^n, \mathcal{R}^n)$ dominada por la medida de Lebesgue m^n, siendo f su función de densidad. Consideremos una biyección $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, tal que φ y φ^{-1} son de clase $\mathcal{C}^1$, y sea J_φ la matriz de las derivadas parciales de dicha transformación. Entonces $|J_\varphi| \cdot f \circ \varphi$ es la densidad de $P^{\varphi^{-1}}$ respecto a m^n.

También es necesario el siguiente resultado previo:

Lema 1.18.

Sean X e Y matrices aleatorias $n \times p$, A y B matrices $n \times n$ y $p \times p$, respectivamente, y D una matriz $n \times p$, tales que $Y = A(XD) - B$. Entonces, el determinante jacobiano de la transformación que lleva $\mathtt{vec}(X)$ a $\mathtt{vec}(Y)$ es $J = |A|^p |B|^n$.

Demostración.
Primeramente, si tenemos una transformación del tipo $Y = XB$, ésta se expresa vectorialmente de la siguiente forma:

$$\mathtt{vec}(Y) = \mathtt{vec}(X) \left(\begin{array}{c|c|c} B & 0 & 0 \\ \hline 0 & \ddots & 0 \\ \hline 0 & 0 & B \end{array} \right),$$

donde la matriz B se repite tantas veces como filas tenga X, es decir, n. El determinante jacobiano de esta transformación es $|B|^n$.

En segundo lugar, si la transformación es del tipo $Y = AX$, entonces, la aplicación que asigna a $\mathtt{vec}(X)$ el vector $\mathtt{vec}(AX)$ puede descomponerse de la siguiente forma:

$$\mathtt{vec}(X) \xrightarrow{\tau_1} \mathtt{vec}(X') \xrightarrow{\phi} \mathtt{vec}(X'A') \xrightarrow{\tau_2} \mathtt{vec}\big((X'A')'\big)$$

Por un razonamiento análogo al anterior, el determinante jacobiano de ϕ es $|A|^p$. Dado que τ_1 y τ_2 son permutaciones opuestas en $\mathbb{R}^{np}$, los determinantes jacobianos correspondientes valen ambos +1 ó -1. Luego, por la regal de la cadena, se concluye. ∎

Teorema 1.19.

Sea $X \sim N_{n,p}(\mu, \Gamma, \Sigma)$ con $\Gamma > 0$ y $\Sigma > 0$. Entonces X admite la siguiente función de densidad respecto a la medida de Lebesgue en $\mathbb{R}^{np}$ [8]:

$$f_X(\mathtt{x}) = \frac{1}{(2\pi)^{np/2}|\Gamma|^{p/2}|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}\left[\Gamma^{-1}(\mathtt{x}-\mu)\Sigma^{-1}(\mathtt{x}-\mu)'\right]\right\}, \qquad x \in \mathcal{M}_{n\times p}.$$

Demostración.

Consideremos Z, A, B como en la demostración del teorema 1.13. Al ser $\Gamma > 0$ y $\Sigma > 0$, podemos considerar A y B invertibles. Como las np componentes de Z son iid $N(0,1)$, $\mathtt{vec}(Z)$ admite la siguiente función de densidad respecto a m^{np} :

$$\begin{aligned} f_Z(\mathtt{z}) &= \prod_{i,j} f_{Z_{ij}}(\mathtt{z}_{ij}) \\ &= \prod_{i,j}\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\mathtt{z}_{ij}^2\right\}\right) \\ &= \frac{1}{(2\pi)^{np/2}}\exp\left\{-\frac{1}{2}\mathtt{tr}(\mathtt{z}\mathtt{z}')\right\}, \end{aligned}$$

donde $\mathtt{z} \in \mathcal{M}_{n\times p}$. Como $X = AZB + \mu$, entonces $Z = A^{-1}(X-\mu)B^{-1}$. El jacobiano de dicha transformación es, según el lema 1.18, $|A^{-1}|^p|B^{-1}|^n = |\Gamma|^{-p/2}|\Sigma|^{-n/2}$. Luego, por el lema 1.17, se tiene que la función de densidad de $\mathtt{vec}(X)$ es la siguiente:

$$\begin{aligned} f(\mathtt{x}) &= \frac{1}{(2\pi)^{np/2}|\Gamma|^{p/2}|\Sigma|^{n/2}}\exp\left\{-\frac{1}{2}\mathtt{tr}\left[A^{-1}(\mathtt{x}-\mu)B^{-1}(B')^{-1}(\mathtt{x}-\mu)'(A')^{-1}\right]\right\} = \\ &= \frac{1}{(2\pi)^{np/2}|\Gamma|^{p/2}|\Sigma|^{n/2}}\exp\left\{-\frac{1}{2}\mathtt{tr}\left[\Gamma^{-1}(\mathtt{x}-\mu)\Sigma^{-1}(\mathtt{x}-\mu)'\right]\right\}, \qquad \mathtt{x} \in \mathcal{M}_{n\times p}. \end{aligned}$$

∎

Veamos qué sucede con las distribuciones condicionales:

[8] Realmente queremos decir que es $\mathtt{vec}(X)$ quien admite función de densidad.

Teorema 1.20.
Sea $(X_1X_2) \sim N_{n,p_1+p_2}\big((\mu_1,\mu_2),\Gamma,\Sigma\big)$, con $\Gamma > 0$ y $\Sigma > 0$. Definamos $\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $B = \Sigma_{22}^{-1}\Sigma_{21}$ y $A = \mu_1 - \mu_2 B$. Entonces, se verifica que

$$X_1|X_2 = \mathtt{x}_2 \sim N_{n,p_1}(A + \mathtt{x}_2 B, \Gamma, \Sigma_{11\cdot 2}).$$

Demostración.
Se trata de demostrar que la función de densidad de la distribución condicional, $f_{X_1|X_2=\mathtt{x}_2}$, corresponde a un modelo de probabilidad $N_{n,p_1}(A + \mathtt{x}_2 B, \Gamma, \Sigma_{11\cdot 2})$, cualquiera que sea $\mathtt{x}_2$. Recordemos que, si $f_{(X_1X_2)}$ es la densidad de (X_1X_2) y f_{X_2} la de X_2, entonces

$$f_{X_1|X_2=\mathtt{x}_2}(\mathtt{x}_1) = \frac{f_{(X_1X_2)}(\mathtt{x}_1,\mathtt{x}_2)}{f_{X_2}(\mathtt{x}_2)}.$$

Vamos a expresar $f_{(X_1X_2)}$ de una manera cómoda. Para ello consideraremos el cambio de variables $(Y_1Y_2) = (X_1X_2)C$, donde

$$C = \begin{pmatrix} \mathtt{Id} & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & \mathtt{Id} \end{pmatrix}.$$

Así, $(Y_1Y_2) \sim N_{n,p_1+p_2}\big((\mu_1,\mu_2)C, \Gamma, C'\Sigma C\big)$. Operando se obtiene

$$(\mu_1,\mu_2)C = \big(\mu_1 - \mu_2\Sigma_{22}^{-1}\Sigma_{21}, \mu_2\big), \quad C'\Sigma C = \begin{pmatrix} \Sigma_{11\cdot 2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}.$$

Luego, $Y_1 \sim N_{n,p_1}\big(\mu_1 - \mu_2\Sigma_{22}^{-1}\Sigma_{21}, \Gamma, \Sigma_{11\cdot 2}\big)$ y $Y_2 \sim N_{n,p_2}(\mu_2,\Gamma,\Sigma_{22})$, siendo independientes entre sí. Nótese que la distribución de Y_2 coincide con la de X_2. En consecuencia, $f_{(Y_1Y_2)}(\mathtt{y}_1,\mathtt{y}_2) = f_{Y_1}(\mathtt{y}_1)\cdot f_{Y_2}(\mathtt{y}_2)$, que es igual a

$$\frac{f_{X_2}(\mathtt{y}_2)}{(2\pi)^{np_1}|\Gamma|^{p_1/2}|\Sigma_{11\cdot 2}|^{n/2}}\, \mathtt{e}^{-\frac{1}{2}\mathtt{tr}\left[\left(\mathtt{y}_1-\mu_1+\mu_2\Sigma_{22}^{-1}\Sigma_{21}\right)\Sigma_{11\cdot 2}^{-1}\left(\mathtt{y}_1-\mu_1+\mu_2\Sigma_{22}^{-1}\Sigma_{21}\right)'\Gamma^{-1}\right]}.$$

Ahora bien, el jacobiano del cambio de variables es, según el lema 1.18, $|C|^n = 1$. Luego, del lema 1.17 se deduce que $f_{(X_1X_2)}(\mathtt{x}_1,\mathtt{x}_2) = f_{(Y_1,Y_2)}\big((\mathtt{x}_1,\mathtt{x}_2)C\big)$, que equivale a

$$\frac{f_{X_2}(\mathtt{x}_2)}{(2\pi)^{np_1}|\Gamma|^{p_1/2}|\Sigma_{11\cdot 2}|^{n/2}}\, \mathtt{e}^{-\frac{1}{2}\mathtt{tr}\left(\left[\mathtt{x}_1-\left(\mu_1+(\mathtt{x}_2-\mu_2)\Sigma_{22}^{-1}\Sigma_{21}\right)\right]\Sigma_{11\cdot 2}^{-1}\left[\mathtt{x}_1-\left(\mu_1+(\mathtt{x}_2-\mu_2)\Sigma_{22}^{-1}\Sigma_{21}\right)\right]'\Gamma^{-1}\right)}$$

Al dividir por $f_{X_2}(\mathtt{x}_2)$ queda la expresión buscada.

∎

Si se descompone la matriz aleatoria por filas, se obtendría una expresión similar con $\Gamma_{11\cdot 2}$. Se deja como ejercicio. $\Sigma_{11\cdot 2}$ es la matriz de las covarianzas parciales de las columnas.

El siguiente resultado da pleno sentido al estudio de la distribución normal matricial.

Teorema 1.21.

Consideremos $\mu = (\mu_1 \dots \mu_n)$ una matriz $p \times n$, Σ una matriz $p \times p$ definida positiva e Y_i, $i = 1, \dots, n$, vectores aleatorios p-dimensionales independientes, con distribución $N_p(\mu_i, \Sigma)$, resp. Entonces, $Y = (Y_1 \dots Y_n)' \sim N_{n,p}(\mu', \mathtt{Id}, \Sigma)$. Recíprocamente, si $(Y_1 \dots Y_n)' \sim N_{n,p}(\mu', \mathtt{Id}, \Sigma)$, entonces $Y_i \sim N_p(\mu_i, \Sigma)$ y son independientes, $i = 1, \dots, n$.

Demostración.
Sea $t = (t_1 \dots t_n)' \in \mathcal{M}_{n\times p}$.

$$\begin{aligned}
\varphi_Y(t) &= \mathtt{E}[\exp\{\mathbf{i}\mathtt{tr}(tY')\}] = \mathtt{E}\left[\exp\{\mathbf{i}\sum_{j=1}^{n} t_j' Y_j\}\right] \\
&= \prod_{j=1}^{n} \mathtt{E}[\exp\{\mathbf{i}t_j' Y_j\}] = \prod_{j=1}^{n} \exp\{\mathbf{i}\mu_j' t_j - \frac{1}{2} t_j' \Sigma t_j\} \\
&= \exp\left\{\sum_{j=1}^{n} (\mathbf{i}\mu_j' t_j - \frac{1}{2} t_j' \Sigma t_j)\right\} = \exp\left\{\mathbf{i}\mathtt{tr}(\mu' t) - \frac{1}{2}\mathtt{tr}(t\Sigma t')\right\} \\
&= \exp\left\{\mathbf{i}\mathtt{tr}(\mu' t) - \frac{1}{2}\mathtt{tr}(t'\mathtt{Id}\; t\Sigma)\right\} = \varphi_{N_{n,p}(\mu, Id, \Sigma)}(t).
\end{aligned}$$

Recíprocamente, si $Y \sim N(\mu', \mathtt{Id}, \Sigma)$, $Y_1, \dots, Y_n$ son independientes, pues $\Gamma = \mathtt{Id}$. Además,

$$Y_i = Y' \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \sim N_p(\mu_i, \Sigma)$$

∎

Proposición 1.22.

Sea $Y \sim N_{n,p}(\mu, \mathtt{Id}, \Sigma)$, con $\Sigma \neq 0$, $A \in \mathcal{M}_{m\times n}$ y $B \in \mathcal{M}_{r\times n}$. Entonces AY y BY son independientes sii $AB' = 0$.

Demostración.
Basta considerar

$$Z = \begin{pmatrix} A \\ B \end{pmatrix} Y \sim N_{m+r,p}\left(\begin{pmatrix} A\mu \\ B\mu \end{pmatrix}, \begin{pmatrix} AA' & AB' \\ BA' & BB' \end{pmatrix}, \Sigma \right)$$

∎

Proposición 1.23.

Sea $Y \sim N_{n,p}(\mu, \mathtt{Id}, \Sigma)$, A y B matrices $n \times n$ simétricas, semidefinidas positivas y tales que $AB = 0$. Entonces $Y'AY$ es independiente de $Y'BY$.

Demostración.
Supongamos que A y B son de rango k y r, respectivamente. Considerando el teorema de diagonalización de una matriz simétrica, podemos descomponer A y B de la forma $A = HH'$ y $B = RR'$, donde H es una matriz $n \times k$ de rango k y R es una matriz $n \times r$ de rango r. Se tiene entonces que $0 = AB$ Por lo tanto, $0 = H'ABR = H'HH'RR'R$. Siendo $H'H$ y $R'R$ matrices invertibles, se tiene que $0 = H'R$. Luego, según la anterior proposición, $H'Y$ y $R'Y$ son independientes. Ahora bien, $Y'AY = (H'Y)'(H'Y)$ y $Y'BY = (R'Y)'(R'Y)$. Al tratarse de funciones medibles, se tiene que $Y'AY$ y $Y'BY$ son, igualmente, independientes.

∎

Proposición 1.24.

Si A y B son matrices $n \times n$, siendo $B \geq 0$ y tales que $AB = 0$, entonces AY y $Y'BY$ son independientes.

Demostración.
Sea k el rango de B y consideremos una matriz R $n \times k$ de rango k tal que $B = RR'$. Entonces $0 = AB$. Luego, $0 = ABR = ARR'R$ Por lo tanto, $0 = AR$. Teniendo en cuenta la proposición 1.22, AY y $R'Y$ son independientes. Como $Y'BY = (R'Y)'R'Y$, acabamos.

∎

El siguiente resultado es consecuencia directa de las dos proposiciones anteriores.

Corolario 1.25.

En particular, si $V_1 \perp V_2$, $Y'P_{V_1}Y$ y $Y'P_{V_2}Y$ son independientes. También son independientes $P_{V_1}Y$ y $Y'P_{V_2}Y$.

Para acabar esta sección vamos a probar un resultado que de gran utilidad en lo sucesivo.

Teorema 1.26.
Si X es una matriz aleatoria $n \times p$ definida sobre un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ que sigue un modelo de distribución $N_{n,p}(\mu, \Gamma, \Sigma)$ [9], con $\Gamma, \Sigma > 0$ y $n \geq p$, entonces $P\big(\mathtt{rg}(X) = p\big) = 1$.

Demostración.
Lo probaremos por inducción sobre p: primeramente, si $p = 1$ y $\Sigma = \sigma^2$, tendremos que X sigue un modelo de distribución $N_n(\mu, \sigma^2\Gamma)$, que está dominada por m^n, la medida de Lebesgue en $\mathbb{R}^n$. Luego $P\left(\mathtt{rg}(X) = 0\right) = P(X = 0) = 0$, pues $m^n(\{0\}) = 0$.

Supongamos que la tesis se verifica para $p < n$ y veamos que es también cierta para $p + 1$: si descomponemos $X = (X_1 \dots X_p X_{p+1}) \sim N_{n,p+1}(\mu, \Gamma, \Sigma)$, se tiene que X_{p+1} condicionada a $(X_1 \dots X_p)$ sigue un modelo de distribución $N_{n,1} = N_n$, con matriz de covarianzas positiva y, por tanto, dominado por m^n. Entonces, se verifica

$$\begin{aligned} P\big(\mathtt{rg}(X_1 \dots X_{p+1}) \leq p\big) &= \int_{\mathbb{R}^{n\times p}} P\big(\mathtt{rg}(X_1 \dots X_{p+1}) \leq p | X_1, \dots, X_p\big) dP^{(X_1 \dots X_p)} \\ &= \int_{\mathbb{R}^{n\times p}} E\left[I_{\{0,\dots,p\}} \circ \mathtt{rg}(X_1 \dots X_{p+1}) | X_1, \dots, X_p\right] dP^{(X_1 \dots X_p)} \end{aligned}$$

Además, en virtud de (1.6), la última expresión equivale a la siguiente

$$\int_{\mathbb{R}^{n\times p}} \left(\int_{\mathbb{R}^n} I_{\{0,\dots,p\}} \circ (\mathtt{rg}(X_1 \dots X_p, X_{p+1}))\, dP^{X_{p+1}|X_1,\dots,X_p} \right) dP^{(X_1 \dots X_p)}$$

Téngase en cuenta que, por hipótesis de inducción, la probabilidad de que $X_1, \dots, X_p$ sean linealmente dependientes es 0. Luego, se verifica que, fuera de un conjunto de medida $P^{(X_1 \dots X_p)}$-nula, el rango de $(X_1 \dots X_p X_{p+1})$ es menor que $p + 1$ sii X_{p+1} pertenece al subespacio $\langle X_1, \dots, X_p \rangle$, de dimensión $p < n$. Como $m^n(H) = 0$, para cualquier hiperplano H de $\mathbb{R}^n$, se concluye que $P(\mathtt{rg}(X_1 \dots X_{p+1}) \leq p) = 0$. ∎

Se puede demostrar de manera totalmente análoga este otro resultado, que queda como ejercicio.

Teorema 1.27.
Si $a \in \mathbb{R}^n \setminus \{0\}$ y $p < n$, entonces $P(\mathtt{rg}(a|X) = p + 1) = 1$.

[9] Realmente, puede demostrarse (cuestión propuesta) que para que la tesis se verifique basta con que la distribución esté dominada por la medida de Lebesgue en $\mathbb{R}^{np}$.

El principal objetivo de la distribución normal matricial es, como se deduce del teorema 1.21, caracterizar una muestra aleatoria simple correspondiente a una distribución normal multivariante. El supuesto de normalidad multivariante será el punto de partida de la gran mayoría de métodos de inferencia a estudiar. De ahí la necesidad de disponer de procedimientos, tanto para contrastar la veracidad de dicho supuesto, como se vio en la sección anterior, como para obtener un ajuste aproximad al modelo normal mediante una transformación adecuada de los datos.

En ese sentido, son bien conocidas las transformaciones de Box-Cox que, aunque se plantean en principio en un contexto univariante, según se ve en el capítulo 4 del volumen 1, pueden extenderse sin problemas al caso multivariante. Recordamos que estas transformaciones se define, en dimensión uno, mediante

$$\phi(\lambda, x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln x & \text{si } \lambda = 0 \end{cases}$$

Si Y al vector n-dimensional compuesto por la muestra estudiada, el método consiste en suponer que existe un valor λ de tal forma que el vector de transformaciones $\Phi(\lambda, Y)$ sigue una distribución de la forma $N_{\mathtt{n}}(\mu \cdot 1_n, \sigma^2 \mathtt{Id})$, para algún $\mu \in \mathbb{R}$ y $\sigma^2 > 0$. El valor adecuado de λ se estima por el método de máxima verosimilitud, es decir, se escogerán los parámetros (es decir, la distribución) λ, μ y σ^2 que hagan más verosímil la observación Y. Tras los cálculos pertinentes, el problema queda reducido a encontrar el valor λ que minimice la varianza muestral de los datos transformados.

La extensión al caso multivariante es obvia: dados dos vectores p-dimensionales $x = (x_1, \ldots, x_p)'$ y $\Lambda = (\lambda_1, \ldots, \lambda_p)'$, consideramos la transformación

$$\phi_M(\Lambda, x) = \big(\phi(\lambda_1, x_1), \ldots, \phi(\lambda_p, x_p)\big)'.$$

Si esta transformación conduce a la p-normalidad de cierta distribución p-dimensional y $X = (x_1, \ldots, x_n)'$ denota una muestra aleatoria simple de tamaño n de dicha distribución, entonces la matriz aleatoria $n \times p$ definida mediante

$$\Phi_M(\Lambda, X) = \big(\phi_M(\lambda_1, x_1), \ldots, \phi_M(\lambda_p, x_p)\big)'$$

debe seguir, en virtud del teorema 1.21, una modelo de distribución $N_{n,p}(\mu, \mathtt{Id}, \Sigma)$, para alguna matriz μ cuyas columnas pertenezcan todas al subespacio $\langle 1_n \rangle$ y alguna matriz $\Sigma > 0$. La función de verosimilitud será entonces de la forma

$$f_\Lambda(\mathcal{X}) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}\left[(\mathcal{X} - \mu)\Sigma^{-1}(\mathcal{X} - \nu)' \right] \right\}.$$

Puede comprobarse que el valor de Λ que maximiza la función anterior es el que minimiza el logaritmo neperiano de la matriz de varianzas-covarianzas muestral de $\Phi_M(\Lambda, X)$.

1.3. Distribuciones de Wishart y Hotelling

Se trata de dos distribuciones asociadas a la normal matricial, que viene a generalizar los modelos χ^2 y t de Student al cuadrado. Consideremos en $(\mathbb{R}^{np}, \mathcal{R}^{np})$ una distribución de probabilidad $N_{n,p}(\mu, \mathtt{Id}, \Sigma)$ y la transformación W que a cada $Y \in \mathcal{M}_{n\times p}$ le asigna la matriz $Y'Y \in \mathcal{M}_{p\times p}$, simétrica y semidefinida positiva. Esta transformación induce una distribución de probabilidad en dicho conjunto, que se identifica con un abierto de $\mathbb{R}^{\frac{p(p+1)}{2}}$. Veamos que depende de μ a través, de $\mu'\mu$, es decir, que si μ_1 y μ_2 son matrices $n \times p$ tales que $\mu_1'\mu_1 = \mu_2'\mu_2$, entonces

$$(N_{n,p}(\mu_1, \mathtt{Id}, \Sigma))^W = (N_{n,p}(\mu_2, \mathtt{Id}, \Sigma))^W .$$

En efecto, dado el experimento estadístico, $Y \sim N_{n,p}(\mu_1, \mathtt{Id}, \Sigma)$, donde μ es cualquier matriz $n \times p$ y Σ cualquier matriz $p \times p$ definida positiva, se verifica que el grupo de transformaciones bimedibles $G = \{g_\Gamma : \Gamma \in \mathcal{O}_{n\times n}\}$ definidas mediante $g_\Gamma(Y) = \Gamma Y$, lo deja invariante. En virtud del teorema 13.9, el estadístico $Y \mapsto Y'Y$ y la aplicación $(\mu, \Sigma) \mapsto (\mu'\mu, \Sigma)$ constituyen invariantes maximales para el espacio de observacones y el de parámetros, respectivamente. Se tiene, por otro lado, que si $n \geq p$ y δ es una matriz $p \times p$ simétrica semidefinida positiva, existe una matriz μ de dimensiones $n \times p$ tal que $\delta = \mu'\mu$ (considerar la diagonalización de δ). Estamos pues en condiciones de definir la distribución de Wishart.

Definición.
Si δ y Σ son matrices $p \times p$ semidefinidas positivas, la distribución $(N_{n,p}(\mu, \mathtt{Id}, \Sigma))^W$ se denomina distribución de Wishart de parámetros p, n, Σ, δ, donde $\delta = \mu'\mu$. Se denota $W \sim W_p(n, \Sigma, \delta)$. Si $\delta = 0$, se denota $W_p(n, \Sigma)$, y si, además, $\Sigma = \mathtt{Id}$, se denota $W_p(n)$.

Veamos algunas propiedades inmediatas:

(a) $\mathtt{E}[W] = n\Sigma + \delta$.

Demostración: Consideremos $\mu = (\mu_1 \dots \mu_n)' \in \mathcal{M}_{n\times p}$ tal que $\delta = \mu'\mu$, $Y_i \sim N_p(\mu_i, \Sigma)$, $i = 1, \dots, n$ independientes e $Y = (Y_1 \dots Y_n)' \sim N(\mu, \mathtt{Id}, \Sigma)$. Entonces $W = Y'Y = \sum_{i=1}^n Y_i Y_i' \sim W_p(n, \Sigma, \delta)$. por tanto,

$$\begin{aligned} \mathtt{E}[W] &= \sum_i \mathtt{E}[Y_i Y_i'] \\ &= \sum_{i=1}^n (\Sigma + \mu_i \mu_i') \\ &= n\Sigma + \mu'\mu = n\Sigma + \delta. \end{aligned}$$

∎

(b) **Caso univariante:** Si $p = 1$, con $\Sigma = \sigma^2$, entonces

$$W_1(n, \sigma^2, \delta) = \sigma^2 \chi^2_n \left(\frac{\delta}{\sigma^2} \right) \tag{1.24}$$

(c) Si $a \in \mathbb{R}^+$, $aW \sim W_p(n, a\Sigma, a\delta)$.

(d) Si $A \in \mathcal{M}_{k\times p}$, $AWA' \sim W_k(n, A\Sigma A', A\delta A')$.

Demostración: Basta considerar $YA' \sim N_{n\times k}(\mu A', \mathtt{Id}, A\Sigma A')$. ∎

(e) Si $p = p_1 + p_2$ y consideramos las correspondientes descomposiciones

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \delta = \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix},$$

Entonces $W_{ii} \sim W_{p_i}(n, \Sigma_{ii}, \delta_{ii})$, $i = 1, 2$.

Demostración: Basta descomponer $Y = (Y_1 Y_2)$ y considerar entonces la correspondiente descomposición de $Y'Y$. ∎

(f) Si $W \sim W_p(n, \Sigma, \delta)$, con $n \geq p$ y $\Sigma > 0$, entonces $P(W > 0) = 1$, es decir, con probabilidad 1 tiene sentido hablar de W^{-1}.

Demostración: Es corolario directo del teorema 1.26. ∎

Teorema 1.28.
Sean $Y \sim N_{n,p}(\mu, \mathtt{Id}, \Sigma)$ y $V \subset \mathbb{R}^n$ de dimensión k. Entonces

$$Y'P_V Y \sim W_p(k, \Sigma, \mu' P_V \mu).$$

Demostración.
Si $X \in \mathcal{M}_{n\times k}$ una base ortonormal de V, $X'Y \sim N_{k,p}(X'\mu, \mathtt{Id}, \Sigma)$. Luego, $Y'XX'Y = Y'P_V Y \sim W_p(k, \Sigma, \mu' P_V \mu)$. ∎

Antes del siguiente resultado vamos a obtener una serie de lemas previos. El primero de ellos trata los conceptos de esperanza condicional e independencia condicional.

Lema 1.29. (a) Si $E[Y|X] = h \circ g$, entonces $\mathtt{E}[Y|g \circ X] = h$.

(b) Si $P^{(Y,Z)|X} = P^{Y|X} \times P^{Z|X}$ entonces $P^{Z|(X,Y)} = P^{Z|X}$.

(c) Supongamos que $P^X << m^k$ y $P^{Y|X=\mathbf{x}} << m^p$, para todo $x \in \mathbb{R}^k$, donde m denota la medida de Lebesgue en $\mathbb{R}$. Entonces, $P^{(X,Y)} << m^{k+p}$. Se verificará, por tanto, que

$$\frac{dP^{(X,Y)}}{dm^{k+p}}(x,y) = \frac{dP^X}{dm^k}(x) \times \frac{dP^{Y|X=\mathbf{x}}}{dm^p}(y)$$

Demostración.
La demostración de (a) es obvia. Veamos la de (b):

$$\begin{aligned}
\int_{A_X \times A_Y} P^{Z|X=\mathbf{x}}(A_Z)\, dP^{(X,Y)} &= \int_{A_X} \left(\int_{A_Y} P^{Z|X=\mathbf{x}}(A_Z)\, dP^{Y|X=\mathbf{x}} \right) dP^X \\
&= \int_{A_X} \left(P^{Z|X=\mathbf{x}}(A_Z) \times P^{Y|X=\mathbf{x}}(A_Y) \right) dP^X \\
&= \int_{A_X} P^{(Y,Z)|X=\mathbf{x}}(A_Y \times A_Z)\, dP^X \\
&= P\left(X \in A_X,\ Y \in A_Y,\ Z \in A_Z\right).
\end{aligned}$$

Se concluye por el teorema de Dynkin.
Probemos por último (c): Dado $A \in \mathcal{R}^{k+p}$ y $x \in \mathbb{R}^k$, definamos $A_x = \{y \in \mathbb{R}^p : (x,y) \in A\} \in \mathcal{R}^p$. Si $m^{k+p}(A) = 0$, entonces $\int_{\mathbb{R}^k} m^p(A_x)\, dm^k = 0$. Luego, existe $N \in \mathcal{R}^k$ tal que $m^k(N) = 0$ (y, por lo tanto, $P^X(N) = 0$) y $m^p(A_x) = 0$ (y, en consecuencia, $P^{Y|X=\mathbf{x}}(A_x) = 0$) si $x \notin N$. Entonces, se verifica que $P^{(X,Y)}(A) = \int_{\mathbb{R}^k} P^{Y|X=\mathbf{x}}(A_x)\, dP^X = 0$.

∎

La propiedad (b) se denomina independencia condicional entre Y y Z dada X. Los siguientes resultados preparan el terreno para definir la distribución T^2 de Hotelling.

Lema 1.30.

Consideremos $W \sim W_{q+p}(n, \Sigma)$, $\Sigma > 0$, $n \geq q+p$. Consideremos, asimismo, las siguientes descomposiciones:

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

donde W_{11} y Σ_{11} son matrices $q \times q$, mientras que W_{22} y Σ_{22} son matrices $p \times p$. Definamos

$$\begin{aligned}
T &= W_{22}, \\
U &= W_{22}^{-1} W_{21}, \\
V &= W_{11} - W_{12} W_{22}^{-1} W_{21}, \\
\Sigma_{11 \cdot 2} &= \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}.
\end{aligned}$$

Se verifica entonces

$$T \sim W_q(n, \Sigma_{22}), \quad V \sim W_q(n-p, \Sigma_{11\cdot 2}), \quad U|T \sim N_{p,q}\left(\Sigma_{22}^{-1}\Sigma_{21}, T^{-1}, \Sigma_{11\cdot 2}\right).$$

Además, V es independiente de (U, T).

Demostración.
Consideremos $(X_1, X_2) \sim N_{n,q+p}(0, \mathtt{Id}, \Sigma)$. En ese caso, se tiene que $W_{ij} = X_i' X_j$, donde $i, j \in \{1, 2\}$. La distribución de $T = X_2' X_2$ se obtiene pues de forma trivial. Además, U se obtiene mediante $(X_2'X_2)^{-1}X_2'X_1$, mientras que $V = X_1'(\mathtt{Id} - X_2(X_2'X_2)^{-1}X_2')X_1$. En virtud del teorema 1.20, sabemos que X_1 condicionada a X_2 sigue un modelo de distribución $N_{n,q}\left(X_2\Sigma_{22}^{-1}\Sigma_{21}, \mathtt{Id}, \Sigma_{11\cdot 2}\right)$. Luego, la distribución de U condicionada a X_2 es $N_{p,q}\left(\Sigma_{22}^{-1}\Sigma_{21}, (X_2'X_2)^{-1}, \Sigma_{11\cdot 2}\right)$. Por el lema 1.29(a), se tiene que U condicionada a T sigue un modelo de distribución $N_{p,q}\left(\Sigma_{22}^{-1}\Sigma_{21}, T^{-1}, \Sigma_{11\cdot 2}\right)$. Teniendo en cuenta que la matriz $\mathtt{Id} - X_2(X_2'X_2)^{-1}X_2$ es idempotente de rango $n-p$ y que, además,

$$\Sigma_{12}\Sigma_{22}^{-1}X_2'[\mathtt{Id} - X_2(X_2'X_2)^{-1}X_2]X_2\Sigma_{22}^{-1}\Sigma_{21} = 0,$$

se deduce de (1.7) que $V|X_2 \sim W_q(n-p, \Sigma_{11\cdot 2})$ y, en consecuencia V es independiente de X_2 (y por lo tanto de T), siguiendo su distribución marginal un modelo $W_q(n-p, \Sigma_{11\cdot 2})$. Por otra parte, teniendo en cuenta que

$$[(X_2'X_2)^{-1}X_2'][\mathtt{Id} - X_2(X_2'X_2)^{-1}X_2] = 0,$$

se sigue de la proposición 1.24 y el lema 1.29(a), que

$$\begin{aligned} P^{(U,V)|X_2=\mathtt{x}_2} &= P^{U|X_2=\mathtt{x}_2} \times P^{V|X_2=\mathtt{x}_2} \\ &= P^{U|T=T(\mathtt{x}_2)} \times P^{V|T=T(\mathtt{x}_2)}, \end{aligned}$$

es decir, $P^{(U,V)|T} = P^{U|T} \times P^{V|T}$. Entonces, por el lema 1.29(b), se verifica que $P^{V|(T,U)} = P^{V|T}$, que, siendo V y T independientes, equivale a la distribución marginal P^V, que se corresponde con el modelo $W_q(n-p, \Sigma_{11\cdot 2})$. Al no depender de (T, U), podemos afirmar que V es independiente de (T, U).

∎

Lema 1.31.
Consideremos $W \sim W_p(n, \Sigma)$, con $n \geq p$ y $\Sigma > 0$. Sea $y \in \mathbb{R}^p \backslash \{0\}$. Entonces

$$\frac{y'\Sigma^{-1}y}{y'W^{-1}y} \sim \chi^2_{n-p+1}.$$

Demostración.
La dividiremos en dos partes:

(a) Consideremos $y_0 = (1, 0, \dots 0)' \in \mathbb{R}^p$ y descompongamos W y Σ de la forma

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

donde W_{11} y Σ_{11} son matrices 1×1 (números). Es obvio que $y_0 W^{-1} y_0$ y $y_0' \Sigma^{-1} y_0$ son las componentes $(1,1)$-ésimas de las matrices W^{-1} y Σ^{-1}, respectivamente. Luego, por el lema 13.6, se tiene que

$$y_0' W^{-1} y_0 = \left(W_{11} - W_{12} W_{22}^{-1} W_{21}\right)^{-1}, \quad y_0' \Sigma^{-1} y_0 = \left(\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}\right)^{-1}.$$

Si se denota

$$V = W_{11} - W_{12} W_{22}^{-1} W_{21}, \quad \sigma^2_{11\cdot 2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21},$$

entonces se tiene, por el lema 1.30, que

$$V \sim W_1(n - p + 1, \sigma^2_{11\cdot 2}) = \sigma^2_{11\cdot 2} \chi^2_{n-p+1},$$

es decir, que

$$\frac{y_0' \Sigma^{-1} y_0}{y_0' W^{-1} y_0} = \frac{V}{\sigma^2_{11\cdot 2}} \sim \chi^2_{n-p+1}.$$

(b) Si $y \in \mathbb{R}^p \backslash \{0\}$ cualquiera y A es una matriz $p \times p$ invertible cuya primera columna es y, entonces $y = A y_0$. Sabemos, por las propiedades generales de la distribución de Wishart, que $A^{-1} W (A')^{-1}$ sigue un modelo de distribución $W_p\left(n, A^{-1} \Sigma (A^{-1})'\right)$. Aplicando la parte (a) a dicha distribución se tiene:

$$\frac{y' \Sigma^{-1} y}{y' W^{-1} y} = \frac{y_0' A' \Sigma^{-1} A y_0}{y_0' A' W^{-1} A y_0} = \frac{y_0' \left(A^{-1} \Sigma (A^{-1})'\right)^{-1} y_0}{y_0' \left(A^{-1} W (A^{-1})'\right)^{-1} y_0} \sim \chi^2_{n-p+1}.$$

∎

El resultado que perseguimos es el siguiente:

Teorema 1.32.

Consideremos un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ y dos variables independientes definidas sobre éste, $Y \sim N_p(\mu, \Sigma)$ y $W \sim W_p(n, \Sigma)$, donde $n \geq p$ y $\Sigma > 0$. Sea la variable

$$F = \frac{n - p + 1}{p} Y' W^{-1} Y.$$

Entonces $F \sim F_{p, n-p+1}\left(\mu' \Sigma^{-1} \mu\right)$.

Demostración.
Consideremos $U = Y'\Sigma^{-1}Y$ y $R = \frac{Y'\Sigma^{-1}Y}{Y'W^{-1}Y}$. Entonces

$$U = \left(\Sigma^{-1/2}Y\right)'\Sigma^{-1/2}Y = \|\Sigma^{-1/2}Y\|^2,$$

donde $\Sigma^{-1/2}Y \sim N_p(\Sigma^{-1/2}\mu, \mathtt{Id})$. Por lo tanto,

$$U \sim \chi^2_p\left(\mu'\Sigma^{-1}\mu\right).$$

Consideremos, para cada $\mathtt{y} \in \mathbb{R}^p$, la función $R(\mathtt{y}, \cdot)$, que asigna a cada matriz W de dimensiones $p \times p$ y definida positiva el número real

$$\frac{\mathtt{y}'\Sigma^{-1}\mathtt{y}}{\mathtt{y}'W^{-1}\mathtt{y}}.$$

Se tiene entonces, en virtud de (1.7), que, para cada $\mathtt{y} \in \mathbb{R}^p$,

$$P^{R|Y=\mathtt{y}} = \left(P^{W|Y=\mathtt{y}}\right)^{R(\mathtt{y},\cdot)}$$

Al ser Y y W independientes, se verifica que $P^{W|Y} = P^W = W_p(n, \Sigma)$. Luego, aplicando el lema 1.31, se deduce que $R|Y = \mathtt{y} \sim \chi^2_{n-p+1}$. Al no depender de y, se deduce que Y y R son independientes y $R \sim \chi^2_{n-p+1}$. Como U es función medible de Y, R y U son independientes. Por lo tanto

$$\frac{U/p}{R/(n-p+1)} \sim F_{p,n-p+1}\left(\mu'\Sigma^{-1}\mu\right).$$

Ahora bien,

$$\frac{U/p}{R/(n-p+1)} = \frac{n-p+1}{p} YW^{-1}Y = F. \qquad \blacksquare$$

Definición.

Consideremos $Y \sim N_p(\mu, \Sigma)$ y $W \sim W_p(n, \Sigma)$ independientes, con $n \geq p$, y sea $\delta = \mu'\Sigma^{-1}\mu \in \mathbb{R}^+$. En esas condiciones, se dice entonces que el estadístico

$$\mathtt{n}Y'W^{-1}Y \tag{1.25}$$

sigue un modelo de distribución T^2-Hotelling de parámetros p, n, δ, denotándose

$$nY'W^{-1}Y \sim T^2_{p,n}(\delta)$$

En el caso $\delta = 0$ se denota $T^2_{p,n}$.

Es fácil comprobar que, dado $\delta \in \mathbb{R}^+$, existen $\mu \in \mathbb{R}^p$ y $\Sigma > 0$ $p \times p$ tales que $\delta = \mu'\Sigma^{-1}\mu$. Por otro lado, el teorema anterior da sentido a esta definición ya que garantiza que la distribución de $Y'W^{-1}Y$ depende de μ y Σ únicamente a través de $\delta = \mu'\Sigma^{-1}\mu$. Además, y en virtud del mismo resultado, deducimos que la distribución T^2 no es esencialmente nueva, sino que se trata (salvo una constante) de un modelo F de Snedecor. Es más, respecto a los cuantiles de la misma, se verifica

$$T^{2,\alpha}_{p,n} = \frac{np}{n-p+1} F^{\alpha}_{p,n-p+1}.$$

Como veremos en los capítulos 2 y 4, la distribución T^2 de Hotelling está ligada al contraste de la media en el modelo Lineal Normal Multivariante, cuando la hipótesis inicial es un hiperplano del espacio de parámetros, por ejemplo, en el contraste de una media, en la comparación de las medias de dos distribuciones o en los contrastes parciales de los coeficientes de regresión. En el caso univariante, todos estos contrastes se resuelven mediante la distribución t de Student. Teniendo en cuenta este comentario junto con el teorema 1.32[10], podemos entender, si se nos permite el abuso, la distribución T^2 de Hotelling como una generalización del cuadrado de la t de Student al caso multivariante.

Cuestiones propuestas

1. Demostrar el teorema 1.27.

2. En el lema 1.30, podemos obtener W_{22}, W_{21} y W_{11} a partir de T, U y V. Además, la matriz de derivadas parciales de dicha transformación verifica:

$$\begin{array}{cccc} & T & U & V \\ W_{22} & \mathtt{Id} & & \\ W_{21} & 0 & W_{22}^{-1} & \\ W_{11} & 0 & 0 & \mathtt{Id} \end{array}$$

   El jacobiano de dicha transformación es $|W_{22}|^{-1}$.

3. ¿Puede estar una distribución W_p dominada por la medida de Lebesgue en $\mathbb{R}^{p^2}$?

4. Considérese $W \sim W_p(n, \mathtt{Id})$, $n \geq p$. Pruébese que admite la siguiente densidad respecto a $m^{\frac{p(p+1)}{2}}$

$$f(w) = \frac{|w|^{(n-p-1)/2} \exp\{(-1/2)tr(w)\}}{2^{np/2}\pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)}, \quad w > 0.$$

[10] Téngase en cuenta que el cuadrado de la distribución t de Student con m grados de libertad es la distribución $F_{1,m}$.

Nota: Consideremos la aplicación siguiente

$$\phi : (w_{11}, \ldots, w_{1p}, w_{22}, \ldots, w_{2p}, \ldots, w_{pp})' \in \mathbb{R}^{\frac{p(p+1)}{2}} \longrightarrow \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1p} \\ w_{12} & w_{22} & \ldots & w_{2p} \\ \vdots & \vdots & & \vdots \\ w_{1p} & w_{2p} & \ldots & w_{pp} \end{pmatrix}$$

Lo que se afirma realmente es que

$$\frac{dP^{\phi^{-1}(W)}}{dm^{\frac{p(p+1)}{2}}}(w^*) = f\left(\phi(w^*)\right), \text{ donde } w^* \in \mathbb{R}^{\frac{p(p+1)}{2}}$$

Indicación: Si razonamos por inducción, el caso $p = 1$ es trivial, teniendo en cuenta que la densidad de la distribución χ^2_n es la siguiente:

$$\left[\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}\right]^{-1} x^{\frac{n}{2}-1} \exp\left\{-\frac{x}{2}\right\} I_{]0,+\infty]}(x).$$

Para demostrar el caso p+1, considerense T, U, V como en el lema 1.30, con q=1. Téngase en cuenta que, según el lema 1.29(c), $(T, U, V) << m^{\frac{p(p+1)}{2}}$. Obténgase la función de densidad $f_{T,U,V}(t, u, v) = f_T(t) \times f_{U|T=t}(u) \times f_{V|(T,U)=(t,u)}(v)$, donde f_T es ya conocida por hipótesis de inducción. Considérese la transformación propuesta en la cuestión 2 y aplíquese el teorema del cambio de variables.

5. Considérese $W \sim W_p(n, \Sigma)$, con $n \geq p$ y $\Sigma > 0$. Entonces admite la siguiente densidad:

$$f(w) = \frac{|w|^{(n-p-1)/2} \exp\left\{-\frac{1}{2}\mathtt{tr}\left(\Sigma^{-1}w\right)\right\}}{2^{np/2}\pi^{p(p-1)/4}|\Sigma|^{n/2}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)}, \quad w > 0.$$

Indicación: Téngase en cuenta que, según el teorema 13.8, existe una matriz $p \times p$ triangular superior C, tal que $\Sigma = CC'$. Se verifica que, si $W^* \sim W_p(n, \mathtt{Id})$, entonces $W = CW^*C' \sim W_p(n, \Sigma)$. El resultado se sigue pues del problema 4 teniendo en cuenta el jacobiano de esta transformación inversa es $|C|^{-(p+1)}$.

6. Demostrar que si $W_1, \ldots, W_k$ son independientes y distribuidas según un modelo $W_p(n_i, \Sigma)$, $i = 1, ..., k$, entonces $\sum_{i=1}^{k} W_i \sim W_p(\sum_{i=1}^{k} n_i, \Sigma)$.

7. Demostrar que si $W \sim W_p(n, \Sigma)$ y $w \in M_{p\times p}$ simétrica, entonces su función característica es la siguiente

$$\varphi_W(w) = |\mathtt{Id} - 2\mathbf{i}\Sigma w|^{-n/2}.$$

Indicación: Considérese $X_1, \dots X_n$ iid $N_p(0, \mathtt{Id})$. Entonces, si $X = (X_1 ... X_n)'$, $X \sim N_{n,p}(0, \mathtt{Id}, \mathtt{Id})$ y $X'X \sim W_p(n)$. Se verifica:

$$\begin{aligned}\varphi_W(w) &= E\left(\exp\{\mathbf{i}tr(X'X)'w\}\right) = \prod_{j=1}^{n} E\left(\exp\{\mathbf{i}X_j'wX_j\}\right) \\ &= \prod_{j=1}^{n} E\left(\exp\{\mathbf{i}Y_j'DY_j\}\right) = \prod_{j=1}^{n}\prod_{k=1}^{p} E\left(\exp\{\mathbf{i}Y_{jk}^2 d_k\}\right) \\ &= \prod_{j=1}^{n}\prod_{k=1}^{p}(1-2\mathbf{i}d_k)^{-1/2} = |\mathtt{Id} - 2\mathbf{i}D|^{-n/2} = |\mathtt{Id} - 2\mathbf{i}\mathtt{Id}w|^{-n/2},\end{aligned}$$

donde $w = \Gamma' D \Gamma$, con Γ ortogonal, $D = \mathsf{diag}(d_1, ..., dk)$, e $Y_j = \Gamma X_j$, $j = 1, ..., n$. Considerando la transformación $R = X\Sigma^{1/2}$ se obtiene el resultado general.

8. Obtener la función generatriz de la distribución de Wishart.

   Indicación: Teniendo en cuenta que $|\Sigma^{-1} - 2w| = |\Sigma|^{-1}|\mathtt{Id} - 2\Sigma w|$ y siguiendo el esquema anterior, concluir que, si w es una matriz simétrica tal que $\Sigma^{-1} - 2w > 0$, entonces $g_W(w) = |\mathtt{Id} - 2\Sigma w|^{-n/2}$.

9. Demostrar que para obtener la tesis del teorema 1.26 es suficiente que la distribución de $\mathtt{vec}(X)$ esté dominada por la medida de Lebesgue en $\mathbb{R}^{np}$, donde $n \geq p$.

10. Obtener la distribución condicional $X_1|X_2 = x_2$ de manera análoga al teorema 1.20, pero estando la matriz aleatoria descompuesta por filas.

11. Consideremos una matriz aleatoria X $m \times r$ y una matriz T $m \times r$. Entonces, $\varphi_{\mathtt{tr}(T'X)}(t) = \varphi_X(tT)$, para todo $t \in \mathbb{R}$.

12. Demostrar que, si $X \sim N_{m,r}(\mu, \Gamma, \Sigma)$ y $T \in M_{m\times r}$, entonces

$$\mathtt{tr}(T'X) \sim N\big(\mathtt{tr}(T'\mu), tr(T'\Gamma T\Sigma)\big).$$

13. Demostrar que, si $Z = (Z_{ij})$ matriz aleatoria con valores en $M_{m\times r}$ y $\mu \in M_{m\times r}$, donde $Z_{ij} \sim N(\mu_{ij}, 1)$ independientes $\forall i, j$, entonces $\mathtt{tr}(Z'Z) \sim \chi^2_{mr}\big(\mathtt{tr}(\mu'\mu)\big)$.

14. Demostrar que, em las condiciones anteriores,

$$\mathtt{tr}\big((X-\mu)'\Gamma^{-1}(X-\mu)\Sigma^{-1}\big) \sim \chi^2_{mr}.$$

15. Demostrar que, en las condiciones anteriores,

$$\mathtt{tr}\big(X'\Gamma^{-1}X\Sigma^{-1}\big) \sim \chi^2_{mr}\big(\mathtt{tr}(\mu'\Gamma^{-1}\mu\Sigma^{-1})\big).$$

16. Demostrar que, si $W \sim W_p(n,\Sigma,\delta)$, entonces $\mathtt{tr}(\Sigma^{-1}W) \sim \chi^2_{np}\big(\mathtt{tr}(\Sigma^{-1}\delta)\big)$.

17. Demostrar que, si $W \sim W_p(n,\Sigma)$, entonces $|\Sigma^{-1}W| \sim \prod_{i=1}^p U_i$, donde U_i sigue un modelo de distribución χ^2_{n-i+1}, $i = 1,...,p$, siendo todos independientes.

    Indicación: Razonar por inducción sobre p. Para ello, descomponer W de la forma

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}.$$

    Aplicar entonces el lema 1.30.

18. Demostrar que $E[|W|] = n(n-1)...(n-p+1)|\Sigma|$.

    Nota 1(varianza generalizada): Si una distribución multivariante posee matriz de covarianza Σ, se define su varianza generalizada como $|\Sigma|$. Debe entenderse, en cierta forma, como el cuadrado del volumen engendrado por las distintas componentes del vector aleatorio. Será de utilidad en distintos aspectos del análisis multivariante, por ejemplo, en el contraste de independencia de las componentes del vector. Según el último resultado, bajo las condiciones del modelo lineal normal multivariante, el estimador insesgado de mínima varianza de $|\Sigma|$ es $\frac{(n-1)|S|}{(n-1)...(n-p)}$. También puede probarse que $|\Sigma| = \prod_{i=1}^p \sigma_i^2|P|$, donde P denota la matriz de correlaciones. Para un estudio más detallado, consultar Anderson(1958) y Johnson (1992).

    Nota 2 (varianza total): dado un vector aleatorio X con matriz de covarianzas Σ, se define la varianza total como

$$\mathtt{var}_T[X] = \mathtt{tr}(\Sigma) \tag{1.26}$$

    . Se hace referencia a este concepto en el capítulo dedicado al análisis de componentes principales. Los dos ejercicios siguientes proporcionan un resultado de interés para dicho capítulo.

19. Sea X un vector aleatorio p-dimensional tal que $\mathtt{E}[X] = \mu$ y $\mathtt{Cov}[X] = \Sigma$. Se verifica entonces

$$\min_{c\in\mathbb{R}^p} \mathtt{E}\left[\|X-c\|^2\right] = \mathtt{var}_T[X],$$

    alcanzándose cuando $c = \mu$.

20. Dado un vector aleatorio

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right),$$

sean

$$\beta = \Sigma_{12}\Sigma_{22}^{-1}, \quad \Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Se verifica entonces

$$\min_{b\in\mathcal{M}_{p\times q}, d\in\mathbb{R}^p} \mathtt{E}\left[\|X_1 - (bX_2 + d)\|^2\right] = \mathtt{tr}\ \Sigma_{11\cdot 2}, \tag{1.27}$$

y se alcanza cuando $b = \beta$, y $d = 0$.

Indicación: Considérese la proposición 1.5 junto con el ejercicio anterior.

21. Probar que la igualdad (1.27) es válida en general (sin suponer normalidad).

Indicación: Ver el Apéndice (sección dedicada a Generalidades sobre Probabilidad) del volumen 1.

Capítulo 2

Modelo lineal normal multivariante

En este capítulo se estudia un modelo estadístico que servirá de marco general para problemas tan importantes como la inferencias acerca de una o varias medias o la regresión lineal multivariantes. Guarda también una estrecha relación con el análisis de correlación canónica y el análisis discriminante. El modelo lineal normal multivariante no es sino una generalización del modelo lineal normal. Podemos encontrar un estudio detallado de este último en el capítulo 3 del volumen 1. Reproduciremos casi por completo la estructura de dicho capítulo, estudiando los problemas de estimación puntual y contraste de hipótesis para la media, para acabar con un estudio asintótico del modelo. Los contrastes relativos a la matriz de varianzas-covarianzas se considerarán en capítulos sucesivos, especialmente en el tercero.

No obstante, en el modelo lineal multivariante no podemos hablar en general de un test comúnmente aceptado como *óptimo* para resolver el contraste de la media, como sucediera en el modelo univariante con el denominado test `F`. De hecho, en la literatura multivariante se manejan hasta cuatro test distintos (Wilks, Lawley-Hotelling, Roy y Pillai), cada uno de los cuales se justifica desde un determinado punto de vista, como veremos a lo largo del capítulo. Estos cuatro tests poseen como denominador común el estar todos construidos a partir de ciertos autovalores, que desempeñan un papel fundamental en el análisis multivariante, de ahí que, siguiendo el esquema de Arnold (1981), se haya hecho un notable esfuerzo para justificar su presencia (por no decir ubicuidad), en virtud de los principios de suficiencia e invarianza[1]. El significado de estos autovalores quedará definitivamente esclarecido en el capítulo dedicado al análisis discriminante.

Las distribuciones nulas de los estadísticos de Wilks, Lawley-Hotelling, Pillai y

[1] En el Apéndice del primer volumen podemos encontrar las nociones básicas relativas al principio de invarianza

Roy son complicadas y no se utilizan en la práctica. En su lugar suelen considerarse aproximaciones a la distribución F Snedecor o bien las aproximaciones asintóticas a la distribución χ^2, en el caso de los tres primeros, y U en el caso de Roy. Lo más interesante de esta convergencia es que sigue siendo válida aún violando la hipótesis de normalidad multivariante de las observaciones, siempre y cuando se verifique la denominada condición de Huber. Todo ello se expone en la séptima sección del capítulo. Por último, podemos encontrar una breve introducción al contraste generalizado para la media que será necesaria para abordar el análisis de perfiles.

El modelo es pues el siguiente: dados $v \in \mathcal{M}_{\mathtt{n}\times p}$, y $V \subset \mathbb{R}^{\mathtt{n}}$, se dice $v \in V$ cuando cada uno de sus vectores columnas pertenece a V. Pues bien, un modelo lineal normal multivariante es una estructura estadística dada por una matriz aleatoria

$$Y \sim N_{\mathtt{n},p}(\mu, \mathtt{Id}, \Sigma), \tag{2.1}$$

donde $\mathtt{n} \geq p$, Σ es una matriz $p \times p$ definida positiva y μ es una matriz $\mathtt{n} \times p$ tal que $\mu \in V$, siendo V un subespacio de $\mathbb{R}^n$ de dimensión menor que $\mathtt{n}$. Si se denota $Y = (Y_1 \dots Y_{\mathtt{n}})'$ y $\mu = (\mu_1 \dots \mu_{\mathtt{n}})'$, se tiene, en virtud del teorema 1.21, que (2.1) es equivalente a afirmar que

$$Y_i \sim N_p(\mu_i, \Sigma), \quad i = 1, \dots, \mathtt{n}$$

siendo además independientes entre sí. El espacio de parámetros para esta estructura estadística es

$$\Theta = \{(\mu, \Sigma) : \mu \in V,\ \Sigma > 0\}.$$

No obstante, si $\mathtt{X}$ es una matriz $\mathtt{n} \times \mathtt{dim}V$ cuyas columnas constituyen una base del subespacio V, μ se expresará como $\mathtt{X}\beta$, para una única matriz β, de tipo $\mathtt{dim}V \times p$. Con lo cual, el espacio de parámetros sería

$$\Theta_{\mathtt{x}} = \{(\beta, \Sigma) : \beta \in \mathcal{M}_{\mathtt{dim}V\times p},\ \Sigma > 0\}.$$

Esta última parametrización del modelo se denomina versión coordenada. Abordamos a continuación los problemas de estimación de los parámetros del modelo y de test de hipótesis sobre la media.

2.1. Estimación

Afrontemos, en primer lugar, el problema de estimar los parámetros μ y Σ. Para ello, consideremos los estadísticos

$$\begin{aligned} \hat{\mu} &= P_V Y = X(X'X)^{-1}X'Y \\ \hat{\Sigma} &= \tfrac{1}{\mathtt{n}-\mathtt{dim}V} Y' P_{V^\perp} Y = \tfrac{1}{\mathtt{n}-\mathtt{dim}V} Y'(\mathtt{Id} - X(X'X)^{-1}X')Y. \end{aligned}$$

Estos estimadores son las traducciones al lenguaje matricial de los estimadores naturales del modelo lineal univariante estudiado en el capítulo 3 del primer volumen.

Teorema 2.1.
Se verifica que $\hat{\mu} \sim N_{n,p}(\mu, P_V, \Sigma)$ y $(\mathtt{n} - \mathtt{dim}V)\hat{\Sigma} \sim W_p(\mathtt{n} - \mathtt{dim}V, \Sigma)$, y son independientes[2].

Demostración.
Se deduce del corolario 1.25 que $P_V Y$ y $Y' P_{V^\perp} Y$ son independientes. Además, teniendo en cuenta que P_V es idempotente, se verifica que

$$P_V Y \sim N_{n,p}(P_V \mu, P_V, \Sigma).$$

Por otro lado,

$$Y' P_{V^\perp} Y \sim W_p(\mathtt{n} - \mathtt{dim}V, \Sigma, \mu' P_{V^\perp} \mu).$$

Como $\mu \in V$, el último parámetro es 0.

∎

Teorema 2.2.
El estadístico $\left(\hat{\mu}, \hat{\Sigma}\right)$ es suficiente y completo.

Demostración.
La función de verosimilitud es la siguiente:

$$\mathcal{L}(y, \mu, \Sigma) = \frac{1}{(2\pi)^{pn/2}|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}\left(\Sigma^{-1}(y-\mu)'(y-\mu)\right)\right\}.$$

Sea $X \in \mathcal{M}_{n\times r}$ una base ortonormal de V (luego, $P_V = XX'$ y $\mu = P_V\mu = XX'\mu$) y h definida de la forma

$$h(\mu, \Sigma) = \frac{1}{(2\pi)^{pn/2}|\Sigma|^{n/2}} \exp\{-\frac{1}{2}\mathtt{tr}\Sigma^{-1}\mu\mu'\}.$$

Entonces

$$\mathcal{L}(y, \mu, \Sigma) = h(\mu, \Sigma) \exp\left\{-\frac{1}{2}\mathtt{tr}(\Sigma^{-1}Y'Y) + \mathtt{tr}(\Sigma^{-1}\mu' XX'Y)\right\}.$$

Siendo $\Sigma > 0$, podemos considerar Σ^{-1}. Utilizaremos la siguiente notación:

$$\Sigma^{-1} = \begin{pmatrix} \Delta_{11} & \dots & \Delta_{1p} \\ \vdots & & \vdots \\ \Delta_{1p} & \dots & \Delta_{pp} \end{pmatrix}, \qquad T = Y'Y = \begin{pmatrix} T_{11} & \dots & T_{1p} \\ \vdots & & \vdots \\ T_{1p} & \dots & T_{pp} \end{pmatrix}.$$

[2]Nótese que se trata de una versión multivariante del teorema de Fisher.

Sea $\theta = X'\mu\Sigma^{-1} \in \mathcal{M}_{r\times p}$, expresándose de la forma $\theta = (\theta_1 \dots \theta_r)'$, con $\theta_i \in \mathbb{R}^p$ $i = 1, \dots, p$. Igualmente, sea $U = X'Y \in \mathcal{M}_{r\times p}$, con $U = (U_1 \dots U_r)'$, donde $U_i \in \mathbb{R}^p$, $i = 1, \dots, p$. Consideremos entonces los siguientes vectores:

$$\tilde{\Delta}_1 = \text{diag}(\Sigma^{-1}) = \begin{pmatrix} \Delta_{11} \\ \vdots \\ \Delta_{pp} \end{pmatrix} \in \mathbb{R}^p, \qquad \tilde{T}_1 = \text{diag}(T),$$

$$\tilde{\Delta}_2 = \text{triangsup}(\Sigma^{-1}) = \begin{pmatrix} \Delta_{12} \\ \vdots \\ \Delta_{p-1,p} \end{pmatrix} \in \mathbb{R}^{p(p-1)/2}, \quad \tilde{T}_2 = \text{triangsup}(T) \in \mathbb{R}^{p(p-1)/2},$$

$$\tilde{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_r \end{pmatrix} \in \mathbb{R}^{pr}, \qquad \tilde{U} = \begin{pmatrix} u_1 \\ \vdots \\ u_r \end{pmatrix} \in \mathbb{R}^{pr}.$$

Entonces, $\text{tr}(\Sigma^{-1}YY') = \tilde{\Delta}_1'\tilde{T}_1 + 2\tilde{\Delta}_2'\tilde{T}_2$ y $\text{tr}(\Sigma^{-1}\mu' XX'Y) = \texttt{tr}(\theta' U) = \tilde{\theta}'\tilde{U}$. Consideremos las siguientes funciones Q y S:

$$Q(\mu, \Sigma) = \begin{pmatrix} \tilde{\Delta}_1 \\ \tilde{\Delta}_2 \\ \tilde{\theta} \end{pmatrix}, \qquad S(Y) = \begin{pmatrix} -\frac{1}{2}\tilde{T}_1 \\ -\tilde{T}_2 \\ \tilde{U} \end{pmatrix}.$$

Se verifica entonces

$$\mathcal{L}(y, \mu, \Sigma) = h(\mu, \Sigma) \exp\left\{ (Q(\mu, \Sigma))' S(y) \right\}.$$

Por lo tanto, estamos hablando de una estructura estadística de tipo exponencial y, aplicando el teorema de factorización de Neyman se deduce que el estadístico S es suficiente. Además, puede comprobarse que el interior de $\{Q(\mu, \Sigma) : \mu \in V,\ \Sigma > 0\}$ es distinto del vacío[3]. En consecuencia, S es completo. Además, podemos encontrar una biyección bimedible ϕ tal que

$$\phi(S) = (\hat{\mu}, \hat{\Sigma}),$$

de manera que $(\hat{\mu}, \hat{\Sigma})$ es, igualmente, suficiente y completo. ∎

[3]Téngase en cuenta que el conjunto de las matrices $p \times p$ simétricas se corresponden, de manera natural, con $\mathbb{R}^{p(p+1)/2}$, y que el subconjunto de las matrices definidas positivas (es decir, aquellas cuyo p-ésimo autovalor es estrictamente positivo) se identifica entonces con un abierto, pues el p-ésimo autovalor es una función continua.

Corolario 2.3.
EIMV $\hat{\mu}$ y $\hat{\Sigma}$ son los estimadores insesgados de mínima varianza de μ y Σ, respectivamente.

Demostración.
Del teorema 2.1 se deduce que

$$\mathtt{E}\left[\hat{\mu}\right] = \mu, \qquad \mathtt{E}[\hat{\Sigma}] = \Sigma.$$

Además, todas las componentes son de cuadrados sumables[4]. Luego, aplicando el teorema 2.2 junto con el de Lehmann-Scheffé, se tiene que

$$\mathtt{E}[\hat{\mu}|(\hat{\mu}, \hat{\Sigma})] \circ (\hat{\mu}, \hat{\Sigma}) = \hat{\mu}$$

es el EIMV (esencialmente único) de μ. Lo mismo sucede con Σ.

∎

Veamos qué podemos decir respecto al principio de máxima verosimilitud.

Teorema 2.4.
$\left(\hat{\mu}, \frac{\mathtt{n-dim}V}{n}\hat{\Sigma}\right)$ es el estimador de máxima verosimilitud de (μ, Σ).

Demostración.
Recordemos que la función de verosimilitud es

$$\mathcal{L}(y, \mu, \Sigma) = \frac{1}{(2\pi)^{pn/2}|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}\left(\Sigma^{-1}(y-\mu)'(y-\mu)\right)\right\}.$$

Tener en cuenta que $y - \hat{\mu} = P_{V^\perp} y$ y, entonces, $(y-\hat{\mu})'(\hat{\mu}-\mu) = 0$. Por lo tanto, se tiene que

$$\mathtt{tr}\left(\Sigma^{-1}(y-\mu)'(y-\mu)\right) = \mathtt{tr}\left(\Sigma^{-1}(y-\hat{\mu})'(y-\hat{\mu})\right) + \mathtt{tr}\left(\Sigma^{-1}(\hat{\mu}-\mu)'(\hat{\mu}-\mu)\right).$$

Puede demostrarse fácilmente que el último sumando no puede ser negativo. Luego, para valores de y y Σ fijos, la anterior expresión alcanza el mínimo (y la función de verosimilitud el máximo) cuando $\mu = \hat{\mu}$. Pues bien, dado y, busquemos entonces el valor de Σ que maximiza

$$\mathcal{L}(y, \hat{\mu}, \Sigma) = \frac{1}{(2\pi)^{pn/2}|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}\left(\Sigma^{-1}(y-\hat{\mu})'(y-\hat{\mu})\right)\right\}.$$

[4] Al ser normales univariantes o sumas de productos de normales univariantes, existen los momentos de todos los órdenes.

Sea $A = (y-\hat{\mu})'(y-\hat{\mu}) = (\mathtt{n}-\mathtt{dim}V)\hat{\Sigma}$, que sabemos se distribuye según un modelo $W_p(\mathtt{n}-\mathtt{dim}V, \Sigma)$ y que, por lo tanto, es definida positiva con probabilidad 1. Aplicando el teorema 13.11, se tiene que el máximo se alcanza en

$$\Sigma = \frac{1}{\mathtt{n}}A = \frac{\mathtt{n}-\mathtt{dim}V}{\mathtt{n}}\hat{\Sigma}$$

Recapitulando, tenemos que, dados y, μ y Σ,

$$\mathcal{L}(y,\mu,\Sigma) \leq \mathcal{L}\left(y,\hat{\mu},\Sigma\right) \leq \mathcal{L}\left(y,\hat{\mu},\frac{\mathtt{n}-\mathtt{dim}V}{\mathtt{n}}\hat{\Sigma}\right).$$

Por cierto, la matriz $\mathtt{n}^{-1}(\mathtt{n}-\mathtt{dim}V)\hat{\Sigma}$, que el el EMV de Σ, no es sino la matriz de varianzas-covarianzas totales muestral $S_{\mathtt{Y}}$, según se define en el Apéndice del primer volumen. Para acabar con el problema de estimación, supongamos que la estructura estadística se ha parametrizado mediante la familia $\Theta_{\mathtt{X}}$, donde $\mathtt{X}$ es una base de V, es decir, $\mu = \mathtt{X}\beta$. En ese caso, se verifica que $\beta = (\mathtt{X}'\mathtt{X})^{-1}X'\mu$. Se define entonces un estimador de β mediante

$$\hat{\beta} = (\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'\hat{\mu} = (\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'Y,$$

con lo cual, se verifican, trivialmente, las siguientes propiedades:

Teorema 2.5. (i) $\hat{\beta} \sim N_{r,p}(\beta, (\mathtt{X}'\mathtt{X})^{-1}, \Sigma)$, y por lo tanto es un estimador centrado.

(ii) El estadístico $\left(\hat{\beta},\hat{\Sigma}\right)$ es suficiente y completo.

(iii) $\hat{\beta}$ es el EIMV de β.

(iv) $\left(\hat{\beta},\frac{\mathtt{n}-\mathtt{dim}V}{\mathtt{n}}\hat{\Sigma}\right)$ es el EMV de (β,Σ).

2.2. Contrastes lineales sobre la media

Afrontaremos en esta sección el problema de contraste de hipótesis acerca del parámetro μ. Concretamente, resolveremos problemas del tipo siguiente: dado un subespacio $W \subset V$, contrastaremos la hipótesis inicial $H_0 : \mu \in W$ contra la alternativa. Por lo tanto, la hipótesis inicial viene dada por el subconjunto de parámetros $\Theta_0 = \{(\mu,\Sigma) : \mu \in W, \Sigma > 0\}$. Supondremos que $n \geq p + \mathtt{dim}V$. Los tests que propondremos para resolver el contraste se justifican en buena medida mediante el principio

de invarianza[5]. El proceso puede descomponerse en tres etapas, al igual que sucediera en el caso univariante[6]:

1. **Paso a forma canónica** En primer lugar, aplicaremos a nuestro modelo una transformación bimedible, basado en un cambio de base de $\mathbb{R}^{\mathtt{n}}$. El objeto del mismo es estructurar el espacio de parámetros de manera natural en función de la hipótesis a contrastar. Para ello consideraremos tres matrices $\mathtt{X}_1$, $\mathtt{X}_2$ y $\mathtt{X}_3$, bases ortonormales de los subespacios ortogonales W, $V|W$ y $V^{\perp}$, respectivamente. Sea entonces la transformación bimedible φ de $\mathcal{M}_{\mathtt{n}\times p}$ en sí mismo, que hace corresponder a cada matriz Y la matriz $Z = \varphi(Y)$ definida mediante

$$Z = \begin{pmatrix} \mathtt{X}_1' \\ \mathtt{X}_2' \\ \mathtt{X}_3' \end{pmatrix} Y.$$

   La columna j-ésima de Z está compuesta por las coordenadas de la columna j-ésima de Y respecto a una base ortonormal de $\mathbb{R}^{\mathtt{n}}$, la cual se descompone a su vez en bases de W, $V|W$ y $V^{\perp}$. Si se denota $Z_i = \mathtt{X}_i' Y$, $\nu_i = \mathtt{X}_i'\mu$, para $i = 1, 2, 3$, se tiene un nuevo modelo, que denominamos canónico, compuesto por tres matrices aleatorios independientes

$$\begin{aligned} Z_1 &\sim N_{\mathtt{dim}W,p}(\nu_1, \mathtt{Id}, \Sigma), \\ Z_2 &\sim N_{\mathtt{dim}V-\mathtt{dim}W,p}(\nu_2, \mathtt{Id}, \Sigma), \\ Z_3 &\sim N_{\mathtt{n}-\mathtt{dim}V,p}(0, \mathtt{Id}, \Sigma). \end{aligned}$$

   La familia de distribuciones puede expresarse pues con la ayuda del parámetro (ν_1, ν_2, Σ), que recorre el espacio

$$\mathcal{M}_{\mathtt{dim}W\times p} \times \mathcal{M}_{(\mathtt{dim}V-\mathtt{dim}W)\times p} \times \mathcal{M}^{+}_{p\times p},$$

   siendo $\mathcal{M}^{+}_{p\times p}$ el conjunto de las matrices cuadradas de orden p, simétricas y definidas positivas. La hipótesis inicial se traduce entonces en $H_0 : \nu_2 = 0$.

2. **Reducción por suficiencia.** En virtud del teorema 2.2, el estadístico $\left(\hat{\mu}, \hat{\Sigma}\right)$ es suficiente y completo. Dado que

$$\hat{\mu} = \begin{pmatrix} \mathtt{X}_1 Z_1 \\ \mathtt{X}_2 Z_2 \end{pmatrix}, \qquad \hat{\Sigma} \propto Z_3' Z_3,$$

[5]Para una introducción a los conceptos de Suficiencia e Invarianza, ver el capítulo prelimianr del volumen 1. Para un estudio más avanzado de Invarianza y su relación con Suficiencia, cf. Lehmann (1986), cap. 6.

[6]Ver el capítulo 3 del volumen 1

se verifica que $S = (Z_1, Z_2, Z_3'Z_3)$ es, a su vez, un estadístico suficiente y completo respecto al modelo canónico. Sabemos que el considerar únicamente la imagen de dicho estadístico, lo cual se denomina reducción por suficiencia, no conlleva pérdida alguna de información en el sentido de Fisher y no afecta, como veremos más adelante, a la búsqueda de un test UMP a nivel α. Además, al ser completo, la reducción por suficiencia es máxima, esto es, una reducción más profunda sí implicaría pérdida de información referente al parámetro. Las distribuciones del nuevo modelo reducido podrán expresarse, igual que en la fase anterior[7], con la ayuda del parámetro (ν_1, ν_2, Σ). La hipótesis a contrastar sigue siendo $\nu_2 = 0$.

Tras estos dos primeros pasos, seguimos contando con un espacio de observaciones demasiado complicado, concretamente

$$\mathbb{R}^{p\cdot \mathtt{dim} V + p(p+1)/2}$$

Ello nos obliga a nuevas reducciones, que implicarán pérdida de información. No obstante, dicha pérdida se verá justificada por el principio de Invarianza.

3. **Reducción por invarianza.** Consideremos el siguiente grupo de transformaciones bimedibles en el modelo canónico

$$G = \left\{ \mathtt{g}_{k,\Gamma,\Lambda} : k \in \mathcal{M}_{\mathtt{dim}W \times p},\ \Gamma \in \mathcal{O}_{\mathtt{dim}V - \mathtt{dim}W},\ \Lambda \in \mathcal{M}_p^* \right\}\ ^{8},$$

siendo

$$g_{k,\Gamma,\Lambda} \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} = \begin{pmatrix} Z_1\Lambda + k \\ \Gamma Z_2 \Lambda \\ Z_3\Lambda \end{pmatrix}.$$

Puede comprobarse fácilmente que G deja invariante tanto el modelo como el problema de contraste de hipótesis considerados. Por ello, el Principio de Invarianza propone restringir la búsqueda de tests a aquellos que sean igualmente invariantes, y entre éstos seleccionar el mejor desde algún criterio establecido. En este caso y dado $\alpha \in (0,1)$, intentaremos encontrar el test UMP-invariante a nivel α.

Dado que previamente hemos efectuado una reducción por suficiencia y que el estadístico suficiente S es trivialmente equivariante[9] respecto a G, podemos

[7] Una reducción por suficiencia no puede implicar simplificación alguna en el espacio de parámetros.

[8] En general, Los términos $\mathcal{O}_m$ y $\mathcal{M}_m^*$ denotan, en general, los conjunto de las matrices cuadradas de orden m y ortogonales, en el primer caso e invertibles en el segundo.

[9] Ver Apéndice del primer volumen.

considerar el grupo de transformaciones G^S que G induce de manera natural sobre el modelo imagen de S y buscar en dicho modelo un test ϕ_S UMP-invariante respecto a G^S a nivel α. De conseguirlo, el test $\phi_S \circ S$, definido sobre el modelo canónico, cumplirá la condición deseada. Vayamos por partes.

En primer lugar, el grupo G^S puede descomponerse en la suma de los subgrupos $G_1 = \{\mathtt{g}_k : k \in \mathcal{M}_{\mathtt{dim}W \times p}\}$, $G_2 = \{\mathtt{g}_\Gamma : O \in \mathcal{O}_{\mathtt{dim}V - \mathtt{dim}W}\}$ y $G_3 = \{\mathtt{g}_\Lambda : \Lambda \in \mathcal{M}_p^*\}$, donde

$$\mathtt{g}_k \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3'Z_3 \end{pmatrix} = \begin{pmatrix} Z_1 + k \\ Z_2 \\ Z_3'Z3 \end{pmatrix}, \quad \mathtt{g}_\Gamma \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3'Z_3 \end{pmatrix} = \begin{pmatrix} Z_1 \\ \Gamma Z_2 \\ Z_3'Z_3 \end{pmatrix},$$

$$\mathtt{g}_\Lambda \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3'Z_3 \end{pmatrix} = \begin{pmatrix} Z_1\Lambda \\ Z_2\Lambda \\ \Lambda' Z_3'Z_3\Lambda \end{pmatrix}.$$

Nuestro primer objetivo es encontrar un estadístico invariante maximal respecto a G^S, así como el correspondiente invariante maximal para el espacio de parámetros. Aprovechando la descomposición de G^S, y dado que los tres grupos satisfacen las propiedades necesarias, dicha búsqueda se realizará en tres etapas, siguiendo primero el orden $1 \to 2 \to 3$ y, a continuación, siguiendo el orden $1 \to 3 \to 2$:

(a) El siguiente estadístico es, trivialmente, G_1-invariante maximal:

$$M_1 \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3'Z_3 \end{pmatrix} = \begin{pmatrix} Z_2 \\ Z_3'Z_3 \end{pmatrix}.$$

Una aplicación $\overline{G}_1$-invariante maximal (es decir, invariante maximal en el espacio de parámetros) es $v_1(\nu_1, \nu_2, \Sigma) = (\nu_2, \Sigma)$. Por tanto, la distribución de M_1 no depende de ν_1. M_1 induce dos grupos de transformaciones sobre el espacio de llegada, G_2^1 y G_3^1, cuyos elementos verifican

$$g_\Gamma^1 \begin{pmatrix} Z_2 \\ Z_3'Z_3 \end{pmatrix} = \begin{pmatrix} \Gamma Z_2 \\ Z_3'Z_3 \end{pmatrix}, \quad g_\Lambda^1 \begin{pmatrix} Z_2 \\ Z_3'Z_3 \end{pmatrix} = \begin{pmatrix} Z_2\Lambda \\ \Lambda' Z_3'Z_3\Lambda \end{pmatrix}.$$

Por el teorema 13.9, se tiene que el estadístico siguiente es G_2^1-invariante maximal:

$$M_2^1 \begin{pmatrix} Z_2 \\ Z_3'Z_3 \end{pmatrix} = \begin{pmatrix} Z_2'Z_2 \\ Z_3'Z_3 \end{pmatrix}.$$

Además, la aplicación $v_2^1(\nu_2, \Sigma) = (\nu_2'\nu_2, \Sigma)$ es $\overline{G}_2^1$-invariante maximal. M_2^1 induce en el espacio de llegada el grupo de transformaciones G_3^{12} cuyos elementos verifican

$$g_\Lambda^{12} \begin{pmatrix} Z_2'Z_2 \\ Z_3'Z_3 \end{pmatrix} = \begin{pmatrix} \Lambda' Z_2'Z_2\Lambda \\ \Lambda' Z_3'Z_3\Lambda \end{pmatrix}.$$

Entonces, el estadístico M_3^{12} que asocia a $(Z_2'Z_2, Z_3'Z_3)$ las raíces ordenadas $(t_1, \ldots, t_p)$ [10] del polinomio en t $|Z_2'Z_2 - tZ_3'Z_3|$, es G_3^{12}-invariante maximal. Efectivamente, efecto, es invariante, pues, para todo Λ, las raíces de $|Z_2'Z_2 - tZ_3'Z_3|$ coinciden con las de

$$|\Lambda' Z_2'Z_2\Lambda - t\Lambda' Z_3'Z_3\Lambda| = |\Lambda|^2 |Z_2'Z_2 - tZ_3'Z_3|.$$

Además, es maximal pues, si U_1, U_2, R_1 y R_2 son matrices de $\mathcal{M}_p^+$ tales que las raíces $t_1, \ldots, t_p$ de $|U_1 - tR_1|$ coinciden con las de $|U_2 - tR_2|$, existen, en virtud del teorema 13.7, dos matrices $C_1, C_2 \in \mathcal{M}_p^*$ invertibles tales que

$$C_1 U C_1' = \begin{pmatrix} t_1 & & 0 \\ & \ddots & \\ 0 & & t_p \end{pmatrix} = C_2 U_2 C_2', \quad C_1 R_1 C_1' = \mathtt{Id} = C_2 R_2 C_2'.$$

Si consideramos $\Lambda = C_2^{-1} C_1$, se tiene que

$$\begin{pmatrix} U_2 \\ R_2 \end{pmatrix} = g_\Lambda^{12} \begin{pmatrix} U_1 \\ R_1 \end{pmatrix},$$

como queríamos demostrar. Análogamente, la aplicación v_3^{12} que asigna al parám tro $(\nu_2'\nu_2, \Sigma)$ las raíces ordenadas $\theta_1, \ldots, \theta_p$ del polinomio en θ

$$|\nu_2'\nu_2 - \theta\Sigma|,$$

es $\overline{G}_3^{12}$-invariante maximal.

(b) Procedamos en distinto orden: tras el primer paso, consideramos el grupo G_3^1, de manera que un estadístico invariante maximal para dicho grupo es el siguiente:

$$M_3^1 \begin{pmatrix} Z_2 \\ Z_3'Z_3 \end{pmatrix} = Z_2 (Z_3'Z_3)^{-1} Z_2'.$$

[10] En virtud del teorema 13.7, todas son reales y no negativas. Es fácil demostrar que el número de raíces positivas coincide con el rango de Z_2, que, bajo nuestras hipótesis es, en virtud del teorema 1.26, el mínimo entre $\mathtt{dim}V - \mathtt{dim}W$ y p.

En efecto, es trivialmente invariante. Además, si $L_1, L_2 \in \mathcal{M}_{(\mathtt{dim}V-\mathtt{dim}W)\times p}$ y $R_1, R_2 \in \mathcal{M}_p^+$ definidas positivas verifican que

$$M_3^1\begin{pmatrix} L_1 \\ R_1 \end{pmatrix} = M_3^1\begin{pmatrix} L_2 \\ R_2 \end{pmatrix},$$

entonces, en virtud del teorema 13.10, existe una matriz $B \in \mathcal{M}_p^*$ invertible tal que $L_2' = BL_1'$ y $R_2 = BR_1B'$. Considerando $A = B'$ se acaba. Por otro lado, la aplicación $v_3^1(\nu_2, \Sigma) = \mu_2\Sigma^{-1}\nu_2'$ es $\overline{G}_3^1$-invariante maximal. Además. M_3^1 induce sobre el espacio de llegada el grupo de transformaciones G_2^{13}, cuyos elementos se definen mediante

$$g_\Gamma^{13}\big(Z_2(Z_3'Z_3)^{-1}Z_2'\big) = \Gamma Z_2(Z_3'Z_3)^{-1}Z_2'\Gamma'.$$

El estadístico M_2^{13} constituido por los autovalores ordenados $(t_1' \ldots, t_{p-k}')$ de la matriz[11]

$$Z_2(Z_3'Z_3)^{-1}Z_2'$$

es G_2^{13}-invariante maximal. En efecto, se tiene que

$$\begin{aligned} |\Gamma Z_2(Z_3'Z_3)^{-1}Z_2'\Gamma' - r\mathtt{Id}| &= |\Gamma Z_2(Z_3'Z_3)^{-1}Z_2'\Gamma' - t'\Gamma\mathtt{Id}\Gamma'| \\ &= |\Gamma|^2|Z_2(Z_3'Z_3)^{-1}Z_2' - t'\mathtt{Id}|. \end{aligned}$$

Por lo tanto, los autovalores de $\Gamma Z_2(Z_3'Z_3)^{-1}Z_2'\Gamma'$ coinciden con los de la matriz $Z_2(Z_3'Z_3)^{-1}Z_2'$. Luego, M_2^{13} es invariante. Por otro lado, si T_1 y T_2 son matrices simétricas semidefinidas positivas de orden $\mathtt{dim}V|W$ y poseen los mismos autovalores, $t_1', \ldots, t_{\mathtt{dim}V-\mathtt{dim}W}'$, entonces, existen $\Gamma_1, \Gamma_2 \in \mathcal{O}_{\mathtt{dim}V|W}$ tales que

$$\Gamma_1T_1\Gamma_1' = \Gamma_2T_2\Gamma_2' = \begin{pmatrix} t_1' & & 0 \\ & \ddots & \\ 0 & & t_{\mathtt{dim}V-\mathtt{dim}W}' \end{pmatrix}.$$

Considerando $\Gamma = \Gamma_2\Gamma_1$ acabamos. Asimismo, la aplicación v_2^{13} que asigna a $\nu_2\Sigma^{-1}\nu_2'$ los autovalores ordenados $(\theta_1', \ldots, \theta_{\mathtt{dim}V-\mathtt{dim}W}')$ de la matriz $\nu_2'\Sigma^{-1}\nu_1$, es $\overline{G}_2^{13}$-invariante maximal.

De esta forma, los estadísticos definidos sobre el modelo canónico

$$(t_1, \ldots, t_p) = M_3^{12} \circ M_2^1 \circ M_1 \circ S,$$

[11] Al ser una matriz semidefinida positiva, son todos no negativos. El número de autovalores positivos coincide con el rango de Z_2.

$$(t'_1,\ldots,t'_{\mathtt{dim}V|W}) = M_2^{13}\circ M_3^1\circ M_1\circ S$$

son G-invariantes maximales. Veamos que el obtener invariantes maximales siguiendo dos secuencias distintas permite simplificar sutilmente el problema. En lo que sigue, b denotará el mínimo entre p y $\mathtt{dim}V|W$. Si $b=p$, se deduce del teorema 13.7 que $t_1,\ldots,t_p>0$, $t'_1,\ldots,t'_p>0$ y $t'_{p+1}=\ldots=t'_{\mathtt{dim}V|W}=0$. Por contra, si $b=\mathtt{dim}V|W$, se tiene que $t'_1,\ldots,t'_{\mathtt{dim}V|W}>0$, $t_1,\ldots,t_{\mathtt{dim}V|W}>0$ y $t_{\mathtt{dim}V|W+1}=\ldots=t_p=0$. Por lo tanto, se deduce de la propia definición de estadístico invariante maximal que $(t_1,\ldots,t_b)$ y $(t'_1,\ldots,t'_b)$ son G-invariantes maximales. Análogamente, las funciones

$$(\theta_1,\ldots,\theta_b),\quad (\theta'_1,\ldots,\theta'_b)$$

son $\overline{G}$-invariantes maximales. Consideremos entonces el siguiente resultado:

Lema 2.6.
Sean $p,k,r,n\in\mathbb{N}$, tales que $n\geq r+p$ y $k\leq r$, y sean $Q\in\mathcal{M}_{(\mathtt{dim}V-\mathtt{dim}W)\times p}$ y $R\in\mathcal{M}_{p\times p}$ definida positiva. Entonces, si $t\neq 0$,

$$|Q'Q-tR| = (-t)^{p-r+k}|R|\cdot|QR^{-1}Q'-t\mathtt{Id}|.$$

Demostración.
Se tiene, en primer lugar,

$$\begin{aligned}\left|\begin{array}{cc}\mathtt{Id} & Q\\ Q' & tR\end{array}\right| &= \left|\begin{array}{cc}\mathtt{Id}-Q\frac{1}{t}R^{-1}Q' & 0\\ Q' & tR\end{array}\right| = |tR|\cdot|\mathtt{Id}-\frac{1}{t}QR^{-1}Q'|\\ &= t^{p-r+k}(-1)^{-r+k}|R|\cdot|QR^{-1}Q'-t\mathtt{Id}|.\end{aligned}$$

Por otra parte,

$$\left|\begin{array}{cc}\mathtt{Id} & Q\\ Q' & tR\end{array}\right| = \left|\begin{array}{cc}\mathtt{Id} & Q\\ 0 & tR-Q'Q\end{array}\right| = |tR-Q'Q| = (-1)^p|Q'Q-tR|.$$

Igualando ambos términos se concluye. ∎

Como consecuencia del lema, se tiene que, si $t\neq 0$,

$$|Z_2'Z_2-tZ_3'Z_3| = (-t)^{p-r+k}|Z_3'Z_3|\cdot|Z_2(Z_3'Z_3)^{-1}Z_2'-t\mathtt{Id}|.$$

Entonces, $(t_1,\ldots,t_b)$ y $(t'_1,\ldots,t'_b)$ coinciden. Lo mismo puede decirse de los parámetros $(\theta_1,\ldots,\theta_b)$ y $(\theta'_1,\ldots,\theta'_b)$. Resumiendo, hemos obtenido el siguiente resultado:

Teorema 2.7.
El estadístico $(t_1, \ldots, t_b)$, donde $b = \min\{\mathtt{dim}V - \mathtt{dim}W, p\}$ y los t_i's son las b raíces positivas ordenadas del polinomio en t $|Z_2'Z_2 - tZ_3'Z_3|$, que coinciden con los b autovalores positivos de la matriz $Z_2(Z_3Z_3)^{-1}Z_2'$, es G-invariante maximal, y su distribución depende del parámetro (ν_1, ν_2, Σ) a través de las b primeras raíces ordenadas $(\theta_1, \ldots, \theta_b)$ del polinomio en θ $|\nu_2'\nu_2 - \theta\Sigma|$, que coinciden con los b primeros autovalores de $\nu_2\Sigma^{-1}\nu_2'$, y depende de las dimensiones $\mathtt{n}$, p, $\mathtt{dim}V$ y $\mathtt{dim}W$ a través de p, $\mathtt{dim}V|W$ y $\mathtt{n} - \mathtt{dim}V$.

Expresando el estadístico y los parámetros en los términos originales, tenemos que $Z_2'Z_2 = Y'X_2X_2'Y = Y'P_{V|W}Y$, que se denotará en lo que sigue por S_2. Igualmente, $Z_3'Z_3 = Y'P_{V^\perp}Y$, que se denota por S_3. Así pues, $(t_1, \ldots, t_b)$ son las raíces positivas ordenadas de $|S_2 - tS_3|$, que coinciden con los autovalores positivos de $Z_2S_3^{-1}Z_2'$, y constituyen un estadístico invariante maximal[12], cuya distribución depende de μ y Σ a través de $(\theta_1, \ldots, \theta_b)$, que son las b primeras raíces ordenadas de $|\mu'P_{V|W}\mu - \theta\Sigma|$ (es decir, los b primeros autovalores de $\nu_2\Sigma^{-1}\nu_2'$). Así pues, la aplicación del principio de Invarianza conduce a considerar únicamente estadísticos de contrastes que sean funciones de $(t_1, \ldots, t_b)$, cuya distribución, $\left(N_{n,p}(\mu, \mathtt{Id}, \Sigma)\right)^{(t_1,\ldots,t_b)}$ se denotará mediante $P_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V}(\theta_1, \ldots, \theta_b)$, siendo $\theta_1 \geq \ldots \geq \theta_b \geq 0$. Es decir, que el principio de Invarianza traslada el problema de constraste de hipótesis al experimento estadístico

$$\left(\mathbb{R}^b, \mathcal{R}^b, \left\{P_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V}(\theta) :\ \theta \in \mathbb{R}^+_{[\cdot]}\right\}\right)\ ^{13},$$

La hipótesis inicial $\mu \in W$, equivalente a $\nu_2 = 0$, queda reducida a $\theta = (0, \ldots, 0)$. En el caso nulo, $P_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V}$ es a distribución de las b primeras raíces ordenadas del polinomio $|S_2 - tS_3|$, donde $S_2 \sim W_p(\mathtt{dim}V|W, \Sigma)$ y $S_3 \sim W_p(\mathtt{n} - \mathtt{dim}V, \Sigma)$, ambas independientes. Se trata de una distribución continua en $\mathbb{R}^b$ cuya función de densidad podemos encontrar en Anderson (1958), capítulo 13 [14]. Sin embargo, dicha distribución resulta excesivamente complicada, siendo su uso poco frecuente en la práctica[15].

Hemos de darnos cuenta de que el espacio de observaciones actual es $\mathbb{R}^b$, con lo que se impone una reducción adicional a las que ya hemos realizado. No obstante, es importante estudiar previamente el caso más favorable, es decir, $b = 1$. Dado que $b = \min\{p, \mathtt{dim}V|W\}$, analizaremos por separado los casos $p = 1$ y $\mathtt{dim}V|W = 1$.

[12] Respecto al grupo $G_\varphi = \{g \circ \varphi : g \in G\}$.

[13] $\mathbb{R}^+_{[\cdot]}$ denota el conjunto de los vectores de $\mathbb{R}^b$ cuyas componentes son no negativas positivas y están ordenadas de mayor a menor.

[14] Asimismo, en Bilodeau (1999) se obtiene la distribución asintótica de las raíces cuando los valores del parámetro son todos distintos.

[15] No obstante, es interesante notar (Arnold, ejercicio 19.C.1) que, $P_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V} = P_{\mathtt{dim}V|W,p,\mathtt{n}-\mathtt{dim}W-p}$.

1. **Caso** $p = 1$**:** bajo las condiciones del modelo lineal normal univariante, el estadístico invariante maximal queda reducido a t, que es el autovalor de $S_2 S_3^{-1}$, cuya distribución depende de θ, el autovalor de $\mu' P_{V|W} \mu \Sigma^{-1}$. En este caso, S_2 y S_3 son números positivos; concretamente

$$S_2 = Y' P_{V|W} Y = \|P_{V|W} Y\|^2, \quad S_3 = Y' P_{V^\perp} Y = \|P_{V^\perp} Y\|^2.$$

Igualmente, $\mu' P_{V|W} \mu = \|P_{V|W} \mu\|^2$ y $\Sigma = \sigma^2$. Por lo tanto, el estadístico F definido mediante

$$F(Y) = \frac{\mathtt{n} - \mathtt{dim}V}{\mathtt{dim}V|W}\, t = \frac{\mathtt{n} - \mathtt{dim}V}{\mathtt{dim}V|W}\, \frac{\|P_{V|W} Y\|^2}{\|P_{V^\perp} Y\|^2}$$

sigue un modelo de distribución

$$F_{\mathtt{dim}V|W, \mathtt{n}-\mathtt{dim}V}\left(\frac{\|P_{V|W}\mu\|^2}{\sigma^2}\right).$$

Se trata pues de contrastar la hipótesis inicial $\theta = 0$ contra la alternativa $\theta > 0$. El experimento estadístico imagen por F posee razón de verosimilitud monótona[16], luego, se deduce del Lema de Neyman-Pearson que el siguiente test es uniformemente más potente a nivel α para dicho contraste

$$\mathtt{F}(\mathtt{y}) = \begin{cases} 1 & \text{si} \quad F(\mathtt{y}) > F^{\alpha}_{\mathtt{dim}V|W, \mathtt{n}-\mathtt{dim}V} \\ 0 & \text{si} \quad F(\mathtt{y}) \leq F^{\alpha}_{\mathtt{dim}V|W, \mathtt{n}-\mathtt{dim}V} \end{cases}$$

Éste es precisamente, el test UMP-invariante a nivel α que se obtiene en el modelo univariante, como podemos ver en el capítulo 3 del volumen 1, lo cual es lógico, pues el grupo de transformaciones consideradas en el modelo multivariante generaliza las que se consideraron en el modelo univariante. Del test `F` sabemos, además, que es insesgado y de razón de verosimilitudes.

2. **Caso** $\mathtt{dim}V|W = 1$**:** si W es un hiperplano de V, t puede definirse como el valor de $Z_2 S_3^{-1} Z_2'$, cuya distribución depende de θ, el valor de $\nu_2 \Sigma^{-1} \nu_2'$. Téngase en cuenta que $Z_2' \sim N_{p,1}(\nu_2', \Sigma, 1) = N_p(\nu_2', \Sigma)$, que $S_3 \sim W_p(\mathtt{n} - \mathtt{dim}V, \Sigma)$ y que ambas son independientes. Dado que $Z_2 S_3^{-1} Z_2'$ es un número, coincide con t y, entonces, el estadístico T^2 definido mediante $T^2(\mathtt{y}) = (\mathtt{n} - \mathtt{dim}V) t(\mathtt{y})$, sigue un modelo de distribución $T^2_{p, \mathtt{n}-\mathtt{dim}V}(\theta)$ [17] o, equivalentemente,

$$\frac{\mathtt{n} - \mathtt{dim}V - p + 1}{p}\, t \sim F_{p, \mathtt{n}-\mathtt{dim}V-p+1}(\theta).$$

[16] Ver Apéndice del volumne1

[17] Este resultado es el que conduce a estudiar a distribución T^2 de Hotelling.

Nuevamente, el problema queda reducido a contrastar la hipótesis inicial $\theta = 0$ contra $\theta > 0$ en el experimento inducido por T^2. Teniendo en cuenta que dicho experimento posee razón de verosimilitudes monótona y aplicando nuevamente el Lema de Neyman-Pearson, se concluye que el test siguiente es UMP-invariante a nivel α :

$$\mathtt{T}^2(y) = \begin{cases} 1 & \text{si} \quad T^2(\mathtt{y}) > T^{2,\alpha}_{p,\mathtt{n}-\mathtt{dim}V} \\ 0 & \text{si} \quad T^2(\mathtt{y}) \leq T^{2,\alpha}_{p,\mathtt{n}-\mathtt{dim}V} \end{cases}$$

Este será el caso, por ejemplo, del contraste de una media, del de contraste de igualdad de dos medias o los contrastes parciales de regresión.

Hemos comprobado entonces que en el caso $b = 1$ las cosas funcionan bastante bien. Sin embargo, si $b > 1$ no existe ningún test UMP ni UMP-invariante[18]. Como ya hemos dicho, en este caso se precisa una reducción adicional. Veremos cuatro opciones que darán lugar a los test de Wilks, Lawley-Hotelling, Roy y Pillay, respectivamente. A continuación vamos a estudiarlos y justificarlos uno a uno, partiendo de principios intuitivos como los de máxima verosimilitud, sustitución o unión-intersección[19].

2.3. Test de Wilks

Veremos en esta sección que el denominado test de Wilks es el de la razón de verosimilitudes a nivel α para contrastar a hipótesis inicial $H_0 : \mu \in W$. Recordemos que éste se define, en un problema general, considerando el el estadístico de razón de verosimilitudes (siempre que exista y sea medible)

$$RV(\mathtt{y}) = \frac{\sup_{\theta\in\Theta_0} \mathcal{L}(\mathtt{y},\theta)}{\sup_{\theta\in\Theta} \mathcal{L}(\mathtt{y},\theta)} \tag{2.2}$$

En esas condiciones y de dado $\alpha \in (0,1)$, un test de la forma

$$\phi(\mathtt{y}) = \begin{cases} 1 & \text{si } RV(\mathtt{y}) < C \\ 0 & \text{si } RV(\mathtt{y}) \geq C \end{cases}$$

donde C es una constante tal que

$$\sup_{\theta\in\Theta_0} P_\theta(RV < C) = \alpha,$$

[18] Cf. Arnold(1981).

[19] En Pillay (1954) Some new test criteria in multivariate analysis, *Ann. Math. Stat.* 26, 117-121, se enumeran otras justificaciones de carácter eminentemente técnico.

se denomina de razón de verosimilitudes a nivel α. Eso ocurre, en particular, si RV posee una distribución $\mathtt{P}$ idéntica para todo los valores de Θ_0, y tomamos $C = \mathtt{P}^{1-\alpha}$. Eso es precisamente lo que ocurre en nuestro caso.

Por otra parte, es un hecho conocido[20], que bajo ciertas condiciones de regularidad (que se verifican en nuestro caso), el estadístico RV es, salvo en un conjunto nulo, invariante. Por lo tanto, dicho estadístico debe ser, salvo en un suceso nulo, función del invariante maximals $(t_1, \ldots, t_b)$. Consideremos, concretamente, el estadístico

$$\lambda_1(y) = \prod_{i=1}^{n} (1 + t_i(y))^{-1}$$

y sea $C^1_{p,\mathtt{dim}V|W,\mathtt{n-dim}V}$ su distribución en el caso nulo $\theta_1 = \ldots = \theta_b = 0$. El test de Wilks se define entonces de la siguiente forma:

$$\phi_1(\mathtt{y}) = \begin{cases} 1 & \text{si} \quad \lambda_1(y) < C^{1,1-\alpha}_{p,\mathtt{dim}V|W,\mathtt{n-dim}V} \\ 0 & \text{si} \quad \lambda_1(y) \geq C^{1,1-\alpha}_{p,\mathtt{dim}V|W,\mathtt{n-dim}V} \end{cases}$$

El estadístico λ_1 puede expresarse más cómodamente así:

$$\lambda_1 = \frac{|S_3|}{|S_2 + S_3|}.$$

Efectivamente, si aplicamos el lema 2.6 con $t = -1$, se tiene que

$$|S_2 + S_3| = |Z_3'Z_3| \cdot |Z_2(Z_3'Z_3)^{-1}Z_2' + \mathtt{Id}|.$$

Denótese

$$U = Z_2(Z_3'Z_3)^{-1}Z_2' + \mathtt{Id}.$$

Como sabemos, $t_1, \ldots, t_b$ son los autovalores positivos de la matriz $Z_2(Z_3'Z_3)^{-1}Z_2'$. Si $b = \mathtt{dim}V - \mathtt{dim}W$, no existen más autovalores; en caso contrario, el resto son nulos. Además, t_i es autovalor de $Z_2(Z_3'Z_3)^{-1}Z_2'$ sii $1 + t_i$ lo es de U. Entonces, si $u_1, \ldots, u_{\mathtt{dim}V-\mathtt{dim}W}$ denotan los autovalores de U, que es simétrica, se sigue del teorema de diagonalización que

$$|U| = \prod_{i=1}^{\mathtt{dim}V-\mathtt{dim}W} u_i = \prod_{i=1}^{b} (1 + t_i) = \lambda_1^{-1}.$$

Despejando se obtiene la expresión deseada. Veamos entonces que λ_1 coincide con RV :

[20] Lehmann (1986), pag. 341.

Teorema 2.8.
ϕ_1 es el test de razón de verosimilitudes.

Demostración.
En este caso se tiene

$$RV(\mathtt{y}) = \frac{\sup_{\mu\in W,\Sigma>0}\mathcal{L}(\mathtt{y},\mu,\Sigma)}{\sup_{\mu\in V,\Sigma>0}\mathcal{L}(\mathtt{y},\mu,\Sigma)}.$$

Respecto al denominador, sabemos que el máximo se alcanza en $\left(\hat{\mu}, \frac{\mathtt{n}-\mathtt{dim}V}{\mathtt{n}}\hat{\Sigma}\right)$, pues es el EMV, siendo

$$\frac{\mathtt{n}-\mathtt{dim}V}{\mathtt{n}}\hat{\Sigma} = \frac{1}{n}(\mathtt{y}-\hat{\mu})'(\mathtt{y}-\hat{\mu}) = \frac{1}{\mathtt{n}}\mathtt{y}'P_{V^\perp}\mathtt{y}.$$

Sustituyendo, se tiene el máximo

$$\frac{\mathtt{n}^{\frac{p\mathtt{n}}{2}}}{(2\pi)^{\frac{p\mathtt{n}}{2}}|\mathtt{y}'P_{V^\perp}\mathtt{y}|^{\frac{\mathtt{n}}{2}}}\,\mathtt{e}^{\left\{-\frac{1}{2}\mathtt{tr}\left[\left(\frac{1}{\mathtt{n}}(\mathtt{y}-\hat{\mu})'(\mathtt{y}-\hat{\mu})\right)^{-1}(\mathtt{y}-\hat{\mu})'(\mathtt{y}-\hat{\mu})\right]\right\}} = \frac{\mathtt{e}^{-\frac{p\mathtt{n}}{2}}\mathtt{n}^{\frac{p\mathtt{n}}{2}}}{(2\pi)^{\frac{p\mathtt{n}}{2}}|\mathtt{y}'P_{V^\perp}\mathtt{y}|^{\frac{\mathtt{n}}{2}}}.$$

Si restringimos el parámetro al caso $\mu \in W$, se obtiene, de manera totalmente análoga, el siguiente máximo:

$$\frac{\mathtt{e}^{-\frac{p\mathtt{n}}{2}}\mathtt{n}^{\frac{p\mathtt{n}}{2}}}{(2\pi)^{\frac{p\mathtt{n}}{2}}|\mathtt{y}'P_{W^\perp}\mathtt{y}|^{\frac{\mathtt{n}}{2}}}.$$

Luego, el cociente buscado es

$$RV(\mathtt{y}) = \left(\frac{|\mathtt{y}'P_{V^\perp}\mathtt{y}|}{|\mathtt{y}'P_{W^\perp}\mathtt{y}|}\right)^{\mathtt{n}/2}. \tag{2.3}$$

Teniendo en cuenta que $P_{W^\perp} = P_{V^\perp} + P_{V|W}$, se tiene que

$$RV(\mathtt{y}) = \left(\frac{|S_3|}{|S_2+S_3|}\right)^{\mathtt{n}/2} = (\lambda_1(y))^{\mathtt{n}/2}.$$

Como la función que asigna a cada número positivo x el valor $x^{n/2}$ es estrictamente creciente, la relación entre los cuantiles de la distribución $\mathtt{P}$ de R_V en el caso nulo y los de $C_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V}$ es la siguiente:

$$\mathtt{P}^{1-\alpha} = \left(C^{1,1-\alpha}_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V}\right)^{n/2}.$$

Por lo tanto, $\lambda_1 > C^{1,1-\alpha}_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V}$ si, y sólo si, $RV > \mathtt{P}^{1-\alpha}$.

∎

En el caso $b = 1$ el test de Wilks coincide con el test UMP-invariante propuesto anteriormente. Como corolario se tiene que dicho test es, además, el de razón de verosimilitudes, cosa que ya sabíamos del volumen 1.

2.4. Tests de Lawley-Hotelling y Pillay

El desconocimiento del parámetro Σ en el contraste de μ supone un serio inconveniente del cual hemos sido víctimas, pues las reducciones por suficiencia e invarianza nos han conducido a una estructura que, salvo en el caso $b = 1$, no es lo suficientemente sencilla. Vamos a comprobar que, si Σ es conocido, podemos llegar, tras aplicar una reducción por suficiencia seguida de otra por invarianza, a un test óptimo. En efecto, vamos a reproducir brevemente el esquema utilizado para la obtención del estadístico $(t_1, \ldots, t_b)$ aplicado a este caso concreto. El experimento estadístico queda, en estas condiciones, parametrizado únicamente por la media μ. Tras pasar a la forma canónica, tendremos

$$\begin{aligned} Z_1 &\sim N_{\mathtt{dim}W,p}(\nu_1, \mathtt{Id}, \Sigma), \\ Z_2 &\sim N_{\mathtt{dim}V|W,p}(\nu_2, \mathtt{Id}, \Sigma), \\ Z_3 &\sim N_{\mathtt{n-dim}V,p}(0, \mathtt{Id}, \Sigma), \end{aligned}$$

siendo la hipótesis inicial $\Theta_0' = \{\nu_2 = 0\}$. Efectuamos entonces el cambio de variables: $w = Z\Sigma^{-1/2}$. Si se denota $W_i = Z_i\Sigma^{-1/2}$ y $\eta_i = \nu_i\Sigma^{-1/2}$, para $i = 1, 2, 3$, se obtiene

$$\begin{aligned} W_1 &\sim N_{\mathtt{dim}W,p}(\nu_1, \mathtt{Id}, \mathtt{Id}), \\ W_2 &\sim N_{\mathtt{dim}V|W,p}(\nu_2, \mathtt{Id}, \mathtt{Id}), \\ W_3 &\sim N_{\mathtt{n-dim}V,p}(0, \mathtt{Id}, \mathtt{Id}), \end{aligned}$$

con hipótesis inicial $\Theta_0'' = \{\eta_2 = 0\}$. Si transformamos todas las matrices en vectores mediante el orden de lecturas por filas $\mathtt{vec}$, se tiene que

$$\begin{aligned} \mathtt{vec}(W_1) &\sim N_{p\mathtt{dim}W}\left(\mathtt{vec}(\eta_1), \mathtt{Id}\right), \\ \mathtt{vec}(W_2) &\sim N_{p\mathtt{dim}V|W}(\eta_2, \mathtt{Id}), \\ \mathtt{vec}(W_3) &\sim N_{p(\mathtt{n-dim}V)}(0, \mathtt{Id}), \end{aligned}$$

siendo todos independientes entre sí. La función de verosimilitud puede expresarse mediante:

$$\mathcal{L}(\mathtt{W}; \eta_1, \eta_2) = \frac{\mathtt{e}^{\left\{-\frac{1}{2}\mathtt{tr}(\mathtt{W}_3'\mathtt{W}_3)\right\}}}{(2\pi)^{(p\mathtt{n})/2}} \mathtt{e}^{\left\{-\frac{1}{2}\mathtt{tr}\left((\mathtt{W}_1-\eta_1)'(\mathtt{W}_1-\eta_1)\right)-\frac{1}{2}\mathtt{tr}\left((\mathtt{W}_2-\eta_2)'(\mathtt{W}_2-\eta_2)\right)\right\}}.$$

Teniendo en cuenta que

$$\mathtt{tr}\big((\mathtt{W}_i - \eta_i)'(\mathtt{W}_i - \eta_i)\big) = \|\mathtt{vec}(H_i) - \mathtt{vec}(\eta_i)\|^2, \qquad i = 1, 2,$$

se deduce que el estadístico

$$S = \begin{pmatrix} \mathtt{vec}(H_1) \\ \mathtt{vec}(H_2) \end{pmatrix}$$

es suficiente (además es completo). Si consideramos la estructura estadística inducida por S, el problema de decisión queda invariante ante la acción de los dos grupos siguientes:

$$G_1 = \{g_k : k \in \mathbb{R}^{p\mathtt{dim}W}\}, \qquad G_2 = \{g_\Gamma : \Gamma \in \mathcal{O}_{\mathtt{dim}V|W}\},$$

donde

$$g_k \begin{pmatrix} \mathtt{vec}(W_1) \\ \mathtt{vec}(W_2) \end{pmatrix} = \begin{pmatrix} \mathtt{vec}(W_1) + k \\ \mathtt{vec}(W_2) \end{pmatrix}, \quad g_\Gamma \begin{pmatrix} \mathtt{vec}(W_1) \\ \mathtt{vec}(W_2) \end{pmatrix} = \begin{pmatrix} \mathtt{vec}(W_1) \\ \Gamma \cdot \mathtt{vec}(W_2) \end{pmatrix}$$

Un estadístico invariante maximal respecto a $G_1 \oplus G_2$ es

$$\|\mathtt{vec}(W_2)\|^2 \sim \chi^2_{p\mathtt{dim}V|W}\big(\|\mathtt{vec}(\eta_2)\|^2\big).$$

Luego, dado que el experimento estadístico imagen posee razón de verosimilitud monótona, estamos en condiciones de aplicar el lema de Neyman-Pearson y concluir que el siguiente test es UMP-invariante a nivel α :

$$\phi(\mathtt{W}) = \begin{cases} 1 & \text{si} \quad \|\mathtt{vec}(\mathtt{W}_2)\|^2 > \chi^{2,\alpha}_{p\mathtt{dim}V|W} \\ 0 & \text{si} \quad \|\mathtt{vec}(\mathtt{W}_2)\|^2 \leq \chi^{2,\alpha}_{p\mathtt{dim}V|W} \end{cases} \tag{2.4}$$

Además,

$$\|\mathtt{vec}(W_2)\|^2 = \mathtt{tr}(W_2'W_2) = \mathtt{tr}(Z_2'Z_2\Sigma^{-1}) = \mathtt{tr}(S_2\Sigma^{-1}),$$

que sigue, en los términos originales del parámetro, un modelo de distribución

$$\chi^2_{p\mathtt{dim}V|W}\big(\mathtt{tr}(\delta\Sigma^{-1})\big), \qquad \delta = \mu' P_{V|W}\mu. \tag{2.5}$$

Por lo tanto, el test consiste en rechazar la hipótesis inicial cuando $\mathtt{tr}(S_2\Sigma^{-1}) > \chi^{2,\alpha}_{p\mathtt{dim}V|W}$. Hemos pues encontrado un test que podemos considerar óptimo en el caso Σ conocida. El método de sustitución propone, en este caso, considerar el mismo estadístico de contraste de dicho test óptimo, sustituyendo Σ (desconocido en realidad) por un buen estimador suyo, por ejemplo el EIMV y confrontar el resultado con el cuantil $(1-\alpha)$ de la distribución nula del nuevo estadístico. El estadístico en cuestiones es

$$\mathtt{tr}\left(S_2\hat{\Sigma}^{-1}\right) = (\mathtt{n} - \mathtt{dim}V)\mathtt{tr}\left(S_2S_3^{-1}\right).$$

Pues bien, se define el estadístico de Lawley-Hotelling de la forma

$$\lambda_2 = \sum_{i=1}^{b} t_i.$$

Denótese por $C^2_{p,\mathtt{dim}V|W,\mathtt{n-dim}V}$ la distribución de λ_2 en el caso nulo. Entonces el test de Lawley-Hotelling se define así:

$$\phi_2(\mathtt{y}) = \begin{cases} 1 & \text{si} \quad \lambda_2(\mathtt{y}) > C^{2,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n-dim}V} \\ 0 & \text{si} \quad \lambda_2(\mathtt{y}) \leq C^{2,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n-dim}V} \end{cases}$$

Teorema 2.9.

El test de Lawley-Hotelling está asociado al método de sustitución por el EIMV.

Demostración.

Se tiene que $\lambda_2 = \sum_{i=1}^{b} t_i$, siendo $t_1, \ldots, t_b$ los autovalores positivos de $Z_2(Z_3'Z_3)^{-1}Z_2'$. Al ser dicha matriz simétrica, se verifica

$$\lambda_2 = \mathtt{tr}\left(Z_2(Z_3'Z_3)^{-1}Z_2'\right) = \mathtt{tr}\left(Z_2'Z_2(Z_3'Z_3)^{-1}\right) = \mathtt{tr}(S_2S_3^{-1}).$$

∎

Podemos considerar el estadístico de contraste

$$\lambda_4 = \sum_{i=1}^{b} \frac{t_i}{1+t_i}.$$

Denótese por $C^4_{p,\mathtt{dim}V|W,\mathtt{n-dim}V}$ su distribución en el caso nulo. Entonces, el test de Pillay a nivel α, se define así:

$$\phi_4(\mathtt{y}) = \begin{cases} 1 & \text{si} \quad \lambda_4(\mathtt{y}) > C^{4,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n-dim}V} \\ 0 & \text{si} \quad \lambda_4(\mathtt{y}) \leq C^{4,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n-dim}V} \end{cases}$$

La siguiente expresión es más manejable:

Proposición 2.10.

$\lambda_4 = \mathtt{tr}\left(S_2(S_2+S_3)^{-1}\right)$

Demostración.

x es raíz no nula del polinomio en t $|Z_2'Z_2 - tZ_3'Z_3|$ sii $\frac{x}{1+x}$ lo es del polinomio en u $|Z_2'Z_2 - u(Z_2'Z_2 + Z_3'Z_3)|$. En virtud del lema 2.6, las raíces no nulas del polinomio $|Z_2'Z_2 - u(Z_2'Z_2 + Z_3'Z_3)|$ coinciden con las de $|Z_2(Z_2'Z_2 + Z_3'Z_3)^{-1}Z_2' - u\mathtt{Id}|$. Luego,

$$\mathtt{tr}\left(S_2(S_2+S_3)^{-1}\right) = \mathtt{tr}\left(Z_2(Z_2'Z_2 + Z_3'Z_3)^{-1}Z_2'\right) = \sum_{i=1}^{b} \frac{t_i}{1+t_i}.$$

∎

La única diferencia entre este test y el de Lawley-Hotelling estriba en que aquí no se consideran los autovalores originales t_i's sino una transformación de los mismos mediante la biyección de $[0,+\infty]$ en $[0,1]$

$$f(x)=\frac{x}{1+x}.$$

Realmente, y en virtud de (5.14), el test de Pillay considera, en vez de los autovalores originales, los denominadas coeficientes de correlación canónica asociados, $r_1^2,\ldots,r_b^2$ [21]. En se caso, se tiene

$$\lambda_4=\sum_{i=1}^{b} r_i^2 \tag{2.6}$$

Desgraciadamente y tras consultar la bibliografía, no estamos en condiciones de justificar este test de manera más precisa. El propio artículo en el cual es introducido[22] no aporta ningún argumento en ese sentido, sino que deja a entrever que se considera este estadístico en concreto porque la suma (2.6) resulta natural.

2.5. Test de Roy

El test de Roy se basa en a idea de considerar únicamente el primer autovalor, t_1, cosa que parece razonable, teniendo en cuenta que la hipótesis inicial se corresponde con el caso $\theta_1=0$. No obstante, probaremos además que este método se deriva de la aplicación del principio de unión-intersección. Además, guarda una estrecha relación con el primer vector discriminante, como se comprobará en su momento.

Así pues, se considera el estadístico $\lambda_3=t_1$. Si se denota por $C^3_{p,\mathtt{dim}V|W,\mathtt{n-dim}V}$ su distribución en el caso nulo, se define el test de Roy mediante

$$\phi_3(\mathtt{y})=\begin{cases}1 & \text{si} \quad \lambda_3(\mathtt{y})>C^{3,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n-dim}V}\\ 0 & \text{si} \quad \lambda_3(\mathtt{y})\leq C^{3,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n-dim}V}\end{cases}$$

Si μ y $\mathtt{d}$ son vectores de W y $V|W$, respectivamente[23], entonces $\mathtt{d}'\mu=(0,\overset{p}{\ldots},0)$. Luego si, además, $\mathtt{q}\in\mathbb{R}^p$, entonces $\mathtt{d}'\mu\mathtt{q}=0$. Por lo tanto, si $\mu\in V$, se verifica la siguiente equivalencia

$$\big[\exists(\mathtt{d},\mathtt{q})\in(V|W)\times\mathbb{R}^p\colon \mathtt{d}'\mu\mathtt{q}\neq 0\big]\Leftrightarrow\big[\mu\notin W\big].$$

[21] Los coeficientes de correlación canónica se estudiarán en el capítulo 6.

[22] Pillay (1954) Some new test criteria in multivariate analysis, *Ann. Math. Stat.* 26, 117-121

[23] En ese caso $\mathtt{d}$ se denomina contraste.

Por ello, el denominado método de la unión-intersección conduce, en este caso, a contrastar, para cada par $(\mathtt{d},\mathtt{q}) \in V|W \times \mathbb{R}^p$, la hipótesis inicial $H_0^{\mathtt{d},\mathtt{q}} : \mathtt{d}'\mu\mathtt{q} = 0$ (el test correspondiente irá ligado a una región de aceptación) y considerar entonces la intersección de todas las regiones de aceptación obtenidas como región de aceptación para el contraste $H_0 : \mu \in W$. El contraste $H_0^{\mathtt{d},\mathtt{q}}$ puede resolverse construyendo un intervalo de confianza para el número $\mathtt{d}'\mu\mathtt{q}$, de manera que optaremos por la hipótesis inicial cuando el 0 quede dentro del correspondiente intervalo. Luego, el test para contrastar H_0 optará por la hipótesis inicial cuando el 0 esté en todos los intervalos de confianza para $\mathtt{d}'\mu\mathtt{q}$, cuando $(\mathtt{d},\mathtt{q})$ recorre $V|W \times \mathbb{R}^p$. Hemos de ingeniarnoslas para que este test tenga el nivel de significación α deseado. Para ello, necesitamos que los intervalos de confianza anteriores constituyan una familia de intervalos de confianza simultáneos a nivel $1-\alpha$. Definimos a continuación este concepto, tratado ya en el capítulo 2 del primer volumen.

Dado $\alpha \in (0,1)$, una familia de intervalos de confianza a nivel $1-\alpha$ para $V|W \times \mathbb{R}^p$ es una función que asigna a cada elemento $(\mathtt{d},\mathtt{q}) \in (V|W) \times \mathbb{R}^p$ una pareja de funciones reales $(a^\alpha_{\mathtt{d},\mathtt{q}}, b^\alpha_{\mathtt{d},\mathtt{q}})$ definidas sobre $\mathcal{M}_{\mathtt{n}\times p}$, de tal forma que, para cada distribución P de la familia (2.1), se verifica

$$P\left(\left\{\mathtt{y} \in \mathbb{M}_{\mathtt{n}\times p} :\ a^\alpha_{\mathtt{d},\mathtt{q}}(\mathtt{y}) \leq \mathtt{d}'\mu\mathtt{q} \leq b^\alpha_{\mathtt{d},\mathtt{q}}(\mathtt{y}), \forall(\mathtt{d},\mathtt{q})\right\}\right) = 1-\alpha.$$

Probaremos a continuación que el test de Roy a nivel α decide la hipótesis alternativa para una muestra $\mathtt{y}$ concreta si, y sólo si, existe algún par $(\mathtt{d},\mathtt{q}) \in (V|W) \times \mathbb{R}^p$ tal que $0 \notin \left(a^\alpha_{\mathtt{d},\mathtt{q}}(\mathtt{y}), b^\alpha_{\mathtt{d},\mathtt{q}}(\mathtt{y})\right)$. En ese sentido decimos que el test de Roy es consistente con los intervalos de confianza simultáneos para $V|W \times \mathbb{R}^p$, y por ello, la relación entre el test de Roy y los intervalos de confianza simultáneos para $V|W \times \mathbb{R}^p$ es la misma que existe entre el test $\mathtt{F}$, los intervalos de confianza simultáneos de Scheffé, estudiado en el modelo lineal univariante.

Un comentario previo: $C^3_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V}$ se ha definido como la distribución de la primera raíz de $|S_2 - tS_3|$ en el caso nulo. En esa situación, S_2 y S_3 son independientes y siguen distribuciones $W_p(\mathtt{dim}V - \mathtt{dim}W, \Sigma)$ y $W_p(\mathtt{n} - \mathtt{dim}V, \Sigma)$, respectivamente. Si s denota la primera raíz del polinomio en x

$$\left|(\hat{\mu}-\mu)'P_{V|W}(\hat{\mu}-\mu) - x(\mathtt{n}-\mathtt{dim}V)\hat{\Sigma}\right|,$$

se tiene que la distribución de s, supuesto que $\mu \in V$ y $\Sigma > 0$, es $C^3_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V}$, dado que, en ese caso,

$$\begin{aligned}(y-\mu)'P_{V|W}(y-\mu) &\sim W_p(\mathtt{dim}V - \mathtt{dim}W, \Sigma),\\ (\mathtt{n}-\mathtt{dim}V)\hat{\Sigma} &\sim W_p(\mathtt{n} - \mathtt{dim}V, \Sigma),\end{aligned}$$

siendo ambas variables independientes y, además,

$$y - \mu = (y - \hat{\mu}) + (\hat{\mu} - \mu)\,,$$

perteneciendo el primer sumando a $V^{\perp}$. Vamos a buscar una expresión alternativa para el estadístico s. El primer resultado es consecuencia directa de la desigualdad de Cauchy-Schwarz:

Lema 2.11.

si $\mathtt{v} \in \mathbb{R}^n$ y $E \subset \mathbb{R}^n$, entonces

$$\sup_{\mathtt{e} \in E \setminus \{0\}} \frac{\langle \mathtt{e}, \mathtt{v} \rangle^2}{\|\mathtt{e}\|^2} = \|P_E \mathtt{v}\|^2.$$

Lema 2.12.

$$s = \sup_{\mathtt{d} \in V|W \setminus \{0\}, \mathtt{q} \in \mathbb{R}^p \setminus \{0\}} \frac{[\mathtt{d}'(\hat{\mu} - \mu)\mathtt{q}]^2}{(\mathtt{n} - \mathtt{dim}V)\|\mathtt{d}\|^2 \mathtt{q}'\hat{\Sigma}\mathtt{q}}.$$

Demostración.
Por el teorema 13.7 se tiene que

$$s = \sup_{\mathtt{q} \in \mathbb{R}^p \setminus \{0\}} \frac{\mathtt{q}'(\hat{\mu} - \mu)' P_{V|W}(\hat{\mu} - \mu)\mathtt{q}}{(\mathtt{n} - \mathtt{dim}V)\mathtt{q}'\hat{\Sigma}\mathtt{q}} = \sup_{\mathtt{q} \in \mathbb{R}^p \setminus \{0\}} \frac{\|P_{V|W}(\hat{\mu} - \mu)\mathtt{q}\|^2}{(\mathtt{n} - \mathtt{dim}V)\mathtt{q}'\hat{\Sigma}\mathtt{q}}.$$

Aplicando el lema anterior se concluye.

∎

Teorema 2.13.

Dado Sea $\alpha \in (0, 1)$, la pareja de funciones siguiente, construida para cada par $(\mathtt{d}, \mathtt{q}) \in (V|W) \times \mathbb{R}^p$, constituye una familia de intervalos de confianza simultáneos a nivel $1 - \alpha$ para los contrastes:

$$\begin{aligned} a^{\alpha}_{\mathtt{d},\mathtt{q}}(\mathtt{y}) &= \mathtt{d}'\hat{\mu}\mathtt{q} - \sqrt{(\mathtt{n} - \mathtt{dim}V) C^{3,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V} \|\mathtt{d}\|^2 \mathtt{q}'\hat{\Sigma}\mathtt{q}} \\ b^{\alpha}_{\mathtt{d},\mathtt{q}}(\mathtt{y}) &= \mathtt{d}'\hat{\mu}\mathtt{q} + \sqrt{(\mathtt{n} - \mathtt{dim}V) C^{3,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V} \|\mathtt{d}\|^2 \mathtt{q}'\hat{\Sigma}\mathtt{q}} \end{aligned}$$

Demostración.
Podemos suponer, sin pérdida de generalidad, que $\mathtt{d}, \mathtt{q} \neq 0$. Si $\mu \in V$ y $\Sigma > 0$, se verifica, para todo par $(\mathtt{d}, \mathtt{q})$,

$$\left[\mathtt{d}'\mu\mathtt{q} \in \mathtt{d}'\hat{\mu}\mathtt{q} \pm \sqrt{(\mathtt{n} - \mathtt{dim}V) C^{3,\alpha} \|d\|^2 \mathtt{q}'\hat{\Sigma}\mathtt{q}}\right] \Leftrightarrow \left[\frac{[\mathtt{d}'(\hat{\mu} - \mu)\mathtt{q}]^2}{(\mathtt{n} - \mathtt{dim}V)\|\mathtt{d}\|^2 \mathtt{q}'\hat{\Sigma}\mathtt{q}} \leq C^{3,\alpha}\right].$$

En consecuencia, se tiene que

$$P\left(\mathtt{d}'\mu\mathtt{q} \in \mathtt{d}'\hat{\mu}\mathtt{q} \pm \sqrt{(\mathtt{n}-\mathtt{dim}V)C^{3,\alpha}\|\mathtt{d}\|^2\mathtt{q}'\hat{\Sigma}\mathtt{q}},\ \forall(\mathtt{d},\mathtt{q})\right) = P\left(s \leq C^{3,\alpha}\right) = 1-\alpha.$$

Dado que t_1 es, por definición, la primera raíz del polinomio en t

$$p(t) = \left|\mathtt{y}'P_{V|W}\mathtt{y} - t(\mathtt{n}-\mathtt{dim}V)\hat{\Sigma}\right|,$$

que coincide con $|\hat{\mu}'P_{V|W}\hat{\mu} - t(\mathtt{n}-\mathtt{dim}V)\hat{\Sigma}|$. Se verifica, por un razonamiento completamente análogo al del lema 2.12,

$$t_1 = \sup_{\mathtt{d}\in V|W\setminus\{0\},\mathtt{q}\in\mathbb{R}^p\setminus\{0\}} \frac{(\mathtt{d}'\hat{\mu}\mathtt{q})^2}{(\mathtt{n}-\mathtt{dim}V)\|\mathtt{d}\|^2\mathtt{q}'\hat{\Sigma}\mathtt{q}}.$$

Entonces, si ϕ_3 es el test de Roy a nivel α, se tiene que ϕ_3 toma el valor 0 si, y sólo si,

$$(\mathtt{d}'\hat{\mu}\mathtt{q})^2 \leq (\mathtt{n}-\mathtt{dim}V)C^{3,\alpha}\|\mathtt{d}\|^2\mathtt{q}'\hat{\Sigma}\mathtt{q},\ \forall(\mathtt{d},\mathtt{q}).$$

La segunda proposición equivale a afirmar que $0 \in \mathtt{d}'\hat{\mu}\mathtt{q} \pm \sqrt{(\mathtt{n}-\mathtt{dim}V)C^{3,\alpha}\|\mathtt{d}\|^2\mathtt{q}'\hat{\Sigma}\mathtt{q}}$. En conclusión, se tiene que $\phi_3 = 1$, es decir, se decide que $\mu \notin W$ si, y solo si, existe un par $(\mathtt{d},\mathtt{q})$ tal que $0 \notin \left(a^{\alpha}_{\mathtt{d},\mathtt{q}}(\mathtt{y}), b^{\alpha}_{\mathtt{d},\mathtt{q}}(\mathtt{y})\right)$. Luego, efectivamente, el test de Roy es consistente con los intervalos de confianza simultáneos para $V|W \times \mathbb{R}^p$. De manera análoga (cuestión propuesta), podemos construir, basándonos en la distribución nula del primer autovalor, una familia de elipsoides de confianza simultáneos para $V|W$, de manera que el test de Roy es también consistente con los mismos.

2.6. Estudio Asintótico del Modelo

En esta sección estudiaremos el comportamiento de los estimadores y tests de hipótesis considerados hasta el momento a medida que el tamaño de la muestra, $\mathtt{n}$, converge a infinito. Una de las ventajas que reportará este estudio será la de poder sustituir los las distribuciones $C^i_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V}$, $i = 1,2,3,4$, por la $\chi^2_{p\mathtt{dim}V|W}$, siempre y cuando la muestra sea los suficientemente grande. Veremos también que estos sucede incluso prescindiendo del supuesto de p-normalidad. Aunque damos por conocidos los elementos y resultados fundamentales de la teoría asintótica, exponemos a continuación las definiciones de convergencias probabilidad y en distribución, así como un compendio de proposiciones relativas a las mismas. En todo caso, el lector

puede recabar información sobre el tema en cualquier referencia clásica sobre Probabilidad[24]. También puede encontrar un breve resumen en el Apéndice del primer volumen.

Sean $(X_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ y X variables aleatorias de $(\Omega, \mathcal{A}, P)$ en $\mathbb{R}^p$. Se dice que $(X_{\mathtt{n}})_{\mathtt{n}}$ converge en probabilidad X (se denota $X_{\mathtt{n}} \xrightarrow{P} X$) cuando para todo $\varepsilon > 0$, $P(\|X_{\mathtt{n}} - X\|) > \epsilon)$ converge a 0. Si $(P_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ y P_0 son probabilidades sobre $(\mathbb{R}^p, \mathcal{R}^p)$, se dice que $(P_{\mathtt{n}})_{\mathtt{n}}$ converge en distribución a P_0, en cuyo caso se denota $P_{\mathtt{n}} \xrightarrow{d} P_0$, cuando $E_{P_{\mathtt{n}}}(f)$ converge a $E_{P_0}(f)$, para toda función f de $\mathbb{R}^p$ en $\mathbb{R}$ medible y continua P_0-c.s. y acotada. Se dice que $(X_{\mathtt{n}})_{\mathtt{n}}$ converge en distribución a X cuando $\left(P^{X_{\mathtt{n}}}\right)_{\mathtt{n}}$ converge en distribución a P^X.

Teorema 2.14. (i) La convergencia en distribución equivale a la convergencia de las respectivas funciones características en todo punto de $\mathbb{R}^p$.

(ii) Si $p = 1$, la convergencia en distribución de $(X_{\mathtt{n}})_{\mathtt{n}}$ a X equivale a la convergencia de las respectivas funciones de distribución $F_{\mathtt{n}}$ a la función de distribución de X en cada punto de continuidad de esta última. En ese caso, si, además, $F_{\mathtt{n}}$ es continua, para cada $\mathtt{n} \in \mathbb{N}$ [25], se da también una convergencia entre las funciones inversas.

(iii) La convergencia en probabilidad implica convergencia en distribución.

(iv) Si dos sucesiones de variables aleatorias convergen en probabilidad a sendas constantes, las sucesiones de las sumas y productos convergen, respectivamente, a la suma y producto de dichas constantes.

(v) La convergencia en distribución a una constante implica convergencia en probabilidad.

(vi) Si $f \in \mathcal{C}(\mathbb{R}^p)$ y $(X_{\mathtt{n}})_{\mathtt{n}}$ converge en distribución a X, $(f(X_{\mathtt{n}}))_{\mathtt{n}}$ converge en distribución a $f(X)$.

(vii) Si f es continua en a y $(X_{\mathtt{n}})_{\mathtt{n}}$ converge en distribución a una constante a, $(f(X_{\mathtt{n}}))_{\mathtt{n}}$ converge en distribución a $f(a)$.

(viii) Si $(X_{\mathtt{n}})_{\mathtt{n}}$, $(U_{\mathtt{n}})_{\mathtt{n}}$ y $(V_{\mathtt{n}})_{\mathtt{n}}$ convergen en distribución a X, a (cte.) y 1, respectivamente,

(a) $X_{\mathtt{n}} + U_{\mathtt{n}} \xrightarrow{d} X + a$.

[24] Ver, por ejemplo Billingsley (1986) o Ash (1972).

[25] En ese caso podemos hablar de la inversas de cada una de ellas

(b) $X_{\mathtt{n}} \cdot U_{\mathtt{n}} \xrightarrow{d} aX$

(c) $\frac{X_{\mathtt{n}}}{V_{\mathtt{n}}} \xrightarrow{d} X$

(ix) Si $(P_{\mathtt{n}})_{\mathtt{n}}$ y $(Q_{\mathtt{n}})_{\mathtt{n}}$ convergen en distribución a P y Q, respectivamente, $(P_{\mathtt{n}} \times Q_{\mathtt{n}})_{\mathtt{n}}$ converge en distribución a $P \times Q$.

Hasta ahora hemos trabajado con modelos en el cual el término $\mathtt{n}$ es fijo. Es lo que se denomina Modelo Exacto. Teniendo en cuenta que la Teoría Asintótica tiene como objeto estudiar la evolución de los distintos estimadores y tests de hipótesis en función de $\mathtt{n}$, es necesario construir un nuevo modelo, denominado Asíntótico, que, por así decirlo, englobe todos los experimentos exactos. En nuestro caso se definiría com sigue. Dada una sucesión $(V_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ de subespacios $\mathtt{v}$-dimensionales de $\mathbb{R}^n$, respectivamente, consideraremos el experimento estadístico constituido por una sucesión $(Z_i)_{i\in\mathbb{N}}$ de vectores aleatorios que se descomponen de la siguiente forma

$$Z_i = \mu(i) + f_i, \quad i \in \mathbb{N},$$

donde $\mu(i) \in \mathbb{R}^p$ y $(f_i)_{i\in\mathbb{N}}$ es una secuencia de vectores aleatorios p-dimensionales independientes e idénticamente distribuidas con media 0 y matriz e varianzas-covarianzas $\Sigma > 0$, y de tal forma que, para cada $\mathtt{n} \in \mathbb{N}$, la matriz $\mu_{\mathtt{n}} = (\mu(1), \ldots, \mu(\mathtt{n}))'$ pertenece al subespacio $V_{\mathtt{n}}$. De esta forma, si se denota $Y_{\mathtt{n}} = (Z_1, \ldots, Z_{\mathtt{n}})'$ y $e_{\mathtt{n}} = (f_1, \ldots, f_{\mathtt{n}})'$, tendremos

$$Y_{\mathtt{n}} = \mu_{\mathtt{n}} + e_{\mathtt{n}}, \quad \mu_{\mathtt{n}} \in V_{\mathtt{n}}, \quad e_{\mathtt{n}} \sim \mathcal{P}_n,$$

siendo $\mathcal{P}_{\mathtt{n}}$ el conjunto de las matriz aleatorias $\mathtt{n} \times p$ cuyas filas constituyen vectores aleatorios independientes, de media 0 y matriz de varianzas-covarianzas común y definida positiva. Nótese que, para cada $n \in \mathbb{N}$, tenemos un Modelo Exacto en dimensión $\mathtt{n}$, en el cual tiene sentido hablar de los estimadores

$$\hat{\mu}_{\mathtt{n}} = P_{V_{\mathtt{n}}} Y_{\mathtt{n}}, \qquad \hat{\Sigma}_{\mathtt{n}} = (\mathtt{n} - \mathtt{v})^{-1} Y_{\mathtt{n}}' P_{V_{\mathtt{n}}^{\perp}} Y_{\mathtt{n}}.$$

Si admitiéramos la normalidad de la distribución de f_1, estaríamos hablando de un modelo lineal normal multivariante. Así mismo y en lo que respecta al problema de contraste de hipótesis, si consideramos una secuencia $(W_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ de subespacios $\mathtt{w}$-dimensionales de $(V_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$, respectivamente, tendrá sentido hablar, para cada $\mathtt{n} \in \mathbb{N}$, del las raíces positivas $t_1(\mathtt{n}), \ldots, t_b(\mathtt{n})$ del polinomio $p_{\mathtt{n}}(t) = |S_2(\mathtt{n}) - tS_3(\mathtt{n})|$, donde $S_2(\mathtt{n}) = Y_{\mathtt{n}}' P_{V_{\mathtt{n}}|W_{\mathtt{n}}} Y_{\mathtt{n}}$, $S_3(\mathtt{n}) = Y_{\mathtt{n}}' P_{V_{\mathtt{n}}^{\perp}} Y_{\mathtt{n}}$ y $b = \text{mín}\{\mathtt{v} - \mathtt{w}, p\}$. $S_2(\mathtt{n})$ y $S_3(\mathtt{n})$ pueden expresarse, al igual que en a sección 2, mediante $Z_2(\mathtt{n})'Z_2(\mathtt{n})$ y $Z_3(\mathtt{n})'Z_3(\mathtt{n})$, respectivamente. De esta forma, $t_1(\mathtt{n}), \ldots, t_b(\mathtt{n})$, son también los autovalores positivos de la

matriz $Z_2(\mathtt{n})S_3(\mathtt{n})^{-1}Z_2(\mathtt{n})'$. El término $\delta(\mathtt{n})$ denotará la matriz $\mu(\mathtt{n})'P_{V_\mathtt{n}|W_\mathtt{n}}\mu(\mathtt{n})$. Por último, consideraremos, para cada $n \in \mathbb{N}$ los siguientes estadísticos:

$$\begin{aligned}
\lambda_1(\mathtt{n}) &= \prod_{i=1}^{b}[1+t_i(\mathtt{n})]^{-1},\\
\lambda_2(\mathtt{n}) &= \sum_{i=1}^{b} t_i(\mathtt{n}),\\
\lambda_3(\mathtt{n}) &= t_i(\mathtt{n}),\\
\lambda_4(\mathtt{n}) &= \sum_{i=1}^{b} \frac{t_i(\mathtt{n})}{1+t_i(\mathtt{n})}.
\end{aligned}$$

Nótese que, al contrario de lo que sucede en el modelo lineal normal exacto, el modelo lineal asintótico no queda parametrizado por un vector media, μ, y una matriz de varianzas-covarianzas Σ. Si acaso, podríamos hablar de una sucesión de medias $(\mu_\mathtt{n})_{\mathtt{n}\in\mathbb{N}}$ y una matriz Σ. Por ello, tiene aquí sentido hablar de una secuencia de estimadores consistente para Σ, pero no para μ. Este problema, que afecta al estudio de Estimación, podría resolverse si consideráramos el modelo asintótico que resulta de imponer a $(\mu_\mathtt{n})_{\mathtt{n}\in\mathbb{N}}$ la siguiente restricción: suponer que existe una sucesión $(\mathtt{X}_\mathtt{n})_{\mathtt{n}\in\mathbb{N}}$ de bases de $(V_\mathtt{n})_{\mathtt{n}\in\mathbb{N}}$, de manera que $(\mu_\mathtt{n})_{\mathtt{n}\in\mathbb{N}}$ verifica

$$\exists \beta \in \mathcal{M}_{\mathtt{v}\times p} :\ \mu_\mathtt{n} = \mathtt{X}_\mathtt{n}\beta,\ \forall \mathtt{n} \in \mathbb{N}. \tag{2.7}$$

De esta forma, sí tendría sentido hablar de una secuencia de estimadores consistente para β. Consideremos, concretamente, la secuencia definida mediante

$$\hat{\beta}_\mathtt{n} = (\mathtt{X}_\mathtt{n}'\mathtt{X}_\mathtt{n})^{-1}\mathtt{X}_\mathtt{n}'Y_\mathtt{n}, \qquad \mathtt{n} \in \mathbb{N}.$$

Hagamos a continuación un inciso sobre una cuestión de carácter matricial. Dada una matriz (se admiten vectores) $A \in \mathcal{M}_{m\times k}$, de componentes a_{ij}, se define $m(A)$ como $\max_{i,j}|a_{ij}|$. Si A es una matriz cuadrada de orden m, simétrica y semi definida positiva, existe, en virtud del teorema 13.5, una matriz B con las misma dimensiones tales que $A = B'B$. Teniendo en cuenta este hecho junto con la desigualdad de Cauchy-Swartz, de puede probar que $m(A) = \max_i |a_{ii}|$. También se verifica, trivialmente, que si $A \in \mathcal{M}_{m\times k}$ y $B \in \mathcal{M}_{k\times r}$,

$$m(AB) \leq km(A)m(B), \qquad (m(A))^2 \leq m(AA'). \tag{2.8}$$

Teniendo en cuenta (2.8) junto con el teorema 13.4, se deduce que, si A es una matriz simétrica de orden k y D es la matriz diagonal constituida por sus autovalores, entonces

$$1/k^2 m(D) \leq m(A) \leq k^2 m(D). \tag{2.9}$$

En primer lugar probaremos que la secuencias de estimadores $(\hat{\beta}_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ y es consistentes. Más adelante se probará lo mismo para el estimador de Σ. Nótese que, hasta el momento, se ha prescindido del supuesto de normalidad.

Teorema 2.15.
Si se verifica (2.7) y $m(\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}}) \to \infty$, la secuencia $(\hat{\beta}_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ es consistente.

Demostración.
Tener en cuenta, primeramente, que

$$\mathtt{E}\left[\hat{\beta}_{\mathtt{n}}\right] = \beta, \qquad \mathtt{Covm}\left[\hat{\beta}_{\mathtt{n}}\right] = (\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}})^{-1} \otimes \Sigma, \quad \forall \mathtt{n} \in \mathbb{N}.$$

Por lo tanto, dado $\varepsilon > 0$, se sigue de la Desigualdad de Chebyshev que

$$P\big(\big\|\mathtt{vec}(\hat{\beta}_{\mathtt{n}}) - \mathtt{vec}(\beta)\big\| > \varepsilon\big) \leq \frac{\sqrt{p \cdot \mathtt{v}} \cdot m(\Sigma) \cdot m((\mathtt{X}_{n}'\mathtt{X}_{\mathtt{n}})^{-1})}{\varepsilon}.$$

Sea $D_{\mathtt{n}}$ la matriz diagonal de los autovalores de $\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}}$, para cada $\mathtt{n} \in \mathbb{N}$. Por el teorema 13.4, la matriz de los autovalores de $(\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}})^{-1}$ será $D_{\mathtt{n}}^{-1}$. Luego, teniendo en cuenta (2.9), se verifica que $m((\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}})^{-1}) \to 0$, lo cual concluye la prueba.

∎

La astucia de Cramer-Wold permite probar la convergencia a la distribución normal multivariante a partir de las convergencias a la normal univariante de cualquier proyección. El resultado se extiende de manera natural al caso matricial.

Teorema 2.16.
Una secuencias $(X_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ de matrices aleatorias de dimensiones $m \times q$ converge a la distribución $N_{m,q}(\theta, \Gamma, \Sigma)$ si, y sólo si, para cada $q \times m$- matriz A, la secuencia $\big(\mathtt{tr}(A'X_{\mathtt{n}})\big)_{\mathtt{n}\in\mathbb{N}}$ converge en distribución a $N\big(\mathtt{tr}(A'\theta), \mathtt{tr}(A'\Xi, A\Sigma)\big)$.

Demostración.
La primera implicación se sigue del problema 12 del capítulo 1 junto con el teorema 2.14-(vi). El recíproco se obtiene valorando las funciones características de la secuencia $\big(\mathtt{tr}(A'X_{\mathtt{n}})\big)_{\mathtt{n}}$ en el punto $t = 1$ y aplicando el teorema 2.14-(i).

∎

En el capítulo 3 del primer volumen se realiza una adaptación del Teorema Límite Central, versión de Lidemberg-Feller, al estudio asintótico del modelo lineal univariante. Veamos una extensión natural de dicho resultado al caso multivariante.

Teorema 2.17.
Si $(A_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ es una secuencia de matrices constantes, de dimensiones $\mathtt{n} \times p$, respectiva-

mente, y tales que $\mathtt{tr}(A_{\mathtt{n}}\Sigma A_{\mathtt{n}}') = 1$, para todo $\mathtt{n} \in \mathbb{N}$ y $m(A_{\mathtt{n}}) \to 0$, entonces

$$\mathtt{tr}(A_{\mathtt{n}}' e_{\mathtt{n}}) \xrightarrow{d} N(0,1).$$

Demostración.
Si, para cada $\mathtt{n} \in \mathbb{N}$, se denota $m_{\mathtt{n}} = m(A_{\mathtt{n}}\Sigma A_{\mathtt{n}}')$, se deduce de (2.8) que $m_{\mathtt{n}} \leq p^2 m(\Sigma) m(A_{\mathtt{n}})^2$ y, por lo tanto, converge a 0. Consideremos la descomposición $A_{\mathtt{n}}' = (a_{\mathtt{n}1}, \ldots, a_{\mathtt{nn}})$ y sea $X_{\mathtt{n}i} = a_{\mathtt{n}i}' f_i$, , para cada $i = 1, \ldots, \mathtt{n}$. De esta forma, se verifica que $\mathtt{tr}(A_{\mathtt{n}}' e_{\mathtt{n}}) = \sum_{i=1}^{\mathtt{n}} X_{\mathtt{n}i}$, siendo todas las variables $X_{\mathtt{n}i}$, $i = 1, \ldots, \mathtt{n}$, independientes, por serlo las f_i. Además,

$$\mathtt{E}[X_{\mathtt{n}i}] = 0, \quad \sum_{i=1}^{\mathtt{n}} \mathtt{var}[X_{\mathtt{n}i}] = \sum_{i=1}^{\mathtt{n}} a_{\mathtt{n}i}' \Sigma\, a_{\mathtt{n}i} = \mathtt{tr}(A_{\mathtt{n}}\Sigma A_{\mathtt{n}}') = 1.$$

Dado $\varepsilon > 0$, denótese, para cada $\mathtt{n} \in \mathbb{N}$,

$$C_{\mathtt{n}} = \mathtt{E}\left[\sum_{i=1}^{\mathtt{n}} X_{\mathtt{n}i}^2 I_{\varepsilon}(X_{\mathtt{n}i})\right].$$

Si demostramos que $(C_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ converge a 0, estaremos en las condiciones del Teorema de Lindemberg-Feller, por lo quedará probada la tesis. Efectivamente, podemos expresar $X_{\mathtt{n}i}^2$ mediante $[(\Sigma^{1/2} a_{\mathtt{n}i})'(\Sigma^{-1/2} f_i)]^2$. Luego, de la desigualdad de Cauchy-Swartz se sigue que $X_{\mathtt{n}i}^2 \leq (a_{\mathtt{n}i}'\Sigma\, a_{\mathtt{n}i})(f_i'\Sigma^{-1} f_i)$, para todo $i = 1, \ldots, \mathtt{n}$. En consecuencia, si se denota a $U = f_1 \Sigma^{-1} f_1$, se verifica

$$0 \leq C_{\mathtt{n}} \leq \mathtt{E}[U I_{\varepsilon^2/m_{\mathtt{n}}}(U)].$$

Dado que, en todo caso $U \cdot I_{\varepsilon^2/m_{\mathtt{n}}}(U)$, siendo ésta una función cuya integralvale p [26], se sigue del Teorema de la Convergencia Dominada que

$$\lim_{\mathtt{n}\to\infty} U I_{\varepsilon^2/m_{\mathtt{n}}}(U) = \mathtt{E}\big[\lim_{\mathtt{n}\to\infty} U I_{\varepsilon^2/m_{\mathtt{n}}}(U)\big].$$

El integrando del segundo término converge puntualmente a 0, pues $m_{\mathtt{n}} \to 0$. Por lo tanto, $(C_{\mathtt{n}})_{\mathtt{n}}$ converge igualmente a 0.

∎

Lema 2.18.

Sea $(\Gamma_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ una secuencia de matrices de dimensión $\mathtt{n} \times u$, respectivamente, tales que $\Gamma_{\mathtt{n}}'\Gamma_{\mathtt{n}} = \mathtt{Id}$, para todo $\mathtt{n} \in \mathbb{N}$, y $m(\Gamma_{\mathtt{n}}\Gamma_{\mathtt{n}}') \to 0$. Entonces $\Gamma_{\mathtt{n}}' e_{\mathtt{n}}$ converge en distribución a $N_{u,p}(0, \mathtt{Id}, \Sigma)$.

[26] Esto se deduce del hecho de que, si X es un vector aleatorio m-dimensional de cuadrado integrable, entonces $\mathtt{E}[\|X\|^2] = \|\mathtt{E}[X]\|^2 + \mathtt{tr}(\mathtt{Cov}[X])$.

Demostración.
En virtud del teorema 2.17, la tesis equivale a que, para cada matriz $A \in \mathcal{M}_{u\times p}$, se verifique que $\mathtt{tr}(A'\Gamma_{\mathtt{n}}'e_{\mathtt{n}})$ converja en distribución a $N\big(0, \mathtt{tr}(A\Sigma A')\big)$. Si se define $A_{\mathtt{n}} = [\mathtt{tr}(A\Sigma A')]^{-1/2}\Gamma_{\mathtt{n}}A$, se tiene que

$$[\mathtt{tr}(A\Sigma A')]^{-1/2}\mathtt{tr}(A'\Gamma_{\mathtt{n}}e_{\mathtt{n}}) = \mathtt{tr}(A_{\mathtt{n}}'e_{\mathtt{n}}), \qquad \mathtt{tr}(A_{\mathtt{n}}\Sigma A_{\mathtt{n}}') = 1.$$

Además, se sigue de (2.8) que

$$m(A_{\mathtt{n}}) \leq p[\mathtt{tr}(A\Sigma A')]^{-1/2}m(A)m(\Gamma_{\mathtt{n}}\Gamma_{\mathtt{n}}')^{1/2},$$

que converge a 0. Luego, por el teorema anterior, se concluye.

∎

De este resultado se sigue un interesante corolario para el estimador de β cuando no se supone normalidad.

Teorema 2.19.

Supongamos que se verifica (2.7) y que, además,

$$m(\mathtt{X}_{\mathtt{n}}(\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}})^{-1}\mathtt{X}_{\mathtt{n}}) \to 0. \tag{2.10}$$

Entonces, $(\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}})^{1/2}\big(\hat{\beta}_{\mathtt{n}} - \hat{\beta}\big) \xrightarrow{d} N_{\mathtt{v}}(0, \mathtt{Id}, \Sigma)$.

Demostración.
Si, para cada $\mathtt{n} \in \mathbb{N}$, consideramos la matriz $\Gamma_{\mathtt{n}} = \mathtt{X}_{\mathtt{n}}(\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}})^{-1/2}$, entonces $(\Gamma_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ satisface las hipótesis del lema anterior con $u = \mathtt{v}$. En consecuencia $\Gamma_{\mathtt{n}}e_{\mathtt{n}}$ converge en distribución a $N_{\mathtt{v},p}(0, \mathtt{Id}, \Sigma)$. Teniendo en cuenta que

$$\hat{\beta}_{\mathtt{n}} - \beta = (\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}})^{-1}\mathtt{X}_{\mathtt{n}}'(Y_{\mathtt{n}} - \mu_{\mathtt{n}}),$$

se deuce que $(\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}})^{1/2}\big(\hat{\beta}_{\mathtt{n}} - \beta\big) = \Gamma_{\mathtt{n}}'e_{\mathtt{n}}$, lo cual acaba la prueba.

∎

Nótese que, del teorema anterior se sigue que, para $\mathtt{n}$ suficientemente grande, el estimador de β sigue *aproximadamente* un modelo de distribución $N_{\mathtt{v}}\big(\beta, (\mathtt{X}_{\mathtt{n}}'\mathtt{X}_{\mathtt{n}})^{-1}, \Sigma\big)$. En ese sentido podemos decir que el la proposición (i) del teorema 2.5 es asintóticamente válida aunque no se verifique el supuesto de normalidad, siempre y cuando se satisfaga la condición (2.10). Veamos otro corolario del anterior lema.

Corolario 2.20.

Si $(E_{\mathtt{n}})_{\mathtt{n}}$ es una secuencia de subespacios k-dimensionales de $\mathbb{R}^{\mathtt{n}}$, respectivamente, tales que $m(P_{E_{\mathtt{n}}}) \to 0$, entonces $e_{\mathtt{n}}'P_{E_{\mathtt{n}}}e_{\mathtt{n}}$ converge en distribución a $W_p(k, \Sigma)$.

Demostración.
basta considera una base ortonormal de cada subespacio y aplicar el lema 2.18 junto cone el teorema 2.14-(vi).

∎

Corolario 2.21.

Si $m(P_{V_{\mathtt{n}}}) \to 0$ y $\mu(\mathtt{n}) \in W_{\mathtt{n}}$, para todo $\mathtt{n} \in \mathbb{N}$, entonces

$$\mathtt{tr}\left(S_2(\mathtt{n})\Sigma^{-1}\right) \xrightarrow{d} \chi^2_{p(\mathtt{v}-\mathtt{w})}.$$

Demostración.
En primer lugar, $m(P_{V_{\mathtt{n}}|W_{\mathtt{n}}}) \to 0$, pues $W_{\mathtt{n}} \subset V_{\mathtt{n}}$, para todo $\mathtt{n} \in \mathbb{N}$. Además, como $\mu_{\mathtt{n}} \in W_{\mathtt{n}}$, se tiene que que $S_2(\mathtt{n}) = e'_{\mathtt{n}} P_{V_{\mathtt{n}}|W_{\mathtt{n}}} e_{\mathtt{n}}$, para todo $\mathtt{n} \in \mathbb{N}$. Luego, del corolario anterior junto con el problema 16 del capítulo 1, se sigue la tesis.

∎

Estamos en condiciones de determinar condiciones que garanticen la consistencia del estimador de Σ.

Teorema 2.22.

Si se verifica la condición

$$m(P_{V_{\mathtt{n}}}) \to 0, \tag{2.11}$$

la secuencia de estimadores $(\hat{\Sigma}_{\mathtt{n}})_{\mathtt{n}\in\mathbb{N}}$ es consistente.

Demostración.
Se verifica

$$\frac{\mathtt{n}-\mathtt{v}}{\mathtt{n}}\hat{\Sigma} = \frac{1}{\mathtt{n}} e'_{\mathtt{n}} P_{V_{\mathtt{n}}^{\perp}} e_{\mathtt{n}} = \frac{1}{\mathtt{n}} e'_{\mathtt{n}} e_{\mathtt{n}} - \frac{1}{\mathtt{n}} e'_{\mathtt{n}} P_{V_{\mathtt{n}}} e_{\mathtt{n}}.$$

Dado que

$$\frac{1}{\mathtt{n}} e'_{\mathtt{n}} e_{\mathtt{n}} = \frac{1}{\mathtt{n}} \sum_{i=1}^{\mathtt{n}} f_i f'_i$$

y teniendo en cuenta que $\mathtt{E}[f_i f'_i] = \Sigma$, para todo $i \in \mathbb{N}$, se sigue de la Ley Débil de los Grandes Números que $\mathtt{n}^{-1} e'_{\mathtt{n}} e_{\mathtt{n}}$ converge en probabilidad a Σ. Por otra parte, sabemos por el corolario 2.20 que $e'_{\mathtt{n}} P_{V_{\mathtt{n}}} e_{\mathtt{n}} \xrightarrow{d} W_p(\mathtt{v}, \Sigma)$. Luego, pr el teorema 2.14-(viii), se deduce que $\mathtt{n}^{-1} e'_{\mathtt{n}} P_{V_{\mathtt{n}}} e_{\mathtt{n}} \xrightarrow{P} 0$, con lo cual se concluye.

∎

Denotense por $s_1(\mathtt{n}), \ldots, s_b(\mathtt{n})$ las raíces positivas ordenadas del polinomio $q_{\mathtt{n}}(s) = |S_2(\mathtt{n}) - s\Sigma|$. Dada una matriz R $p \times p$ semidefinida positiva y dos matrices U, V de dimensión $p \times p$ y definidas positivas, se define $q_i(R, U, V)$ como la i-ésima raíz del

polinomio $p_1(x) = |R - xU|$ [27], y $r_i(R, U, V)$ como la i-ésima raíz del polinomio $p_2(x) = |R - xV|$ [28], para $i = 1, \ldots, b$. Se define entonces, para $i = 1, \ldots, b$, las funciones $h_i = q_i - r_i$, que son continuas. Se verifica entonces:

Teorema 2.23.
Si se verifica a condición (2.11) y $\mu(\mathtt{n}) \in W_{\mathtt{n}}$, para todo $\mathtt{n} \in \mathbb{N}$, entonces,

$$(\mathtt{n} - \mathtt{v})t_i(\mathtt{n}) - s_i(\mathtt{n}) \xrightarrow{P} 0, \qquad \forall i = 1, \ldots, b. \tag{2.12}$$

Demostración.
En primer lugar, dado que $P_{V_{\mathtt{n}}|W_{\mathtt{n}}} \leq P_{V_{\mathtt{n}}}$, se verifica que $m(P_{V_{\mathtt{n}}|W_{\mathtt{n}}}) \to 0$. Además, dado que $\mu(\mathtt{n}) \in W_{\mathtt{n}}$, se tiene que $e'_{\mathtt{n}} P_{V_{\mathtt{n}}|W_{\mathtt{n}}} e_{\mathtt{n}} = S_2(\mathtt{n})$. Luego, aplicando el corolario 2.20 por una parte y el teorema 2.22 por otra, se tiene que

$$S_2(\mathtt{n}) \xrightarrow{d} W_p(\mathtt{v} - \mathtt{w}, \Sigma), \qquad \frac{1}{\mathtt{n} - \mathtt{v}} S_3(\mathtt{n}) \xrightarrow{P} \Sigma.$$

Dado que $(\mathtt{n} - \mathtt{v})t_i(\mathtt{n}) = q_i\big(S_2(\mathtt{n}), \frac{1}{\mathtt{n}-\mathtt{v}} S_3(\mathtt{n}), \Sigma\big)$ y $s_i(\mathtt{n}) = r_i\big(S_2(\mathtt{n}), \frac{1}{\mathtt{n}-\mathtt{v}} S_3(\mathtt{n}), \Sigma\big)$, se verifica que $(\mathtt{n} - \mathtt{v})t_i(\mathtt{n}) - s_i(\mathtt{n}) = h_i\big(S_2(\mathtt{n}), \frac{1}{\mathtt{n}-\mathtt{v}} S_3(\mathtt{n}), \Sigma\big)$ y, en virtud del teorema 2.14(ix)+(vii)+(v),

$$(\mathtt{n} - \mathtt{v})t_i(\mathtt{n}) - s_i(\mathtt{n}) \xrightarrow{P} h_i\left(W_p(\mathtt{v} - \mathtt{w}, \Sigma), \Sigma, \Sigma\right) = 0,$$

lo cual acaba la prueba. ∎

Hasta ahora hemos supuesto que $(f_i)_{i \in \mathbb{N}}$ es una muestra aleatoria simple infinita correspondiente a una distribución p-dimensional, cuyas componentes poseen media 0 y cuadrado integrable. En el caso p-normal, los estimadores considerados son de máxima verosimilitud, lo cual confiere una excelente justificación desde el punto de vista asintótico. Lo mismo puede decirse del test de Wilks. No obstante, veamos cuál es la distribución asintótica de los cuatro tests considerados para el contraste de la media. Empezaremos obteniendo un resultado ligeramente más preciso que el del teorema anterior. En lo que sigue, distinguiremos entre las dos hipótesis siguientes:

$$m(V_{\mathtt{n}}) \to 0, \qquad \mu(\mathtt{n}) \in W_{\mathtt{n}}, \ \forall \mathtt{n} \in \mathbb{N}. \tag{2.13}$$

$$f_1 \text{ Normal}, \qquad \delta(\mathtt{n}) = \delta, \ \forall \mathtt{n} \in \mathbb{N}. \tag{2.14}$$

Nótese que la hipótesis $\mu_(\mathtt{n}) \in W_{\mathtt{n}}$ equivale a $\delta(\mathtt{n}) = 0$. Por lo tanto, la segunda proposición de la hipótesis (2.13) puede expresarse mediante $\delta(\mathtt{n}) = \delta$, para todo $\mathtt{n} \in \mathbb{N}$ con $\delta = 0$.

[27] Si $R = Q'Q$, donde Q es una matriz $(\mathtt{v} - \mathtt{w}) \times p$, se trata del i-ésimo autovalor de $QU^{-1}Q'$.
[28] Idem.

Lema 2.24.
(2.14) implica (2.12).

Demostración.
La demostración es completamente análoga a la del teorema anterior, con la salvedad de que $S_2(\mathtt{n})$ sigue, en todo caso, una distribución $W_p(\mathtt{v}-\mathtt{w},\Sigma,\delta)$. Por lo tanto, se verifica, para todo $i=1,\ldots,b$, que

$$(\mathtt{n}-\mathtt{v})t_i(\mathtt{n})-s_i(\mathtt{n})\xrightarrow{P}h_i\left(W_p(\mathtt{v}-\mathtt{w},\Sigma,\delta),\Sigma,\Sigma\right),$$

que sigue siendo igual a 0.

∎

Definamos una nueva distribución: dados $m,n\in\mathbb{N}$, y $\eta\in\mathcal{M}_{n\times n}$ semidefinida positiva, respectivamente, se define $U^i_{n,m}(\eta)$, para cada $i=1,\ldots,\text{mín}\{n,m\}$, como la distribución del i-ésimo autovalor de una matriz aleatoria $n\times n$ que sigue un modelo de distribución $W_n\big(m,\mathtt{Id},\eta\big)$. En el caso $\eta=0$ se denotará $U^i_{n,m}$. Por lo tanto, se deduce del problema 16 del capítulo 1, que, tanto (2.13) como (2.14) implican

$$s_i(\mathtt{n})\xrightarrow{d}U^i_{p,\mathtt{v}-\mathtt{w}}(\eta),\quad\forall i=1,\ldots,b,\tag{2.15}$$

con $\eta=\Sigma^{-1/2}\delta\Sigma^{-1/2}$. En el caso (2.13), se verifica, en particular,

$$s_i(\mathtt{n})\xrightarrow{d}U^i_{p,\mathtt{v}-\mathtt{w}},\qquad i=1,\ldots,b.\tag{2.16}$$

Lema 2.25.
Tanto (2.13) como (2.14) implican las siguientes proposiciones:

(i) $(\mathtt{n}-\mathtt{v})\lambda_2(\mathtt{n})-\mathtt{tr}\left(S_2(\mathtt{n})\Sigma^{-1}\right)\xrightarrow{P}0$

(ii) $(\mathtt{n}-\mathtt{v})\lambda_4(\mathtt{n})-\mathtt{tr}\left(S_2(\mathtt{n})\Sigma^{-1}\right)\xrightarrow{P}0$

(iii) $-(\mathtt{n}-\mathtt{v})\log\lambda_1(\mathtt{n})-\mathtt{tr}\left(S_2(\mathtt{n})\Sigma^{-1}\right)\xrightarrow{P}0$

Demostración.
(i) Se verifica que

$$(\mathtt{n}-\mathtt{v})\lambda_2(\mathtt{n})-\mathtt{tr}\left(S_2(\mathtt{n})\Sigma^{-1}\right)=(\mathtt{n}-\mathtt{v})\lambda_2(\mathtt{n})-\mathtt{tr}\left(Z_2(\mathtt{n})\Sigma^{-1}Z_2(\mathtt{n})'\right),$$

que , au vez, equivale a

$$(\mathtt{n}-\mathtt{v})\sum_{i=1}^{b}t_i(\mathtt{n})-\sum_{i=1}^{b}s_i(\mathtt{n}).$$

Este término convergen probabilida a 0, por hipótesis.

(ii) Se verifica que

$$(\mathtt{n}-\mathtt{v})t_i(\mathtt{n}) - U^i_{p,\mathtt{v}-\mathtt{w}}(\eta) = [(\mathtt{n}-\mathtt{v})t_i(\mathtt{n}) - s_i(\mathtt{n})] - [U^i_{p,\mathtt{v}-\mathtt{w}}(\eta) - s_i(\mathtt{n})]\ ^{29}.$$

Luego, teniendo en cuenta el lema 2.23, junto con (2.15) y el teorema 2.14, se sigue que $(\mathtt{n}-\mathtt{v})t_i(\mathtt{n}) \xrightarrow{d} U^i_{p,\mathtt{v}-\mathtt{w}}(\eta)$. Luego, aplicando nuevamente el teorema 2.14, se deduce que $t_i(\mathtt{n})$ converge en probabilidad a 0 y, en consecuencia, $1 + t_i(\mathtt{n}) \xrightarrow{P} 1$. También se sigue del teorema 2.14 que $t_i(\mathtt{n})s_i(\mathtt{n}) \xrightarrow{P} 0$. Dado que

$$(\mathtt{n}-\mathtt{v})\lambda_4(\mathtt{n}) - \mathtt{tr}\left(S_2(\mathtt{n})\Sigma^{-1}\right) = \sum_{i=1}^{b} \frac{(\mathtt{n}-\mathtt{v})t_i(\mathtt{n}) - s_i(\mathtt{n})}{t_i(\mathtt{n})+1} - \sum_{i=1}^{b} \frac{t_i(\mathtt{n})s_i(\mathtt{n})}{1+t_i(\mathtt{n})},$$

se concluye, pues, en virtud del teorema 2.14-(viii), el segundo término converge en probabilidad a 0.

(iii) Veamos una cuestiones previas. Por un desarrollo de Taylor en 0, se tiene que $\log(1+x) = x - \frac{1}{2}(1+h)^{-2}x^2$, con $0 \le h \le x$. Supongamos que $U_{\mathtt{n}} \xrightarrow{d} F$ y $V_{\mathtt{n}} \xrightarrow{P} 0$. Entonces

$$(\mathtt{n}-\mathtt{v})\log\left(1 + \frac{U_{\mathtt{n}}+V_{\mathtt{n}}}{\mathtt{n}-\mathtt{v}}\right) - U_{\mathtt{n}} = V_{\mathtt{n}} - \frac{1}{(1+H_{\mathtt{n}})^2}\frac{(U_{\mathtt{n}}+V_{\mathtt{n}})^2}{\mathtt{n}-\mathtt{v}},$$

donde $0 \le H_{\mathtt{n}} \le (\mathtt{n}-\mathtt{v})^{-1}(U_{\mathtt{n}}+V_{\mathtt{n}})$ Por su parte, $(\mathtt{n}-\mathtt{v})^{-1}(U_{\mathtt{n}}+V_{\mathtt{n}}) \xrightarrow{P} 0$. Luego, $H_{\mathtt{n}} \xrightarrow{P} 0$ y, por lo tanto, $(1+H_{\mathtt{n}})^{-2} \xrightarrow{P} 1$. Además, $(\mathtt{n}-\mathtt{v})^{-1}(U_{\mathtt{n}}+V_{\mathtt{n}})^2 \xrightarrow{P} 0$. En consecuencia,

$$(\mathtt{n}-\mathtt{v})\log\left(1 + \frac{U_{\mathtt{n}}+V_{\mathtt{n}}}{\mathtt{n}-\mathtt{v}}\right) - U_{\mathtt{n}} \xrightarrow{P} 0. \tag{2.17}$$

Pues bien, teniendo en cuenta el lema 2.23 junto con (2.15), consideremos $U_{\mathtt{n}} = s_i(\mathtt{n})$, $F = U^i_{p,\mathtt{v}-\mathtt{w}}(\eta)$ y $V_{\mathtt{n}} = (\mathtt{n}-\mathtt{v})t_i(\mathtt{n}) - s_i(\mathtt{n})$. Se sigue entonces de (2.17) que

$$(\mathtt{n}-\mathtt{v})\log(1+t_i(\mathtt{n})) - s_i(\mathtt{n}) \xrightarrow{P} 0.$$

Teniendo en cuenta que

$$-(\mathtt{n}-\mathtt{v})\log\lambda_1(\mathtt{n}) - \mathtt{tr}\left(S_2(\mathtt{n})\Sigma^{-1}\right) = \sum_{i=1}^{b}\left[(\mathtt{n}-\mathtt{v})\log(1+t_i(\mathtt{n})) - s_i(\mathtt{n})\right],$$

se concluye. ∎

[29] Tomar $\delta = 0$ si se considera (2.13).

Teorema 2.26.
Supongamos que se satisface (2.13) o (2.14). Entonces,se verifica:

(i) Las secuencias $-(\mathtt{n}-\mathtt{v})\log\lambda_1(\mathtt{n})$, $(\mathtt{n}-\mathtt{v})\lambda_2(\mathtt{n})$ y $(\mathtt{n}-\mathtt{v})\lambda_4(\mathtt{n})$ convergen en distribución a $\chi^2_{p(\mathtt{v}-\mathtt{w})}\left(\mathtt{tr}(\delta\Sigma^{-1})\right)$, con $\delta=0$ en el caso (2.13).

(ii) Si $\delta=0$, $(\mathtt{n}-\mathtt{v})\lambda_3(\mathtt{n})$ converge en distribución a $U^1_{p,\mathtt{v}-\mathtt{w}}$.

Demostración.
(i) Lo probaremos únicamente para λ_1, pues en el caso de λ_2 y λ_4 el razonamiento es completamente análogo. Consideremos la siguiente descomposición:

$$-(\mathtt{n}-\mathtt{v})\log\lambda_1(\mathtt{n})=\left[-(\mathtt{n}-\mathtt{v})\log\lambda_1(\mathtt{n})-\mathtt{tr}\left(S_2(\mathtt{n})\Sigma^{-1}\right)\right]+\mathtt{tr}\left(S_2(\mathtt{n})\Sigma^{-1}\right).$$

En virtud del lema anterior, el primer sumando converge a 0 en probabilidad y, por tanto, en distribución. La distribución del segundo es, bajo las condiciones de (2.14), la que corresponde a la traza de una matriz aleatoria $p\times p$ siguiendo un modelo de distribución $W_p\big(\mathtt{v}-\mathtt{w},\mathtt{Id},\Sigma^{-1/2}\delta\Sigma^{-1/2}\big)$. Teniendo en cuenta de nuevo el problema 16 del capítulo 1, sabemos que se trata de la distribución $\chi^2_{p(\mathtt{v}-\mathtt{w})}\big(\mathtt{tr}(\delta\Sigma^{-1})\big)$. Aplicando el teorema 2.14(viii) se concluye. Por otra parte, bajo las condiciones de (2.13), la tesis se sigue del corolario 2.21.

(ii) Se sigue directamente de (2.16).

∎

Recordemos que en la sección dedicada al test de Lawley-Hotelling se obtuvo un test UMP-invariante en el caso de que Σ fuese conocida. Bajo las condiciones de (2.14), la misma función potencia asintótica es igual a la función potencia exacta de dicho test óptimo. Además, dado que el test de Wilks es, bajo la hipótesis (2.14), el test de razón de verosimilitudes, la convergencia a la distribución $\chi^2_{p,\mathtt{v}-w}$ en el caso normal y nulo podría haberse deducido directamente de las propiedades del mismo.

Por otra parte, téngase en cuenta que, si se verifica la condición (2.11), la distribución asintótica de los estadísticos de contraste correspondientes los test de Wilks, Lawlley-Hotelling, Roy y Pillay son iguales a las que correspondería el caso de que f_1 fuera normal. Por lo tanto, hemos probado el siguiente resultado:

Corolario 2.27.
Si se verifica la condición (2.11), los test de Wilks, Lawlley-Hotelling, Roy y Pillay son asintóticamente válidos.

Que sean asintóticamente válidos significa que el límite cuando el tamaño de muestra tiende a infinito de los niveles de significación es el valor α deseado en

principio para el test, y estamos afirmando que esto es cierto aunque no se satisfaga el supuesto de p-normalidad. Es decir, que si el tamaño de muestra es suficientemente grande, el nivel de significación del test será aproximadamente correcto. En definitiva, del los teoremas 2.15, 2.19 y 2.22, junto con el corolario 2.27, que todos los métodos de Inferencias considerados en este capítulo son asintóticamente válidos aunque no se verifique la normalidad. No obstante, recordemos que lo dicho se verifica cuando se satisface la condición (2.11) o, equivalentemente en el modelo con coordenadas, (2.10). Esta hipótesis, que ya apareció en el studio asintótico del modelo lineal univariante, se conoce como condición de Huber, y puede considerarse como una traducción de la condición de Lindemberg al caso del modelo lineal. En Arnold (1981), pp. 144, 145 y 146, podemos encontrar diversos comentarios sobre la misma. Se tiene, por ejemplo, que en un diseño completamente aleatorizado del análisis de la varianza, la condición de Huber se satisface cuando el número de datos de cada una de las casillas tiende a infinito (en términos prácticos, que tengamos muchos datos en cada casilla). No obstante, estas cuestiones se detallarán mejor más adelante.

Por otra parte, podemos obtener otra interesante ventaja del corolario 2.27: sustituir los cuantiles $C^{i,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n-dim}V}$, para $i = 1, 2, 3, 4$, correspondientes a distribuciones multivariantes muy complejas y que aparecen en la expresión de los test de Wilks, Lawlley-Hotelling, Roy y Pillay, por otros correspondientes a distribuciones bien conocidas y perfectamente tabuladas, como son la χ^2 central y la distribución $U^1_{p,\mathtt{dim}V|W}$. Recordamos que esta última es la distribución del primer autovalor de una matriz aleatoria que sigue un modelo de distribución $W_p(\mathtt{dim}V|W)$. Se encuentra tablada en Pearson & Heartley (1976). Concretamente, los tests mencionados a nivel α pueden expresarse, respectivamente, mediante

$$\phi_1(\mathtt{y}) = \begin{cases} 1 & \text{si } \prod_{i=1}^b (1+t_i) > \mathtt{e}^{\frac{\chi^{2,\alpha}_{p(\mathtt{V-W})}}{\mathtt{n-V}}} \\ 0 & \text{si } \prod_{i=1}^b (1+t_i) \leq \mathtt{e}^{\frac{\chi^{2,\alpha}_{p(\mathtt{V-W})}}{\mathtt{n-V}}} \end{cases}$$

$$\phi_2(\mathtt{y}) = \begin{cases} 1 & \text{si } \sum_{i=1}^b t_i > \frac{\chi^{2,\alpha}_{p(\mathtt{V-W})}}{\mathtt{n-V}} \\ 0 & \text{si } \sum_{i=1}^b t_i \leq \frac{\chi^{2,\alpha}_{p(\mathtt{V-W})}}{\mathtt{n-V}} \end{cases}$$

$$\phi_3(\mathtt{y}) = \begin{cases} 1 & \text{si } t_1 > U^{1,\alpha}_{p,\mathtt{V-W}} \\ 0 & \text{si } t_1 \leq U^{1,\alpha}_{p,\mathtt{V-W}}. \end{cases}$$

$$\phi_4(\mathtt{y}) = \begin{cases} 1 & \text{si } \sum_{i=1}^b \frac{t_i}{1+t_i} > \frac{\chi^{2,\alpha}_{p(\mathtt{V-W})}}{\mathtt{n-V}} \\ 0 & \text{si } \sum_{i=1}^b \frac{t_i}{1+t_i} \leq \frac{\chi^{2,\alpha}_{p(\mathtt{V-W})}}{\mathtt{n-V}} \end{cases}$$

No obstante, el programa estadístico SPSS hace uso también de otras aproximaciones a la distribución F-Snedecor para los tests ϕ_1, $\phi 2$ y ϕ_4. Son las siguientes[30]: en el caso del test de Wilks puede considerarse la siguiente aproximación:

$$F = \frac{\lambda_1^{\frac{1}{t}}}{\lambda_1^{\frac{1}{t}-1}} \frac{df_1}{df_2},$$

donde

$$\begin{aligned} df_1 &= p(\mathtt{v}-\mathtt{w}) \\ df_2 &= wt - \frac{1}{2}\big(p(\mathtt{v}-\mathtt{w})-2\big) \\ w &= \mathtt{n}-\mathtt{w}-\frac{1}{2}\big(p+(\mathtt{v}-\mathtt{w})+1\big) \\ t &= \sqrt{\frac{p^2(\mathtt{v}-\mathtt{w})^2-4}{p^2+(\mathtt{v}-\mathtt{w})^2-5}}. \end{aligned}$$

En los casos $b = 1$ ó 2 ($b = \text{mín}\{p, \mathtt{v}-\mathtt{w}\}$), Se tiene que $F \sim F(df_1, df_2)$. En el caso $b = 1$, el test sería el mismo que ya hemos estudiado para tal situación, que es UMP-invariante. En general, se verifica que $F \sim F(df_1, df_2)$, aproximadamente. Esta es la la aproximación que utiliza SPSS.

Respecto al test de Lawley-Hotelling SPSS considera

$$F = \frac{2(SN+1)\lambda_2}{S^2(2M+S+1)},$$

donde $S = b$, $M = \frac{1}{2}(|\mathtt{v}-\mathtt{w}-p|-1)$ y $N = \frac{1}{2}(\mathtt{n}-\mathtt{v}-p-1)$. Entonces F sigue, aproximadamente, una distribución $F(df_1, df_2)$, donde

$$\begin{aligned} df_1 &= S(2M+S+1) \\ df_2 &= 2(SN+1) \end{aligned}$$

Respecto al test de Pillay, SPSS considera

$$F = \frac{(2N+S+1)\lambda_4}{(2M+S+1)(S-\lambda_4)},$$

que sigue, aproximadamente, una distribución $F(df_1, df_2)$, donde df_1 como antes y $df_2 = S(2N+S+1)$.

[30] Por desgracia, no estamos en condiciones de probar la validex de las mismas.

2.7. Contraste de hipótesis generalizado.

Consideremos una generalización del problema de decisión estudiado, consistente en contrastar la hipótesis inicial $\mu A \in W$, siendo A una matriz $p \times s$ de rango s. Este problema se trata en Arnold (1981), cap. 19, ejercicio C4. Resumidamente, una reducción por suficiencia y cinco por invarianza nos conducen a considerar las raíces positivas $t_1^* \geq \ldots \geq t_{b^*}^*$ del polinomio $|A'S_2A - t^*A'S_3A|$, donde $b^* = \text{mín}\{s, \texttt{dim}V|W\}$, que sigue un modelo de distribución $Q_{s,\texttt{dim}V|W,\texttt{n}-\texttt{dim}V}\left(\theta_1^*, \ldots, \theta_{b^*}^*\right)$, siendo $\theta_1^* \geq \ldots \geq \theta_{b^*}^*$ las primeras raíces del polinomio $|A'\mu'P_{V|W}\mu A - \theta^*A'\Sigma A| = 0$. Se trata, en definitiva, de los estadísticos que obtendríamos según el procedimiento ya estudiado una vez sustituida la matriz de datos originales Y por $Y^* = YA$. En el caso $b^* = 1$ se obtendrá un test UMP-invariante a nivel α. Por analogía al estudio anterior, en el caso $b^* > 1$ proponemos los test asociados a las siguientes transformaciones:

- $\lambda_1^* = \prod_{i=1}^{b^*}(1+t_i^*)^{-1} = \frac{|A'S_2A+A'S_3A|}{|A'S_3A|}$
- $\lambda_2^* = \sum_{i=1}^{b^*} t_i^* = \texttt{tr}\left(A'S_2A(A'S_3A)^{-1}\right)$
- $\lambda_3^* = t_1^*$
- $\lambda_4^* = \sum_{i=1}^{b^*} \frac{t_i^*}{1+t_i^*} = \texttt{tr}\left(A'S_2A(A'S_2A+A'S_3A)^{-1}\right)$.

El tercer test es consistente con los intervalos de confianza para $V|W \times \mathbb{R}^p$. Al igual que sucedía antes, no podemos encontrar un test UMP-invariante. Pero además, en este caso, no podemos siquiera justificar los tests propuestos en términos semejantes a los anteriores para los tres tests restantes. Un contraste de este tipo se aplicará en el análisis de perfiles.

Cuestiones propuestas

1. Demostrar que $\{\Sigma^{-1}\colon \Sigma > 0\ p \times p\}$ se identifica con un abierto de $\mathbb{R}^{p(p+1)/2}$.
2. Demostrar que, en el teorema 2.2, el interior de la imagen de Q es no vacío.
3. En el mismo teorema, encontrar una biyección bimedible ϕ tal que
$$\phi(S) = \left(\hat{\mu}, \hat{\Sigma}\right).$$
4. Demostrar que los estadísticos M_1, M_2^1, M_3^{12} y M_2^{13} son invariantes maximales.
5. Demostrar que $\hat{\beta}$ es el EIMV de β, y que $\left(\hat{\beta}, \hat{\Sigma}\right)$ es el EMV de (β, Σ).

6. Si $Y \in \mathcal{M}_{n\times p}$ y $a, b \in \mathcal{R}^p$, demostrar que la covarianza muestral entre los vectores Ya e Yb puede expresarse como sigue:

$$S_{Ya,Yb} = \mathtt{n}^{-1}(a'S_3b + a'S_2b)$$

7. Demostrar que los autovalores de $S_3^{-1}S_2$ coinciden con los de $S_2S_3^{-1}$, que a su vez son las raíces del polinomio en t $|S_2 - tS_3|$. Demostrar que a lo sumo cuenta con b autovalores positivos.

8. Demostrar que, bajo las condiciones del modelo lineal normal multivariante, t_b, la b-ésima raíz de $S_3^{-1}S_2$, ha de ser estrictamente positiva con probabilidad 1.

9. Demostrar que, si $b = 1$, $\phi_1 = \phi_2 = \phi_3 = \phi_4$.

10. Demostrar que $\hat{\Sigma}$ es un estimador consistente de Σ, es decir, que converge en probabilidad a este último cuando $\mathtt{n}$ tiende a infinito.

11. Demostrar que, para cada probabilidad de la familia (2.1), se verifica:

$$P\left(\frac{(\mathtt{d}'\hat{\mu} - \mathtt{d}'\mu)\hat{\Sigma}^{-1}\big(\hat{\mu}'\mathtt{d} - \mu'\mathtt{d}\big)}{(\mathtt{n} - \mathtt{dim}V)\|\mathtt{d}\|^2} \leq C^{3,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V},\ \forall \mathtt{d} \in V|W\right) = 1 - \alpha. \tag{2.18}$$

De esta forma podemos establecer una familia de regiones (elipsoides) de confianza simultáneos a nivel $1 - \alpha$ para el subespacio de los contrastes $V|W$. (**Indicación:** razonar de manera análoga a la obtención de los intervalos para $V|W \times \mathbb{R}^p$, pero considerando t_1 como el primer autovalor de $Z_2S_3^{-1}Z_2'$ y teniendo en cuenta que $Z_2 = X_2'Y$, donde X_2 es una base ortonormal de contrastes.

12. Demostrar que en el caso $p = 1$, los elipsoides coinciden con los intervalos de confianza obtenidos por el método de Scheffé.

13. Probar que el test de Roy a nivel $1-\alpha$ es consistente con la familia de elipsoides de confianza simultáneos a nivel $1-\alpha$ para los contrastes, es decir, que se decide $\mu \notin W$ si, y sólo si, existe un contraste $\mathtt{d} \in V|W$ tal que 0 no pertenece al elipsoide

$$\mathcal{E}^{\alpha}_{\mathtt{d}}(Y) = \left\{x \in \mathbb{R}^p\colon \frac{(x' - \mathtt{d}'\hat{\mu})\hat{\Sigma}^{-1}\big(x - \hat{\mu}'\mathtt{d}\big)}{(\mathtt{n} - \mathtt{dim}V)\|\mathtt{d}\|^2} \leq C^{3,\alpha}_{p,\mathtt{dim}V|W,\mathtt{n}-\mathtt{dim}V}\right\}. \tag{2.19}$$

14. Si $\mathtt{d}$ es un contraste, establecer una región de confianza a nivel $1 - \alpha$ para el vector $\mu'\mathtt{d} \in \mathbb{R}^p$ y un test UMP-invariante y de razón de verosimilitudes a nivel α para contrastar la hipótesis inicial $\mu'\mathtt{d} = 0$.

15. Demostrar que, para cada $\alpha \in [0,1]$ y $m \in \mathbb{N}$, se verifica

$$\lim_{n\to\infty} mF_{m,n}^{\alpha} = \chi_m^{2,\alpha}.$$

16. Probar que la familia de intervalos de confianza simultáneos de Roy es asintóticamente válida. ¿En qué sentido?

Capítulo 3

Contrastes para la matriz de covarianzas.

En este capítulo nos proponemos resolver una serie de contrastes relativos a la matriz de varianzas-covarianzas de una distribución normal multivariante. Algunos de estos estudios, aunque no todos, pueden enmarcarse en un modelo lineal. En ese sentido, este capítulo complementa el anterior. Realmente, en capítulos posteriores, concretamente en los dedicados a los análisis de correlación canónica y componentes principales, se expondrán otros contrastes referentes a la matriz de covarianzas. En este aspecto hemos preferido seguir el esquema de Rencher (1995). Algunos de estos tests, como la prueba de esfericidad de Barlett o el test M de Box, suelen desempeñar la función de pruebas intermedias en un estudio más amplio, como puede ser análisis factorial o discriminante. En todo caso se utilizará el test de la razón de verosimilitudes partiendo de la hipótesis de p-normalidad. En estas condiciones, el estadístico de contraste sigue una distribución asintótica χ^2, según se estudia en el apéndice de volumen 1, dedicado a los Modelos Lineales en dimensión 1. Se efectuará en una ocasiones una pequeña corrección del mismo para obtener un test insesgado[1] y además, en cada caso, se multiplicará el estadístico de contraste por el denominado coeficiente de corrección ρ de Barlett, con el objeto de conseguir una aproximación satisfactoria a la distribución límite con muestras de tamaño pequeño o moderado. En Bilodeau (1999) o Anderson (1958) se expone el procedimiento para determinar el valor de ρ.

Ya hemos dicho que estos métodos parten de la normalidad multivariante de las observaciones. Este problema no es exclusivo del análisis multivariante, sino que se hace ya patente en dimensión 1. En tal caso[2], para aplicar resultados asintóticos

[1] Ver Lehmann (1986).

[2] Tal y como podemos comprobar en Arnold (1981) cap. 9.

análogos a los considerados en el capítulo anterior se precisa que el cociente entre la momento de orden 4 de la distribución considerada y el cuadrado de su varianza sea igual al cociente que se obtendría en caso normal, a saber, 3. Los tests acerca de la varianza (y, con mayor razón, los test acera de la matriz de varianzas y covarianzas) son bastante sensibles ante la violación de este supuesto. De hecho, el coeficiente de Kurtosis o aplastamiento, definido mediante

$$\delta_2 = \frac{\mu_4}{\sigma^4} - 3$$

puede considerarse una medida de la validez asintótica de dichos métodos. Ello nos obliga a ser muy precavidos a la hora de extraer conclusiones de carácter general.

También en este caso hemos optado por realizar reducciones previas por suficiencia e invarianza antes de dar con el estadístico del test de la razón de verosimilitudes. Este esfuerzo nos permite, entre otras cosas, establecer una clara vinculación entre el contraste de independencia de dos vectores aleatorios y los coeficientes de correlación canónica, que aparecen aquí por vez primera. Al final del capítulo incluimos algunos ejemplos.

3.1. Test de correlación.

Empezaremos contrastando la independencia de dos vectores aleatorios. Posteriormente consideraremos el contraste, más general, de la independencia de varios vectores aleatorios. Partimos de $\mathtt{n}$ observaciones

$$\begin{pmatrix} Y_1 \\ X_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_{\mathtt{n}} \\ X_{\mathtt{n}} \end{pmatrix},$$

que constituyen una muestra aleatoria simple de una distribución

$$N_{p+q}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right),$$

siendo la matriz de covarianzas definida positiva. Nótese que, en el caso $p = 1$, éste es el mismo modelo estudiado en el capítulo 5 del volumen 1, denominado modelo de correlación. Si se denota $V = \langle 1_{\mathtt{n}} \rangle$ y consideramos la matriz aleatoria

$$(Y\,X) = \begin{pmatrix} Y_1' X_1' \\ \vdots \\ Y_{\mathtt{n}}' X_{\mathtt{n}}' \end{pmatrix},$$

la familia de probabilidades del modelo estadístico es la siguiente:

$$\left\{ N_{n,p+q}\left(\mu, \mathtt{Id}, \Sigma\right) : \mu \in V,\ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} > 0 \right\}.$$

En dicho modelo se desea contrastar la hipótesis nula de independencia de ambos componentes, es decir, se plantea la hipótesis inicial $H_0 : \Sigma_{12} = 0$[3]. En primer lugar, si se consideran el vectores media muestral y la matriz de covarianzas total muestral

$$\begin{pmatrix} \overline{y} \\ \overline{x} \end{pmatrix}, \qquad S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

se tiene que el estadístico $(\overline{y}, \overline{x}, S_{11}, S_{12}, S_{22})$ es suficiente y completo. Además, el grupo de transformaciones

$$G = \left\{ g_{a,b,A,B} \colon (a, b, A, B) \in \mathcal{M}_{n\times p} \times \mathcal{M}_{n\times q} \times \mathcal{M}_p^* \times \mathcal{M}_q^* \right\}\ ^{4},$$

definidas mediante

$$g_{a,b,A,B}(YX) = (YA + a, XB + b),$$

deja invariantes tanto el modelo estadístico como el problema de contraste de hipótesis. Entonces, reduciendo por suficiencia e invarianza, se llega[5] al estadístico

$$\{r_1^2, \ldots, r_p^2\} \tag{3.1}$$

de los autovalores ordenados[6] de la matriz

$$S_{11}^{-1} S_{12} S_{22}^{-1} S_{21}, \tag{3.2}$$

cuya distribución depende de Σ a través de $\{\rho_1^2, \ldots, \rho_p^2\}$, que denotan los autovalores ordenados de la matriz

$$\Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \tag{3.3}$$

Los parámetros $\rho_1, \ldots, \rho_p$ se denominan coeficientes de correlación canónica poblacionales, mientras que $r_1, \ldots, r_p$ son los muestrales. Pueden considerarse como la generalización del coeficiente de regresión múltiple. De hecho, si $p = 1$, se tiene que

$$\rho_1^2 = \rho_{Y,X}^2, \qquad r_1^2 = R_{Y,X}^2$$

[3]Recordemos que, en estas condiciones $Y|X = x \sim N_{n,p}\left(\alpha + x\beta, \Sigma_{11,2}\right)$, con $\beta = \Sigma_{22}^{-1}\Sigma_{21}$. Luego el contraste de independencia se traduce en un contraste $\beta = 0$ en el modelo condicionado de regresión lineal.

[4]Recordemos que $\mathcal{M}_m^*$ denota el conjunto de las matrices cuadradas de orden m e invertibles.

[5]Para más detalles, consultar Arnold (1981).

[6]Realmente, el número de autovalores positivas no excederá en ningún caso de $b = \text{mín}\{p, q\}$. Por tanto, si se considera tan sólo los b primeras autovalores, el estadístico sigue siendo invariante maximal.

El capítulo 5 se dedica exclusivamente al estudio de estos parámetros. No existe test UMP invariante para este problema. En estas condiciones[7], el estadístico de contraste del test de razón de verosimilitudes debe expresarse a través del estadístico invariante maximal anterior. De hecho, puede comprobarse[8] que admite la expresión siguiente:

$$\begin{aligned} \lambda &= \left(\frac{|S_{11}||S_{22}|}{|S|}\right)^{-\mathtt{n}/2} && (3.4)\\ &= \left|\mathtt{Id} - S_{11}^{-1/2}S_{12}S_{22}^{-1}S_{21}S_{11}^{-1/2}\right|^{-\mathtt{n}/2} && (3.5)\\ &= \left[\prod_{i=1}^{p}(1-r_i^2)\right]^{-\mathtt{n}/2}. && (3.6) \end{aligned}$$

El problema es encontrar la distribución nula de este estadístico y calcular entonces el cuantil $1-\alpha$ de la distribución nula de $\lambda^{-2/\mathtt{n}}$, Q^α, de manera que el test de la razón de verosimilitudes a nivel α será pues el siguiente:

$$TRV = \begin{cases} 1 \text{ si } \prod_{i=1}^{p}(1-r_i^2) > Q^\alpha \\ 0 \text{ si } \prod_{i=1}^{p}(1-r_i^2) \leq Q^\alpha \end{cases} \tag{3.7}$$

Este test resuelve el contraste de la hipótesis inicial $H_0 : \Sigma_{12} = 0$, equivalente a $\beta = 0$. Nótese que la información de la muestra que se utiliza en la resolución del problema es la que contienen los coeficientes de correlación canónica al cuadrado. En el caso $p = 1$, lo único que interesa de la muestra es el valor del coeficiente de correlación múltiple o, para ser más exactos, del de determinación. En el caso general, es decir, cuando se contrastan r vectores aleatorios de dimensiones $p_1, \ldots, p_r$, respectivamente, y siendo $p = \sum_i p_i$, el estadístico de contraste sería, trivialmente, el siguiente:

$$\lambda = \left(\frac{|S_{11}| \times \ldots \times |S_{rr}|}{|S|}\right)^{-\mathtt{n}/2} \tag{3.8}$$

Nótese que el test relaciona la varianza generalizada del vector global con el producto de las varianzas generalizadas de los vectores individuales, lo cual nos permite una interesante interpretación geométrica[9]. En el caso nulo, es bien conocido [10] que el estadístico $-2\log\lambda$ se distribuye asintóticamente según un modelo χ^2_f, donde f es igual a pq en el caso de dos vectores y a $\frac{1}{2}\eta_2$ en general, donde η_i se define, para

[7] Cf. Lehmann (1986), pag. 341.
[8] Cuestión propuesta.
[9] Cf. Anderson (1951), sec. 7.5.
[10] Cuestión propuesta.

$i = 2, 3$, mediante

$$\eta_i = p^i - \sum_{j=1}^{r} p_j^i.$$

No obstante, se prueba en Bilodeau (1999) que la convergencia puede mejorarse considerando el estadístico $-2\rho \log \lambda$, siendo ρ el coeficiente de corrección de Barlett

$$\rho = 1 - \frac{2\eta_3 + 9\eta_2}{6\mathtt{n}\eta_2}.$$

En el caso de dos vectores, se tiene

$$\rho = 1 - \frac{p + q + 3}{2\mathtt{n}}.$$

3.2. Test M de Box.

En esta sección se estudia el test M de Box para contrastar la igualdad de r matrices de covarianzas a partir de sendas muestras independientes. Se trata de una generalización multivariante del conocido test univariante de Barlett (que compara r varianzas). Puede constituir un trámite previo al manova o al análisis discriminante. No obstante y dado que requiere del supuesto de normalidad, debe emplearse con precaución.

Consideraremos r muestras aleatorias simples independientes de tamaños $\mathtt{n}_i$ correspondientes a sendas distribuciones $N_p(\nu_i, \Sigma_i)$, para $i = 1, \ldots, r$. Sea $\mathtt{n} = \sum_i \mathtt{n}_i$. Si $V_i = \langle 1_{\mathtt{n}_i} \rangle \subset \mathbb{R}^{\mathtt{n}_i}$, para $i = 1, \ldots, r$, y $V = \langle 1_{\mathtt{n}} \rangle \subset \mathbb{R}^{\mathtt{n}}$, entonces, para cada $i = 1, \ldots, r$, la muestra i-ésima está asociada al modelo estadístico

$$\mathcal{E}_i = (\mathbb{R}^{\mathtt{n}_i p}, \mathcal{R}^{\mathtt{n}_i p}, \{N_{\mathtt{n}_i, p}(\mu_i, \mathtt{Id}, \Sigma_i) \colon \mu_i \in V_i, \Sigma_i > 0\}).$$

Luego, al ser las muestras independientes, la matriz $\mathtt{n} \times p$ de las observaciones está asociada al modelo producto $\mathcal{E} = \prod_{i=1}^{r} \mathcal{E}_i$. Además, la hipótesis nula $\Sigma_1 = \ldots = \Sigma_r$ es correcta si, y sólo si, dicho modelo equivale a

$$\mathcal{E}_0 = (\mathbb{R}^{\mathtt{n}p}, \mathcal{R}^{\mathtt{n}p}, \{N_{\mathtt{n},p}(\mu, \mathtt{Id}, \Sigma) \colon \mu \in V, \Sigma > 0\}).$$

Denótense por $S_i^{\mathtt{I}}$, $S_i^{\mathtt{mv}}$ y $S^{\mathtt{I}}$, $S^{\mathtt{mv}}$ el EIMV y el EMV, respectivamente, de Σ_i en el modelo $\mathcal{E}_i$ y de Σ en $\mathcal{E}_0$. En el caso $r = 2$, obtenemos[11], a partir de una reducción por suficiencia y otra por invarianza, el estadístico invariante maximal $\{t_1, \ldots, t_p\}$ de las

[11] Cf. Arnold, (1981)

raíces ordenadas de $|S_1^{\mathtt{mv}} - tS_2^{\mathtt{mv}}|$[12]. La distribución del este estadístico depende del parámetro de la estructura estadística a través de las raíces $\{\tau_1, \ldots, \tau_p\}$ del polinomio $p(\tau) = |\Sigma_1 - \tau\Sigma_2|$. Puede comprobarse por argumentos ya conocidos que el estadístico de razón de verosimilitudes es el siguiente:

$$\begin{aligned}\lambda &= \frac{|S_1^{\mathtt{mv}}|^{\mathtt{n}_1/2}\,|S_2^{\mathtt{mv}}|^{\mathtt{n}_2/2}}{|S^{\mathtt{mv}}|^{(\mathtt{n}_1+\mathtt{n}_2)/2}} \\ &= (\mathtt{n}_1+\mathtt{n}_2)^{\frac{1}{2}p(\mathtt{n}_1+\mathtt{n}_2)} \prod_{i=1}^{p} \left[t_i^{\frac{1}{2}\mathtt{n}_1} (\mathtt{n}_1 t_i + \mathtt{n}_2)^{\frac{1}{2}(\mathtt{n}_1+\mathtt{n}_2)} \right],\end{aligned}$$

que se contrastará con el valor de cuantil correspondiente a la distribución nula. Puede comprobarse también que, en el caso general (r covarianzas), se obtiene como estadístico de razón de verosimilitudes:

$$\lambda = \frac{\prod_{i=1}^{r} |S_i^{\mathtt{mv}}|^{\mathtt{n}_i/2}}{|S^{\mathtt{mv}}|^{\mathtt{n}/2}}.$$

No obstante, suele efectuarse la corrección siguiente (para que el test sea insesgado):

$$\lambda^* = \frac{|S^{\mathtt{I}}|^{(\mathtt{n}-r)/2}}{\prod_{i=1}^{r} |S_i^{\mathtt{I}}|^{(\mathtt{n}_i-1)/2}}.$$

Se tiene entonces que la distribución asintótica del estadístico de contraste $2\rho \log \lambda$ es, en el caso nulo, χ_f^2, donde

$$\begin{aligned} f &= \frac{1}{2}p(p+1)(r-1), \\ \rho &= 1 - \frac{2p^2+3p-1}{6(p+1)(r-1)} \left(\sum_{i=1}^{r} \frac{1}{n_i - 1} - \frac{1}{n-r} \right). \end{aligned}$$

En el caso $r = 2$, se tiene

$$f = p(p+1)/2, \quad \rho = 1 - \left(\frac{2p^2+3p-1}{6(p+1)} \right) \left(\frac{1}{n_1-1} + \frac{1}{n_2-1} - \frac{1}{n_1+n_2-2} \right).$$

3.3. Contraste para una matriz de covarianza.

En esta ocasión, partiremos de una muestra aleatoria simple de tamaño $\mathtt{n}$ de una distribución $N_p(\nu, \Sigma)$. Por consiguiente, el modelo estadístico será el siguiente:

$$Y \sim N_{\mathtt{n},p}\left(\mu, \mathtt{Id}_n, \Sigma\right), \qquad \mu \in \langle 1_{\mathtt{n}} \rangle,\ \Sigma > 0.$$

[12] Nótese que en el caso univariante ($p = 1$) estaríamos hablando del cociente de las varianzas, lo cual nos conduciría al test de Snedecor.

Se trata de contrastar la hipótesis inicial $\Sigma = \Sigma_0$, siendo Σ_0 una matriz de valores conocidos. Tras reducir por suficiencia e invarianza, se obtiene el estadístico invariante maximal de los autovalores ordenados $\{t_1, \ldots, t_p\}$ de la matriz $S^{\mathtt{I}}$ (el EIMV de Σ), cuya distribución depende del parámetro a través de los autovalores ordenados $\{\theta_1, \ldots, \theta, p\}$ de Σ. El estadístico de razón de verosimilitudes es el siguiente:

$$\begin{aligned}\lambda &= \frac{\left(\frac{\mathtt{n}-1}{\mathtt{n}}\right)^{\mathtt{n}/2} |S^{\mathtt{I}}|^{n/2}}{\exp\{p\mathtt{n}/2\}|\Sigma_0|^{\mathtt{n}/2}} \exp\left\{-\frac{1}{2}(\mathtt{n}-1)\mathtt{tr}\left(\Sigma_0^{-1}S^{\mathtt{I}}\right)\right\} \\ &= \exp\{-p\mathtt{n}/2\}\left(\frac{n-1}{n}\right)^{\mathtt{n}/2}\left(\prod t_i\right)^{\mathtt{n}/2}\exp\left\{\frac{1}{2}(\mathtt{n}-1)\sum t_i\right\}.\end{aligned}$$

Suele considerarse una leve modificación para conseguir que sea insesgado:

$$\lambda^* = \frac{\exp\{-p(\mathtt{n}-1)/2\}|S^{\mathtt{I}}|^{(\mathtt{n}-1)/2}}{|\Sigma_0|^{(\mathtt{n}-1)/2}}\exp\left\{-\frac{1}{2}(\mathtt{n}-1)\mathtt{tr}(\Sigma_0^{-1}S^{\mathtt{I}})\right\}.$$

En este caso, los parámetros de la distribución asintótica son los siguientes:

$$f = p(p+1)/2, \quad \rho = 1 - \frac{2p+1-2/(p+1)}{6(\mathtt{n}-1)}.$$

3.4. Test de esfericidad de Barlett.

Para terminar supongamos que, en las mismas condiciones de la sección anterior, deseamos contrastar la hipótesis inicial $\Sigma = \sigma^2\mathtt{Id}$, donde $\sigma^2 > 0$. Se trata pues de decidir si la distribución normal multivariante estudiada es esférica. En ese caso, tras reducir por suficiencia e invarianza se obtiene el invariante maximal $\{t_p^{-1}t_1, \ldots, t_p^{-1}t_{p-1}\}$, cuya distribución depende de $\{\theta_p^{-1}\theta_1, \ldots, \theta_p^{-1}\theta_{p-1}\}$. El estadístico de razón de verosimilitudes es el siguiente:

$$\lambda = \frac{[\mathtt{tr}(S^{\mathtt{I}})/p]^{p\mathtt{n}/2}}{|S^{\mathtt{I}}|^{\mathtt{n}/2}} = \frac{\prod(t_p^{-1}t_i)^{\mathtt{n}/2}}{\left(\frac{1+\sum(t_p^{-1}t_i)}{p}\right)^{p\mathtt{n}/2}}.$$

Modificado para que sea insesgado:

$$\lambda^* = \frac{|S^{\mathtt{I}}|^{(\mathtt{n}-1)/2}}{[\mathtt{tr}(S^{\mathtt{I}})/p]^{p(\mathtt{n}-1)/2}}.$$

Los parámetros de la distribución asintótica son los siguientes:

$$f = \frac{1}{2}p(p+1) - 1, \quad \rho = 1 - \frac{2p^2+p+2}{6p(\mathtt{n}-1)}.$$

Téngase en cuenta que, si la hipótesis inicial es verdadera, entonces estaremos en condiciones de aplicar métodos de análisis univariante, que al contar, no con $\mathtt{n}$ sino con $p\mathtt{n}$ observaciones independientes, serán más potentes que los métodos multivariantes. También se utiliza el test para contrastar la validez del modelo de análisis factorial. No obstante, dado que este test depende de la normalidad de la variable y que la esfericidad parte como hipótesis inicial, hemos de ser muy precavidos a la hora de extraer conclusiones.

3.5. Ejemplos

Ejemplo 3.1.
Se considera una muestra de 30 observaciones de vino japonés Seishu. Se recogen las siguientes variables: Y_1=textura; Y_2=olor; X_1=Ph; X_2=acidez 1; X_3=acidez 2; X_4=sake meter; X_5=azucar reducido; X_6=azucar total; X_7=alcohol; X_8= nitrógeno. Los resultados se presentan en la tabla 7.1 de Rencher (1995). Suponiendo que el vector aleatorio se distribuye según un modelo 10-normal, se desea contrastar al 1 %la hipótesis inicial de independencia de los vectores $(Y_1Y_2)'$, $(X_1X_2X_3)'$, $(X_4X_5X_6)'$ y $(X_7X_8)'$. La matriz de covarianzas muestrales S se expresa a continuación:

$$\left(\begin{array}{cc|ccc|ccc|cc}
,16 & ,08 & ,01 & ,006 & ,02 & -,03 & ,02 & ,01 & -,02 & ,28 \\
 & ,22 & -,01 & ,003 & ,01 & -,07 & ,03 & ,06 & ,04 & -1,83 \\
\hline
 & & ,03 & ,004 & ,03 & -,11 & -,03 & -,03 & -,01 & 4,73 \\
 & & & ,024 & ,020 & -,009 & -,009 & ,0004 & ,038 & 1,76 \\
 & & & & ,07 & -,18 & -,03 & -,03 & ,05 & 8,97 \\
\hline
 & & & & & 4,67 & -,33 & -,63 & -,14 & -21,15 \\
 & & & & & & ,13 & ,15 & ,05 & -5,05 \\
 & & & & & & & ,26 & ,13 & -4,93 \\
\hline
 & & & & & & & & ,35 & 3,43 \\
 & & & & & & & & & 1948
\end{array}\right)$$

Se tiene entonces

$$\lambda = \left(\frac{|S|}{|S_{11}||S_{22}||S_{33}||S_{44}|}\right)^{\mathtt{n}/2} = 0{,}02937^{30/2}.$$

Tras los cálculos pertinentes, se obtiene $f = 37$, $\rho = 0{,}81$. Entonces, calculamos $-2\rho \log \lambda = 85{,}72$, que se contrasta con $\chi^{2,0,001}_{37} = 69{,}35$. La decisión que corresponde es rechazar la hipótesis inicial. Parece pues bastante claro que existe correlación entre los vectores considerados.

Ejemplo 3.2.
Consideremos los datos de la tabla 5.1 de Rencher (1995). Supongamos que constituyen dos muestras aleatorias simples de sendas distribuciones 4-normales. Contrastemos entonces la igualdad de las matrices de covarianza al 5 %. Debemos contrastar $-2\rho \log \lambda$ con $\chi^{2,0,05}_f$. En este caso,

$$-2\rho \log \lambda = \rho\big((\mathtt{n}_1 + \mathtt{n}_2 - 2)\log |S^{\mathtt{I}}| - (\mathtt{n}_1 - 1)\log |S^{\mathtt{I}}_1| - (\mathtt{n}_2 - 1)\log |S^{\mathtt{I}}_2|\big).$$

Tras los cálculos pertinentes, se obtiene $f = 10$ y $\rho = 0{,}931$. Ya conocemos $S^{\mathtt{I}}_1$, $S^{\mathtt{I}}_2$ y $S^{\mathtt{I}}$. Sus determinantes son 7917.7, 58019.2 y 26977.7, resp. Luego el valor que toma el estadístico de contraste de arriba es 13.275; al contrastarlo con $\chi^{2,0,05}_{10} = 18{,}307$, aceptamos la hipótesis inicial de igualdad de matrices de covarianzas. Este resultado nos sitúa en las condiciones del modelo lineal normal multivariante.

Ejemplo 3.3.
Considerense los datos de la tabla 3.7 de Rencher (1995) y supóngase que corresponde a una muestra aleatoria simple de una distribución 5-normal con media ν y matriz de varianzas-covarianzas Σ. Nuestro propósito es contrastar la hipótesis inicial de esfericidad al 5 % de significación. Hemos de aplicar, por lo tanto, el test de esfericidad de Barlett. Se obtienen los siguientes resultados:

$$f = 14, \quad \rho = 0{,}81, \quad |S^{\mathtt{I}}| = 27236586, \quad \mathtt{tr}\, S^{\mathtt{I}} = 292{,}891.$$

Luego, el valor del estadístico de contraste es 26.177. Al compararlo con $\chi^{2,0,05}_{14} = 23{,}68$, se rechaza la hipótesis inicial.

Cuestiones propuestas

1. Demostrar que el el test propuesto en la primera sección es, en efecto, el de la razón de verosimilitudes. Obtener el grado de libertad de de la distribución χ^2 aplicando los resultados de Teoría Asintótica que se encuentran en el apéndice del volimen dedicado a los Modelos Lineales..

2. Demostrar que el el test propuesto en la segunda sección es, en efecto, el de la razón de verosimilitudes. Obtener el grado de libertad de de la distribución χ^2.

3. Idem con la tercera y cuarta sección.

4. Obtener, por cierto, el grado de libertad de la distribución asintótica del test de Wilks.

5. Demostrar que en el caso univariante, el test M de Box coincide con el test de Barlett de comparación de varianzas, y que, en el caso de dos varianza, este último coincide con el test de Snedecor.

Capítulo 4

Análisis Multivariante de la Varianza

En el segundo capítulo se definió el modelo lineal normal multivariante y se desarrolló el procedimiento general para abordar los problemas de estimación y tests de hipótesis referentes al parámetro media. En el presente, vamos a aplicar estas técnicas con el objeto de determinar si uno o varios factores cualitativos influyen el la distribución de un vector aleatorio respuesta. En principio, se supondrá la p-normalidad e igualdad de matrices de covarianzas de dicho vector para todos los niveles del factor. En ese caso, la igualdad de las distribuciones equivale a la igualdad de sus correspondientes medias, y eso es, en esencia, lo que nos proponemos a contrastar. Empezaremos estudiando las inferencias para una única muestra y para dos muestras independientes. Estos problemas pueden considerarse casos particulares, especialmente sencillos, del modelo lineal estudiado en el capítulo 2, si bien han sido históricamente estudiados al margen del modelo lineal. Posteriormente se considerará el estudio de un factor con un número arbitrario de niveles.

Sería muy conveniente que el lector o abordara este capítulo con un conocimiento previo del análisis de la varianza univariante (anova), que puede encontrar, por ejemplo, en el capítulo 6 del volumen dedicado a los Modelos Lineales. Nuestro objetivo no es repetir punto por punto el estudio de los diversos modelos del anova (modelo de clasificación simple, de bloques aleatorizados, bifactorial, anidado, etc) desde un punto de vista multidimensional, sino dejar bien claro qué cambio hay que realizar en el test univariante para obtener el análogo multivariante (manova). Es decir, que el lector debe tener la garantía de que, si domina un determinado diseño del anova, por sofisticado que sea, domina automáticamente su generalización al caso multivariante. En el desarrollo del capítulo se presta también especial atención al aspecto asintótico

y al impacto de la violación de los supuestos del modelo (normalidad y homocedasticidad). En ese sentido, los tests que presentamos tienen un comportamiento bastante satisfactorio. No obstante, recordamos aquí que en Puri and Sen (1971) podemos encontrar alternativas no paramétricas, aunque escasamente implementadas en los programas estadísticos.

Además, en la última sección hemos incluido el análisis de perfiles de una, dos o más distribuciones multivariantes, que se encuadra en el contraste generalizado de la media (ver sección 2.9), y que carecería de sentido en un estudio unidimensional.

4.1. Contraste de una media

Se trata del problema mas simple de entre todos aquellos que pueden formalizarse mediante el modelo lineal normal multivariante. Partimos de una muestra aleatoria simple (m.a.s.) $Y_1, \ldots, Y_{\mathtt{n}}$ de una distribución $N_p(\nu, \Sigma)$. En ese caso, contamos con una matriz de datos $Y = (Y_1, \ldots, Y_{\mathtt{n}})'$ cuya esperanza es $\mu = (\nu, \ldots, \nu)'$. Concretamente, el modelo estadístico es el siguiente

$$Y \sim N_{\mathtt{n},p}(\mu, \mathtt{Id}, \Sigma), \quad \mu \in \langle 1_{\mathtt{n}} \rangle, \ \Sigma > 0. \tag{4.1}$$

Sabemos por el teorema 2.2 que el par $[P_{\langle 1_{\mathtt{n}} \rangle} Y, Y' P_{\langle 1_{\mathtt{n}} \rangle^\perp} Y]$ constituye un estadístico suficiente y completo. Teniendo en cuenta que

$$P_{\langle 1_{\mathtt{n}} \rangle} = \frac{1}{\mathtt{n}} \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix} \tag{4.2}$$

se tiene que

$$P_{\langle 1_{\mathtt{n}} \rangle} Y = (\overline{y}, \ldots, \overline{y})' \qquad Y' P_{\langle 1_{\mathtt{n}} \rangle^\perp} Y \propto S,$$

donde S denotará en este caso a la varianza muestral dividiendo por $\mathtt{n} - 1$ en lugar de $\mathtt{n}$. En consecuencia todas nuestras inferencias se realizarán a partir de $(\overline{y}, S)$. De hecho, tanto $(\overline{y}, \ldots, \overline{y})'$ cono S son los EIMV de μ y Σ, respectivamente. Dado que, en virtud del teorema 2.1,

$$\overline{y} \sim N_p(\nu, \mathtt{n}^{-1}\Sigma), \qquad (\mathtt{n} - 1)S \sim W_p(\mathtt{n} - 1, \Sigma) \tag{4.3}$$

siendo ambos independientes, se sigue de (1.25) que

$$\mathtt{n}(\overline{y} - \nu)' S^{-1} (\overline{y} - \nu) \sim T^2_{p,\mathtt{n}-1} \tag{4.4}$$

Luego, dado $\alpha \in (0,1)$ el elipsoide de $\mathbb{R}^p$ definido mediante

$$\mathcal{E}_\alpha(Y) = \{x \in \mathbb{R}^p \colon (x-\overline{y})' S^{-1} (x-\overline{y}) \leq \mathtt{n}^{-1} T^{2,\alpha}_{p,\mathtt{n}-1}\} \tag{4.5}$$

constituye una región de confianza a nivel $1-\alpha$ para ν, es decir, que ν pertenece a $\mathcal{E}_\alpha(Y)$ con probabilidad $1-\alpha$. Esto puede servir para construir un test a nivel α para contrastar la hipótesis inicial $H_0 : \nu = \nu_0$, siendo ν_0 un valor conocido. Efectivamente, la hipótesis inicial se aceptaría si, y sólo si, ν_0 pertenece al elipsoide $\mathcal{E}_\alpha(Y)$. El test puede expresarse pues mediante el estadístico de contraste $T^2(Y) = \mathtt{n}(\overline{y}-\nu_0)' S^{-1} (\overline{y}-\nu_0)$, de manera que se define así:

$$\phi(Y) = \begin{cases} 1 & \text{si} \quad T^2(Y) > T^{2,\alpha}_{p,\mathtt{n}-1} \\ 0 & \text{si} \quad T^2(Y) \leq T^{2,\alpha}_{p,\mathtt{n}-1} \end{cases} \tag{4.6}$$

El hecho de que el nivel de significación de este test sea exactamente α, no es argumento suficiente para justificar su uso. Recordemos que se ha construido teniendo en cuenta que la distribución del estadístico (4.4) es bien conocida, lo cual no es una casualidad. Efectivamente, si consideramos los datos trasladados $Y_1 - \nu_0, \ldots, Y_{\mathtt{n}} - \nu_0$, que componen una matriz de datos $Y^* \sim N_{\mathtt{n},p}(\mu^*, \mathtt{Id}, \Sigma)$, la hipótesis inicial $\nu = \nu_0$ se traduce a $\mu^* = 0$. En ese caso, $\mathtt{dim} V|W = 1$, por lo tanto, existe un único autovalor positivo de $S_3^{-1} S_2$, que coincide con el número $t = Z_2 S_3^{-1} Z_2'$ (ver sección 2.2). Concretamente, puede comprobarse fácilmente (cuestión propuesta) que $(\mathtt{n}-1)t = T^2$. Del Lema Fundamental de Neyman-Pearson se sigue que el test (4.6) es UMP-invariante a nivel α. Además, es el test de la razón de verosimilitudes. Así pues, el estudio de la distribución T^2 de Hotelling podría justificarse, al igual que el de la t de Student, en virtud de los principios de Suficiencia, Invarianza y Máxima Verosimilitud, tal y como se explica en la sección 2.2. Más aún, nótese que el estadístico (4.4) no es más que una distancia de Mahalanobis entre la media muestral y la poblacional. Este tipo de distancia se halla presente en la propia función de densidad de la distribución normal multivariante, y es lo que realmente determina el camino que van tomando las reducciones por suficiencia e invarianza.

Si no suponemos la normalidad de la distribución considerada, es decir, si los datos $Y_1, \ldots, Y_n$ constituyen una m.a.s. de una distribución p-dimensional con momentos de orden 2 finitos, cabe preguntarse por el comportamiento asintótico del test anterior. En primer lugar, hemos de percatarnos de que la condición de Huber (2.11) es, en este caso, completamente vacua, supuesto que el tamaño de muestra $\mathtt{n}$ tiende a infinito. En consecuencia, se sigue del teorema 2.22 que la matriz de varianzas-covarianzas total muestral S converge en probabilidad a la matriz de varianzas-covarianzas de la distribución. Por otra parte, se sigue del Teorema Central del Límite Multivariante

(versión iid), la convergencia en distribución

$$\sqrt{\mathtt{n}} \cdot \Sigma^{-1/2}(\overline{y} - \nu) \longrightarrow N_p(0, \mathtt{Id}) \tag{4.7}$$

y, por lo tanto, $\mathtt{n}(\overline{y}-\nu)'\Sigma^{-1}(\overline{y}-\nu)$ converge en distribución a χ^2_p. En definitiva, dado que $\mathtt{n}(\overline{y} - \nu)'S^{-1}(\overline{y} - \nu)$ se descompone en el producto

$$\mathtt{n}(\overline{y} - \nu)'\Sigma^{-1}(\overline{y} - \nu) \cdot \frac{\mathtt{n}(\overline{y} - \nu)'S^{-1}(\overline{y} - \nu)}{\mathtt{n}(\overline{y} - \nu)'\Sigma^{-1}(\overline{y} - \nu)},$$

y teniendo en cuenta que el segundo factor converge en probabilidad a 1, se sigue la convergencia en distribución

$$\mathtt{n}(\overline{y} - \nu)'S^{-1}(\overline{y} - \nu) \longrightarrow \chi^2_p \tag{4.8}$$

Luego, tanto el elipsoide de confianza como el test de hipótesis anteriores tiene validez asintótica si se reemplaza el término $T^{2,\alpha}_{p,\mathtt{n}-1}$ por $\chi^{2,\alpha}_p$ [1]. A esta conclusión podría haberse llegado aplicando directamente el corolario 2.27. No obstante, dada la sencillez del problema, hemos optado por utilizar una versión más sencilla del Teorema Central del Límite.

Ejemplo 4.1.
Se mide en $n = 10$ ocasiones las variables Y_1 =Calcio disponible en tierra, Y_2 =Calcio intercambiable en tierra e Y_3 =Calcio en nabos verdes. Los resultados aparecen en la tabla 3.5 de Rencher (1995). Supongamos que los diez datos constituyen una m.a.s. de una distribución 3-Normal. Queremos contrastar al 5 % de significación la hipótesis inicial de que la media de dicha distribución es el vector $(15,6{,}0,2{,}85)'$. Tras los cálculos pertinentes se obtiene:

$$\overline{y} = \begin{pmatrix} 28{,}1 \\ 7{,}18 \\ 3{,}09 \end{pmatrix} \; ; \qquad S = \begin{pmatrix} 140{,}54 & & \\ 49{,}68 & 72{,}25 & \\ 1{,}94 & 3{,}68 & 0{,}25 \end{pmatrix}.$$

Luego, $T^2 = 24{,}559$, que se contrasta con $T^{2,0,05}_{3,9} = 3{,}85 \times F^{0,05}_{3,7} = 16{,}766$. Por lo tanto, la decisión correspondiente es rechazar la hipótesis inicial.

4.2. Contraste de dos medias

El siguiente paso en nuestro estudio es la comparación de la media de dos distribuciones normales multivariantes con matriz de covarianzas común, a partir de

[1] Si no se reemplaza también poseen validez asintótica, aunque el nivel de significación exacto puede diferir más del asintótico α.

sendas muestras aleatorias e independientes de las mismas. Sean $Y_{11},\dots,Y_{1\mathtt{n}_1}$ m.a.s. de $N_p(\nu_1,\Sigma)$ e $Y_{21},\dots,Y_{2\mathtt{n}_2}$ m.a.s. $N_p(\nu_2,\Sigma)$. Se supone que ambas muestras son independientes. La hipótesis de igualdad de las matrices de covarianzas debería ser, en principio, contrastada mediante el test M de Box. Los datos de ambas muestras componen la matriz aleatoria Y:

$$Y=(Y_{11}\dots Y_{1n_1}Y_{21}\dots Y_{2n_2})'\sim N_{\mathtt{n},p}(\mu,\mathtt{Id},\Sigma),\qquad \mu\in V,\ \Sigma>0,$$

donde $\mathtt{n}=\mathtt{n}_1+\mathtt{n}_2$ y V es el subespacio bidimensional de $\mathbb{R}^{\mathtt{n}}$ generado por los vectores $\mathtt{v}_1=(1,\dots1,0,\dots,0)'$ y $\mathtt{v}_2=(0,\dots,0,1,\dots,1)'$. Por tanto,

$$P_V=\left(\begin{array}{ccc|ccc} \frac{1}{n_1} & \dots & \frac{1}{n_1} & 0 & \dots & 0\\ \vdots & & \vdots & \vdots & & \vdots\\ \frac{1}{n_1} & \dots & \frac{1}{n_1} & 0 & \dots & 0\\ \hline 0 & \dots & 0 & \frac{1}{n_2} & \dots & \frac{1}{n_2}\\ \vdots & & \vdots & \vdots & & \vdots\\ 0 & \dots & 0 & \frac{1}{n_2} & \dots & \frac{1}{n_2} \end{array}\right) \tag{4.9}$$

En consecuencia, si, para $i=1,2$, $\overline{y}_i$ y S_i denotan la media muestral y matriz de varianzas-covarianzas total muestral (dividiendo por $\mathtt{n}_i-1$ en lugar de $\mathtt{n}_i$), respectivamente, de la muestra i-ésima, y S_c se define mediante $(\mathtt{n}-2)^{-1}[(\mathtt{n}_1-1)S_1+(\mathtt{n}_2-1)S_2]$, el estadístico $(\overline{y}_1,\overline{y}_2,S_c)$ es suficiente y completo. De hecho, $(\overline{y}_1,\dots,\overline{y}_1,\overline{y}_2,\dots,\overline{y}_2)'$ y $(\mathtt{n}-2)^{-1}\mathtt{n}S_c$ constituyen los EIMV de μ y Σ, respectivamente. Razonando como en la sección anterior, se tiene que $\overline{y}_1-\overline{y}_2\sim N_p(\nu_1-\nu_2,(\mathtt{n}_1^{-1}+\mathtt{n}_2^{-1})\Sigma)$ y $(\mathtt{n}-2)S_c\sim W_p(\mathtt{n}-2,\Sigma)$, siendo ambas distribuciones independientes. Por lo tanto,

$$\frac{\mathtt{n}_1\mathtt{n}_2}{\mathtt{n}_1+\mathtt{n}_2}\big[\overline{y}_1-\overline{y}_2-(\nu_1-\nu_2)\big]'S_c^{-1}\big[\overline{y}_1-\overline{y}_2-(\nu_1-\nu_2)\big]\sim T^2_{p,n-2} \tag{4.10}$$

Ello nos permite obtener un elipsoide de confianza a nivel $1-\alpha$ para $\nu_1-\nu_2$, con centro en $\overline{y}_1-\overline{y}_2$. Concretamente

$$\mathcal{E}_\alpha(Y)=\left\{x\in\mathbb{R}^p\colon \frac{\mathtt{n}_1\mathtt{n}_2}{\mathtt{n}_1+\mathtt{n}_2}\big[x-(\overline{y}_1-\overline{y}_2)\big]'S_c^{-1}\big[x-(\overline{y}_1-\overline{y}_2)\big]\le T^{2,\alpha}_{p,\mathtt{n}_1+\mathtt{n}_2-2}\right\} \tag{4.11}$$

Supongamos que deseamos contrastar a un nivel de significación α la hipótesis inicial de que las dos muestras corresponden a un mismo modelo de distribución, es decir, $\nu_1=\nu_2$. Teniendo en cuenta lo anterior, el test definido mediante

$$\phi(Y)=\left\{\begin{array}{lll} 1 & \text{si} & T^2(Y)>T^{2,\alpha}_{p,\mathtt{n}_1+\mathtt{n}_2-2}\\ 0 & \text{si} & T^2(Y)\le T^{2,\alpha}_{p,\mathtt{n}_1+\mathtt{n}_2-2}\end{array}\right. , \tag{4.12}$$

donde $T^2(Y) = \mathtt{n}_1^{-1}\mathtt{n}_2^{-1}(\mathtt{n}_1+\mathtt{n}_2)(\overline{y}_1-\overline{y}_2)'S_c^{-1}(\overline{y}_1-\overline{y}_2)$, tiene ciertamente un nivel de significación α. No obstante y al igual que en el caso de una media, podemos aportar una justificación muy convincente para el mismo: la hipótesis inicial equivale a que la matriz μ pertenece al subespacio unidimensional $W = \langle 1_{\mathtt{n}}\rangle$. Siendo $\mathtt{dim}V|W = 1$, el test que se obtiene en la teoría es UMP-invariante y de razón de verosimilitudes. Su estadístico de contraste, multiplicado por $\mathtt{n}-2$, es precisamente T^2. Las razones de esta coincidencia son las mismas que expusimos en la sección anterior. Puede comprobarse fácilmente (cuestión propuesta) que este test es una generalización multivariante del conocido test de Student para dos muestras independientes.

Este test presenta también un buen comportamiento asintótico. en primer lugar, la condición de Huber se traduce, teniendo en cuenta (4.9), en que tanto $\mathtt{n}_1$ como $\mathtt{n}_2$ converjan a infinito. En ese caso, queda garantizada la convergencia en probabilidad de S_c a Σ, aunque se viole el supuesto de normalidad. Teniendo en cuenta el Teorema Central de Límite (versió iid) y razonando como en la sección anterior se prueba la validez asintótica del test al reemplazar $T^{2,\alpha}_{p,\mathtt{n}-2}$ por $\chi^{2,\alpha}_p$. Este resultado podría obtenerse también como consecuencia del corolario 2.27. Así pues, podemos decir, en términos heurísticos, que el supuesto de normalidad puede ser obviado si ambas muestras son suficientemente grandes.

Más delicado resulta ser el supuesto de igualdad de matrices de covarianzas, en primer lugar porque el método del que disponemos para contrastar esta hipótesis es el tes M de Box, que requiere de la p-normalidad de las variables. Es conocido que en el caso $p = 1$ (univariante) disponemos de una variante del test de Student conocida como test de Welch. También se han propuestos tests alternativos al nuestro en el caso multivariante pero es más importante tener en cuenta la robustez del método y, sobre todo, el siguiente resultado. En el mismo se considera $\mathtt{n}$ como tamaño de muestra, siendo $\mathtt{n}_1$ el número de datos de ésta correspondientes a la primera distribución y $\mathtt{n}_2 = \mathtt{n} - \mathtt{n}_1$, los correspondientes a la segunda. En ese caso y prescindiendo de los supuestos de normalidad e igualdad de las matrices de covarianzas, se afirma lo siguiente:

Teorema 4.1.
Si $\mathtt{n}_1, \mathtt{n}_2 \to \infty$ y $\frac{\mathtt{n}_1}{\mathtt{n}_2} \to 1$ entonces el test (4.12) es asintóticamente válido para contrastar a nivel α la hipótesis inicial $\nu_1 = \nu_2$.

Demostración.
En primer lugar, tengamos en cuenta, que, en virtud del Teorema Central del Límite, se verifica que

$$\sqrt{\mathtt{n}_1}(\overline{y}_1-\nu_1) \xrightarrow{d} N_p(0,\Sigma_1), \qquad \sqrt{\mathtt{n}_2}(\overline{y}_2-\nu_2) \xrightarrow{d} N_p(0,\Sigma_2),$$

siendo ambas secuencias independientes. Por lo tanto,

$$\sqrt{\mathtt{n}_1}[(\overline{y}_1-\nu_1)-\sqrt{\frac{\mathtt{n}_2}{\mathtt{n}_1}}\cdot(\overline{y}_2-\nu_2)\xrightarrow{d}N_p(0,\Sigma_1+\Sigma_2).$$

Luego, si $\nu_1=\nu_2$, se verifica que que el término $\tau_{\mathtt{n}_1,\mathtt{n}_2}$, definido mediante $\tau_{\mathtt{n}_1,\mathtt{n}_2}=\mathtt{n}_1(\overline{y}_1-\overline{y}_2)'(\Sigma_1+\Sigma_2)^{-1}(\overline{y}_1-\overline{y}_2)$, converge en distribución a χ^2_p. Consideremos entonces la siguiente descomposición

$$\frac{\mathtt{n}_1+\mathtt{n}_2}{\mathtt{n}_1\mathtt{n}_2}(\overline{y}_1-\overline{y}_2)'S_c^{-1}(\overline{y}_1-\overline{y}_2)=\tau_{\mathtt{n}_1,\mathtt{n}_2}\cdot\frac{\frac{\mathtt{n}_1+\mathtt{n}_2}{\mathtt{n}_1\mathtt{n}_2}(\overline{y}_1-\overline{y}_2)'S_c^{-1}(\overline{y}_1-\overline{y}_2)}{\tau_{\mathtt{n}_1,\mathtt{n}_2}}.$$

Dado que el numerador $\mathtt{n}_1^{-1}\mathtt{n}_2^{-1}(\mathtt{n}_1+\mathtt{n}_2)(\overline{y}_1-\overline{y}_2)'S_c^{-1}(\overline{y}_1-\overline{y}_2)$ puede expresarse trivialmente mediante $\mathtt{n}_1(\overline{y}_1-\overline{y}_2)'\left(\mathtt{n}_1\mathtt{n}_2^{-1}S_1+S_2\right)^{-1}(\overline{y}_1-\overline{y}_2)$ y teniendo en cuenta que S_1 y S_2 convergen en probabilidad a Σ_1 y Σ_2, respectivamente, se concluye. ∎

Es decir, prescindiendo de los supuestos de normalidad e igualdad de matrices de covarianzas, podemos considerar que el método anterior (que engloba al test de hipótesis y al elipsoide de confianza) es válido si $\mathtt{n}_1$ y $\mathtt{n}_2$ son suficientemente grandes y suficientemente parecidos[2]. No obstante, se demuestra también en Lehmann (1998) que, si los tamaños de muestra $\mathtt{n}_1$ y $\mathtt{n}_2$ son grandes aunque no similares, el test consistente en comparar el estadístico de contraste

$$(\overline{Y}_1-\overline{Y}_2)'\left(\frac{1}{\mathtt{n}_1}S_1+\frac{1}{\mathtt{n}_2}S_2\right)^{-1}(\overline{Y}_1-\overline{Y}_2) \tag{4.13}$$

con el cuantil $\chi^{2,\alpha}_p$, tiene nivel de significación asintótico α. Sin embargo, para poder aplicar este resultado conviene que las muestras sean verdaderamente grandes.

Ejemplo 4.2.
Se miden un total de cuatro variables psicológicas en 32 hombres y otras tantas mujeres. Los resultados se exponen en la tabla 5.1 de Rencher (1995). Suponiendo que ambos grupos de datos constituyen muestras aleatorias simples independientes de sendas distribuciones 4-Normales con matriz de covarianza común, decidir al 1 % si, por término medio, existen diferencias entre hombres y mujeres en lo que respecta a este grupo de variables. Los cálculos son los siguientes:

$$\overline{y}_1=\begin{pmatrix}15{,}97\\15{,}91\\27{,}19\\22{,}75\end{pmatrix};\quad S_1=\begin{pmatrix}5{,}192&&&\\4{,}545&13{,}18&&\\6{,}522&6{,}760&28{,}67&\\5{,}250&6{,}266&14{,}47&16{,}65\end{pmatrix}$$

[2]Además, puede comprobarse (ver cuestión 4) que el test coincide con el test óptimo que se obtendría si las distribuciones fuesen normales con matriz de covarianzas común y conocida.

$$\overline{y}_2 = \begin{pmatrix} 12{,}34 \\ 13{,}91 \\ 16{,}59 \\ 21{,}94 \end{pmatrix}; \quad S_2 = \begin{pmatrix} 9{,}136 & & & \\ 7{,}549 & 18{,}60 & & \\ 5{,}531 & 10{,}73 & 30{,}25 & \\ 4{,}151 & 5{,}446 & 13{,}55 & 28{,}00 \end{pmatrix}$$

$$S_c = \begin{pmatrix} 7{,}164 & & & \\ 6{,}047 & 15{,}89 & & \\ 6{,}027 & 8{,}737 & 29{,}46 & \\ 4{,}701 & 5{,}856 & 14{,}01 & 22{,}32 \end{pmatrix}.$$

Entonces, $T^2 = 96{,}603$, $T^{2,0,01}_{4,62} = 15{,}442$, luego, la decisión es que los grupos son distintos. Dado que se trata de muestras de igual tamaño y éste es grande, la conclusión final sería la misma si prescindimos de los supuestos de normalidad e igualdad de las matrices de covarianzas.

En este tipo de problema de tests de hipótesis, puede seguirse el método consistente en comparar una a una las cuatro variables (o, lo que es lo mismo, proyectar sobre cada uno de los ejes de coordenadas). Este procedimiento es, desde luego, menos potente, es decir, posee una menor capacidad de discriminar los dos grupos, caso de ser distintos. Si la diferencia entre éstos es debida, fundamentalmente, a una variable (un eje de coordenadas), ambos procedimientos conducirán a conclusiones similares. Suele suceder, sin embargo, que lo que diferencia lo dos vectores aleatorios se deba a cierta combinación lineal de las componentes. El problema del análisis discriminante consiste en encontrar la combinación lineal que mejor distingue los grupos, es decir, el eje sobre el que debemos proyectar para obtener una máxima discriminación. Se trata, concretamente, de la dirección determinada por el vector $S_c^{-1}(\overline{y}_1 - \overline{y}_2)$, como veremos en el capítulo dedicado al Análisis Discriminante I.

4.3. Manova con un factor

El siguiente paso en nuestro estudio es considerar el contraste de igualdad de medias para los distintos niveles de un factor cualitativo, supuesto que se satisfacen la p-normalidad e igualdad de matrices de varianzas-covarianzas para los distintos nivees del factor. El número de niveles, que se denotará por la letra r, es arbitrario. Partiremos de r muestral aleatorias simples e independientes entre sí, $Y_{i1}, \ldots, Y_{i\mathtt{n}_i}$, correspondientes sendas distribuciones $N_p(\nu_i, \Sigma)$, donde $i = 1, \ldots, r$. La hipótesis de igualdad de matrices de varianzas-covarianzas puede ser contrasta, en principio, mediante el test M de Box. Si $\mathtt{n}$ denota la suma de los tamaños muestrales, los datos

componen na matriz aleatoria $Y = (Y_{11}, \ldots, Y_{1\mathtt{n}_1}, \ldots, Y_{r1}, \ldots, Y_{r\mathtt{n}_r})'$ de dimensiones $\mathtt{n} \times p$. El modelo estadístico correspondiente es

$$Y \sim N_{\mathtt{n},p}(\mu, \mathtt{Id}, \Sigma), \quad \mu \in V,\ \Sigma > 0,$$

siendo V el subespacio r-dimensional de $\mathbb{R}^{\mathtt{n}}$ definido mediante $\langle \mathtt{v}_1, \ldots, \mathtt{v}_r \rangle$, donde, para cada $i = 1, \ldots, r$, $\mathtt{v}_i$ se define mediante

$$\mathtt{v}_i = (0'_{\mathtt{n}_1}, \ldots, 0'_{\mathtt{n}_{i-1}}, 1'_{\mathtt{n}_i}, 0'_{\mathtt{n}_i+1}, \ldots, 0'_{\mathtt{n}_r})' \tag{4.14}$$

En lo que sigue, dados i entre 1 y r, j entre 1 y $\mathtt{n}_i$, y k entre 1 y p, se denotará por $\overline{y}_{i\cdot}$ la media aritmética de la i-ésima muestra, mientras que Y_{ij}^k e $\overline{y}_{i\cdot}^k$ denotarán respectivamente las componente k-ésimas de los vectores Y_{ij} e $\overline{y}_{i\cdot}$. Se denotará por $\overline{y}_{\cdot\cdot}$ la media aritmética global de los $\mathtt{n}$ datos. Por otra parte, la siguiente matriz corresponde, trivialmente, a la proyección ortogonal sobre V:

$$P_V = \left(\begin{array}{ccc|c|ccc} \frac{1}{n_1} & \cdots & \frac{1}{n_1} & & & & \\ \vdots & & \vdots & & & 0 & \\ \frac{1}{n_1} & \cdots & \frac{1}{n_1} & & & & \\ \hline & & & \ddots & & & \\ \hline & & & & \frac{1}{n_r} & \cdots & \frac{1}{n_r} \\ & 0 & & & \vdots & & \vdots \\ & & & & \frac{1}{n_r} & \cdots & \frac{1}{n_r} \end{array}\right) \tag{4.15}$$

Luego,

$$P_V Y = \sum_{i=1}^{r} \overline{y}_{i\cdot} \cdot 1_{\mathtt{n}_i}, \qquad Y' P_{V^\perp} Y = \begin{pmatrix} SCE_{11} & \ldots & SCE_{1p} \\ \vdots & & \vdots \\ SCE_{1p} & \ldots & SCE_{pp} \end{pmatrix}, \tag{4.16}$$

donde

$$SCE_{hk} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(Y_{ij}^h - \overline{y}_{i\cdot}^h\right)\left(Y_{ij}^k - \overline{y}_{i\cdot}^k\right), \quad h, k = 1, \ldots, p. \tag{4.17}$$

Si se desea contrastar a nivel α la hipótesis inicial de que todas las muestras corresponden a un mismo modelo de distribución, es decir, $\nu_1 = \ldots = \nu_r$, la hipótesis inicial se expresará mediante $\mu \in W$, siendo $W = \langle 1_{\mathtt{n}} \rangle$. En ese caso, se verifica que $P_W Y = \overline{y}_{\cdot\cdot} \cdot 1_{\mathtt{n}}$.Tenemos pues la siguiente descomposición de Y:

$$Y = \begin{pmatrix} \overline{y}_{\cdot\cdot} \\ \vdots \\ \overline{y}_{\cdot\cdot} \end{pmatrix} + \begin{pmatrix} \overline{y}_{1\cdot} - \overline{y}_{\cdot\cdot} \\ \vdots \\ \overline{y}_{r\cdot} - \overline{y}_{\cdot\cdot} \end{pmatrix} + \begin{pmatrix} Y_{11} - \overline{y}_{1\cdot} \\ \vdots \\ Y_{r\mathtt{n}_r} - \overline{y}_{r\cdot} \end{pmatrix}. \tag{4.18}$$

Dado que el primer sumando es $P_W Y$ y el tercero $P_{V^\perp} Y$, el segundo ha de ser, necesariamente, $P_{V|W} Y$. Por lo tanto,

$$Y' P_{V|W} Y = \begin{pmatrix} SCH_{11} & \dots & SCH_{1p} \\ \vdots & & \vdots \\ SCH_{1p} & \dots & SCH_{pp} \end{pmatrix}, \tag{4.19}$$

donde

$$SCH_{hk} = \sum_{i=1}^{r} n_i \left(\overline{y}_{i\cdot}^h - \overline{y}_{\cdot\cdot}^h\right)\left(\overline{y}_{i\cdot}^k - \overline{y}_{\cdot\cdot}^k\right),\ h,k = 1,\dots,p. \tag{4.20}$$

Así pues, ya tenemos las matrices S_2 y S_3. El test para resolver el contraste se construirá, en todo caso, a partir de los autovalores positivos $t_1,\dots,t_b$ de $S_3^{-1}S_2$, siendo $b = \text{mín}\{p, r-1\}$. No existirá un test UMP-invariante a menos que el número autovalores positivos sea 1, es decir, a menos que $p = 1$ (lo cual se corresponde con el caso univariante, que conduciría al test `F`), o $r = 2$ (lo cual corresponde al estudio abordado en la sección anterior, que conduce al test T^2). Excluidos ambos caso, es decir, si $p > 1$ y $r > 2$, utilizaremos alguna de las cuatro transformaciones propuestas (Wilks, Lawley-Hotelling, Roy o Pillay). Si `n` es grande y se ha escogido la transformación λ_1, λ_2 o λ_4, debemos contrastar el valor obtenido con el cuantil $\chi^{2,\alpha}_{p(r-1)}$. Si hemos escogido la transformación λ_3, lo contrastaremos con $U_{p,r-1}$.

Respecto al comportamiento asintótico de estos tests puede demostrarse fácilmente que la condición de Huber (2.11) se traduce, en este caso, en que todos los tamaños muestrales $n_1,\dots,n_r$ converjan a infinito (condición ésta muy natural). En ese caso y en virtud del corolario 2.27, los cuatro tests serán el test correspondiente será asintóticamente válidos, aun prescindiendo la hipótesis de p-normalidad. También pueden considerarse aproximaciones a la distribución F-Snedecor, como se vio en el capítulo 2. Si se escoge $\lambda_3 = t_1$, confrontaremos con la tabla de Pearson y Heartley. En general el procedimiento seguido se conoce como manova de 1 vía.

En el caso $p = 1$, se tendrán dos únicos valores SCE y SCH, y el único autovalor de $S_3^{-1}S_2$ será entonces $t = SCE^{-1} \cdot SCH$. Se verifica que, en el caso nulo

$$\frac{\mathtt{n} - r}{r-1}\frac{SCH}{SCE} \sim F_{r-1,\mathtt{n}-r},$$

luego debe contrastarse dicho estadístico con el cuantil correspondiente. El procedimiento se denomina anova de 1 vía. Es importante percatarse de que la única diferencia entre un anova ($p = 1$) y un manova ($p > 1$) radica en que el primero considera el cociente de los números positivos

$$SCH = \|P_{V|W} Y\|^2, \qquad SCE = \|P_{V^\perp} Y\|^2.$$

En el manova ($p > 1$), por el contrario, no se puede hablar de números positivos sino de las matrices $p \times p$ definidas positivas $S_2 = Y'P_{V|W}Y$ y $S_3 = Y'P_{V^\perp}Y$. Los elementos de estas matrices se obtienen en (4.19) y (4.16) de manera muy análoga a SCH y SCE, de hecho, los p elementos de las diagonales coinciden con los valores de éstos, componente a componente. El resto se obtiene considerando productos cruzados entre distintas componentes, según se indica en (4.20) y (4.17). Por último, como $S_3^{-1}S_2$ no es un número, consideramos sus autovalores positivos $t_1, \ldots, t_b$. Ésta es la clave para extender al análisis multivariante cualquier diseño del análisis de la varianza univariante[3], como el de bloques al azar, diseños multifactoriales, anidados, por cuadrados latinos, etc.

Al igual que en el caso univariante, podemos realizar también contrastes hipótesis del tipo $\nu_i = \nu_j$, donde $i \neq j$. En ese caso, la hipótesis inicial se corresponde con un hiperplano de V, por lo que el contraste se resuelve, como en las secciones 1 y 2, mediante la distribución T^2 de Hotelling. Nótese que la condición $\nu_i = \nu_j$ equivale a $\mathtt{d}'\mu = 0$, siendo $\mathtt{d} = \mathtt{n}_i^{-1}1_{\mathtt{n}_i} - \mathtt{n}_j^{-1}1_{\mathtt{n}_j}$, que pertenece a $V|W$. Recordemos que se denominan contrastes a los elementos de $V|W$, y que (2.19) constituye una familia de elipsoides de confianza simultáneos para los contrastes a nivel $1-\alpha$, que además es consistente con test de Roy ϕ_3 a nivel α, es decir, que el test decide la hipótesis alternativa si, y sólo si, el vector 0 queda fuera de $\mathcal{E}_{\mathtt{d}}^{\alpha}$ para algún contraste $\mathtt{d} \in V|W$. Ello va asociado de forma natural a un método de comparaciones múltiples, basado en la distribución $C_{p,r-1,\mathtt{n}-r}$ de Roy, que generaliza el método de las comparaciones múltiples de Scheffé, basado a su vez en la distribución $F_{r-1,\mathtt{n}-r}$.

De todas formas y caso de que el manova resulte significativo, suele dilucidarse en primer lugar qué componentes del vector observado presentan diferencias en cuanto a sus medias para realizar posteriormente las comparaciones múltiples componente a componente, mediante los métodos ya conocidos del caso univariante (Scheffé, Tuckey, Bonferroni,...). Así, SPSS muestra junto con el resultado del manova los resultados de los p anovas y las p comparaciones múltiples por el método que escojamos.

Ejemplo 4.3.
Se tomaron 6 muestras aleatorias independientes de manzanos correspondientes a seis tipos de terrenos, midiéndose en cada caso cuatro variables referentes al tamaño de los mismos. Los resultados se encuentran en la tabla 6.2 de Rencher (1995). Suponiendo que se verifican las condiciones requeridas[4], se desea contrastar al 0.1 % de significación la hipótesis inicial de que el tipo de terreno no influye en el desarrollo

[3] Ver capítulo 6 del volumen dedicado a los Modelos Lineales.
[4] ¿Cuáles son y cómo se contrastan? ¿Es estrictamente necesario su cumplimiento?

del manzano. Tras los cálculos pertinentes, se obtuvo:

$$S_2 = \begin{pmatrix} ,074 & ,537 & ,332 & 208 \\ & 4,200 & 2,355 & 1,637 \\ & & 6,114 & 3,781 \\ & & & 2,493 \end{pmatrix}, \quad S_3 = \begin{pmatrix} ,320 & 1,697 & ,554 & ,217 \\ & 12,143 & 4,364 & 2,110 \\ & & 4,291 & 2,482 \\ & & & 1,723 \end{pmatrix}.$$

Los autovalores de la matriz $S_3^{-1}S_2$ son 1.876, 0.791, 0.229 y 0.026. Por lo tanto, el estadístico de Wilks (test de la razón de verosimilitudes) toma el valor $\lambda_1 = 0{,}154$ La aproximación a la F-Snedecor correspondiente a este caso es $F = 4{,}937$, que se contrasta con $F^{0,001}_{20,130} = 2{,}53$. Luego, queda rechazada la hipótesis inicial.

4.4. Análisis de perfiles

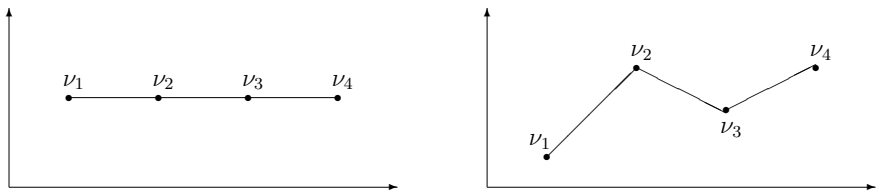
Este estudio, que se encuadra en el contraste generalizado para la media, carecería de sentido en un análisis univariante. Está estrechamente relacionado co el modelo de medidas repetidas y puede ser de gran utilidad cuando se analiza el crecimiento por etapas de una determinada especie animal o vegetal. Comenzaremos con el análisis de perfiles para una muestra. Consideremos $Y_1, \ldots, Y_n$ una muestra aleatoria simple de una distribución

$$N_p\left(\begin{pmatrix} \nu_1 \\ \vdots \\ \nu_r \end{pmatrix}, \Sigma\right)$$

Deseamos contrastar la hipótesis inicial $H_0 : \nu_1 = \ldots = \nu_p$ o, lo que es lo mismo,

$$H_0 : \nu_1 - \nu_2 = \ldots = \nu_{p-1} - \nu_p = 0.$$

Gráficamente, las hipótesis inicial y alternativa pueden ilustrarse de la forma siguiente:



La hipótesis inicial se expresa mediante $H_0 : (\nu_1, \ldots, \nu_p)A' = 0$, donde

$$A = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -1 & 1 & \ldots & 0 \\ 0 & -1 & \ldots & \vdots \\ \vdots & \vdots & \ldots & 1 \\ 0 & 0 & \ldots & -1 \end{pmatrix} \tag{4.21}$$

Tener en cuenta que A es una matriz $p \times (p-1)$ de rango $p-1$. Nuestros datos compondrán una matriz aleatoria $Y = (Y_1, \ldots, Y_{\mathtt{n}})'$, de manera que

$$Y \sim N_{\mathtt{n},p}(\mu, \mathtt{Id}, \Sigma), \qquad \mu \in \langle 1_{\mathtt{n}} \rangle, \ \Sigma > 0.$$

Se trata entonces de contrastar la hipótesis inicial $H_0 : \mu A = 0$, lo cual se puede clasificar, según hemos comentado, como contraste generalizado para la media. En este caso, $b^* = \text{mín}\{p-1, 1-0\} = 1$, con lo cual, con la notación considerada en el capítulo 2, hemos de tomar como estadístico de contraste la única raíz positiva del polinomio $|A'S_2A - t^*A'S_3A|$ o, lo que es lo mismo, el número $Z_2A(A'S_3A)^{-1}A'Z_2'$. Si se multiplica por $\mathtt{n}-1$, se obtiene el estadístico de contraste

$$T^2(Y) = \mathtt{n}\overline{y}'A\big(A'SA\big)^{-1}A'\overline{y}, \tag{4.22}$$

que sigue, en el caso nulo, una distribución $T^2_{p-1,\mathtt{n}-1}$. Luego, el test siguiente es UMP-invariante al nivel α :

$$\phi(Y) = \begin{cases} 1 & \text{si} \quad T^2(Y) > T^{2,\alpha}_{p-1,n-1} \\ 0 & \text{si} \quad T^2(Y) \leq T^{2,\alpha}_{p-1,n-1} \end{cases}$$

Este test es el que corresponde al contraste de la hipótesis inicial $\mu = 0$ para los datos trasformados $A'Y_1, \ldots, A'Y_{\mathtt{n}}$, que se estudia en la primera sección.

El siguiente paso del estudio consiste en considerar dos muestras aleatorias simples e independientes, $Y_{i1}, \ldots, Y_{i\mathtt{n}_i}$, $i = 1, 2$, correspondientes a sendas distribuciones p-normales con la misma matriz de varianzas-covarianzas y medias $\nu_1 = (\nu_{11} \ldots \nu_{1p})'$ y $\nu_2 = (\nu_{21} \ldots \nu_{2p})'$, respectivamente. Podemos plantearnos, primeramente, la hipótesis inicial (1) de paralelismo, es decir,

$$\begin{aligned} \nu_{11} - \nu_{12} &= \nu_{21} - \nu_{22}, \\ \vdots &\quad \vdots \\ \nu_{1,p-1} - \nu_{1p} &= \nu_{2,p-1} - \nu_{2p}. \end{aligned}$$

Si consideremos la matriz A (4.21), la hipótesis inicial se expresa mediante $H_0^{(1)}$: $\nu_1'A = \nu_2'A$. Consideremos la matriz aleatoria $Y = (Y_{11}, \ldots, Y_{1\mathtt{n}_1}, Y_{21}, \ldots, Y_{2\mathtt{n}_2})'$, la cual está asociada al siguiente modelo estadístico

$$Y \sim N_{\mathtt{n}_1+\mathtt{n}_2,p}(\mu, \mathtt{Id}, \Sigma), \quad \mu \in \langle \mathtt{v}_1, \mathtt{v}_2 \rangle, \ \Sigma > 0.$$

La hipótesis inicial puede expresarse mediante $H_0^{(1)} : \mu A \in \langle 1_{\mathtt{n}_1+\mathtt{n}_2} \rangle$. Luego, estamos nuevamente ante un contraste generalizado, con $\mathtt{dim}V|W = 1$. Por lo tanto, el test UMP-invariante se obtiene, como en el caso anterior, considerando el estadístico invariante maximal $Z_2A(A'Z_3'Z_3A)^{-1}A'Z_2'$. Multiplicando por $\mathtt{n}_1 + \mathtt{n}_2 - 2$, obtenemos el estadístico de contrastes

$$T^2(Y) = \frac{\mathtt{n}_1\mathtt{n}_2}{\mathtt{n}_1 + \mathtt{n}_2}(\overline{y}_1 - \overline{y}_2)'A\big(A'S_cA\big)^{-1}A'(\overline{y}_1 - \overline{y}_2), \tag{4.23}$$

que sigue un modelo de distribución $T^2_{p-1,\mathtt{n}_1+\mathtt{n}_2-2}$ en el caso nulo. Luego, el test UMP-invariante a nivel α es el siguiente:

$$\phi(Y) = \begin{cases} 1 & \text{si} \quad T^2 > T^{2,\alpha}_{p-1,\mathtt{n}_1+\mathtt{n}_2-2} \\ 0 & \text{si} \quad T^2 \leq T^{2,\alpha}_{p-1,\mathtt{n}_1+\mathtt{n}_2-2} \end{cases},$$

que es el que corresponde al contraste de la hipótesis $\mu \in \langle 1_{\mathtt{n}_1+\mathtt{n}_2} \rangle$ para los datos trasformados $A'Y_{ij}$, $i = 1, 2$ y $j = 1, \ldots, \mathtt{n}_i$. Este contraste se estudió en la sección 2. También podemos plantear el contraste de la hipótesis inicial (2) siguiente:

$$H_0^{(2)} : \sum_{j=1}^{p} \mu_{1j} = \sum_{j=1}^{p} \mu_{2j},$$

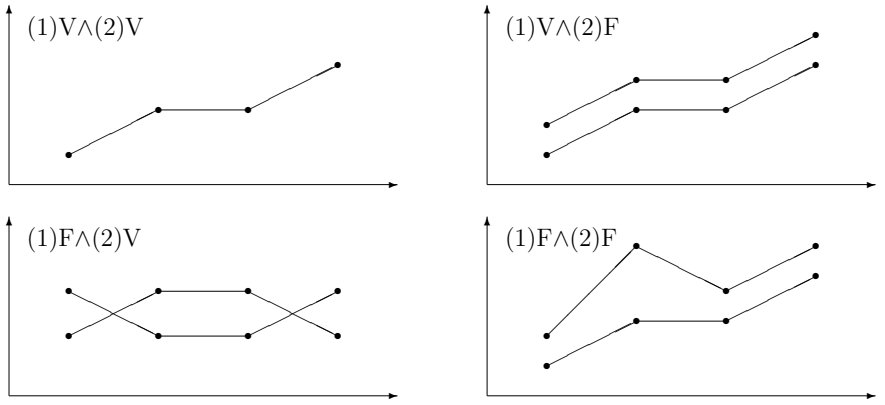
que puede expresarse también mediante $H_0^{(2)} : \mu 1_p \in \langle 1_{\mathtt{n}_1+\mathtt{n}_2} \rangle$. Nuevamente, se trata de un contraste generalizado para la media, en este caso con $s = 1$ y $\mathtt{dim}V|W = 1$. Luego, conviene de considerar el número $Z_2 1_p(1_p'S_3 1_p)^{-1}1_p'Z_2'$, que, multiplicado por $\mathtt{n}_1 + \mathtt{n}_2 - 2$ conduce al estadístico de contraste

$$t(Y) = \frac{\left|\sum_{k=1}^{p}(\overline{y}_1^k - \overline{y}_2^k)\right|}{\sqrt{\left(\frac{1}{\mathtt{n}_1} + \frac{1}{\mathtt{n}_2}\right)\sum_{k,h=1}^{p} s_c^2(k,h)}}, \tag{4.24}$$

siendo $s_c^2(k,h)$ la componente (k,h) de la matriz S_c. El estadístico (4.24) sigue un modelo de distribución $t_{n_1+n_2-2}$ en el caso nulo. Luego, el test UMP-invariante a nivel α es el siguiente:

$$\phi(Y) = \begin{cases} 1 & \text{si} \quad t > t^{\alpha}_{1,n_1+n_2-2} \\ 0 & \text{si} \quad t \leq t^{\alpha}_{1,n_1+n_2-2} \end{cases}$$

La visión gráfica de las hipótesis $H_0^{(1)}$ y $H_0^{(2)}$ es la siguiente:



Por último, el caso de r perfiles se sigue de manera análoga: dadas $Y_{i1}, \ldots, Y_{in_i}$, $i = 1, \ldots, r$, muestras aleatorias simples e independientes de sendas distribuciones p-normales con idénticas matrices de varianzas-covarianzas, podemos plantear considerar también contrastes tipo (1) y (2). En ambos casos habrá que considerar las b^* raíces positivas del polinomio $|C'S_2C - tC'S_3C|$, donde $C = A$ o $C = 1_{\sum_i \mathtt{n}_i}$, según se contraste la hipótesis (1) ó (2), respectivamente. Se obtiene $b^* = \text{mín}\{p-1, r-1\}$, las matriz S_2 y S_3 como en el manova de una vía. Según la transformación considerada, tendremos el test de Wilks, Lawley-Hotelling, Roy o Pillay.

Cuestiones propuestas

1. Probar que el estadístico T^2 definido en (4.4) verifica $(\mathtt{n}-1)t = T^2$, siendo $t = Z_2 S_3^{-1} Z_2'$.

2. Probar. igualmente, que el, estadístico T^2 para la comparación de dos medias verifica $(\mathtt{n}-2)t = T^2$.

3. Consideremos una muestra aleatoria simple de tamaño n de una distribución p-Normal con matriz de covarianzas conocida. Construir una región de confianza a nivel $1-\alpha$ para la media de la distribución. ¿De qué figura geométrica se trata? ¿Bajo qué condiciones se obtiene una esfera?

4. Consideremos dos muestras aleatorias simples independientes de tamaños n_1 y n_2 de sendas distribuciones p-normales con la misma matriz de covarianzas, que se supone conocida. Construir una región de confianza a nivel $1-\alpha$ para

la diferencia de las medias y un test a nivel α para contrastar la igualdad de las mismas.

5. Demostrar que el test propuesto para contrastar la diferencia de dos medias en la sección primera es una generalización multivariante del test de Student.

6. Situémonos en el contraste de las medias de dos distribuciones p-normales con matriz de covarianza común a partir de sendas muestras independientes de las mismas. Relacionar el concepto de distancia de Mahalanobis con el estadístico de contraste y la potencia del test correspondiente.

7. Demostrar (4.16).

8. ¿Qué requisitos deben exigirse antes de aplicar un manova de una vía? ¿En qué medida son realmente necesarios?

9. Si realizas mediante SPSS un manova de una vía para dos muestras independientes, observarás que los resultados correspondientes a los tests de Wilks, Lawley-Hotelling, Roy y Pillay son idénticos. ¿Puedes explicarlo? Además, se indica en la salida que el valor F es exacto ¿Qué quiere decir?

10. Obtener un elipsoide de confianza a nivel $1-\alpha$ para la diferencia $\nu_i - \nu_j$ en el manova de una vía.

11. Obtener el estadístico de contraste (4.24) así como su distribución en el caso nulo.

12. Los datos que se presentan en la tabla 5.7 de Rencher (1995) corresponden a la medición de seis parámetros relativos a la atrofia muscular, llevados a cabo en dos muestras aleatorias independientes. La primera de ellas, de tamaño 39, corresponde a un grupo control, mientras que la segunda, de tamaño 34, a un grupo de transportistas. Contrastar si existe diferencia significativa entre las medias de ambos grupos. ¿Se dan las condiciones de validez del test?

Capítulo 5

Regresión Lineal Multivariante

Continuamos estudiando las aplicaciones del modelo lineal normal multivariante o, mejor dicho, los problemas que pueden formalizarse mediante dicho modelo. En este caso, se trata de explicar cierta cantidad (en total p) de variables denominadas *respuesta* mediante una relación de tipo lineal con otras (en total q) variables denominadas *explicativas*. Este estudio generaliza el de regresión lineal múltiple[1], donde existe una única variable respuesta y, por supuesto, al de regresión lineal simple, donde, contamos únicamente con una variable respuesta y otra explicativa.

El esquema a seguir es el siguiente: empezaremos por presentar el modelo; seguidamente, lo compararemos con el modelo de correlación lineal[2]; a continuación, abordaremos los problemas de estimación y contraste de hipótesis, así como algunos aspectos asintóticos de interés; finalmente, explicaremos qué se entiende por regresión con variables ficticias, con lo cual esperamos que quede claro que tanto el manova como el mancova pueden considerarse casos particulares de la regresión lineal multivariante.

El problema de regresión múltiple puede considerarse tema propio del análisis multivariante pues considera diversas variables explicativas. No obstante, dado que se formaliza mediante el modelo lineal univariante, consideramos que debe darse por conocida. Partiendo de esta premisa, el objetivo fundamental de este capítulo es bastante modesto: clarificar las pequeñas modificaciones que hay que realizar para adaptar las técnicas del análisis de regresión múltiple a las del análisis de regresión multivariante. Al igual que sucediera en el anterior capítulo, el lector debe tener claro que conociendo la regresión múltiple se domina, casi automáticamente, la regresión multivariante. Por el mismo razonamiento, la técnica multivariante hereda el mismo

[1] Ver capítulo 4 del volumen1 dedicado a los Modelos Lineales.
[2] Ver capítulo 5 del volumen 1.

problema de su análoga univariante: la posible violación de los supuestos del modelo, uno de cuyos aspectos, la no normalidad de las variables respuesta, se estudiará en el presente capítulo. El problema de multicolinealidad, que no es en rigor una violación de los supuestos del modelo, se tratará e un capítulo posterior mediante el análisis de componentes principales.

5.1. El modelo de Regresión

Un modelo de regresión lineal multivariante[3] no es sino un modelo lineal normal expresado mediante coordenadas, es decir, un estructura estadística del tipo

$$Y \sim N_{n,p}(\mathbf{X}\beta, \mathtt{Id}, \Sigma), \quad \beta \in \mathcal{M}_{(q+1)\times p},\ \Sigma > 0,$$

donde $\mathbf{X}$ es una matriz $n \times (q+1)$ de rango $q+1$ cuya primera columna, denominada término independiente, es el vector $1_{\mathtt{n}}$. En lo que sigue que $\mathbf{X}$ se expresará mediante

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{z}_1[1] & \dots & \mathbf{z}_1[q] \\ \vdots & \vdots & & \vdots \\ 1 & z_{\mathtt{n}}[1] & \dots & z_{\mathtt{n}}[q] \end{pmatrix} \tag{5.1}$$

Se denotará por $\mathbf{Z}$ la matriz anterior desprovista de del término independiente, es decir $\mathbf{X} = (1_{\mathtt{n}}|\mathbf{Z})$. El modelo puede expresarse también de la siguiente forma:

$$Y = (1_{\mathtt{n}}|\mathbf{Z})\beta + \mathcal{E}, \qquad \mathcal{E} \sim N_{n,p}(0, \mathtt{Id}, \Sigma). \tag{5.2}$$

Nótese que, en el caso $p = 1$, tendremos un modelo de regresión lineal múltiple. Si, además, $q = 1$, se trata de un modelo de regresión lineal simple. La matriz aleatoria Y correspondiente a la observación de las variables respuestas en $\mathtt{n}$ individuos, y la matriz β se expresarán, respectivamente, mediante

$$Y = \begin{pmatrix} y_1[1] & \dots & y_1[p] \\ \vdots & & \vdots \\ y_{\mathtt{n}}[1] & \dots & y_{\mathtt{n}}[p] \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_{01} & \dots & \beta_{0p} \\ \vdots & & \vdots \\ \beta_{q1} & \dots & \beta_{qp} \end{pmatrix}.$$

De esta forma, (5.2) equivale a que, para cada $i = 1, \dots, \mathtt{n}$ y cada $j = 1, \dots, p$, se verifique

$$y_i[j] = \beta_{0j} + \beta_{1j}\mathbf{z}_i[1] + \dots + \beta_{qj}\mathbf{z}_i[q] + \varepsilon_{ij}, \tag{5.3}$$

[3]En este capítulo sólo consideraremos modelos de regresión de rango completo.

de manera que, si se denota $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})'$, entonces $\varepsilon_1, \ldots, \varepsilon_{\mathtt{n}}$ constituyan una muestra aleatoria simple de una distribución $N_p(0, \Sigma)$. En lo sucesivo, $y[1], \ldots, y[p]$ y $\mathtt{z}[1], \ldots, \mathtt{z}[q]$ denotarán las columnas de Y y Z, respectivamente. por otra parte, las filas de Z se denotarán por $\mathtt{z}_1, \ldots, \mathtt{z}_{\mathtt{n}}$. Así mismo, $\beta[1], \ldots, \beta[p]$ y $\beta_0, \ldots, \beta_q$ denotarán, respectivamente, las columnas y filas de β. Por último $\underline{\beta}$ y $\underline{\beta}[j]$, $j = 1, \ldots, p$, denotarán, respectivamente la matriz $q \times p$ y el vector de $\mathbb{R}^q$ que se obtienen excluyendo de β y $\beta[j]$ la fila β_0 y el número β_{0j}, respectivamente. Por lo tanto, se trata de la submatriz compuesta por los coeficientes de las variables explicativas (o, mejor dicho, vectores explicativos).

A partir de la matrices Y y Z podemos construir, siguiendo la notación empleada en el apéndice del volumen dedicado a los Modelos Lineales, las medias muestrales $\overline{y} \in \mathbb{R}^p$ y $\overline{\mathtt{z}} \in \mathbb{R}^q$; las matrices $\overline{Y}$ y $\overline{\mathsf{Z}}$ que se obtienen reemplazando cada fila de Y y Z por $\overline{y}'$ y $\overline{z}'$; las matrices $Y_0 = Y - \overline{Y}$ y $\mathsf{Z}_0 = \mathsf{Z} - \overline{\mathsf{Z}}$; las matrices de varianzas-covarianzas totales muestrales

$$S_{yy} = \mathtt{n}^{-1} Y_0' Y_0, \quad S_{\mathtt{ZZ}} = \mathtt{n}^{-1} \mathsf{Z}_0' \mathsf{Z}_0, \quad S_{y\mathtt{Z}} = \mathtt{n}^{-1} Y_0' \mathsf{Z}_0,$$

que componen la matriz $S_{(y\mathtt{Z})(y\mathtt{Z})}$. Por último, consideraremos la matriz de varianzas-covarianzas parciales muestral

$$S_{yy \cdot \mathtt{Z}} = S_{yy} - S_{y\mathtt{Z}} S_{\mathtt{ZZ}}^{-1} S_{\mathtt{Z}y}.$$

5.2. Regresión y correlación

En la propia formulación del modelo de Regresión se determina que, mientras que los valores correspondientes a las variables respuesta son aleatorios, los de las variables explicativas, las $\mathtt{z}$'s, no lo son, es decir, quedan determinados de antemano, de ahí que debamos hablar más bien de vectores explicativos. Aunque el control de las variables (o vectores) explicativas pueda reportar interesantes ventajas, no cabe de duda de que, en muchos estudios, por no decir la mayoría, los valores de las variables explicativas no han sido determinados de antemano sino que son tan aleatorios como las respuestas. En ese tipo de estudio, se seleccionan aleatoriamente individuos de una población dada y se anotan los valores de las distintas variables, tanto explicativas como respuestas. Es lo que se denomina un problema de Correlación. Estos estudios son simétricos, en el sentido de que los papeles que desempeñan las variables pueden intercambiarse. La pregunta que nos formulamos es la siguiente: ¿pueden aplicarse técnicas de regresión en sentido estricto, a problemas de correlación? Afortunadamente, la respuesta será positiva, como veremos a continuación.

En primer lugar, debemos formular de manera precisa el modelo de Correlación. Concretamente, vendrá dado simplemente por una muestra aleatoria simple de tamaño $\mathtt{n}$

$$\begin{pmatrix} Y_1 \\ Z_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_{\mathtt{n}} \\ Z_{\mathtt{n}} \end{pmatrix}$$

correspondiente a una distribución

$$N_{p+q}\left(\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix} \right),$$

siendo la matriz de covarianzas definida positiva y $\mathtt{n} \geq p+q$. Por lo tanto, se trata del modelo en el cual se resuelve el contraste de independencia (capítulo 3). Consideremos las matrices aleatorias $Y = (Y_1 \ldots Y_{\mathtt{n}})'$ y $Z = (Z_1 \ldots Z_{\mathtt{n}})'$. En ese caso, si se denota $\mu_1 = (\nu_1, \ldots, \nu_1)'$ y $\mu_2 = (\nu_2, \ldots, \nu_2)'$, se verifica, en virtud del teorema 1.21, que

$$(Y|Z) \sim N_{n,p+q}\left((\mu_1|\mu_2), \mathtt{Id}, \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix} \right).$$

Denótese

$$\Sigma = \Delta_{11} - \Delta_{12}\Delta_{22}^{-1}\Delta_{21}, \quad \beta = \begin{pmatrix} \nu_1' - \nu_2'\Delta_{22}^{-1}\Delta_{21} \\ \Delta_{22}^{-1}\Delta_{21} \end{pmatrix}.$$

En ese caso, dada una matriz $\mathtt{Z} \in \mathcal{M}_{n\times q}$, se verifica por el teorema 1.20

$$Y|Z = \mathtt{Z} \sim N_{n,p}\big((1_{\mathtt{n}}|\mathtt{Z})\beta, \mathtt{Id}, \Sigma\big).$$

Por lo tanto, se sigue de (1.7) que

$$Y = (1_{\mathtt{n}}|Z)\beta + \mathcal{E}, \quad Z \sim N_{n,q}(\mu_2, \mathtt{Id}, \Delta_{22}), \quad \mathcal{E} \sim N_{n,p}(0, \mathtt{Id}, \Sigma), \tag{5.4}$$

siendo Z independiente de $\mathcal{E}$ y tal que el rango de $(1_{\mathtt{n}}|Z)$ es $q+1$ con probabilidad 1 (esto último se deduce del teorema 1.27). Puede probarse sin dificultad (se deja como ejercicio) que, recíprocamente, (5.4) implica que $(Y_i'Z_i')'$, $i = 1, \ldots, \mathtt{n}$, constituyen una muestra aleatoria simple de una distribución p-normal no degenerada. Por lo tanto, el modelo de correlación puede expresarse también mediante (5.4).

Nótese pues que, a diferencia del modelo de regresión, Z esa su vez una matriz aleatoria que sigue una distribución normal matricial. La relación entre ambos modelos estriba en que el modelo de regresión se obtiene *condicionando* en el de correlación para un valor concreto de las variables explicativas, $\mathtt{Z}$. Este hecho es de vital importancia pues, teniendo en cuenta nuevamente (1.7), las distribuciones de los estimadores y estadísticos de contrastes que se obtienen para el modelo de regresión

son igualmente correctas para el de correlación, siempre que éstas no dependan de la matriz $\mathtt{Z}$. De ahí que los tests que propondremos a continuación para el modelo de regresión sean válidos para el de correlación. Además, la justificación teórica de los mismos es muy similar, al menos en el caso $p = 1$: si los tests para el modelo de regresión se obtienen tras reducir por suficiencia e invarianza (ver sección 2.2), desde el punto de vista del modelo de correlación pueden obtenerse los mismo tests reduciendo por suficiencia e invarianza, pero respecto a grupos de transformaciones distintos[4]. Todo ello es francamente importante teniendo en cuenta que, en la práctica y como ya hemos comentado, la matriz $\mathtt{Z}$ considerada no suele ser predeterminada con anterioridad a la ejecución del experimento (modelo de regresión), sino que es a su vez el valor observado de una matriz aleatoria.

Además, de todo esto se obtiene una justificación de todos los supuestos del modelo de Regresión a partir de la hipótesis de normalidad, pues si las observaciones constituyen una una muestra de una distribución $(p+q)$-normal, el modelo condicionado verifica automáticamente los supuestos de normalidad, independencia, homocedasticidad y linealidad. Desde el punto de vista práctico, esta afirmación no aporta demasiado, puesto que la hipótesis de $(p+q)$-normalidad es muy estricta. No obstante, lo dicho debe servirnos para que nos percatemos del íntimo vínculo existente entre normalidad y linealidad. Restringiremos nuestro estudio en lo que resta del capítulo al modelo de regresión, propiamente dicho

5.3. Estimación de los parámetros

Supuesto que se verifican las condiciones del modelo de regresión lineal, los problemas de Estimación y Tests de hipótesis se resuelven según los métodos expuestos en el capítulo 2, como veremos a continuación. Respecto al problema de Estimación, ya sabemos que los EIMV de β y Σ son[5], respectivamente,

$$\begin{aligned}\hat{\beta} &= (\mathtt{X}'\mathtt{X})^{-1}\mathtt{X}'Y,\\ \hat{\Sigma} &= \frac{1}{\mathtt{n}-(q+1)}Y'P_{\langle \mathtt{X}\rangle^{\perp}}Y = \frac{1}{\mathtt{n}-q-1}(Y'Y-\hat{\beta}'\mathtt{X}'Y).\end{aligned}$$

Se verifica entonces, en virtud del teorema 2.5, que

$$\begin{aligned}\hat{\beta} &\sim N_{q+1,p}(\beta,(\mathtt{X}'\mathtt{X})^{-1},\Sigma),\\ (\mathtt{n}-(q+1))\hat{\Sigma} &\sim W_p(\mathtt{n}-r,\Sigma)\end{aligned}$$

[4]Para más detalles, ver Arnold(1981), cap. 16 o bien el capítulo 5 del volumen 1 dedicado a los Modelos Lineales.

[5]Tener en cuenta, en la última expresión, que $Y'P_VY = \hat{\beta}'\mathtt{X}'Y$.

siendo ambos independientes. Se observa que el estimador de la columna $\beta[j]$, $j = 1, \ldots, p$, es el que se obtendría si consideramos una regresión lineal múltiple de la componente $y[j]$ respecto a las variables explicativas $\mathtt{z}[1], \ldots, \mathtt{z}[q]$ [6]. En ese sentido, precisamente, podemos afirmar que una regresión multivariante es una *composición* de p regresiones múltiples. Concretamente, se verifica

$$\underline{\hat{\beta}}[j] = S_{\mathtt{ZZ}}^{-1} S_{\mathtt{Z}y[j]}, \qquad \hat{\beta}_{0j} = \overline{y}[j] - \overline{\mathtt{z}}' \underline{\hat{\beta}}[j]. \tag{5.5}$$

Por lo tanto, el estimador de β puede expresarse como sigue.

$$\underline{\hat{\beta}} = S_{\mathtt{ZZ}}^{-1} S_{\mathtt{Z}y}, \qquad \hat{\beta}_0 = \overline{y}' - \overline{\mathtt{z}}' \underline{\hat{\beta}}. \tag{5.6}$$

Por otro lado, es fácil verificar que

$$P_{\langle \mathtt{1_n} \rangle^\perp} Y = Y_0, \qquad P_{\langle \mathtt{x} \rangle | \langle \mathtt{1_n} \rangle} Y = \mathtt{Z}_0 \underline{\hat{\beta}}. \tag{5.7}$$

En consecuencia,

$$\begin{aligned} S_3 &= Y' P_{\langle \mathtt{x} \rangle^\perp} Y \\ &= Y' P_{\langle \mathtt{1_n} \rangle^\perp} Y - Y' P_{\langle \mathtt{x} \rangle | \langle \mathtt{1_n} \rangle} Y \\ &= Y_0' Y_0 - \left(Z_0 \underline{\hat{\beta}} \right)' Z_0 \underline{\hat{\beta}} \\ &= {}_{\mathtt{n}}S_{yy} - S_{y\mathtt{Z}} S_{\mathtt{ZZ}}^{-1} S_{\mathtt{Z}y} \end{aligned}$$

Por lo tanto, el estimador de la matriz de covarianza puede expresarse también de forma más intuitiva mediante

$$\hat{\Sigma} = \frac{\mathtt{n}}{\mathtt{n} - (q+1)} \left(S_{yy} - S_{y\mathtt{Z}} S_{\mathtt{ZZ}}^{-1} S_{\mathtt{Z}y} \right). \tag{5.8}$$

Por otra parte, se sigue del teorema 1.15 que, para cada $j = 1, \ldots, p$, la distribución del estimador de la columna $\beta[j]$, correspondiente a la componente $y[j]$, es la siguiente

$$\hat{\beta}[j] \sim N_{q+1} \left(\beta[j], \Sigma_{jj} \cdot (\mathtt{X}'\mathtt{X})^{-1} \right), \tag{5.9}$$

donde Σ_{jj} denota el j-ésimo elemento de la diagonal de Σ. Así mismo, para cada $j = 0, 1, \ldots, q$, la distribución del estimador de la fila β_j, correspondiente a la variable explicativa j-ésima (o altérmino independiente cuando $k = 0$) es

$$\hat{\beta}_j' \sim N_p \left(\beta_j', \left[(\mathtt{X}'\mathtt{X})^{-1} \right]_{jj} \cdot \Sigma \right). \tag{5.10}$$

A partir de (5.9) y (5.10), se obtienen los elipsoides de confianza a nivel $1 - \alpha$ (5.17) y (5.18), para los vectores $\beta[j]$ y β_j', respectivamente (ver cuestiones propuestas).

[6] Ver capítulo 4 del volumen 1.

5.4. Tests de hipótesis

Como ya sabemos el modelo de regresión no es sino un modelo lineal en el que el subespacio V que recorre el parámetro media viene generado por las columnas de la matriz $\mathtt{X}$. En el capítulo 2 hemos estudiado cómo contrastar una hipótesis inicial del tipo $H_0 : \mu \in W$. Teniendo en cuenta que el parámetro β lo componen las coordenadas de μ respecto de la base $\mathtt{X}$, es fácil demostrar que H_0 puede expresarse en términos de β mediante $H_0 : A\beta = 0$, siendo $A = C'\mathtt{X}$ para cualquier base C del subespacio $\langle \mathtt{X} \rangle | W$. En ese caso, A es una matriz de dimesnsiones $(q+1-\mathtt{dim}W) \times (q+1)$ y rango $(q+1) - \mathtt{dim}W$.

Recíprocamente, dada una hipótesis inicial del tipo $A\beta = 0$, siendo A una matriz de orden $m \times (q+1)$ y rango m (lo cual implica que $m \leq q+1$), existe un subespacio lineal $W \subset \langle \mathtt{X} \rangle$ de dimensión $q+1-m$ tal que la hipótesis inicial $H_0 : A\beta = 0$ equivale a $H_0 : \mu \in W$. Concretamente, se trata de la imagen del subespacio $\overline{W}$ de dimensión $q+1-m$, constituido por los vectores b de $\mathbb{R}^{q+1}$ tales que $Ab = 0$, por la aplicación lineal inyectiva que a cada b en $\mathbb{R}^{q+1}$ le asigna el vector $\mathtt{X}b$ de $\langle \mathtt{X} \rangle$.

En definitiva, contrastar hipótesis del tipo $\mu \in \langle \mathtt{X} \rangle$ equivale, en términos de β, a contrastar hipótesis del tipo $A\beta = 0$, siendo A una matriz de orden $m \times (q+1)$ y rango completo. De hecho, en regresión lineal expresaremos así las hipótesis iniciales. escogida. Conviene pues expresar el estadísticos $S_2 = Y' P_{\langle \mathtt{X} \rangle | W_A} Y$, estudiados en el capítulo 2, en función de la matriz A correspondiente (el valor del estadístico $S_3 = Y' P_{\langle \mathtt{X} \rangle^\perp} Y$ no depende de A). Para ello es conveniente encontrar una base adecuada de $\langle \mathtt{X} \rangle | W_A$.

Lema 5.1.
Dada una matriz A orden $m \times (q+1)$ y rango m, las columnas de la matriz $C = \mathtt{X}(\mathtt{X}'\mathtt{X})^{-1}A'$ constituyen una base del subespacio $\langle \mathtt{X} \rangle | W_A$, de dimensión m.

Demostración.
Veamos que las columnas de C son linealmente independientes. En efecto, si existe un vector $g \in \mathbb{R}^m$, tal que $Cg = 0$, entonces, $A\mathtt{X}'Cg = 0$. Dado que AA' es una matriz cuadrada de orden m y rango m, podemos afirmar que

$$0 = (AA')^{-1}AX'Cg = (AA')^{-1}A\mathtt{X}'\mathtt{X}(\mathtt{X}'\mathtt{X})^{-1}A'g = g.$$

Por lo tanto, el rango de C es m. Falta probar que las columnas de C son ortogonales a $W_{\mathtt{X},A}$, es decir, que dado $b \in \mathbb{R}^m$ tal que $Ab = 0$, se verifica $(\mathtt{X}b)'C = (0, \ldots, 0)$. Efectivamente,

$$(\mathtt{X}b)'C = b'\mathtt{X}'\mathtt{X}(\mathtt{X}'\mathtt{X})^{-1}A' = b'A' = (0, \ldots, 0).$$

Teorema 5.2.
Un estadístico invariante maximal para el contraste $A\beta = 0$ es $t_1, \ldots, t_b$, las raíces positivas del polinomio $|S_2 - tS_3|$, donde $b = \min\{p, q+1-k\}$,

$$\begin{aligned} S_2 &= \hat{\beta}' A' \left(A(\mathtt{X}'\mathtt{X})^{-1} A'\right)^{-1} A\hat{\beta} \\ S_3 &= \mathtt{n} S_{yy} - S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y} \end{aligned}$$

Demostración.
La expresión de S_2 se obtiene sin más que sustituir de acuerdo con el lema anterior, mientras que la de S_3, se sigue directamente de (5.8).

Una vez obtenidos las valores $t_1, \ldots, t_b$, se procederá a aplicar uno de los cuatro tests propuestos en la teoría (Wilks, Lawley-Hotelling, Roy o Pillay). Destacaremos tres tipo de contrastes:

(a) $H_0 : \beta = \beta^*$, $\beta^* \in \mathcal{M}_{(q+1)\times p}$. En este caso, se considera $Y - X\beta^*$. Entonces, se trata de contrastar la hipótesis inicial $\beta = 0$, es decir, $\mathtt{Id}\beta = 0$, a partir de los nuevos datos obtenidos. En ese caso, $\dim(\langle \mathtt{X} \rangle | W) = q + 1$.

(b) $H_0 : \underline{\beta} = 0$. Se corresponde con $H_0 : A\beta = 0$, donde

$$A = \begin{pmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \end{pmatrix}.$$

Por lo tanto, en este caso $W = \langle 1_n \rangle$ y la dimensión de $\langle \mathtt{X} \rangle | W$ es q. En el caso nulo, estaremos afirmando que los parámetros predictores z no tienen ninguna influencia sobre las variables dependientes. En el modelo de correlación, esta hipótesis equivale a la indepedencia entre las variables respuesta y las explicativas.

El cálculo de S_2 puede efectuarse teniendo en cuenta el teorema anterior o bien, directamente, a partir de (5.7). En todo caso, se tiene que

$$S_2 = \mathtt{n} S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y}. \tag{5.11}$$

De (5.7) se sigue también que

$$Y' P_{\langle 1_{\mathtt{n}} \rangle^\perp} Y = \mathtt{n} S_{yy}.$$

Teniendo en cuenta (2.3), se deduce que el test de Wilks para contrastar la hipótesis inicial $H_0 : \underline{\beta} = 0$ tiene por estadístico de contraste

$$\lambda_1(Y)(\mathbf{Z}) = \frac{|S_{yy} - S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y}|}{|S_{yy}|}.$$

Luego, aplicando el teorema 13.6, se tiene

$$\lambda_1(Y)(\mathbf{Z}) = \frac{|S_{(\mathbf{Z}y)(\mathbf{Z}y)}|}{|S_{\mathbf{ZZ}}||S_{yy}|}.$$

Como podemos ver, se trata de un algoritmo bastante sencillo. Se denominan coeficientes de correlación canónica al cuadrado, denotándose por $r_1^2, \ldots, r_p^2$, a los autovalores[7] de la matriz

$$S_{yy}^{-1} S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y}.$$

Suponen una generalización del coeficiente de correlación múltiple R^2 al caso multivariante. Se puede comprobar fácilmente que, en el caso $p = 1$, que se corresponde con la regresión múltiple, se obtiene un único coeficiente de correlación canónica r_1^2, concretamente

$$\frac{S_{yz} S_{zz}^{-1} S_{zy}}{s_y^2} = R^2 = r_1^2.$$

La relación entre los coeficientes de correlación canónica y los autovalores $t_1, \ldots, t_b$ correspondientes al contraste de la hipótesis $\underline{\beta} = 0$ (por lo tanto, $b = \text{mín}\{p, q\}$) es la siguiente: $t_1, \ldots, t_b$ son las raíces positivas de

$$\begin{aligned} |S_2 - tS_3| &= |S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y} - t[S_{yy} - S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y}]| \\ &= |(t+1) S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y} - tS_{yy}|, \end{aligned}$$

es decir, $\frac{t_i}{1+t_i}$ son las raíces positivas de $|S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y} - xS_{yy}|$, que coinciden con los b autovalores positivos de $S_{yy}^{-1} S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y}$. Por lo tanto, la relación obtenida es la siguiente

$$t_i = \frac{r_i^2}{1 - r_i^2}, \quad i = 1, \ldots, b. \tag{5.12}$$

Así, el estadístico de contraste del test de Wilks, por ejemplo, puede expresarse en función de los coeficientes de correlación canónica como sigue:

$$\lambda_1(Y)(Z) = \prod_{i=1}^{b} (1 - r_i^2) \tag{5.13}$$

[7] Realmente, se considerarán sólo los b primeros, donde $b = \text{mín}\{p, q\}$, pues el resto son nulos.

Nótese la relación existente entre (3.6) y (5.13). Por su parte, el estadístico de contraste del test de Pillay, se expresa de esta forma

$$\lambda_4(Y)(Z) = \sum_{i=1}^{b} r_i^2. \tag{5.14}$$

(c) Consideremos una descomposición de las variables explicativas en dos grupos

$$\mathsf{Z}_D = \{\mathtt{z}_{[1]}^D, \ldots, \mathtt{z}_{[d]}^D\}, \quad Z_R = \{\mathtt{z}_{[1]}^R, \ldots, \mathtt{z}_{[q-d]}^R\},$$

y sean β_R y β_D las submatrices de $\underline{\beta}$ asociadas a los grupos Z_R y Z_D, respectivamente. Consideremos entonces el contraste parcial $H_0 : \beta_D = 0$. Se corresponde con $A\beta = 0$, siendo

$$A = \begin{pmatrix} 0 & \overset{q+1-d}{\cdots} & 0 & 1 & & 0 \\ \vdots & & \vdots & & \ddots & \\ 0 & \ldots & 0 & 0 & & 1 \end{pmatrix}.$$

La dimensión de $\langle \mathsf{X} \rangle | W$ es, en este caso, d. Se trata pues de contrastar si el grupo de variables explicativas Z_D tiene influencia sobre Y. En caso negativo habrá que considerar su eliminación. En las condiciones del modelo de correlación, esta hipótesis equivale a la independencia condicional entre Y y Z_D dado Z_R.

Especial interés tiene el caso $d = 1$, es decir, cuando se plantea la posible eliminación de una variable explicativa. Es lo que se denomina un contraste parcial. En ese caso, tendremos un único autovalor positivo que se contrastará con un cuantil de la distribución $T^2_{p,\mathtt{n}-(q+1)}$ (que, recordemos, es, salvo una constante, un modelo F-Snedecor). Además, puede comprobarse (cuestión propuesta) que el test a nivel α decide rechazar la hipótesis inicial si, y sólo si, el vector 0 queda fuera del elipsoide (5.18). También puede plantearse un contraste parcial para el término independiente $1_{\mathtt{n}}$.

El test de Wilks proporciona un interesante algoritmo para los contrastes parciales: en general, se tiene que

$$\lambda_1 = \frac{|S_3|}{|S_2 + S_3|}.$$

Denótese por $(Y)(\mathsf{Z})$ el modelo completo y por $(Y)(\mathsf{Z}_R)$ el modelo reducido, donde se ha eliminado el grupo Z_D. La hipótesis inicial $H_0^D : \mu \in W_R$ determinada por $\beta_D = 0$ coincide con la hipótesis estructural del modelo reducido,

$V_R = (1_{\mathtt{n}}|\mathtt{Z}_R)$. Si se desea contrastar la hipótesis $H_0 : \mu \in \langle 1_{\mathtt{n}} \rangle$, equivalente a $\underline{\beta} = 0$ (es decir, se trata de un contraste tipo (b)), en el modelo completo V, se tendrá el estadístico de Wilks

$$\lambda_1(Y)(\mathtt{Z}) = \frac{|Y' P_{\langle 1_{\mathtt{n}} \rangle^\perp} Y|}{|Y' P_{V^\perp} Y|}.$$

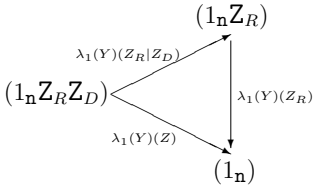
Si se desea contrastar la hipótesis inicial $H_0 \mu \in \langle 1_{\mathtt{n}} \rangle$ (contraste tipo (b)) en el modelo V_R reducido, se tiene

$$\lambda_1(Y)(\mathtt{Z}_R) = \frac{|Y' P_{V_R^\perp} Y|}{|Y' P_{\langle 1_{\mathtt{n}} \rangle^\perp} Y|}.$$

Luego, si se desea realizar el contraste parcial $H_0^D : \beta_D = 0$ en el modelo completo, se tiene

$$\lambda_1(Y)(\mathtt{Z}_R|\mathtt{Z}_D) = \frac{|Y' P_{V^\perp} Y|}{|Y' P_{V_R^\perp} Y|} = \frac{\lambda_1(Y)(\mathtt{Z}_R)}{\lambda_1(Y)(\mathtt{Z})}.$$

Por lo tanto, el estadístico del test de Wilks para resolver un contraste tipo (c) puede obtenerse como cociente entre los estadísticos de Wilks para los contrastes tipo (b) en los modelos completo y reducido.



Dado que $\lambda_1(Y)(\mathtt{Z}_R) = \frac{|S_{(\mathbf{Z}_R y)(\mathbf{Z}_R y)}|}{|S_{\mathbf{Z}_R \mathbf{Z}_R}||S_{yy}|}$, el estadístico queda como sigue

$$\lambda_1(Y)(\mathtt{Z}_R|\mathtt{Z}_D) = \frac{|S_{\mathbf{Z}_R \mathbf{Z}_R}| \cdot |S_{(\mathbf{Z}y)(\mathbf{Z}y)}|}{|S_{\mathbf{Z}\mathbf{Z}}| \cdot |S_{(\mathbf{Z}_R y)(\mathbf{Z}_R y)}|}.$$

Nótese que, conocida la matriz de covarianzas completa $S_{(\mathbf{Z}y)(\ zy)}$, cada contraste parcial se reduce a escoger las submatrices apropiadas, $S_{\mathbf{Z}_R \mathbf{Z}_R}$ y $S_{(\mathbf{Z}_R y)(\mathbf{Z}_R y)}$ y efectuar la división anterior.

En lo referente al vector respuesta Y podemos también contrastar la utilidad de un subgrupo de componentes, de tamaño d, en el modelo de regresión. Ello

puede hacerse desde dos puntos de vista: el primero sería contrastar la nulidad de las columnas de β asociadas a dichas componentes. Ésta sería una hipótesis del tipo

$$H_0 : \beta A = 0,$$

lo cual se correspondería con un contraste generalizado[8]. El test que se deriva de la teoría coincide, en el caso $d = 1$, con el test F para un contraste tipo (a) en el modelo de regresión múltiple.

No obstante, si consideramos un modelo de correlación, resulta más congruente (por pura simetría), a tenor de lo visto anteriormente, optar por desechar un subgrupo de variables dependientes Y_D frente a otro Y_R cuando se da la independencia condicional entre Y_D y Z dado Y_R. Si trasladamos este argumento al modelo de regresión, contrastaremos si, en el modelo $(Y_D)(\mathsf{Z}, Y_R)$, la matriz de los coeficientes correspondientes a Z vale 0. Ello significa que, conocido Y_R, Z no aporta ninguna información adicional respecto a Y_D, es decir, que podemos limitar nuestro estudio a la relación entre Y_R y Z, siendo Y_D redundante. El estadístico del test de Wilks para contrastar dicha hipótesis es el siguiente:

$$\begin{aligned}\lambda_1(Y_R|Y_D)(\mathsf{Z}) &= \frac{\lambda_1(Y_D)(Y_R)}{\lambda_1(Y_D)(\mathsf{Z}, Y_R)} \\ &= \frac{|S_{(\mathbf{z}y)(\mathbf{z}y)}| \cdot |S_{y_R y_R}|}{|S_{(\mathbf{z}y_R)(\mathbf{z}y_R)}| \cdot |S_{yy}|} \\ &= \frac{\lambda_1(Y_R)(\mathsf{Z})}{\lambda_1(Y_R, Y_D)(\mathsf{Z})}.\end{aligned}$$

Para acabar con esta sección vamos a considerar el problema selección de variables, consistente en la eliminación de aquellas que no tengan relevancia en el modelo, con la idea de obtener una estructura lo más sencilla posible. Distinguimos entre selección de variables explicativas y selección de variables respuesta. En ambos casos podrían considerarse todos los posibles modelos reducidos y considerar los contrastes correspondientes. No obstante, esta técnica no se utilizan normalmente, pues el número de posibles modelos reducidos suele ser demasiado grande. En su lugar suele recurrirse a métodos de eliminación hacia atrás (backward) o de introducción hacia adelante (forward).

En la selección de variables explicativas, el método backward consiste en considerar inicialmente todas las variables $Z's$ en el modelo y considerar todos los test F's parciales. Se elimina aquella variable que aporte el resultado menos significativo.

[8] Ver sec. 2.7

A continuación se vuelve a aplicar todos los test parciales en el modelo (reducido) resultante y se elimina otra. El proceso de eliminación finaliza cuando todos los tests parciales dan resultado significativo (se suele subir la cota de significación a 0.10).

El método forward consiste en contrastar todas las q regresiones de Y sobre las variables individuales de Z e introducir la variable que presente un resultado más significativo. A continuación se consideran los $q-1$ modelos que tienen como variables independientes la ya introducida y otra, considerándose los contrastes parciales para decidir la eliminación de la otra. Ingresa en el modelo la que aporte un resultado más significativo. El proceso continúa hasta que no se da ningún resultado significativo.

El método stepwise es una mezcla de los anteriores: se basa en el método forward con la añadidura siguiente: cada vez que se introduce una nueva variable, se realiza un método backward para las variables introducidas hasta ese momento. De esta forma, la introducción de una nueva variable puede suponer la eliminación de otra que ya había sido introducida previamente. En el caso se selección de variables dependientes, los métodos funcionan de manera análoga.

Los algoritmos para la ejecución de estos métodos son muy sencillos teniendo en cuenta las expresiones anteriores del estadístico de Wilks para los contrastes parciales.

5.5. Estudio asintótico.

A continuación, vamos a realizar algunas consideraciones asintóticas acerca de los modelos de regresión y correlación. Supongamos que no se verifica la normalidad de la matriz residual $\mathcal{E}$. La cuestión es determinar el comportamiento de los estimadores[9] y tests de hipótesis considerados a medida que vamos añadiendo datos a la muestra, es decir, cuando agregamos filas a las matrices Y y $\mathtt{X}$. En ese caso, si $\mathtt{X_n}$, $\mathtt{n} \in \mathbb{N}$, denota la matriz estructural de valores predictores en la fase $\mathtt{n}$-ésima (es decir, con $\mathtt{n}$ filas) la condición de Huber (2.11) puede expresarse de la siguiente forma

$$m(\mathtt{X}_n(\mathtt{X}_n'\mathtt{X}_n)^{-1}\mathtt{X}_n') \longrightarrow 0 \tag{5.15}$$

Recordemos que, dada una matriz A semidefinida positiva, $m(A)$ es el máximo valor de su diagonal. En nuestro caso y según se demuestra en el capítulo 3 del volumen

[9] Realmente, respecto al comportamiento asintótico de los estimadores no vamos a añadir nada a lo que se dijo en el capítulo 2.

dedicado a los Modelos Lineales, estos valores, que se denotan por ν_{ii}, $i = 1, \ldots, \mathtt{n}$, se obtienen mediante

$$\nu_{ii} = \frac{1}{\mathtt{n}}\big(1 + (\mathtt{z}_i - \overline{\mathtt{z}})' S_{\mathtt{ZZ}}^{-1} (\mathtt{z}_i - \overline{\mathtt{z}})'\big).$$

Dicho de otra forma

$$\nu_{ii} = \frac{1}{\mathtt{n}}(1 + d^2(\mathtt{z}_i, \overline{\mathtt{z}})),$$

donde d^2 denota la distancia de Mahalanobis entre cada vector $\mathtt{z}_i$, $i = 1, \ldots, \mathtt{n}$, y la media de todos ellos. En consecuencia, la condición de Huber queda reducida a

$$\sup_{i \leq \mathtt{n}} \frac{d^2(\mathtt{z}_i, \overline{\mathtt{z}})}{\mathtt{n}} \longrightarrow 0. \tag{5.16}$$

Desde un punto de vista práctico esto se traduce en que, si el tamaño de muestra es *suficientemente grande* y no existen *outliers* importantes en las filas de $\mathtt{z}$, la violación del supuesto de normalidad de los residuos no afecta apenas al nivel de significación del test. En ese sentido, la validez asintótica del modelo está condicionada a la ausencia de valores *potencialmente influyentes*.

Si la matriz explicativa constituyen una muestra aleatoria simple Z, de tamaño $\mathtt{n}$, de un vector aleatorio q-dimensional, la conclusión es más interesante, si cabe. Puede demostrarse[10] que, si el vector aleatorio explicativo está dominado por la medida de Lebesgue en $\mathbb{R}^q$ y todas sus componentes tienen momentos de orden 2 finitos, la condición (5.16) se verifica con probabilidad 1, en cuyo caso podemos obviar el supuesto de normalidad desde un punto de vista asintótico.

No obstante, no conviene ser excesivamente optimistas en cuanto a estos resultados pues, en la práctica, la violación de la normalidad de los residuos en los modelos (5.2) y (5.4), suele venir asociada al incumplimiento del supuesto de linealidad o de homocedasticidad, o incluso de ambos.

5.6. Regresión con variables ficticias. Mancova

Como hemos visto, los problema de comparación de medias y de regresión lineal pueden formalizarse mediante un mismo modelo: el modelo lineal. En el caso de la regresión, esto se consigue considerando el parámetro $\mu = \mathtt{X}\beta$, que, según se deduce de (5.2), debe recorrer el subespacio $V = \langle \mathtt{X} \rangle$. Recíprocamente, un problema de comparación de medias (manova), puede transformarse en otro de regresión respecto

[10] Arnold, *Asymptotic Validity of F Test for the Ordinary Linear Model and Multiple Correlation Model*, Journal of the American Statistical Association, Dec. 1980, Vol. 75, 890-894.

a ciertas variables explicativas creadas al efecto, junto con un término independiente. Veremos únicamente cómo se hace en el manova con un factor. En el capítulo 6 del volumen 1 dedicado a los Modelos Lineales se consideran diseños con más factores aunque, eso sí, equilibrados.

Así pues, asumiendo las notaciones y condiciones de la sección 4.3, se trata de contrastar la hipótesis inicial de igualdad de medias $H_0 : \nu_1 = \ldots = \nu_r$, partiendo de sendas muestras independientes de distribuciones p-normales con matriz de covarianzas común. En ese caso, el parámetro media μ recorres el subespacio V generado por los vectores $\mathtt{v}_i$, definidos en (4.14). Dado que $1_{\mathtt{n}} \subset V$, el problema se reduce a considerar una base X de V que contenga el término independiente. De esta forma, si $\mathsf{X} = (1_{\mathtt{n}}|\mathsf{Z})$ es una base de V, nuestro problema puede entenderse como una regresión lineal multivariate respecto a Z, y la hipótesis $H_0 : \underline{\beta} = 0$ equivaldrá a la igualdad de medias. A la hora de elegir Z y si las muestras son de tamaño idéntico número de datos de cada muestra es el mismo , puede resultar natural considerar la descomposición ortogonal

$$V = \langle 1_{\mathtt{n}} \rangle \oplus V|\langle 1_{\mathtt{n}} \rangle$$

y considerar, precisamente, una base de $V|\langle 1_{\mathtt{n}} \rangle$. En ese caso, el coeficiente del término independiente coincidirá con la media aritmética de las r medias, $\overline{\nu}$, mientras que los coeficientes de las variables explicativas $\mathtt{z}[j]$, $j = 1, \ldots, r-1$, equivaldrán, respectivamente, a las diferencias $\nu_i - \overline{\nu}$, $i = 1, \ldots, r-1$.

También puede considerarse, y así lo hace el programa SPSS, la matriz Z compuesta por las columnas $(\mathtt{v}_1, \ldots, \mathtt{v}_{r-1})$. En ese caso, el término independiente equivaldrá a la media ν_r, mientras que que el coeficiente de $\mathtt{z}[j]$ será igual a $\nu_j - \nu_r$, $j = 1, \ldots, r-1$.

En todo caso, el contraste de igualdad de medias se convierte en una contraste tipo (b) respecto a ciertas variables $\mathtt{z}[1], \ldots \mathtt{z}[r-1]$, denominadas *ficticias*, que indican en definitiva el grupo al que pertenece cada individuo. Por ello, el estadístico de contraste, que depende exclusivamente de los autovalores $t_1, \ldots, t_b$ puede expresarse en función de los coeficientes de correlación canónica entre las variables observadas y las ficticias, $r_1^2, \ldots, r_b^2$. La relación entre ambos es la expresada en (5.12). En el caso univariante, es decir, cuando $p = 1$, el contraste se resolverá a través del coeficiente de correlación múltiple al cuadrado.

Ésta es, realmente, la forma de proceder del programa SPSS para resolver una comparación de media, es decir, en el momento que introducimos un factor de variabilidad, se generan tantas variables ficticias como niveles tenga el factor menos 1 y se realiza una regresión lineal respecto a las mismas. El análisis de la covarianza (ancova en el caso univariante y mancova en el multivariante) combina los métodos de regresión y comparación de medias pues, dadas q variables explicativas y un

factor cualitativo, realiza una regresión lineal diferente para cada nivel del factor. Esto se consigue mediante una regresión respecto a las variables ficticias asociadas al factor, las variables explicativas consideradas y los productos (interacciones) entre ambos tipos de variables. Si no se consideran las interacciones, las ecuaciones de regresión de los distintos niveles del factor podrán diferir únicamente en el coeficiente del término independiente. Todo lo dicho no es sino una generalización del caso univariante ($p = 1$), que se estudia, insistimos, en el capítulo 6 del volumen 1.

Cuestiones propuestas

1. Comparar (1.5) con (5.6).

2. Hemos visto en la teoría que bajo la hipótesis de normalidad matricial se verifican los supuestos del modelo de correlación lineal. Demostrar la implicación recíproca, es decir, que si Y y X son dos matrices aleatorias en las condiciones del modelo de correlación lineal, entonces la matriz (YX) es normal matricial.

3. Demostrar que el elipsoide siguiente es una región de confianza al nivel $1-\alpha$ para el parámetro $\beta[j]$, $j = 1, \ldots, p$:

$$\mathcal{E}_\alpha(Y) = \left\{ x \in \mathbb{R}^{q+1} \colon (q+1)^{-1} \hat{\Sigma}_{jj}^{-1} \big(x - \hat{\beta}[j]\big)' \mathbf{X}'\mathbf{X} \big(x - \hat{\beta}[j]\big) \leq F_{q+1,\mathtt{n}-q-1}^{\alpha} \right\} \tag{5.17}$$

4. Demostrar que el elipsoide siguiente es una región de confianza al nivel $1-\alpha$ para el parámetro β_j, $j = 0, \ldots, q$:

$$\mathcal{E}_\alpha(Y) = \left\{ x \in \mathbb{R}^{p} \colon \frac{1}{\big[(\mathbf{X}'\mathbf{X})^{-1}\big]_{jj}} \big(x' - \hat{\beta}_j\big) \hat{\Sigma}^{-1} \big(x - \hat{\beta}_j'\big) \leq T_{p,\mathtt{n}-q-1}^{\alpha} \right\} \tag{5.18}$$

5. Demostrar que test UMP invariante a nivel α y de razón de verosimilitudes para el contraste parcial de una única variable explicativa es consistente con el elipsoide de confianza (5.18), en el sentido de que el test rechaza la hipótesis inicial si, y sólo si, el vector 0 queda fuera del elipsoide.
   Indicación: considerar el teorema 5.2 junto con el lema 2.6.

6. Demostrar (4.16).

7. Demostrar que, en el estadístico de contraste de la hipótesis inicial $H_0 : \underline{\beta} = 0$, se verifica

$$S_2 = S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y}.$$

Capítulo 6

Análisis de correlación canónica

Los coeficientes de correlación canónica constituyen la generalización natural del coeficiente de correlación múltiple R al caso multivariante. En el presente capítulo veremos en qué sentido podemos afirmar esto. Ya hemos hecho referencia a estos coeficientes en dos ocasiones: en el contraste de independencia y en contraste de la hipótesis $H_0 : \underline{\beta} = 0$. Realmente, se trata de un mismo problema, sólo que el primer contraste se realiza en el modelo de correlación mientras que el segundo se efectúa en el de regresión, que se obtiene a a partir del anterior, recordemos, condicionando en las variables explicativas.

Aquí aparece por primera una vez una técnica típica, quizás la más característica, del análisis multivariante, consistente en el cálculo de lo que algunos autores denominan valores teóricos, que son combinaciones en lineales (es decir, sumas ponderadas) de las variables consideradas originalmente. Desde un punto de vista geométrico podemos interpretar los valores teóricos como proyecciones de las observaciones originales sobre determinados ejes, que varían según la finalidad del estudio (correlación canónica, componentes principales, análisis discriminante). Esta forma de proceder está orientada a estructurar los datos de manera canónica o natural, dependiendo del propósito perseguido, y puede dar pie a una profunda reducción en la dimensión verdadera del problema.

6.1. Definición

Empezaremos recordando los coeficientes de correlación simple y múltiple. Dadas dos variables aleatorias reales Y y Z con varianzas finitas σ_y^2 y σ_z^2 y covarianza σ_{zy},se

define el coeficiente de correlación lineal simple mediante

$$\rho_{zy} = \frac{\sigma_{zy}}{\sigma_z \sigma_y} \in [-1, 1].$$

Se prueba en el apéndice del volumen 1, dedicado a los Modelos Lineales, que la varianza parcial, definida mediante

$$\sigma_{yy \cdot z} = \sigma_y^2 - \sigma_{yz} \sigma_z^{-2} \sigma_{zy},$$

es la varianza total de la distribución de $P_{\langle 1, Z \rangle^\perp} Y$, y por lo tanto expresa la parte de varianza de Y no explicada mediante la mejor regresión lineal posible sobre Z. Dado que

$$\sigma_{yy \cdot z} = \sigma_y^2 (1 - \rho_{zy}^2),$$

el coeficiente ρ_{zy}^2, denominado de determinación, se interpreta como la proporción de variabilidad Y explicada mediante la regresión lineal respecto a Z. Bajo las condiciones del modelo de correlación lineal simple, es decir, cuando $(Y, Z)'$ sigue un modelo 2-normal, ρ_{zy}^2 será la proporción de varianza de Y explicada, a secas, por Z, puesto que la función que mejor explica Y a partir de Z (es decir, $E(Y|Z)$) es, en este caso, una recta. Así pues, un valor nulo del coeficiente de correlación equivale a la independencia entre Y y Z. Además, en este modelo, podemos aplicar conocidos métodos de inferencia respecto al parámetro ρ_{zy}. Así el EMV de ρ_{zy} es

$$r_{zy} = \frac{s_{zy}}{s_z s_y},$$

donde s_{zy}, s_z y s_y son los EMV de σ_{zy}, σ_z y σ_y, respectivamente. En Anderson (1958) podemos encontrar la distribución asintótica de r_{zy}^2 y un test a nivel α para contrastar la hipótesis $\rho_{zy} = 0$, todo ello bajo la hipótesis de normalidad bivariante.

Si en vez de una variable explicativa Z contamos con q variables explicativas Z_j, donde $j = 1, \ldots, q$ (se denota $Z = (Z_1, \ldots, Z_q)'$) definimos el coeficiente de correlación múltiple (al cuadrado) como sigue

$$\rho_{y \cdot Z}^2 = \frac{\Sigma_{yz} \Sigma_{zz}^{-1} \Sigma_{zy}}{\sigma_y^2}.$$

Puede demostrarse[1] la siguiente proposición:

$$\rho_{y \cdot Z}^2 = \max_{\beta \in \mathbb{R}^q} \rho_{y, \beta' Z}^2. \tag{6.1}$$

[1] Ver el capítulo 4 del volumen 1. No obstante, probaremos aquí un resultadomásgeneral.

Es decir, que $\rho_{y\cdot Z}$ es la maxima correlación lineal simple entre Y y cualquier combinación lineal de las variables explicativas. Por cierto que el vector β donde se alcanza dicho máximo es el que se obtiene como solución al problema de regresión lineal. El coeficiente de correlación múltiple puede interpretarse también como la parte de la variabilidad de Y explicada por la regresión lineal sobre $Z_1, \ldots, Z_q$. En las condiciones del modelo de Correlación, el EMV de $\rho^2_{y\cdot Z}$ es el siguiente

$$R^2_{y\cdot Z} = \frac{S_{yz} S_{zz}^{-1} S_{zy}}{s_y^2}, \tag{6.2}$$

o R^2 para abreviar. Admite otras expresiones alternativas, como puede verse en el capítulo 4 del volumen 1. También podemos encontrar en Anderson (1959) la distribución asintótica del R^2, bajo las condiciones del modelo de correlación, así como un test de hipótesis a nivel α para contrastar la hipótesis $\rho^2_{y\cdot Z} = 0$ [2].

Si, además, consideramos no una sino p variables dependientes, $Y_1, \ldots, Y_p$, siendo $Y = (Y_1, \ldots, Y_p)'$, necesitamos un concepto que generalice el anterior, es decir, un número entre 0 y 1 que suponga la máxima correlación lineal simple entre una combinación lineal de las Y's y otra combinación lineal de las Z's. Conviene también analizar dichas combinaciones con el objeto de explorar la estructura de correlación lineal entre las variables explicativas y respuesta. Todo ello nos conducirá a la definición de los coeficientes de correlación canónica.

Pues bien, consideremos dos vectores Y y Z, de dimensiones p y q, respectivamente, tales que $p \leq q$ y

$$\mathtt{E}\left[\begin{pmatrix} Y \\ Z \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \mathtt{Cov}\left[\begin{pmatrix} Y \\ Z \end{pmatrix}\right] = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{pmatrix} > 0.$$

El suponer que el vector tiene media nula tampoco implicará ninguna pérdida de generalidad. Lo mismo puede decirse respecto a la hipótesis $p \leq q$, puesto que Y y Z van a jugar papeles simétricos. Si $\alpha \in \mathbb{R}^p$ y $\beta \in \mathbb{R}^q$, entonces $\alpha' Y$ y $\beta' Z$ son variables aleatorias reales tales que

$$\mathtt{var}[\alpha' Y] = \alpha' \Sigma_{yy} \alpha, \quad \mathtt{var}[\beta' Z] = \beta' \Sigma_{zz} \beta, \quad \mathtt{cov}[\alpha' Y, \beta' Z] = \alpha' \Sigma_{yz} \beta.$$

Nuestro propósito inicial, insistimos, es buscar la máxima correlación entre una combinación lineal de las Z's y otra de las Y's, así como conocer para qué combinaciones

[2] Obviamente, estas inferencias extienden las ya comentadas en el caso de correlación simple y será a su vez generalizadas por las correspondientes a los coeficientes de correlación canónica que estudiaremos a continuación.

concretas se da dicha correlación. El siguiente paso será encontrar la máxima correlación entre combinaciones lineales incorreladas con las anteriores, y así sucesivamente. En definitiva, obtenemos el siguiente resultado, cuya demostración podemos encontrar en la sección (B) del Apéndice.

Teorema 6.1.
Sean $\rho_1^2 \ldots, \rho_b^2$ son los b primeros autovalores de la matriz $\Sigma_{yy}^{-1}\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}$, contados con su multiplicidad, y consideremos, para cada $i = 1, \ldots, b$, las variables reales $U_i = \alpha_i' Y$ y $V_i = \beta_i' Z$, donde α_i es el autovector asociado al autovalor ρ_i^2 para la matriz anterior tal que $\mathtt{var}(U_i) = 1$, mientras que β_i es el autovector asociado al mismo autovalor para la matriz $\Sigma_{zz}^{-1}\Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz}$, y tal que $\mathtt{var}(V_i) = 1$. Se verifica entonces:

(i) $\mathtt{var}[U_i] = \mathtt{var}[V_i] = 1,\ i = 1, \ldots, b,$

(ii) $\mathtt{Cov}\left[\begin{pmatrix} U_i \\ V_i \end{pmatrix}, \begin{pmatrix} U_j \\ Vj \end{pmatrix}\right] = 0,\ \forall i \neq j.$

(iii) $\rho_{U_i,V_i} = \rho_i,\ i = 1, \ldots, b.$

(iv) ρ_1 es la máxima correlación entre una variable del tipo $\alpha' Y$ y otra del tipo $\beta' Z$.

(v) Si $1 < i \leq b$, ρ_i es la máxima correlación entre entre una variable del tipo $\alpha' Y$ y otra del tipo $\beta' Z$, si imponemos la condición

$$\mathtt{Cov}\left[\begin{pmatrix} \alpha' Y \\ \beta' Z \end{pmatrix}, \begin{pmatrix} U_j \\ Vj \end{pmatrix}\right] = 0,\ \forall j < i.$$

Pues bien , en las condiciones anteriores se dice que $(U_1, V_1), \ldots, (U_b, V_b)$, es la serie de pares de variables canónicas. Los coeficientes $\rho_1, \ldots, \rho_b$, cuyos cuadrados son los b primeros autovalores de la matriz (13.4), se denominan coeficientes de correlación canónica. Lo que hemos hecho pues es reorganizar de una manera natural (canónica) la estructura de correlación inicial, tal y como se ilustra en la figura:

$$\begin{pmatrix} Z_1 \\ \vdots \\ Z_q \end{pmatrix} \longrightarrow \left(\begin{array}{c} V_1 \\ \hline \vdots \\ \hline V_b \end{array}\right) \begin{array}{c} \overset{\rho_1}{\longleftrightarrow} \\ \\ \overset{\rho_b}{\longleftrightarrow} \end{array} \left(\begin{array}{c} U_1 \\ \hline \vdots \\ \hline U_b \end{array}\right) \longleftarrow \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}$$

En el caso $p = 1$, tendremos una única correlación canónica ρ_1, de manera que, si $\Sigma_{yz} = (\sigma_{yZ_1}, \ldots, \sigma_{yZ_q})$,

$$\rho_1^2 = \frac{\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}}{\sigma_y^2}.$$

Es decir, $\rho_1 = \rho_{y \cdot Z_1 \ldots Z_q}$. Por lo tanto, los coeficientes de correlación canónica suponen una generalización del coeficiente de correlación múltiple. Así, $\rho_{y \cdot Z_1 \ldots Z_q}$ debe entenderse como la máxima correlación[3] entre la variable Y y una variable del tipo $\beta' Z$. Puede demostrarse sin dificultad (cuestión propuesta) que dicha correlación se alcanza en

$$\beta = \Sigma_{zz}^{-1} \Sigma_{zy},$$

es decir, en los coeficiente de regresión múltiple. En consecuencia, se verifica que, en el caso general, ρ_1 es el máximo coeficiente de correlación múltiple entre una combinación lineal de las $Y's$ y todas las $Z's$, que se alcanza con α_1.

Además, teniendo en cuenta (13.8), puede demostrarse (cuestión propuesta) el siguiente resultado:

Teorema 6.2.
ρ_1^2 es la máxima correlación múltiple entre una variable aleatoria real de la forma $\alpha' Y$ y el vector aleatorio Z, que se alcanza con $\alpha \in \langle \alpha_1 \rangle$. Se denota $U_1 = \alpha_1' Y$. Si $i = 2, \ldots, b$, ρ_i^2 es la máxima correlación múltiple entre una variable aleatoria real de la forma $\alpha' Y$ incorrelada con U_j, $j = 1, \ldots, i-1$, con el vector aleatorio Z, que se alcanza en $\langle \alpha_i \rangle$.

Por último, en el caso $p = q = 1$, se tiene que

$$\rho_1^2 = \frac{\sigma_{yz}^2}{\sigma_{zz}\sigma_{yy}} = \rho^2.$$

6.2. Inferencias

Supongamos que partimos de una muestra aleatoria de tamaño $\mathtt{n}$ de un modelo de distribución con dimensión $p + q$. Se trata pues de dos matrices de datos, Y, de dimensión $\mathtt{n} \times p$ y Z, de dimensión $\mathtt{n} \times q$. Podemos estimar los coeficientes de correlación canónica poblacionales, $\rho_1, \ldots, \rho_b$, mediante los coeficientes $r_1, \ldots, r_b$, tales que $r_1^2, \ldots, r_b^2$ son los b primeros autovalores de la matriz

$$S_{\mathsf{yy}}^{-1} S_{\mathsf{yz}} S_{\mathsf{zz}}^{-1} S_{\mathsf{zy}} \tag{6.3}$$

donde $S_{\mathsf{yy}}, S_{\mathsf{zz}}$ y S_{yz} son las matrices de covarianza totales muestrales. Se verifica también que los autovalores positivos de ésta matriz coinciden con los de la matriz

$$S_{\mathsf{zz}}^{-1} S_{\mathsf{zy}} S_{\mathsf{yy}}^{-1} S_{\mathsf{yz}}. \tag{6.4}$$

Bajo la hipótesis de $(p+q)$-normalidad, r_b será positivo con probabilidad 1 y el estadístico $(r_1^2, \ldots, r_b^2)$, será el EMV de $(\rho_1, \ldots, \rho_b)$.

[3] Nótese que queda pues demostrada la igualdad (6.1).

Seguidamente, para cada $i = 1, \ldots, b$ se consideran los autovectores a_i y b_i correspondientes a los autovalores r_i^2 para las matrices (6.3) y (6.4), respectivamente, sujetos a la condición de que los vectores ya_i y zb_i, de $\mathbb{R}^{\mathtt{n}}$, tengan varianza 1 [4]. Por un razonamiento completamente análogo al que se efectúa en la sección (B) del Apéndice, obtenemos una una interpretación de los coeficientes s $r_1, \ldots, r_b$ y de los vectores a_i's y b_i's equivalente, en términos muestrales, a la que se expresa en el teorema 6.1 para los parámetros probabilísticos. Es decir, r_1 es la máxima correlación entre una transformación lineal del tipo $\mathsf{Z}b$, donde $b \in \mathbb{R}^q$, con otra del tipo $\mathsf{Y}a$, donde $a \in \mathbb{R}^p$, que se alcanza con a_1 y b_1; r_2 es la maxima correlación lineal entre una transformación del tipo $\mathsf{Y}a$ y otra del tipo $\mathsf{Z}b$ sujetas a la condición de que $r_{\mathsf{Y}a,\mathsf{Y}a_1} = r_{\mathsf{Z}b,\mathsf{Z}b_1} = 0$, etc. Además, en el caso $p = 1$, se tiene que el único coeficiente de correlación canónica muestral, r_1, coincide con el coeficiente de correlación multiple muestral R. Se denotan por (u_i, v_i), $i = 1, \ldots, b$ los pares (canónicos) de vectores de datos obtenidos mediante

$$u_i = \mathsf{Y}a_i, \quad v_i = \mathsf{Z}b_i, \quad i = 1, \ldots, b.$$

En todo caso, la independencia entre los vectores de variables dependientes e independientes implica que el primer coeficiente de correlación canónica (y, por lo tanto, todos) es nulo. Además, bajo la hipótesis de (p+q)-normalidad, basta considerar la distribución condicional para probar la implicación recíproca.

Asimismo, si partimos de la hipótesis de normalidad , podemos conocer[5] la distribución asintótica de los coeficientes de correlación canónica muestrales. Concretamente, si ρ_i es de multiplicidad 1, se tiene

$$\sqrt{\mathtt{n}}(r_i^2 - \rho_i^2) \xrightarrow{d} N\left(0, 4\rho_i^2(1-\rho_i^2)^2\right).$$

Si todos los coeficientes poblacionales son distintos, entonces

$$\sqrt{\mathtt{n}} \begin{pmatrix} r_1^2 - \rho_1^2 \\ \vdots \\ r_b^2 - \rho_b^2 \end{pmatrix} \xrightarrow{d} N_b \left(0, 4 \begin{pmatrix} \rho_1^2(1-\rho_1^2)^2 & & 0 \\ & \ddots & \\ 0 & & \rho_b^2(1-\rho_b^2)^2 \end{pmatrix} \right).$$

Por otra parte, puede comprobarse (cuestión propuesta) que, si $\rho_1 = 0$, entonces el estadístico

$$-n \sum_{i=1}^{b} \ln(1 - r_i^2)$$

[4] Aquí sí influye (mínimamente) el hecho de dividir por n o por $n-1$ en las matrices de covarianzas muestrales.

[5] Cf. Bilodeau (1999).

sigue una distribución asintótica χ^2_{pq} [6]. Así mismo, se verifica para cada $k = 1, \ldots, b$, que, si $\rho_{k+1} = \ldots = \rho_b = 0$, entonces

$$-n \sum_{i=k+1}^{b} \ln(1 - r_i^2) \tag{6.5}$$

sigue una distribución asintótica $\chi^2_{(p-k)(q-k)}$. También podemos encontrar en Rencher (1995) aproximaciones a la distribución F-Snedecor. Todos estos resultados generalizan a los ya comentados sobre el coeficiente de correlación múltiple. Además, se han obtenido[7] distribuciones asintóticas de los coeficientes de correlación canónica bajo ciertas hipótesis más débiles que la de normalidad.

6.3. Relación con el test de correlación

Como se vio en la sección 3.1, los coeficientes $r_1^2, \ldots, r_b^2$, constituyen un estadístico invariante maximal en el contraste de independencia entre los vectores aleatorios Y y Z. Por lo tanto, todo test invariante para resolver el contraste de independencia puede expresarse en función de los coeficientes de correlación canónica muestrales. Por su parte, $\rho_1, \ldots, \rho_b$ constituye un estadístico invariante maximal para el espacio de parámetros. Luego, la distribución del estadístico de contraste correspondiente a un test invariante depende exclusivamente del valor de los coeficientes de correlación canónica poblacionales. De hecho, así sucede con el test de la razón de verosimilitudes (3.8). La hipótesis inicial de independencia equivale a $\rho_1 = 0$.

En particular, si $p = 1$, es decir, si sólo existe una variable dependiente (es el caso de la correlación múltiple), el test para contrastar la independencia de ésta respecto a las variables explicativas dependerá de la muestra exclusivamente a través del coeficiente de correlación múltiple al cuadrado. Concretamente, se reduce a comparar el valor del estadístico[8]

$$\frac{n-q-1}{q} \frac{R^2}{1-R^2}$$

con el cuantil $F^{\alpha}_{q,n-q-1}$. En el caso de la correlación simple ($p = q = 1$), se confrontarán el cuantil $F^{\alpha}_{1,n-2}$ y el estadístico

$$(n-2)\frac{r^2}{1-r^2}.$$

[6] De aquí se obtienen, en particular, la distribuciones asintóticas del coeficiente de correlación múltiple R y del de correlación simple, r, bajo la hipótesis nula de independencia, suponiendo normalidad.

[7] Bilodeau (1999).

[8] Cf. Anderson (1958).

6.4. Relación con regresión y manova

Como vimos en el capítulo anterior, si $t_1, \ldots, t_b$ es el invariante maximal correspondiente al contraste $\beta_1 = 0$, entonces

$$t_i = \frac{r_i^2}{1 - r_i^2},$$

o, equivalentemente,

$$r_i^2 = \frac{t_i}{1 + t_i}.$$

Por lo tanto, los test de Wilks, Lawley-Hotelling, Roy y Pillay pueden expresarse en términos de las correlaciones canónicas. Por ejemplo,

$$\lambda_1(Y)(Z) = \prod_{i=1}^{b}(1 - r_i^2).$$

Por su parte, el test de Pillay puede expresarse así:

$$\lambda_4(Y)(Z) = \sum_{i=1}^{b} r_i^2.$$

Nótese, que el valor del coeficiente r_i^2 ha de estar comprendido entre 0 y 1, que se corresponden, respectivamente, con los casos $t_i = 0$ y $t_i = +\infty$. El hecho de que los coeficientes de correlación canónica estén acotados supone una gran ventaja respecto a los autovalores $t_1, \ldots, t_b$, lo cual otorga sentido al test de Pillay.

Además, dado que un manova puede considerarse como una regresión multivariante sobre variables explicativas ficticias, el estadístico de contraste del manova puede expresarse también mediante las correlaciones canónicas entre las variables respuestas y las explicativas, lo cual será de gran importancia en el análisis discriminante (capítulo 9), donde $\langle a_1 \rangle, \ldots, \langle a_b \rangle$ serán los ejes discriminates y los coeficientes de correlación canónica $r_1^2, \ldots, r_b^2$ determinarán el poder de discriminación de los mismos.

Por la misma razón, el test F para la comparación de r medias univariante (anova) lleva asociado el coeficiente de determinación R^2 de la única variable respuesta respecto a las $r - 1$ variables ficticias asociadas.

6.5. Reducción de dimensión en correlación lineal

La principal virtud de los coeficientes de correlación canónica es que explican la estructura de de correlación lineal entre p variables dependientes y q variables predictores. En ese sentido, hemos de establecer en primer lugar la verdadera dimensión

del problema de correlación, es decir, el numero de pares de variables canónicas que presentan una correlación (canónica) realmente significativa. Posteriormente, hemos de interpretar dichas variables canónicas para determinar la relevancia de cada una de las variables observadas, tanto dependientes como predictores, en la relación lineal entre ambos grupos de variables.

Si se verifica que $\rho_{k+1} = \ldots = \rho_b = 0$, cabría pensar en una reducción a dimensión k del problema, considerando tan sólo los pares de variables canónicas asociadas a los k primeros coeficientes de correlación, es decir

$$(u_1, v_1), \ldots, (u_k, v_k).$$

Bajo la hipótesis de $(p+q)$-normalidad, se conoce la distribución asintótica del estadístico (6.5) en el caso $\rho_{k+1} = \ldots = \rho_b = 0$, lo cual nos permite proponer un test de hipótesis a nivel α para resolver esta cuestión. No obstante hemos de advertir dos aspectos: en primer lugar, insistimos en que se precisa del supuesto de normalidad y que la muestra sea de gran tamaño; por otra parte, y contrariamente al espíritu del análisis de correlación canónica (eliminar pares de variables canónicas con correlaciones *pequeñas*), se pretende probar en este caso la hipótesis inicial, lo cual nos obliga a considerar un nivel de significación mayor de lo habitual. Por todo ello, en la práctica, suele prevalecer sobre el test el método descriptivo consistente en analizar, para cada $j = 1, \ldots, b$, el cociente

$$p_j = \frac{\sum_{i=1}^{j} r_i^2}{\sum_{i=1}^{b} r_i^2}.$$

Si k es el primer valor tal que $p_k \simeq 1$, lo cual equivale a a afirmar que r_{k+1} es el primer coeficiente próximo a 0, cabrá pensar en la nulidad de ρ_{k+1} y, con mayor razón, de los coeficientes siguientes, lo cual implicaría considerar únicamente los k primeros pares de variables canónicas.

Una vez resueltos a analizar los k primeros pares de variables canónicas, disponemos de tres métodos distintos para tal fin.

(a) **Ponderaciones canónicas:** consiste en analizar, para todo $j = 1, \ldots, k$, la magnitud y el signo de las p componentes del vector a_j (correspondiente a la variable canónica u_j dependiente considerada) y de las q componentes del vector b_j (correspondiente a la variable canónica predictor). En ese sentido, una ponderación pequeña para la variable Y_i, se interpreta como una pobre aportación de dicha variable a a la correlación lineal entre los grupos de variables. Lo mismo puede decirse de las variables explicativas. No obstante, las ponderaciones deben ser interpretadas con mucha cautela pues están sujetas a gran

variabilidad, especialmente ante la presencia de multicolinealidad (ver capítulo 6). Además, es conveniente, a la hora de interpretar estos coeficientes, tipificar todas las variables observadas. Razonemos esta última afirmación:

Es evidente que una traslación de las variables no afecta al análisis de correlación canónica. Veamos qué sucede al realizar un cambio de escala: consideremos las matrices diagonales $D_1 = \text{diag}(d_1, \ldots, d_p)$ y $D_2 = \text{diag}(d'_1, \ldots, d'_q)$, y realicemos los cambios de escala $y^* = yD_1$ y $z^* = zD_2$ (es decir, multiplicamos los datos correspondiente a cada variable dependiente Y_i por d_i y los correspondientes a Z_i por d'_i). Se verifica entonces

$$S^{-1}_{y^*y^*} S_{y^*z^*} S^{-1}_{z^*z^*} S_{z^*y^*} = D_1 S^{-1}_{yy} S_{yz} S^{-1}_{zz} S_{zy} D_1,$$

cuyos b primeros autovalores son (cuestión propuesta) $r_i^2, \ldots, r_b^2$, asociados a los autovectores $D_1^{-1}a_1, \ldots, D_1^{-1}a_b$. Igualmente sucede con los b's, obteniéndose los autovectores $D_2^{-1}b_1, \ldots, D_2^{-1}b_b$. Por tanto, el cambio de escala no afecta a las correlaciones canónicas pero sí a las variables canónicas, lo cual ha de tenerse en cuenta a la hora de analizar las ponderaciones. Así, por ejemplo, si las variables dependientes expresan medidas en metros, el hecho de pasar a centímetros la primera variable no afecta a los coeficientes de correlación pero sí supone dividir por 100 la ponderación (y por lo tanto la importancia) de dicha variable. Por ello, es conveniente que las variables del estudio sean *conmensurables*, y una manera artificial de garantizar este requisito es tipificarlas. Podemos tipificar todas o sólo las variables respuesta, según convenga. En el caso de la regresión respecto a variables ficticias sólo se tipificarán las variables respuesta. Tras la tipificación, la componente j-ésima de los autovectores i-ésimos quedan de la forma siguiente

$$a^*_{ij} = s_{y_j} a_{ij}, \qquad b^*_{ij} = s_{z_j} b_{ij}.$$

La tipificación tiene como ventaja el conseguir un mismo grado de dispersión para todas las variables. No obstante, supone en todo caso una alteración de los datos originales y de la distribución de los mismos, lo cual ha de tenerse en cuenta si se desea aplicar métodos de inferencia estadística. La disyuntiva de tipificar o no tipificar se nos presentará en numerosas ocasiones y, por desgracia, no estamos en condiciones de proponer un criterio general para resolver este dilema.

(b) **Cargas canónicas:** consiste en analizar las correlaciones entre cada variable respuesta observada Y_i, $i = 1, \ldots, p$ y cada variable canónica U_j, $j = 1, \ldots, p$, así como las correlaciones entre las variables explicativas observadas, Z_i, $i =$

$1, \ldots, q$, y las variables canónicas V_j, $j = 1, \ldots, b$. Teniendo en cuenta que contamos con muestra aleatorias simples de tamaño $\mathtt{n}$ (vectores de $\mathbb{R}^{\mathtt{n}}$) de dichas variables, que se denotan por $y_{(i)}$, $z_{(i)}$, u_j y v_j, respectivamente, estimaremos las correlaciones anteriores mediante los coeficientes de correlación muestral, calculados a partir de dichos vectores. Estas correlaciones pueden obtenerse (cuestión propuesta) mediante

$$r_{y_{(i)},u_j} = (0, \ldots, \overset{i}{1}, \ldots, 0) R_{yy} a_j \tag{6.6}$$

$$r_{z_{(i)},v_j} = (0, \ldots, \overset{i}{1}, \ldots, 0) R_{zz} b_j \tag{6.7}$$

Dado que el coeficiente de correlación simple es invariante ante cambios de escala, el hecho de tipificar las variables no afecta en absoluto el análisis de las cargas, de ahí que sea el método más extendido a la hora de interpretar las contribuciones de las variables dependientes y predictores a la correlación lineal conjunta.

(c) **Cargas cruzadas canónicas y análisis de redundancia:** las cargas canónicas cruzadas son las correlaciones entre cada variable repuesta observada y cada la variable canónica explicativa, así como las correlaciones entre las variables explicativas observadas y las variables canónicas respuesta. El análisis de las cargas cruzadas ha sido sugerido como una alternativa al de las cargas canónicas convencional.

Por otra parte, los coeficientes de correlación canónica al cuadrado pueden entenderse como medidas de la proporción de variabilidad de cada variable canónica respuesta U_j explicada por la correspondiente variable canónica explicativa V_j, $j = 1, \ldots, b$, o, dicho de otra forma, son medidas del grado de redundancia existente en cada par (U_i, V_j). Cuando estos son altos los pares de variables canónicas se seleccionan para su posterior estudio. No obstante, conviene también que se dé un alto grado de redundancia entre las variables respuesta observadas Y_i, $i = 1, \ldots, p$ y las variables canónicas explicativas V_j, $j = 1, \ldots, b.$, es decir, que éstas últimas sean realmente últiles a la hora de explicar las primeras. Igualmente, conviene un alto grado de redundancia entre las variables observadas explicativas y las variables canónicas respuesta. Ello puede determinarse mediante el análisis de cargas canónicas cruzadas al cuadrado. No obstante, contamos con otro método denominado análisis de redundancia. Consiste en definir los índices de redundancia j-ésimos poblacionales , $j = 1, \ldots, b$

mediante

$$Rd(Y|V_j) = \frac{\sum_{i=1}^{p} \rho^2_{Y_i,U_j}}{p}\ \rho_j^2, \qquad Rd(Z|U_j) = \frac{\sum_{i=1}^{q} \rho^2_{Z_i,V_j}}{q}\ \rho_j^2.$$

Se estiman, respectivamente, mediante

$$\hat{Rd}(Y|V_j) = \frac{\sum_{i=1}^{p} r^2_{\mathbf{Y}_i,u_j}}{p}\ r_j^2, \qquad \hat{Rd}(Z|U_j) = \frac{\sum_{i=1}^{q} r^2_{\mathbf{Z}_i,v_j}}{q}\ r_j^2.$$

El primer coeficiente del producto, denominado proporción[9] de varianza explicada, es una media aritmética de las correlaciones al cuadrado entre cada variable respuesta (respectivamente explicativa) observada y la variable canónica respuesta (respectivamente explicativa) j-ésima, y pretende expresar la proporción de varianza del vector de variables observadas respuesta (resp. explicativas) que es explicada por la variable canónica respueta (resp. explicativa) j-ésima. El segundo coeficiente de la expresión es el cuadrado la j-ésima correlación canónica, es decir, del coeficiente de correlación entre la j-ésima variable canónica respuesta y la j-ésima variable canónica explicativa. En definitiva, el índice de redundancia pretende expresar la proporción de varianza del vector observado respuesta (resp. explicativo) que es explicada por la variable canónica explicativa (resp. respuesta) j-ésima. Por cierto, existe un test para contrastar la nulidad de los distintos índice de redundancia aunque su uso es poco frecuente.

Cuestiones propuestas

1. Demostrar que en el modelo de correlación lineal la hipótesis $\beta_1 = 0$ equivale a $\rho_1 = 0$.

2. Demostrar que, si $\rho_1 = 0$, entonces el estadístico

$$-n \sum_{i=1}^{b} \ln(1 - r_i^2)$$

sigue una distribución asintótica χ^2_{pq}. Indicación: considerar el test de la razón de verosimilitudes (Wilks) para contrastar $\beta_1 = 0$.

3. Demostrar que los autovalores positivos de la matriz $\Sigma_{yy}^{-1}\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}$ coinciden con los autovalores positivos de $\Sigma_{zz}^{-1}\Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz}$.

[9] En la cuestión 8 se explica en qué sentido puede hablarse de proporción.

4. Demostrar el teorema 6.2.

5. Obtener (6.6) y (6.7).

6. Demostrar que los autovalores de $D_1 S_{yy}^{-1} S_{yz} S_{zz}^{-1} S_{zy} D_1$, son $r_1^2, \ldots, r_b^2$, y sus autovectores $D_1^{-1} a_1, \ldots, D_1^{-1} a_b$.

7. Consideremos los vectores aleatorios $Y = (Y_1, \ldots, Y_p)'$ y $Z = (Z_1, \ldots, Z_q)'$, donde las q últimas son incorreladas. Probar que

$$\rho^2_{Y_i \cdot Z} = \rho^2_{Y_i, Z_1} + \ldots + \rho^2_{Y_i, Z_q}.$$

   Probar también que, si existe una matriz A $p \times q$ tal que

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} = A \begin{pmatrix} Z_1 \\ \vdots \\ Z_q \end{pmatrix},$$

   se verifica que $\rho^2_{Y_i, Z} = 1$, para todo $i = 1, \ldots, p$.

8. Demostrar que, si $p \leq q$, se verifica

$$\sum_{i,j=1}^{p} \rho^2_{Y_i, U_j} = p.$$

   Luego, en ese caso, la suma de las proporciones de la varianza para las distintas variables canónicas dependientes es 1.

9. Suponiendo normalidad multivariante, construir un intervalo de confianza asintótico a nivel $1-\alpha$ para el primer coeficiente de correlación canónica ρ_1. Así mismo, construir una región de confianza asintótica para el vector $(\rho_1, \ldots, \rho_b)'$.

10. A partir de resultados obtenidos en este capítulo, demostrar que, en una regresión múltiple, el coeficiente de correlación múltiple es la máxima correlación simple entre la variable respuesta y una combinación lineal de las explicativas, y que esta combinación resulta de multiplicar escalarmente las mismas por sus respectivos coeficientes de regresión.

11. Razonar en qué medida puede afectar un cambio de escala de cualquier variable en el cálculo de los coeficientes de correlación canónica.

12. Se realiza un estudio de regresión multiple de una variable respuesta Y respecto de q variables explicativas $Z_1, \ldots, Z_q$. Razonar si puede disminuir el coeficiente de correlación múltiple R^2 al introducir una nueva variable predictor Z_{q+1}. ¿Qué tendría que suceder exactamente para que R^2 permaneciera en ese caso invariante?

Capítulo 7

Análisis de componentes principales

En este capítulo continuamos la línea iniciada en el análisis de correlación canónica consistente en el estudio de valores teóricos (sumas ponderadas de las variables originales). Nos proponemos ahora transformar los datos de manera que la matriz de varianzas y covarianzas quede expresada de una forma canónica (diagonal), todo ello con el objeto de lograr una reducción en la dimensión de las observaciones cuando se dé una fuerte correlación entre las variables. La reducción de la dimensión tiene interés en sí misma. No obstante, en capítulos posteriores consideraremos otras aplicaciones, más concretas, del análisis de componentes lineales, como pueden ser el estudio de multicolinealidad en regresión, el análisis de correspondencias o el análisis factorial.

La primera sección se encuadra en un marco probabilístico. En el mismo se definen las componentes principales y se pretende razonar geométricamente por qué las componentes con menor varianza son las de menor interés probabilístico. En la siguiente sección se realiza un estudio análogo desde un punto de vista muestral, es decir, partiendo de los datos, y se establece también un resultado que será clave para el análisis de correspondencias. Se proponen distintos contrastes de hipótesis de especial interés que, no obstante, podrían haber sido estudiados sin ningún problema en el capítulo 3. En la última sección, se analiza la relación entre las componentes principales y las variables originales.

7.1. Punto de vista probabilístico

Consideremos un vector aleatorio $X : (\Omega, \mathcal{A}, P) \longrightarrow \mathbb{R}^p$, cuyas componentes son $X_1, \ldots, X_p$, con media μ y matriz de covarianzas Σ. El procedimiento a seguir consistirá en proyectar X sobre una subvariedad afín de la menor dimensión posible, k, siempre y cuando el error consiguiente (en términos de la distancia euclídea) se mantenga dentro de unos márgenes que consideremos aceptables. Esta proyección quedará entonces determinada por un nuevo vector k-dimensional. Posteriormente veremos que esta técnica equivale a eliminar las proyecciones con menor variabilidad.

Nuestra teoría se basa en el teorema (13.4) del Apéndice, que denominaremos teorema de diagonalización. En lo que sigue y dada una matriz de covarianzas poblacional $\Sigma \in \mathcal{M}_{p\times p}$, $\delta_1, \ldots, \delta_p$ denotarán sus autovalores ordenados, mientras que $\gamma_1, \ldots, \gamma_p$ denotará una base de autovectores asociados a los mismos. Por tanto, si se denota $\Gamma = (\gamma_1 \ldots \gamma_p)$ y $\Delta = \mathsf{diag}(\delta_1, \ldots \delta_p)$, se verifica

$$\Sigma = \Gamma\Delta\Gamma'$$

Análogamente, dada una matriz de covarianzas muestral S, $d_1, \ldots, d_p$ y $g_1, \ldots, g_p$ denotarán, respectivamente, sus autovalores ordenados y una base de autovectores asociados, de tal forma que, si se denota $D = \mathsf{diag}(d_1, \ldots, d_p)$ y $G = (g_1 \ldots g_p)$, se verifica

$$D = GDG'$$

Dado $k \leq p$, $\mathcal{P}_k^{\perp}$ denotará el conjunto de todas las proyecciones ortogonales sobre cualquier subespacio k-dimensional de $\mathcal{R}^p$. Asimismo, se denotará por Γ_k y G_k las matrices $p \times k$ compuestas por las k primeras columnas de Γ y G, respectivamente. El siguiente lema es consecuencia del teorema de diagonalización.

Lema 7.1.
Sea $A \in \mathcal{M}_{p\times p}$ simétrica, siendo sus autovalores ordenados λ_i, $i = 1, \ldots, p$, y h_i, $i = 1, \ldots, p$, sus respectivos autovectores. Entonces

$$\begin{aligned} \max_{P\in\mathcal{P}_k^{\perp}} \mathtt{tr}\ AP &= \sum_{i=1}^{k} \lambda_i, \\ \min_{P\in\mathcal{P}_k^{\perp}} \mathtt{tr}\ A(\mathtt{Id} - P) &= \sum_{i=k+1}^{p} \lambda_i, \end{aligned}$$

alcanzándose ambos extremos en $P = \sum_{i=1}^{k} h_i h_i'$.

Demostración.
Sea $P \in \mathcal{P}_k^{\perp}$ y $A = (a_1 \dots a_k)$ una matriz $p \times k$ cuyas columnas constituyen una base ortonormal de $\mathsf{Im}P$, de tal forma que $P = AA'$. En consecuencia, si H y Λ denotan respectivamente las matrices $(h_1 \dots h_p)$ y $\mathsf{diag}(\lambda_1 \dots \lambda_p)$, se verifica

$$\mathtt{tr}AP = \mathtt{tr}H\Lambda H'AA' = \mathtt{tr}\Lambda(H'A)(H'A)'.$$

Nótese que las columnas $f_1, \dots, f_k$ de la matriz $F = H'A$ constituyen un sistema ortonormal. Por lo tanto, aplicando el teorema de diagonalización a la matriz Λ se deduce que

$$\begin{aligned} \mathtt{tr}\ AP &= \mathtt{tr}\Lambda \sum_{i=1}^{k} f_i f_i' = \sum_{i=1}^{k} f_i' \Lambda f_i \\ &\leq \max_{\|f\|=1} f'\Lambda f + \sum_{i=2}^{k} \max_{\|f=1\|, f \in \langle f_1, \dots, f_{i-1}\rangle^{\perp}} f'\Lambda f = \sum_{i=1}^{k} \lambda_i, \end{aligned}$$

alcanzándose los máximos anteriores en los vectores $(1, 0 \dots 0)', \dots, (0, \dots \overset{k}{1} \dots 0)'$, respectivamente. Despejando se deduce que el máximo se obtiene cuando A es la matriz $(h_1 \dots h_k)$, en cuyo caso $P = \sum_{i=1}^{k} h_i h_i'$. Dado que

$$\min_{P \in \mathcal{P}_k^{\perp}} \mathtt{tr}\ A(\mathtt{Id} - P) = \mathtt{tr}A - \max_{P \in \mathcal{P}_k^{\perp}} \mathtt{tr}\ AP,$$

se sigue que el mínimo se alcanza para el mismo valor de P, siendo su valor $\sum_{i=k+1}^{p} \lambda_i$. ∎

La siguiente proposición resulta interesante a la hora de interpretar las componentes principales desde un punto de vista probabilístico.

Proposición 7.2.

Dados un vector aleatorio p-dimensional X no degenerado y $k \leq p$, se verifica

$$\min \left\{ \mathtt{E}[\|X - (CBX + a)\|^2] : \ a \in \mathbb{R}^p, \ B, C' \in \mathcal{M}_{k \times p}, \ \mathtt{rg}(B) = k \right\} = \sum_{i=k+1}^{p} \delta_i,$$

alcanzándose cuando $CB = \Gamma_k \Gamma_k'$ y $a = \mathtt{Id} - CB\mu$.

Demostración.
Empezaremos suponiendo $\mu = 0$ y consideremos fija una matriz B en las condiciones del enunciado. En ese caso, se tiene que el vector $(X', X'B')'$ posee media 0 y matriz de varianzas-covarinzas

$$\begin{pmatrix} \Sigma & \Sigma B' \\ B\Sigma & B\Sigma B' \end{pmatrix}.$$

Luego, se sigue de (1.27) que

$$\begin{aligned}\min_{C\in\mathcal{M}_{p\times k},a\in\mathbb{R}^p}\mathtt{E}[\|X-(CBX+a)\|^2] &= \mathtt{tr}\,[\Sigma-\Sigma B'(B\Sigma B')^{-1}B\Sigma]\\ &= \mathtt{tr}\,\Sigma[\mathtt{Id}-\Sigma^{1/2}B'(B\Sigma B')^{-1}B\Sigma^{1/2}]\\ &= \mathtt{tr}\,\Sigma(\mathtt{Id}-P_B),\end{aligned}$$

donde $P_B = \Sigma^{1/2}B'(B\Sigma B')^{-1}B\Sigma^{1/2}$, alcanzándose dicho extremo en $a = 0$ y en la matriz $C = (\Sigma B)^{-1}(B\Sigma B')^{-1}$. Puede comprobarse que P_B es una matriz $p \times p$ idempotente de rango k. De hecho, se sigue del corolario 13.5-(vi) que $\mathcal{P}_k^{\perp} = \{P_B : B \in \mathcal{M}_{k\times p}, \mathtt{rg}(B) = k\}$. Por lo tanto, se deduce del lema anterior que $\mathtt{tr}\,\Sigma(\mathtt{Id}-P_B)$ toma un valor mínimo $\sum_{i=k+1}^{p}\delta_i$ en $P_B = \Gamma_k\Gamma_k'$, de lo cual se sigue que el mínimo buscado se alcanza en cuando $a = 0$ y B y C verifican

$$\Sigma^{1/2}B'(B\Sigma B')^{-1}B\Sigma^{1/2} = \Gamma_k\Gamma_k', \quad C = (\Sigma B)^{-1}(B\Sigma B')^{-1}.$$

Por consiguiente, teniendo en cuenta que $\Sigma^{1/2} = \Gamma\Delta^{1/2}\Gamma'$, se concluye que $CB = \Gamma_k\Gamma_k'$. En general, para cualquier $\mu \in \mathbb{R}^p$, se obtiene la tesis razonando con el vector aleatorio $Y = X - \mu$.

∎

Como hemos indicado en la introducción del capítulo, nuestro propósito es reducir al máximo la dimensión del vector aleatorio p-dimensional X, siempre y cuando ello suponga un error *aceptable*. Para ello, reemplazaremos en un primer paso el vector aleatorio original X por otro contenido en una subvariedad afín H_k de dimensión k, X^k. Proponemos como medida del error cometido en esta reducción el siguiente

$$\mathtt{E}[\|X - X^k\|^2]. \tag{7.1}$$

X_*^k denotará el vector aleatorio contenido en una subvariedad afín k-dimensional que minimiza dicho error (probaremos a continuación su existencia). Desde luego, dada una subvariedad afín H_k, el vector X^k contenido en la misma que minimiza dicho error ha de ser la proyección ortogonal sobre ésta, $P_{H_k}X$. La cuestión a dilucidar es pues, dado k, sobre qué subvariedad hemos de proyectar para minimizar el error. El siguiente resultado nos ofrece la respuesta.

Teorema 7.3.

Dados un vector aleatorio p-dimensional X, de media μ y matriz de varianzas-covarianzas Σ, y $k \leq p$, se verifica que $\min_{H_k}\mathtt{E}[\|X - P_{H_k}X\|^2] = \sum_{i=k+1}^{p}\delta_i$, alcanzándose en $H_k^* = \mu + V_k^*$, donde V_k^* es el subespacio generado por los k primeros autovectores de Σ.

Demostración.
Dada un subvariedad afín k-dimensional $H_k \subset \mathbb{R}^p$, existen un vector $x \in \mathbb{R}^p$ y un subespacio lineal V_k, tales que $H_k = y + V_k$. En ese caso, si F_k denota una matriz $p \times k$ cuyas columnas constituyen una base ortonormal de V_k, se tiene que

$$P_{H_k}X = y + P_{V_k}(X - y) = F_kF_k'X + y - F_kF_k'y.$$

Aplicando la proposición anterior, se deduce que el mínimo buscado es, efectivamente, $\sum_{i=k+1}^{p} \delta_i$, alcanzándose cuando $F_kF_k' = P_{V_k^*}$ y $\mu - y = P_{V_k^*}(\mu - y)$. La última igualdad equivale a afirmar que $\mu - y \in V_k^*$, luego, por (13.3), acabamos.

∎

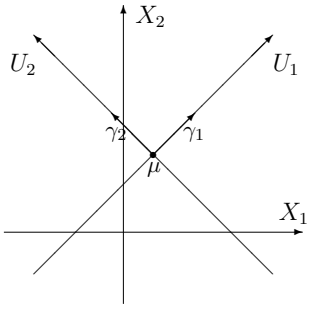
Una vez ha quedado patente la trascendencia de los autovectores y autovalores de Σ en el problema de reducción de la dimensión, estamos en condiciones de definir las componentes principales. El vector aleatorio p-dimensional

$$U = \Gamma'(X - \mu) \tag{7.2}$$

se denomina vector de componentes principales, mientras que sus distintas componentes U_i, $i = 1, \ldots, p$, que se obtienen mediante $U_i = \gamma_i'(X - \mu_i), \quad i = 1, \ldots, p$, serán, respectivamente, las i-ésimas componentes principales[1]. Se verifica entonces

$$\mathtt{Cov}[U] = \Gamma'\Sigma\Gamma = \Delta, \quad \mathtt{E}[U] = 0.$$

Luego, las componentes principales U_i's son incorreladas entre sí y con varianzas δ_i's, respectivamente. El vector X se recuperaría, por tanto, mediante $X = \mu + \Gamma U$. Hemos de tener en cuenta que la transformación (7.2) no es sino una traslación (la media pasa a ser el origen de coordenadas) seguida de una rotación (los ejes perpendiculares determinados por los autovectores de Σ pasan a ser los ejes de coordenadas). La gráfica ilustra estas consideraciones:



[1]Realmente, es más usual, dado que el problema sólo atañe a la matriz de covarianzas, suponer centrado el vector aleatorio, en cuyo caso se definiría $U = \Gamma'X$. De la misma forma, se tendría $U_i = \gamma_i'X$.

En el caso de que se verifique la hipótesis de p-normalidad del vector X, deberíamos dibujar una elipse [2] centrada en μ cuyos ejes vienen determinados por los autovectores de Σ, es decir, son los nuevos ejes rotados, dependiendo la magnitud de los mismos de los autovalores correspondientes, es decir, que el eje principal queda determinado por el primer autovector y así sucesivamente. La hipótesis de normalidad no ha sido necesaria para obtener la tesis del teorema anterior. No obstante, está indirectamente presente en este estudio pues, recordemos, el objetivo planteado, al menos inicialmente, es la reducción de la dimensión cuando se observan fuertes correlaciones lineales entre las componentes del vector aleatorio, situación ésta estrechamente vinculada con el supuesto de normalidad del mismo. Por otra parte, se sigue del teorema anterior que

$$X_*^k = \mu + \Gamma_k \begin{pmatrix} U_1 \\ \vdots \\ U_k \end{pmatrix}, \qquad \mathtt{E}[\|X - X_*^k\|^2] = \sum_{i=k+1}^{p} \delta_i. \tag{7.3}$$

Por lo tanto, X_*^k se ha obtenido desechando las $p-k$ últimas componentes principales, es decir, las $p-k$ últimas componentes de (7.2), de ahí que, para reducir a k la dimensión de X, resulte razonable considerar en un segundo paso el vector constituido por las k primeras componentes principales. Además, (7.3) ofrece el error cometido en esta reducción, de manera que podemos, antes que nada, seleccionar el valor de k adecuado (es decir, aquél que, dando lugar a un error *aceptable*, sea lo más pequeño posible) para después construir las componentes principales de interés.

Por otra parte, podemos analizar el problema desde otra perspectiva y llegar a una conclusión análoga a la del teorema 7.3. Recordemos que, si α es un vector de $\mathbb{R}^p$ de norma euclídea 1, $\alpha' X$ es una variable aleatoria real que se corresponde con la longitud de la proyección del vector X sobre el eje que determina el vector α. Se verifica, trivialmente, que

$$\mathtt{var}[\alpha' X] = \alpha' \Sigma \alpha, \quad \mathtt{cov}[\alpha' X, \beta' X] = \alpha' \Sigma \beta, \quad \forall \alpha, \beta \in \mathbb{R}^p.$$

En ese sentido, las componentes principales pueden interpretarse como las proyecciones de $X - \mu$ sobre los ejes perpendiculares que determinan los autovectores de Σ. El siguiente resultado precisa más en línea de esta interpretación.

Teorema 7.4.

Dado un vector aleatorio p-dimensional X de media μ y matriz de covarianzas Σ se

[2] Pues, si $X \sim N_p(\mu, \Sigma)$, el lugar geométrico de los puntos de $\mathbb{R}^p$ con verosimilitud constante es un elipsoide con centro en μ y cuya forma está determinada por Σ^{-1}.

verifica

$$\delta_1 = \sup\{\mathtt{var}[\alpha'(X-\mu)]\colon \|\alpha\| = 1\},$$

alcanzándose dicho supremo en $\alpha = \gamma_1$, es decir, con la variable real U_1. Además,

$$\delta_i = \sup\{\mathtt{var}[\alpha'(X-\mu)]\colon \|\alpha\| = 1,\ \mathtt{cov}[\alpha' X, U_j] = 0,\ \forall j < i\}, \quad \forall i = 2, \ldots, p,$$

alcanzándose dicho supremos con $\alpha = \gamma_i$, respectivamente[3].

Demostración.
La demostración es consecuencia inmediata del teorema de diagonalización. No obstante y respecto a la última afirmación, téngase en cuenta que, dados $j = 1, \ldots, p-1$, y $\alpha \in \mathbb{R}^p$,

$$\mathtt{cov}[\alpha'(X-\mu), U_j] = \alpha' \Sigma \gamma_j = \delta_j \alpha' \gamma_j.$$

Luego,

$$\alpha \perp \gamma_j \Rightarrow \mathtt{cov}[\alpha'(X-\mu), U_j] = 0.$$

Si $\delta_j > 0$ el recíproco es también cierto. Si $\delta_j = 0$, con mayor razón lo son los sucesivos, luego la tesis se obtiene más fácilmente.

∎

Por lo tanto, el eje $\langle \gamma_1 \rangle$ proporciona la máxima varianza para las proyecciones, δ_1. Si $i \in \{2, \ldots, p\}$, el eje $\langle \gamma_i \rangle$ proporciona la máxima varianza entre todos los ejes perpendiculares a $\langle \gamma_1, \ldots, \gamma_{i-1} \rangle$. Podemos considerar como caso extremo aquél en el cual $\delta_i = 0$, $i = k+1, \ldots, p$. Ello implicaría que $\mathtt{var}[U_i] = 0$, es decir, $\gamma_i'(X-\mu) \equiv 0$, para todo $i = k+1, \ldots, p$. Lo cual equivale a afirmar que $X = X_*^k$, por lo que X puede construirse enteramente sin tener en cuenta las $p-k$ últimas componentes principales. Es decir, que la imagen de X está contenida en una subvariedad afín k-dimensional, concretamente

$$\mathtt{Im}(X) \subset H_k^*.$$

En este caso, se verifica para cada $\varepsilon > 0$ que $P(\|X - X_*^k\|^2 > \varepsilon) = 0$, lo cual, insistimos, es una situación extrema. En general, dada una constante positiva ε, se sigue de la desigualdad de Chebichev que

$$P\left(\gamma_i'(X-\mu)^2 > \frac{\varepsilon}{p}\right) < \frac{p\delta_i}{\varepsilon}, \quad i = 1, \ldots, p \tag{7.4}$$

[3]Nótese que en el enunciado puede sustituirse $\alpha'(X-\mu)$ por $\alpha' X$, pues dicho cálculo no afecta al cálculo de covarianzas.

Luego, de la desigualdad de Bonferroni[4] se deduce que

$$P(\|X - X_*^k\|^2 > \varepsilon) \leq \frac{p}{\varepsilon} \sum_{i=k+1}^{p} \delta_i. \tag{7.5}$$

Por lo tanto, si los autovalores δ_i, $i = k+1, \ldots, p$ son lo suficientemente pequeños, podemos considerar X_k^* como una aceptable aproximación a X, no sólo desde el punto de vista del error medio expresado en (7.3), sino también en el sentido análogo expresado en (7.5). De esta forma, ambas ecuaciones invitan a considerar $\sum_{i=1}^{k} \delta_i$ como una medida de la *pérdida* que conlleva la reducción óptima a dimensión k, que es la que se consigue considerando únicamente las k primeras componentes principales. Luego, en definitiva, la dimensión k será aquélla lo más pequeña posible siempre y cuando suponga un valor suficientemente próximo a 1 para

$$\frac{\sum_{i=1}^{k} \delta_i}{\sum_{i=1}^{p} \delta_i}.$$

Recordemos a su vez la varianza total de X, definida en (1.26), que puede calcularse mediante $\mathtt{var}_T[X] = \sum_{i=1}^{p} \delta_i$, y cuya interpretación se encuentra en el problema (19) del capítulo 1. Por lo tanto el cociente anterior equivale a

$$\frac{\sum_{i=1}^{k} \delta_i}{\mathtt{var}_T[X]}$$

y se denominada proporción de varianza total explicada[5]. Nótese que, en el extremos opuesto al caso analizado anteriormente, si todos los autovalores son idénticos, entonces Σ es de la forma $\sigma^2 \mathtt{Id}$. Este caso es el menos apropiado a la hora de reducir dimensiones. Desde un punto de vista geométrico y bajo la hipótesis de p-normalidad, la relación entre los distintos autovalores determina la excentricidad del elipsoide. Así, si todos los autovalores son idénticos tendremos una distribución normal esférica, es decir, las regiones de $\mathbb{R}^p$ con verosimilitud constante son esferas. Por contra y según hemos visto, si los últimos $p-k$ autovalores son *muy pequeños*, el elipsoide estará *casi* contenido en la subvariedad afín k-dimensional que determinan la media y los k primeros autovectores.

7.2. Punto de vista muestral

Obviamente, el método de obtención de las componentes principales expuesto en la sección anterior no tiene viabilidad práctica, debido al desconocimiento de la matriz

[4] $P(\cap A_i) \geq 1 - \sum_i P(A_i)$

[5] Se entiende que explicada por las k primeras componentes principales.

de covarianzas poblacional Σ. En la práctica hemos de partir pues de una estimación de Σ mediante una muestra aleatoria de P^X.

Supongamos que $Y_1, \ldots, Y_{\mathtt{n}}$ es una muestra aleatoria simple de P^X o, más general, $\mathtt{n}$ vectores de $\mathbb{R}^p$. Como es costumbre, los expresamos matricialmente mediante $Y = (Y_1, \ldots, Y_{\mathtt{n}})'$ de dimensiones $\mathtt{n} \times p$. En ese caso, consideraremos la matriz de varianzas-covarianzas totales muestral [6]

$$S = \frac{1}{\mathtt{n}} \left[Y'Y - \frac{1}{\mathtt{n}} Y' \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{pmatrix} Y \right],$$

la cual será diagonalizada utilizando la matriz diagonal de sus autovalores D y la matriz ortogonal de sus autovectores G, mediante $S = GDG'$. Si la distribución P^X es p-normal, entonces S y $\frac{n-1}{n}S$ son los EIMV y EMV de Σ, respectivamente. Bajo esas condiciones, podemos encontrar en Anderson (1958), cap. 13, las distribución del estadístico $(d_1, \ldots, d_p)$. Si, además, $\delta_1 > \ldots > \delta_p$, $\frac{n-1}{n}D$ y G serán los EMV de Δ y Γ, respectivante[7], verificándose también[8] la siguiente propiedad asintótica:

$$\sqrt{\mathtt{n}} \begin{pmatrix} d_1 - \delta_1 \\ \vdots \\ d_p - \delta_p \end{pmatrix} \xrightarrow{d} N_p \left[0, \begin{pmatrix} 2\delta_1^2 & & 0 \\ & \ddots & \\ 0 & & 2\delta_p^2 \end{pmatrix} \right].$$

Por tanto, se verifica[9]

$$\sqrt{\frac{\mathtt{n}}{2}} \begin{pmatrix} \log \frac{d_1}{\delta_1} \\ \vdots \\ \log \frac{d_p}{\delta_p} \end{pmatrix} \xrightarrow{d} N_p(0, \mathtt{Id}). \tag{7.6}$$

Este resultado permite construir intervalos de confianza para los autovalores si partimos de la hipótesis de p-normalidad y la muestra es de gran tamaño (cuestión propuesta). El hecho de considerar S o $\frac{\mathtt{n}-1}{\mathtt{n}}S$ no afecta asintóticamente a la matriz de autovalores D, pues $\lim_{n\to\infty} \frac{\mathtt{n}-1}{\mathtt{n}} = 1$.

[6] Cosideramos aquí la división entre $\mathtt{n}$. Ello no afectará al los autovectores aunque sí a los autovalores. No obstante, éstos serán porporcionales a los que se obtendrían dividiendo por $\mathtt{n}-1$, por lo que este hecho no afecta al análisis de componentes principales. La razón del mismo es puramente estética y quedará patente, creemos, un poco más adelante.

[7] Nótese que, sólo si la multiplicidad de todos los autovalores de Σ es 1, podremos determinar los autovectores correspondientes.

[8] Flury (1997).

[9] Aplicando el conocido método Delta (ver el apéndice del primer volumen).

A la hora de calcular las componentes principales suele suponerse que la media aritmética de las observaciones es 0 [10], lo cual se traduce en que las componentes principales *verdaderas* son una traslación de la que de hecho se construyen, que se calculan de la siguiente forma[11]: se considera

$$u_i = G'Y_i, \quad i = 1, \ldots, n.$$

Sea entonces u la matriz $n \times p$ cuyas filas son, respectivamente, $u_1', \ldots, u_n'$, es decir,

$$u = YG.$$

Si $u[j]$, $j = 1, \ldots, p$ denotan las columnas de u, entonces $u[j] = Yg_j$, es decir, se trata del vector de $\mathbb{R}^{\mathtt{n}}$ constituido por las j-ésimas componentes principales de cada uno de los $\mathtt{n}$ datos. Por lo tanto, si se denota $u = (u_{ij}) \in \mathcal{M}_{n\times p}$, entonces u_{ij} es la j-ésima componente principal del i-ésimo dato, es decir, $u_{ij} = g_j'Y_i$. Análogamente al caso poblacional[12] se verifica que

$$s^2_{u[j]} = d_j, \qquad s_{u[j],u[k]} = 0, \quad \text{si } j \neq k.$$

Además,

$$d_1 = \sup\{s^2_{Y\alpha} \colon \|\alpha\| = 1\},$$

alcanzándose dicho supremo con $\alpha = g_1$, es decir, con el vector $u[1]$. También se verifica

$$d_j = \sup\{s^2_{Y\alpha} \colon \|\alpha\| = 1,\ s_{Y\alpha,u^k} = 0,\ \forall k < j\}, \quad \text{para todo } j = 2, \ldots, p,$$

alcanzádose dicho supremo con $u[j]$. Como en el caso poblacional, se define la varianza total muestral mediante $\sum_{j=1}^p d_j = \mathtt{tr}(S)$. Dado que estamos considerando los autovalores y autovectores muestrales como estimaciones puntuales de los poblaciones, el objetivo a perseguir será, nuevamente, encontrar la dimensión k lo más pequeña posible para la cual sea suficientemente próximo a 1 el cociente

$$\frac{\sum_{j=1}^k d_j}{\sum_{j=1}^p d_j},$$

que, en términos de la varianza total muestral $s_T^2 = \mathtt{tr}[S]$ equivale a

$$\frac{\sum_{j=1}^k d_j}{s_T^2},$$

[10] Más aún, es bastante habitual e suponer, como veremos más adelante, que los datos están tipificados.

[11] Si no se supusiera nula la media aritmética de los datos, se tendría $u_i = G'(Y_i - \overline{y})$.

[12] Tener en cuenta que $s^2_{Y\alpha} = \alpha' S_Y \alpha$ y $s_{Y\alpha,Y\beta} = \alpha' S_Y \beta$, para todo $\alpha, \beta \in \mathbb{R}^p$.

por lo que se denomina proporción de la varianza total (muestral) explicada. Una vez seleccionado, nos quedaremos únicamente con las k primeras componentes principales.

Como vemos, éste es un método puramente descriptivo. Sería interesante poder hacer inferencias respecto al parámetro poblacional

$$\Psi_k = \frac{\sum_{i=1}^{k} \delta_i}{\mathtt{tr}\ \Sigma}, \quad k = 1, \ldots, p-1.$$

En Mardia et al. (1979) se demuestra que, bajo la hipótesis de p-normalidad, el estadístico

$$\hat{\Psi}_k = \frac{\sum_{i=1}^{k} d_i}{\mathtt{tr}\ S},$$

sigue una distribución asintótica $N(\Psi_k, \tau^2)$, donde

$$\begin{aligned} \Psi &= \frac{\sum_{i=1}^{k} \delta_i}{\mathtt{tr}\ \Sigma}, \\ \tau^2 &= \frac{2\mathtt{tr}\ \Sigma}{(\mathtt{n}-1)(\mathtt{tr}\ \Sigma)^2}(\Psi_k^2 - 2\alpha\Psi_k + \alpha), \\ \alpha &= \frac{\sum_{i=1}^{k} \delta_i^2}{\mathtt{tr}\ \Sigma}. \end{aligned}$$

El principio de sustitución propone, en este caso, reemplazar Σ, Ψ_k y δ_i, $i = 1, \ldots, p$, por S, $\hat{\Psi}_k$ y d_i, $i = 1, \ldots, p$, de manera que obtendríamos una estimación puntual $\hat{\tau}^2$ de τ^2. En ese caso, se verifica[13]

$$\frac{\hat{\Psi} - \Psi_k}{\hat{\tau}} \sim N(0, 1),$$

lo cual puede servir tanto para construir un intervalo de confianza para Ψ_k como para resolver un contraste del tipo

$$H_0 : \frac{\sum_{i=1}^{k} \delta_i}{\mathtt{tr}\ \Sigma} = \Psi_0, \qquad k = 1, \ldots, p-1, \quad , \psi \in]0, 1[. \tag{7.7}$$

También bajo el supuesto de p-normalidad podemos contrastar la siguiente hipótesis inicial

$$H_0 : \delta_1 \geq \ldots \geq \delta_k \geq \delta_{k+1} = \ldots = \delta_p, \qquad k = 1, \ldots, p-1. \tag{7.8}$$

Téngase en cuenta que, si la distribución es no degenerada, no tiene sentido contrastar una hipótesis del tipo $\delta_{k+1} = \ldots = \delta_p = 0$. El test para resolver el contraste consiste en rechazar la hipótesis inicial cuando

$$\mathtt{n}(p-k) \log \frac{\frac{1}{p-k}\sum_{i=k+1}^{p} d_i}{\left(\prod_{i=k+1}^{p} d_i\right)^{\frac{1}{p-k}}} > \chi^{2,\alpha}_{(p-k)(p-k+1)/2-1}.$$

[13] Aproximadamente, se entiende.

Se entiende que este test tiene validez asintótica y, recordamos, precisa del supuesto de normalidad multivariante. En Bilodeau (1999) podemos encontrar también inferencias respecto a los autovalores de la matriz de correlaciones.

Las componentes principales, vistas desde un punto de vista muestral, admiten una clara interpretación geométrica en la línea del teorema 7.3, que veremos a continuación. Dado $k \leq p$, se trata de sustituir los datos originales Y_i, $i = 1, \ldots, n$, por otros vectores Y_i^k, $i = 1, \ldots, p$, contenidos en una subvariedad afín de dimensión k, H_k, de tal manera que el error que implica la sustitución sea *razonable*, en el sentido de que, si Y^k denota la matriz cuyas filas son Y_i^k, $k = 1, \ldots, n$, sea pequeña la distancia

$$\mathtt{d}(Y, Y^k) = \frac{1}{n} \sum_{i=1}^{n} \|Y_i - Y_i^k\|^2.$$

Resolveremos el problema desde un punto de vista más general, lo cual facilitará enormemente las cosas a la hora de afrontar la sección dedicada al análisis de correspondecias[14]. Ello nos obliga a introducir algunas definiciones. Concretamente, definiremos distintas métricas en el espacio $\mathcal{M}_{\mathtt{n}\times p}$. Primeramente, la distancia anterior, que puede definirse mejor de esta forma:

$$\mathtt{d}(X, Y) = \frac{1}{\mathtt{n}} \mathtt{tr}[(X - Y)(X - Y)'], \qquad X, Y \in \mathcal{M}_{I\times J}. \tag{7.9}$$

En este caso, si d denota la distancia en $\mathbb{R}^J$ y X_i' y Y_i', $i = 1, \ldots I$, denotan las filas de X e Y, respectivamente, se verifica que $\mathtt{d}(X, Y) = \frac{1}{\mathtt{n}} \sum_{i=1}^{I} d(X_i, Y_i)$. Sea $\Omega = \mathsf{diag}(\omega_1, \ldots, \omega_n)$, donde $\omega_i \geq 0$ para todo $i = 1, \ldots, \mathtt{n}$ y $\sum_i \omega_i = 1$. Sea a su vez Φ una matriz definida positiva de orden p. En ese caso, para cada matriz X se define el centroide

$$\overline{x}^{\Omega} = \sum_{i=1}^{\mathtt{n}} \omega_i X_i \tag{7.10}$$

Se trata pues de la media aritmética ponderada por los pesos de Ω. Si $1_{\mathtt{n}}$ denota el vector constante 1 en $\mathbb{R}^{\mathtt{n}}$, se denota

$$\overline{X}^{\Omega} = 1_{\mathtt{n}} \big(\overline{x}^{\Omega}\big)', \tag{7.11}$$

que es la matriz de $\mathcal{M}_{\mathtt{n}\times p}$ cuyas filas son todas iguales a la traspuesta de la media aritmética ponderada. Definimos también la matriz de covarianzas mediante

$$\begin{aligned} S_{\Omega,\Phi} &= \Phi^{1/2} \left(Y - \overline{Y}^{\Omega}\right)' \Omega \left(Y - \overline{Y}^{\Omega}\right)' \Phi^{1/2} \\ &= \sum_{i=1}^{\mathtt{n}} \omega_i \left(\Phi^{1/2} Y_i - \Phi^{1/2} \overline{y}^{\Omega}\right) \left(\Phi^{1/2} Y_i - \Phi^{1/2} \overline{y}^{\Omega}\right)'. \end{aligned}$$

[14] Dicho capítulo puede considerarse en buena parte continuación del presente.

Es importante notar que el rango de la matriz $Y - \overline{Y}^{\Omega}$, y por lo tanto el de $S_{\Omega,\Phi}$, es igual al rango de Y menos 1. Definimos entonces las siguientes métricas sobre $\mathbb{R}^{\mathtt{n}}$ y $\mathbb{R}^p$:

$$d_\Omega(a,b) = (a-b)'\Omega(a-b), \qquad a,b \in \mathbb{R}^{\mathtt{n}}. \tag{7.12}$$

$$d_\Phi(u,v) = (u-v)'\Phi(u-v), \qquad u,v \in \mathbb{R}^p. \tag{7.13}$$

la aplicación lineal $x \in \mathbb{R}^p \mapsto \Phi^{1/2}x$ supone una isometría de $(\mathbb{R}^p, d)$ en $(\mathbb{R}^p, d_\Phi)$. En la sección primera del Apéndice podemos encontrar diversos comentarios respecto al concepto de proyección ortogonal utilizando la métrica euclídea d, que pueden generalizarse a la métrica d_Φ teniendo en cuenta la isometría anterior. De esta forma, si $a \in \mathbb{R}^p$ y V es un subespacio k-dimensional de $\mathbb{R}^p$, P^{Φ}_{a+V} denotará la proyección ortogonal sobre la subvariedad afín $a+V$ considerando d_Φ, que se obtiene mediante

$$P^{\Phi}_{a+V}(x) = a + \Phi^{-1/2} P_{\Phi^{1/2}(V)}\left(\Phi^{1/2}(x-a)\right), \quad x \in \mathbb{R}^p. \tag{7.14}$$

Definimos la siguiente distancia en $\mathcal{M}_{I\times J}$:

$$\mathtt{d}_{\Omega,\Phi}(X,Y) = \mathtt{tr}[\Omega(X-Y)\Phi(X-Y)'], \qquad X,Y \in \mathcal{M}_{I\times J}. \tag{7.15}$$

Se verifica entonces

$$\mathtt{d}_{\Omega,\Phi}(X,Y) = \mathtt{d}_{\Phi,\Omega}(X',Y'), \tag{7.16}$$

$$\mathtt{d}_{\Omega,\Phi}(X,Y) = \sum_{i=1}^{I} \omega_i d_\Phi(X_i,Y_i). \tag{7.17}$$

Además,

$$\frac{1}{\mathtt{n}}\mathtt{d}_{\Omega,\Phi}(X,Y) = \mathtt{d}\big(\Omega^{1/2}X\Phi^{1/2}, \Omega^{1/2}Y\Phi^{1/2}\big), \tag{7.18}$$

$$\mathtt{d}(X,Y) = \mathtt{d}_{\frac{1}{\mathtt{n}}\mathtt{Id}_{\mathtt{n}},\mathtt{Id}_p}(X,Y). \tag{7.19}$$

Por último, se tiene que

$$\overline{y} = \overline{y}^{\frac{1}{\mathtt{n}}\mathtt{Id}_{\mathtt{n}}}, \quad S = S_{\frac{1}{\mathtt{n}}\mathtt{Id}_{\mathtt{n}},\mathtt{Id}_p}.$$

Teorema 7.5.

Sean $\{d_1,\ldots,d_p\}$ los autovalores ordenados de $S_{\Omega,\Phi}$ y $\{h_1,\ldots,h_p\}$ sus respectivos autovectores. Dado $k < p$, la mínima distancia $\mathtt{d}_{\Omega,\Phi}(Y,Y^k)$ cuando las filas de Y^k se encuentran en una subvariedad afín k-dimensional vale $\sum_{i=k+1}^{p} d_i$, y se alcanza cuando, para cada $i = 1,\ldots,\mathtt{n}$, $Y_i^k = P^{\Phi}_{H_k^*}(Y_i)$, siendo $H_k^* = \overline{y}^{\Omega} + \left\langle \Phi^{-1/2}h_1,\ldots,\Phi^{-1/2}h_k \right\rangle$.

Demostración.
Dada una subvariedad afín k-dimensional $H_k = a + V_k$, e Y^k una matriz $\mathtt{n} \times p$ tal que $Y_i^k \in H_k$, para todo $i = 1, \ldots, \mathtt{n}$, se verifica

$$\begin{aligned}
\mathtt{d}_{\Omega,\Phi}(Y, Y^k) &= \sum_{i=1}^{\mathtt{n}} \omega_i d_\Phi(Y_i, Y_i^k) \\
&\geq \sum_{i=1}^{\mathtt{n}} \omega_i d_\Phi\left(Y_i, P_{H_k}^\Phi Y_i\right) \\
&= \sum_{i=1}^{\mathtt{n}} \omega_i \left\|\Phi^{1/2}\left(Y_i - P_{H_k}^\Phi Y_i\right)\right\|^2 \\
&= \sum_{i=1}^{\mathtt{n}} \omega_i \left\|\Phi^{1/2}(Y_i - a) - P_{\Phi^{1/2}V_k}\Phi^{1/2}(Y_i - a)\right\|^2 \\
&= \sum_{i=1}^{\mathtt{n}} \omega_i \left\|P_{\left(\Phi^{1/2}V_k\right)^\perp}\Phi^{1/2}(Y_i - a)\right\|^2 \\
&= \sum_{i=1}^{\mathtt{n}} \omega_i \left\|P_{\left(\Phi^{1/2}V_k\right)^\perp}\Phi^{1/2}\left(Y_i - \overline{y}^\Omega\right) + P_{\left(\Phi^{1/2}V_k\right)^\perp}\Phi^{1/2}\left(\overline{y}^\Omega - a\right)\right\|^2
\end{aligned}$$

Teniendo en cuenta que $\sum_{i=1}^{\mathtt{n}} \omega_i \left(Y_i - \overline{y}^\Omega\right) = 0$, el último término puede descomponerse así

$$\sum_{i=1}^{\mathtt{n}} \omega_i \left\|P_{\left(\Phi^{1/2}V_k\right)^\perp}\Phi^{1/2}\left(Y_i - \overline{y}^\Omega\right)\right\|^2 + \left\|P_{\left(\Phi^{1/2}V_k\right)^\perp}\Phi^{1/2}\left(\overline{y}^\Omega - a\right)\right\|^2.$$

El segundo sumando es, en todo caso, mayor o igual que 0, siendo 0 sii $\overline{y}^\Omega \in H_k$, es decir, que la subvariedad afín que minimice $\mathtt{d}_{\Omega,\Phi}$ debe contener a la media aritmética ponderada por Ω. Supuesto que eso sucede, veamos en qué subespacio V_k se alcanza el mínimo. El primer sumando es igual a

$$\sum_{i=1}^{\mathtt{n}} \left[\omega_i^{1/2}\left(Y_i - \overline{y}^\Omega\right)\right]' \Phi^{1/2} P_{\left(\Phi^{1/2}V_k\right)^\perp}\Phi^{1/2}\left[\omega_i^{1/2}\left(Y_i - \overline{y}^\Omega\right)\right],$$

o, lo que es lo mismo,

$$\mathtt{tr}\left[\Omega^{1/2}\left(Y - \overline{Y}^\Omega\right)\Phi^{1/2} P_{\left(\Phi^{1/2}V_k\right)^\perp}\Phi^{1/2}\left(Y - \overline{Y}^\Omega\right)'\Omega^{1/2}\right],$$

que equivale a

$$\mathtt{tr}\left[S_{\Omega,\Phi}\left(\mathtt{Id} - P_{\Phi^{1/2}V_k}\right)\right].$$

Se deduce entonces del lema 7.1 que el valor mínimo buscado es $\sum_{i=k+1}^{p} d_i$, alcanzándose cuando $\Phi^{1/2}V_k = \langle h_1, \ldots, h_k \rangle$.

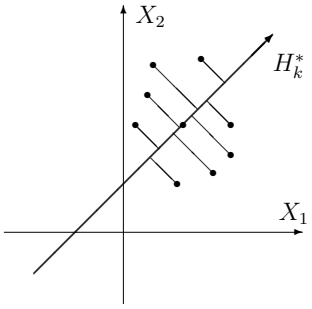
∎

El siguiente resultado, que es el que realmente interesa en este capítulo, se deduce directamente del teorema anterior.

Corolario 7.6.

Sean $\{d_1, \ldots, d_p\}$ los autovalores ordenados de S y $\{h_1, \ldots, h_p\}$ sus respectivos autovectores. Dado $0 \leq k < p$, la mínima distancia $\mathrm{d}(Y, Y^k)$ cuando las filas de Y^k se encuentran en una subvariedad afín k-dimensional vale $\mathtt{n} \sum_{i=1}^{k} d_i$, y se alcanza cuando, para cada $i = 1, \ldots, \mathtt{n}$, $Y_i^k = P_{H_k^*}(Y_i)$, siendo $H_k^* = \overline{y} + \langle h_1, \ldots, h_k \rangle$.

Este resultado da pues una justificación geométrica de las componentes principales muestrales. El gráfico siguiente ilustra la idea anteriormente descrita:



Así pues, a la luz del teorema 7.3 y del corolario 7.6, entendemos que, a la hora de reducir la dimensión, las proyecciones de menor varianza resultan ser las de menor *interés estadístico.*

El siguiente resultado, que suele denominarse teorema de la descomposición en valores singulares (singular value descomposition, abreviadamente SVD), se obtiene como corolario del teorema de diagonalización de una matriz simétrica.

Teorema 7.7.

Dadas $A \in \mathcal{M}_{n\times p}$ de rango r y Ω y Φ como antes, existen una matriz $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$ con elementos positivos y ordenados de mayor a menor, y otras dos matrices $N \in \mathcal{M}_{n\times n}$ y $M \in \mathcal{M}_{p\times p}$, verificando

(i) $A = N \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array} \right) M'$

(ii) $M'\Phi M = \mathtt{Id}_p$

(iii) $N'\Omega N = \mathtt{Id}_n$

Demostración.
Sea $\Delta = \mathsf{diag}(d_1, \ldots, d_r, 0)$ la matriz diagonal de orden p de los autovalores ordenados de $A'A$ y H una matriz $p \times p$ cuyas columnas $h_1, \ldots, h_p$ constituyen una base ortonormal de autovectores respectivos. El teorema de diagonalización permite afirmar afirma que

$$A'A = H\Delta H'.$$

Consideremos Δ_r y H_r las submatrices de Δ y H constituidas respectivamente por los r primeros autovalores y sus correspondientes autovectores. Definamos

$$G_r = AH_r\Delta_r^{-1/2}.$$

Se verifica entonces que $G_r'G_r = \mathtt{Id}_r$. Por lo tanto, sus columnas pueden completarse hasta obtener una matriz ortogonal de orden n que se denota por G. En ese caso, si se denota $D = \Delta_r^{1/2}$, se tiene que

$$G'AH = \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right),$$

de lo cual se sigue que

$$A = G\left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right)H'.$$

Por lo tanto, la tesis queda demostrada en le caso $\Omega = \mathtt{Id}_n$ y $\Phi = \mathtt{Id}_p$. En el caso general bastaría considerar la descomposición

$$\Omega^{1/2}A\Phi^{1/2} = G\left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right)H'$$

en las condiciones anteriores. En ese caso, si definimos $N = \Omega^{-1/2}G$ y $M = \Phi^{-1/2}H$, se verifican (i), (ii) y (iii).

∎

La expresión (i) en las condiciones (ii) y (iii) se denomina descomposición SVD de A para la métrica $\mathtt{d}_{\Omega,\Phi}$. Nótese que las columnas de M, $m_1, \ldots, m_p$, y las de N, $n_1, \ldots, n_n$, constituyen sendas bases ortonormales respecto a las métricas $\mathtt{d}_\Omega$ y d_Φ, respectivamente. Si se denota $F = ND$ (siendo f_{ij} sus componentes), y $A_1', \ldots, A_n'$ son las filas de A, se tiene entonces que

$$A_i = \sum_{j=1}^{K} f_{ij}m_j, \qquad i = 1, \ldots, n. \tag{7.20}$$

Por lo tanto, la matriz F está constituida por las coordenadas de las filas de A respecto a la base ortonormal M. Igualmente, dado que $A' = MDN'$, la matriz $F^* = MD$ está constituida por las coordenadas de las columnas de A respecto a la base ortonormal N. Respecto a la unicidad de las descomposición SVD, podemos afirmar trivialmente que la única matriz diagonal D en las condiciones del teorema es la compuesta por las raíces de los autovalores de $\Phi^{1/2}A'\Omega A\Phi^{1/2}$. Si la multiplicidad de todos ellos es 1, las columnas (autovectores) de M quedan unívocamente determinadas salvo reflexiones. En general, se verifica que los subespacios generados por las columnas asociadas a un determinado autovalor son únicos[15]. Lo mismo puede decirse de N, razonando en este caso con $\Omega^{1/2}A\Phi A'\Omega^{1/2}$. En ocasiones la descomposición SVD se presenta de la forma

$$A = \overline{N}D\overline{M}', \text{ donde } \overline{N} \in \mathcal{M}_{n\times r},\ \overline{M} \in \mathcal{M}_{p\times r}, \quad \overline{N}'\Omega\overline{N} = \overline{M}'\Phi\overline{M} = \mathtt{Id}_r. \tag{7.21}$$

Se trata simplemente de eliminar los autovectores asociados a autovalores nulos que, a la postre, no tendrán utilidad práctica alguna. Todo lo dicho anteriormente acerca de la unicidad de la descomposición sigue siendo igualmente válido aquí. Estamos ya en condiciones de enunciar el resultado clave en el análisis de correspondencias.

Teorema 7.8.

Dada una matriz de datos $Y \in \mathcal{M}_{\mathtt{n}\times p}$ de rango $r+1$, existen una matriz $D = \mathsf{diag}(\lambda_1, \ldots, \lambda_r)$ con elementos positivos y ordenados de mayor a menor, y otras dos matrices $N \in \mathcal{M}_{\mathtt{n}\times\mathtt{n}}$ y $M \in \mathcal{M}_{p\times p}$, verificando

(i) $Y = \overline{Y}^{\Omega} + N\left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right)M'$

(ii) $M'\Phi M = \mathtt{Id}_p$

(iii) $N'\Omega N = \mathtt{Id}_n p$

(iv) $\mathsf{d}_{\Omega,\Phi}\left(Y, \overline{Y}^{\Omega}\right) = \sum_{i=1}^{r}\lambda_i$

(v) Si $0 \leq k < r$, D_k denota la submatriz diagonal de D formada por los k primeros valores de ésta, y N_k y M_k denotan las submatrices constituidas por las k primeras columnas de N y M, respectivamente, se verifica que la matriz Y^{k*} de $\mathcal{M}_{\mathtt{n}\times p}$ que minimiza $\mathsf{d}_{\Omega,\Phi}(Y, Y^k)$ entre todas aquellas cuyas filas pertenecen a una subvariedad afín k-dimensional $\mathbb{R}^p$, es

$$Y^{k*} = \overline{Y}^{\Omega} + N_k D_k M_k'.$$

[15] Que, en definitiva, será lo que interese.

Se verifica además

$$\mathsf{d}_{\Omega,\Phi}\left(Y, Y^{k*}\right) = \sum_{i=k+1}^{r} \lambda_i, \qquad \mathsf{d}_{\Omega,\Phi}\left(Y^{k*}, \overline{Y}^{\Omega}\right) = \sum_{i=1}^{k} \lambda_i.$$

Demostración.
Se trata de seguir el mismo razonamiento de la demostración anterior, pero aplicado a la matriz de rango r

$$\Omega^{1/2}\left(Y - \overline{Y}^{\Omega}\right)\Phi^{1/2},$$

es decir, considerando la matriz diagonal Δ de los autovalores de $S_{\Omega,\Phi}$, la matriz H de sus respectivos autovectores y, por último, la matriz

$$G_r = \Omega^{1/2}\left(Y - \overline{Y}^{\Omega}\right)\Phi^{1/2} H_r \Delta_r^{-1/2},$$

que se completa hasta conseguir una base ortonormal G. Haciendo $D = \Delta^{1/2}$, se verifica

$$G'\Omega^{1/2}\left(Y - \overline{Y}^{\Omega}\right)\Phi^{1/2} H = \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right),$$

de lo cual se sigue que

$$\Omega^{1/2}\left(Y - \overline{Y}^{\Omega}\right)\Phi^{1/2} = G\left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right) H'.$$

Con lo cual, basta considerar $M = \Phi^{-1/2}H$ y $N = \Omega^{-1/2}G$ para obtener (i), (ii) y (iii). Para obtener (iv) considerar la definición de $\mathsf{d}_{\Omega,\Phi}$ teniendo en cuenta (i), (ii) y (iii). Respecto a (v), basta sustituir para obtener

$$Y^{k*} = \overline{Y}^{\Omega} + \left(Y - \overline{Y}^{\Omega}\right)\Phi^{1/2} H_k H_k' \Phi^{-1/2}.$$

Es decir, para cada $i = 1, \ldots, \mathtt{n}$, se verifica

$$\begin{aligned} Y_i^{k*} &= \overline{y}^{\Omega} + \Phi^{-1/2} H_k H_k' \Phi^{1/2}\left(Y_i - \overline{y}^{\Omega}\right) \\ &= P^{\Phi}_{\overline{y}^{\Omega} + \left\langle \Phi^{-1/2} h_1, \ldots, \Phi^{-1/2} h_k \right\rangle}(Y_i). \end{aligned}$$

La tesis se sigue entonces del teorema 7.5. La distancia entre Y^{k*} e $\overline{Y}^{\Omega}$ se obtiene como en (iv).

∎

Al igual que sucediera antes, las columnas de M constituyen una base ortonormal de $(\mathbb{R}^p, d_\Phi)$. Además, si $F = ND$, se verifica

$$Y_i = \overline{y}^\Omega + \sum_{j=1}^{K} f_{ij} m_j, \qquad i = 1, \ldots, \mathtt{n}. \tag{7.22}$$

Por lo tanto, la matriz F está constituida por las coordenadas de las filas de $Y - \overline{Y}^\Omega$ respecto a la base ortonormal M. La submatriz compuesta por las k primeras columnas de F determinan pues la proyección de las filas de $Y - \overline{Y}^\Omega$ en el subespacio generado por el sistema reducido $\{m_1, \ldots, m_k\}$, lo cual permite, como sabemos, minimizar la distancia $\mathsf{d}_{\Omega,\Phi}$. Respecto a la unicidad de la descomposición, podemos afirmar que la única matriz diagonal D en las condiciones del teorema es la compuesta por las raíces de los autovalores de $S_{\Omega,\Phi}$, y que los subespacios generados por las columnas de M asociadas a un determinado autovalor son únicos. En el problema de reducción a dimensión k es esto es lo que en definitiva importa, dado que el objetivo que se persigue es proyectar sobre dichos subespacios. No obstante, encontraríamos un serio problema si se diera el caso $d_k = d_{k+1}$. Las columnas de M se denominan ejes principales de las filas de Y.

7.3. Relación con las variables originales

Aunque el análisis de componentes principales pueda ser correcto desde un punto de vista matemático, la interpretación estadística del mismo puede verse seriamente falseada si las variables en juego no son *conmensurables*. Ilustremos esta afirmación con un ejemplo.

Supongamos que se miden dos variables, una correspondiente a la medida de una longitud en metros y otra correspondiente a una anchura en metros. Si pasamos a considerar la longitud en centímetros, la varianza de dicha variable se multiplicará por 10.000, con lo cual, en términos relativos, la variable anchura será prácticamente constante (y, por tanto, sin interés estadístico) en relación con la longitud (desde un punto de vista geométrico, los ejes de la elipse coincidirán, prácticamente, con los ejes de coordenadas, y su excentricidad será enorme).

Por ello, hemos de analizar previamente la naturaleza de las variables antes de buscar una reducción de dimensiones mediante componentes principales. El problema es complejo y, evidentemente, no siempre se resuelve expresando los resultados en las mismas unidades pues las distintas variables pueden corresponder a diferentes magnitudes. Proponemos dos soluciones a este problema:

(a) Considerar una transformación que conduzca a la homocedasticidad. Puede darse el caso de que dicha transformación genere también normalidad; de esta forma, estaríamos en condiciones de aplicar los métodos inferenciales de la sección anterior.

(b) Dado que, en la mayoría de los casos, los análisis se resuelven de manera meramente descriptiva, el método más común consiste en tipificar las variables (se obtendría una matriz Z de variables tipificadas), o equivalentemente, trabajar con la matriz de correlaciones R en vez de la de covarianzas. Se tiene entonces

$$R = GDG', \quad u = ZG.$$

Esta es la opción que suele tomarse más a menudo. El programa SPSS la asume por defecto.

Cuando se ha reduce la dimensión del vector por el método de componentes principales, las variables perdidas no son, en principio, variables originales, sino combinaciones lineales de éstas. Al igual que sucede en el análisis de correlación canónica, existen distintas formas de evaluar la influencia de las variables originales en las componentes que permanecen o en las que se eliminan. La más inmediata puede ser analizando la matriz de ponderaciones G, cuyas columnas determinan las direcciones de las distintas componentes principales. No obstante, es más extendido el estudio de la matriz de cargas principales o de componentes . Se trata de una matriz $\Lambda \in \mathcal{M}_{p\times p}$, cuya columna j-ésima, $j = 1, \ldots, p$, está compuesta por los coeficientes de correlación entre cada variable original y $u[j]$. Si se ha procedido a una tipificación previa de las variables observadas, la relación entre la matriz de ponderaciones y la de cargas principales es la siguiente:

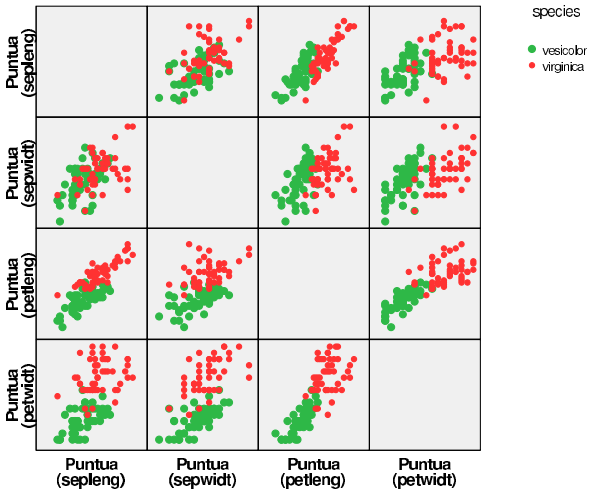
$$\begin{aligned} \Lambda &= Z'uD^{-1/2} && (7.23)\\ &= Z'ZGD^{-1/2} && (7.24)\\ &= RGD^{-1/2} && (7.25)\\ &= GDG'GD^{-1/2} && (7.26)\\ &= GD^{1/2}. && (7.27) \end{aligned}$$

Es decir, si λ_j denota la j-ésima columna de Λ,

$$g_j = \frac{1}{\sqrt{d_j}}\lambda_j, \ j = 1, \ldots, p. \tag{7.28}$$

Para acabar este capítulo veamos un ejemplo. Consideremos las cuatro variables tipificadas (nos referimos a la longitud y anchura de pétalos y sepalos) para las

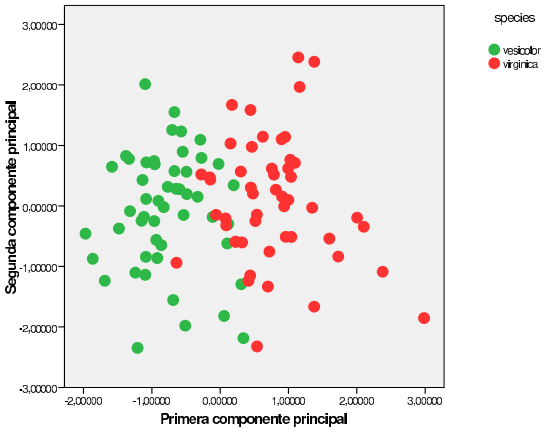
especies vesicolor y virginica del conocido archivo `irisdata` de Fisher. Se ha excluido la especie setosa pues su matriz de varianzas-covarianzas es netamente diferente. Vamos a representarlas mediante un diagrama de dispersión matricial.



Cada observación constituye pues un punto de $\mathbb{R}^4$ que intentamos representar parcialmente confrontado por pares sus diferentes componentes. Supongamos que queremos identificar cada observación con un punto de $\mathbb{R}^2$ desechando la menor variabilidad posible. Para ello calculamos las dos primeras componentes principales que, en este caso, presentan varianzas 2.958 y 0.555, respectivamente, con lo que vienen a explicar el 87.82 % de la varianza total. La matriz de componentes o cargas principales es la siguiente.

	U_1	U_2
long-sep	0.872	-0.164
anch-sep	0.748	0.662
long-pet	0.935	-0.283
anch-pet	0.874	-0.100

Podemos apreciar que todas las variables originales participan en la construcción de las componentes principales. Veamos la representación de dichas componentes mediante un diagrama de dispersión simple:



Cuestiones propuestas

1. Construir, a partir del resultado (7.6), un intervalo de confianza asintótico al 95 % para el autovalor δ_i, partiendo de la hipótesis de p-normalidad. Construir una región de confianza asintótica al 95 % para el vector $(\delta_1, \ldots, \delta_p)' \in \mathbb{R}^p$.

2. Demostrar que, si X sigue una distribución p-dimensional con matriz de covarianzas Σ, siendo δ_p el último autovalor de la misma, existe una subvariedad afín H de dimensión $p-1$ tal que, para todo $\varepsilon > 0$, se verifica que

$$P\left(\mathtt{d}(X, H) < \varepsilon\right) \geq 1 - \frac{\delta_p}{\varepsilon^2},$$

donde $\mathtt{d}$ denota la distancia euclídea. ¿Qué corolario podemos deducir si Σ es singular?

3. El análisis de componentes principales está enfocado a la reducción de la dimensión en el caso de que exista una clara relación lineal entre las variables observadas. ¿Puedes sugerir alguna forma de actuación en el caso de que la relación exista pero no sea de tipo lineal?

4. Si $Y \sim N_p(\mu, \Sigma)$ y $\mathtt{rg}(\Sigma) = p - 1$, demostrar que existe una única dirección en $\mathbb{R}^p$ tal que la proyección sobre la misma del vector aleatorio Y es una variable real constante. ¿Qué crees que debemos hacer exactamente ante una muestra aleatoria simple de una distribución normal multivariante degenerada?

5. Demostrar que, si $n \geq p$, el p-ésimo autovalor d_p es estrictamente positivo con probabilidad 1.

6. Demostrar que que el último autovalor de Σ es la mínima varianza de una proyección de X sobre cualquier dirección.

Capítulo 8

Aplicaciones de componentes principales

En términos generales, podemos decir que el análisis de componentes principales sirve para reducir dimensiones en un problema multivariante. No obstante, consideraremos en este capítulo dos aplicaciones concretas de esta técnica: la solución al problema de multicolinealidad en regresión lineal y el análisis de correspondencias. También consideraremos la aplicación del análisis de componentes principales a la hora de resolver un análisis factorial, aunque eso será en otro capítulo.

8.1. Multicolinealidad

Hemos visto en el capítulo anterior que la utilidad del análisis de componentes principales estriba en que nos permite reducir la dimensión de las observaciones cuando se observa un fuerte grado de correlación lineal entre las variables consideradas. Aunque este hecho tiene por qué ser necesariamente perjudicial (es más, puede resultar en algunos casos hasta positivo, según Hair et al. (1999)), sí que puede llegar a tener un impacto negativo cuando se da entre las variables explicativas en el modelo de regresión lineal múltiple o multivariante. A este problema hace comúnmente referencia el término *multicolinealidad*. El análisis de componentes principales puede aportarnos en ocasiones la solución al mismo. La sección queda pues dividida en tres partes: la primera se dedica a estudiar el impacto que puede tener el fenómeno de multicolinealidad; la segunda, a establecermétodos de diagnóstico de la misma; en la última parte proponemos una posible solución mediante el uso de las componentes principales. Nuestro estudio se restringirá al problema de regresión múltiple, de tal forma que ha sido ya estudiada con cierta extensión en el capítulo 4 del primer volu-

men. De todas formas, volveremos a desarrollar el tema, en unas ocasiones de manera más sucinta y en otras, más extensa.

Supongamos que estamos en las condiciones del modelo de regresión lineal múltiple, es decir, $Y = \mathtt{X}\beta + \mathcal{E}$, donde $\mathcal{E} \sim N_{\mathtt{n}}(0, \sigma^2 \mathtt{Id})$, y

$$\mathtt{X} = \begin{pmatrix} 1 & z_1[1] & \dots & z_1[q] \\ \vdots & \vdots & & \vdots \\ 1 & z_{\mathtt{n}}[1] & \dots & z_{\mathtt{n}}[q] \end{pmatrix},$$

de rango $q+1$. En ese caso, $(\mathtt{X}'\mathtt{X})^{-1} \in \mathcal{M}_{(q+1)(q+1)}$ tendrá determinante no nulo y será invertible. Si embargo, si se da un alto grado de multicolinealidad, podemos deducir, razonando por continuidad, que $|\mathtt{X}'\mathtt{X}|$ tomará un valor próximo a 0 y, entonces, la matriz $(\mathtt{X}'\mathtt{X})^{-1}$ puede tener valores muy altos. Ello afecta al estimador de β, dado que

$$\hat{\beta} \sim N_{q+1}\big(\beta, \sigma^2(\mathtt{X}'\mathtt{X})^{-1}\big).$$

Entonces, para todo $j = 0, 1, \dots, q$,

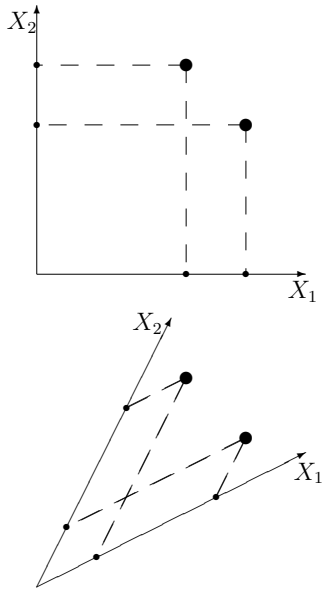
$$\mathtt{var}\big[\hat{\beta}_j\big] = \big[(\mathtt{X}'\mathtt{X})^{-1}\big]_{jj}\sigma^2,$$

donde $\big[(\mathtt{X}'\mathtt{X})^{-1}\big]_{jj}$ denota el j-ésimo elemento de la diagonal de $(\mathtt{X}'\mathtt{X})^{-1}$. Por ello, la varianza del estimador tomará un valor alto cuando dicho elemento sea elevado. De esta forma, la varianza del estimador depende, en primer lugar, del error implícito al modelo de regresión, expresado mediante la varianza residual σ^2, y en segundo lugar, de la estructura de la matriz $\mathtt{X}$ de variables explicativas a través de de $(\mathtt{X}'\mathtt{X})^{-1}$. Para entender mejor por qué la multicolinealidad incrementa esa varianza, basta demostrar (cuestión propuesta) que

$$\mathtt{var}\big[\hat{\beta}_j\big] = \sigma^2 \cdot \frac{1}{\mathtt{n}} \cdot \frac{1}{s^2_{\mathtt{z}[j]}} \cdot \frac{1}{1 - R_j^2}, \quad j = 1, \dots, q, \tag{8.1}$$

donde R_j denota el coeficiente de correlación múltiple de $\mathtt{z}[j]$ respecto al resto de variables explicativa. Como podemos observar, una fuerte correlación entre las variables explicativas repercutirá en un incremento de la varianza de los estimadores. De hecho, el término $(1 - R_j^2)^{-1}$ suele denominarse factor de inflación de la varianza, FIV.

El extremo opuesto a la multicolinealidad es la incorrelación (es decir, perpendicularidad entre las variabilidades totales) de las variables o vectores explicativos. En ese caso minimizamos las varianzas de los estimadores. Presentamos una visión geométrica que puede terminar de aclarar el problema:



La interpretación de los gráficos es la siguiente: las observaciones de los datos están sometidas a un cierto grado de variabilidad que viene determinado σ^2. Las componentes del estimador de β serán las coordenadas de la proyección de la observación Y sobre el subespacio generado por las columnas de $\mathbf{X}$. Podemos observar que, cuando se da la perpendicularidad entre las columnas, pequeñas variaciones en la observación se corresponden con pequeñas variaciones en las componentes del estimador. Sin embargo, conforme nos vamos aproximando a la dependencia lineal, una pequeña variación en la observación puede suponer una gran variación en la estimación de β, es decir, que el estimador de β está sometido a una fuerte variabilidad.

Que las varianzas de los estimadores sean elevadas podría considerarse, por principio, perjudicial para el estudio de regresión, pues se traduce en una escasa fiabilidad de las estimaciones obtenidas. No obstante, esta afirmación merecería diversas matizaciones. Desde luego, puede resultar especialmente dañino si el propósito del análisis de regresión es evaluar el grado influencia de cada variable explicativa en la variable dependiente, es decir, si se trata de un estudio de "causa-efecto". Un problema de multicolinealidad puede llevarnos a sobreestimar la influencia de ciertas variables o subestimar la de otras. Más aún, en muchas ocasiones se utilizan métodos de selección de variables explicativas que eliminan aquellas cuyo coeficiente no difiere significativamente de 0. Una varianza elevada del estimador de un coeficiente entorpece el

estudio pues resta potencia al test de significación, lo cual puede suponer la eliminación de variables realmente influyentes. De todo ello se sigue la importancia de la detección y solución del problema de multicolinealidad.

El problema de multicolinealidad puede resolverse, como veremos a continuación, mediante la eliminación de componentes principales, siempre y cuando las componentes desechadas no desempeñen un papel relevante en la regresión lineal. Hemos de recalcar pues que el análisis de las componentes principales no tiene por qué solucionar necesariamente un problema de este tipo.

Insistimos: el uso de componentes principales en un problema de regresión lineal está orientado a facilitar un estudio de causa-efecto, donde se pretende determinar la trascendencia de cada variable predictor en la variable respuesta a través de la magnitud del correspondiente coeficiente de regresión, pues este expresa la razón o cociente entre el incremento de la segunda y el de la primera. Desde luego, la importancia de cada variable explicativa no debería depender de la escala utilizada ni del valor 0 de referencia para la misma. Respecto al segundo punto, notar que los coeficiente de regresión de las variables explicativas (no así el del término independiente) son invariantes ante traslaciones de las mismas (es decir, que, el sumar una constante a cualquiera de estas variables no afecta a los correspondientes coeficientes). Sin embargo y respecto al primer punto, hemos de tener en cuenta que un cambio de escala de cualquier variable explicativa repercute de manera inversa en el coeficiente correspondiente. De todo ello puede deducirse la conveniencia de utilizar una trasformación de las variables explicativas invariante ante traslaciones y cambios de escala. Por ello, es usual trabajar con las variables explicativas tipificadas. Así pues, nuestro modelo queda de la forma siguiente

$$Y = \beta_0 1_{\mathtt{n}} + \mathtt{Z}\underline{\beta} + \mathcal{E}, \qquad \mathcal{E} \sim N(0, \sigma^2 \mathtt{Id}),$$

donde $\beta_0 \in \mathbb{R}$, $\underline{\beta} \in \mathbb{R}^q$, y $\mathtt{Z}$ es una matriz $n \times q$ de rango q cuyas columnas son linealmente independientes con media muestral 0, varianza muestral 1, y perpendiculares al vector $1_{\mathtt{n}}$. Centraremos por lo tanto nuestro análisis en la matriz $\mathtt{Z}'\mathtt{Z}$, que es igua a la matriz R de correlaciones entre las variables explicativas, multiplicada escalarmente por $\mathtt{n}$. En virtud de (5.6), se verifica

$$\hat{\beta}_0 = \overline{y}, \quad \underline{\hat{\beta}} = R^{-1} S_{\mathtt{z}y}. \tag{8.2}$$

El diagnóstico de multicolinealidad consiste en la búsqueda de indicios de un fuerte impacto en la varianza de los estimadores de los β_j's derivado de una fuerte multicolinealidad entre las variables explicativas $\mathtt{z}[1], \ldots, \mathtt{z}[q]$. Una de las más extendidas es el análisis de los FIV asociados a cada variable explicativa, de modo que la presencia

de algún FIV alto (mayor que 10) suele asociarse a un problema de multicolinealidad. No obstante, este procedimiento no arroja luz sobre el número de dimensiones redundantes en e problema.

Otro método de diagnóstico consiste en considerar la diagonalización de la matriz de correlaciones :

$$R = GDG', \tag{8.3}$$

siendo $D = \text{diag}(d_1, \ldots, d_q)$, con $d_1 \geq \ldots \geq d_q$, y $G = (g_1, \ldots, g_q)$, donde $g_j = (g_{1j}, \ldots, g_{qj})'$, $j = 1, \ldots, q$. Estando las variables explicativas tipificadas, se verifica

$$\mathtt{X'X} = \mathtt{n} \begin{pmatrix} 1 & 0 \\ 0 & R \end{pmatrix}. \tag{8.4}$$

En ese caso, la varianza total del estimador $\hat{\beta}$, cuya interpretación podemos encuentrar en el problema (19) del capítulo 1, es la siguiente

$$\mathtt{var}_T \left[\hat{\beta}\right] = \frac{\sigma^2}{\mathtt{n}} \left(1 + \sum_{j=1}^{q} d_j^{-1}\right) \tag{8.5}$$

Por lo tanto, una inflación de la varianza total se asocia a la existencia de autovalores pequeños. Dado que la suma de los mismo es en todo caso q, ello equivale a una gran desproporción entre el primero y alguno o varios de ellos. Esto nos lleva a considerar los denominados índices de condicionamiento, definidos mediante

$$IC_j := \sqrt{\frac{d_1}{d_j}}, \quad j = 1, \ldots, k.$$

Obviamente,

$$1 = IC_1 \leq \ldots \leq IC_q.$$

Teniendo en cuenta que

$$\sum_{j=1}^{q} d_j = \mathtt{tr}R = q,$$

y, por lo tanto, $d_1 \geq 1$, valores altos (suele entenderse por alto un valor mayor de 30) de $IC_{k+1}, \ldots, IC_q$ se corresponden, necesariamente, con valores muy bajos de $d_{k+1}, \ldots, d_q$ (de hecho, $IC_j > 30$ implica $d_j < 1/900$), lo cual es signo de fuerte multicolinealidad y podrá repercutir negativamente en las varianzas de los estimadores.

Analicemos individualmente los estimadores. En virtud de (8.3), se verifica

$$\mathtt{var}\left[\hat{\beta}_j\right] = \frac{\sigma^2}{\mathtt{n}} \left[R^{-1}\right]_{jj} = \frac{\sigma^2}{\mathtt{n}} \sum_{k=1}^{q} \frac{g_{jk}^2}{d_k}, \quad j = 1, \ldots, q. \tag{8.6}$$

Hemos de tener en cuenta que $\sum_{k=1}^{q} g_{jk}^2 = 1$, y, por lo tanto, $g_{jk}^2 \leq 1$, para todo j, k. Así pues, la existencia un al menos un autovalor d_k no es condición suficiente para garantizar un valor alto de la varianza del estimador de β_j, pues dicho autovalor puede corresponderse con un valor también bajo de g_{jk}^2. Lo importante pues, para que exista un impacto un la varianza del estimador de β_j, es que el cociente de ambos sea grande. Una forma de medir la magnitud de dicho cociente es considerar la proporción de la varianza , definida mediante

$$PV_{jk} = \frac{g_{jk}^2/d_k}{\sum_{i=1}^{q} g_{ji}^2/d_i}, \ j, k = 1, \ldots, q.$$

De esta forma, PV_{jk} debe entenderse como la proporción de la varianza del estimador de β_j debida a la k-ésima componente principal $u[k]$ de las variables explicativas. Se conviene que una componente principal responsable de más del 80 % de la varianza de dos o más estimadores[1] es candidata a ser eliminada de la regresión. Todos los PV_{jk}'s constituyen una matriz $q \times q$, donde el índice j denota la columna y k la fila. Obviamente, la suma de la columna j-ésima, donde $j = 1, \ldots, q$, vale, en todo caso, 1. Se trata pues de detectar filas con al menos dos componentes mayores que 0.80.

En definitiva, a la hora de diagnosticar multicolinealidad con impacto en las varianzas de los estimadores, debemos analizar los IC's y la matriz PV y calibrarlos respecto a unas cotas convencionales (sobre las que no existe unanimidad). Así, si $IC_{r+1}, \ldots, IC_q > 30$ y, para cada $k \in \{r+1, \ldots, q\}$ se verifica que existen al menos dos PV_{jk}'s mayores que 0.80, entenderemos que existe un problema de multicolinealidad con repercusiones en las varianzas de los coeficientes de regresión, que nos invita a reducir a dimensión r.

Vayamos entonces con la tercera parte del estudio dedicada a la solución del problema. Supongamos que los autovalores d_k, $k = r+1, \ldots, q$, presentan valores elevados de IC_k y al menos dos valores altos de $PV_{j,k}$ cada uno. El método en sí consiste en transformar los datos de manera que los factores de inflación de la varianza desaparezcan en favor de las varianzas de las vectores explicativos, que aumentan. Para ello, debemos encontrar una transformación en las variables explicativas (rotación) que las haga incorreladas. Ello nos lleva a calcular las componentes principales

$$u = \mathsf{Z}G, .$$

Esta transformación puede deshacerse mediante $\mathsf{Z} = uG'$. La ventaja que presentan las componentes principales es que son incorreladas. Concretamente, $S_u = D$

[1] Esta condición se establece para marcar diferencias con un diseño ortogonal, en el que no cabe mejora posible.

Así pues, la regresión lineal respecto a $\mathtt{Z}$ puede convertirse en una regresión respecto a u si consideramos el parámetro $\gamma = G'\underline{\beta}$

$$\begin{aligned} Y &= \beta_0 1_{\mathtt{n}} + \mathtt{Z}\underline{\beta} + \mathcal{E} \\ &= \beta_0 1_{\mathtt{n}} + uG + \mathcal{E}, \end{aligned}$$

donde $\mathcal{E}$ sigue un modelo de distribución $N_{\mathtt{n}}(0, \sigma^2 \mathtt{Id})$. El EIMV de γ es

$$\hat{\gamma} = (u'u)^{-1}u'Y = G'\hat{\underline{\beta}},$$

de manera que el estimador de $\underline{\beta}$ puede reconstruirse mediante

$$\hat{\underline{\beta}} = G\hat{\gamma}. \tag{8.7}$$

Sin embargo,

$$\hat{\gamma} \sim N_q\left(\gamma, \frac{\sigma^2}{\mathtt{n}}\, D^{-1}\right).$$

En consecuencia, los estimadores γ_j, $j = 1, \ldots, q$ son independientes, siendo su varianza

$$\mathtt{var}\left[\hat{\gamma}_j\right] = \frac{\sigma^2}{\mathtt{n}}\, d_j^{-1}. \tag{8.8}$$

Además, puede comprobarse que los estimadores $\hat{\gamma}_j$ coinciden con los que se obtendrían en cada caso con una regresión simple. Un diseño de este tipo, en el que los vectores explicativos tienen media aritmética nula y son incorreladas, se denomina ortogonal. Podemos observar que la varianza del estimador es inversamente proporcional a la varianza de la correspondiente componente principal, sin que en este caso exista un factor de inflación de la varianza. Esto no debe inducirnos a pensar que hemos conseguido reducir la matriz de varianzas-covarianzas de los estimadores. De hecho, puede demostrarse fácilmente que, tanto la varianza generalizada[2] como la varianza total[3], permanecen invariantes cuando se consideran las componentes principales.

Consideremos una división de D en dos submatrices diagonales Δ_1 y Δ_2, lo cual induce una división análoga en la matriz G, en vector γ y en su estimador. De esta forma, se verifica

$$\mathtt{Cov}\left[\hat{\underline{\beta}}\right] = \frac{\sigma^2}{\mathtt{n}}\, (G_1 G_2) \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}^{-1} \begin{pmatrix} G_1' \\ G_2' \end{pmatrix} \tag{8.9}$$

$$= \frac{\sigma^2}{\mathtt{n}}\, G_1 D_1^{-1} G_1' + \frac{\sigma^2}{\mathtt{n}}\, G_2 D_2^{-1} G_2'. \tag{8.10}$$

[2]Nos referimos al determinante de la matriz de varianza-covarianzas.

[3]Es decir, la suma de las varianzas de $\hat{\beta}_1, \ldots, \hat{\beta}_q$. o, lo que es lo mismo, la traza de la matriz de varianzas-covarianzas.

Más concretamente, la varianza total de $\hat{\underline{\beta}}$ puede expresarse mediante

$$\mathtt{var}_T\left[\hat{\underline{\beta}}\right] = \frac{\sigma^2}{\mathtt{n}}\left[\mathtt{tr}(D_1) + \mathtt{tr}(D_2)\right] \tag{8.11}$$

Además, $\hat{\underline{\beta}}$ descompone en

$$\hat{\underline{\beta}} = G_1\hat{\gamma}_1 + G_2\hat{\gamma}_2.$$

Si consideramos un nuevo estimador $\hat{\underline{\beta}}^*$ de $\underline{\beta}$ que se obtiene depreciando los coeficientes correspondientes a las componentes principales asociadas a D_2, es decir,

$$\hat{\underline{\beta}}^* = G_1\hat{\gamma}_1, \tag{8.12}$$

se verificará lo siguiente:

$$\mathsf{Sesgo}\left[\hat{\underline{\beta}}^*\right] = G_2\gamma_2, \qquad \mathtt{Cov}\left[\hat{\underline{\beta}}^*\right] = \mathtt{Cov}\left[\hat{\underline{\beta}}\right] - \frac{\sigma^2}{\mathtt{n}}\, G_2 D_2^{-1} G_2',$$

$$\mathtt{var}_T\left[\hat{\underline{\beta}}^*\right] = \mathtt{var}_T\left[\hat{\underline{\beta}}^*\right] - \frac{\sigma^2}{\mathtt{n}}\mathtt{tr}(D_2^{-1}).$$

Así pues, si D_2 contiene los $q - r$ autovalores menores (que son las varianzas de las últimas componentes principales), detectados en el diagnóstico, al considerar este nuevo estimador de $\underline{\beta}$ conseguiremos una gran reducción en la matriz de varianzas-covarianzas. Por contra, el estimador obtenido será sesgado. Este procedimiento resultará rentable cuando el sesgo introducido es pequeño en relación con reducción en las varianzas, cosa que sucede cuando γ_2 es próximo a 0. Por lo tanto, la estrategia consiste en despreciar las componentes principales de menor varianza siempre y cuando su correspondiente coeficiente sea próximo a 0. Una decisión de este tipo puede basarse en los resultados de los test parciales. Mucho autores coinciden en considerar un nivel de significación mayor de lo habitual, por ejemplo 0.20, a la hora de aplicarlos. Por desgracia, no podemos garantizar que los tests parciales aporten resultados no significativos para las componentes principales de menor varianza, pero si esto sucede, cabrá confiar en una sustancial reducción de la matriz de varianzas-covarianzas y, por lo tanto, en una clara mejoría del análisis.

8.1.1. Ejemplo

Para ilustrar en qué medida puede llegar a afectar la correlación lineal entre los vectores explicativos a las estimaciones de los parámetros de regresión, así como la utilidad del uso de componentes principales a la hora de mejorar la precisión de las

estimaciones, haremos uso de un ejemplo extremo con datos simulados mediante el programa SPSS.

Concretamente, generamos `n` = 100 datos correspondientes a dos variables, Z1 y Z2 ya tipificadas que presentan un coeficiente de correlación lineal r=0.969, lo cual supone un factor de inflación de la varianza de 385.42 si actúan como variables explicativas en un problema de regresión. Respecto a las componentes principales, la primera de ellas, que se obtiene proyectando sobre el eje $\langle(1,1)\rangle$, presenta varianza 1.999, lo cual supone el 99.9 % de la varianza total (recordamos que se trata de un caso extremo).

Generamos una variable respuesta Y mediante la fórmula

$$Y = 1.98 \cdot \mathsf{Z1} + 2.02 \cdot \mathsf{Z2} + \varepsilon, \tag{8.13}$$

donde $\varepsilon \sim N(0,1)$. En ese caso se obtiene R^2=0.938 y, tras ejecutar la regresión lineal, obtenemos $\hat{\sigma}^{2,\mathrm{I}}$=1.027, $\hat{\beta}_1$=1.08 no significativo y $\hat{\beta}_2$=2.83 no significativo. Obsérvese que, aunque los coeficientes de regresión estimados distan mucho de los verdaderos, si obviamos los resultados de los test parciales la ecuación obtenida es muy válida para predecir Y.

El diagnóstico de multicolinealidad se puede basar en la no significación de ambos coeficientes, en la presencia de FIV altos y también en el análisis de los IC y de las proporciones de la varianza que presentamos a continuación:

Dimensión	Autovalor	IC	PV-Z1	PV-Z2
1	1.999	1	0	0
2	0.001	39.239	1	1

Todo parece indicar que resultaría conveniente eliminar, si es que es posible, el efecto de la segunda componente principal de la regresión. Para ello efectuamos la regresión respecto a las componentes principales U1 y U2, obteniendo, efectivamente, un resultado en absoluto significativo para el coeficiente de U2. El coeficiente para la primera componentes resulta ser 1.95. Luego, la ecuación quedaría de la forma $Y \simeq 1.95 \cdot \mathtt{U}_1$. Según (8.12), obtenemos la siguiente ecuación de regresión en términos de las variables originales, mucho más ajustada a la realidad del modelo (8.13):

$$Y = 1.95 \cdot \mathsf{Z1} + 1.95 \cdot \mathsf{Z2} + \mathtt{e}, \qquad \mathtt{var}[\mathtt{e}] = 1.08$$

Nótese que, dado que el primer eje principal ha resultado ser en este caso $\langle(1,1)\rangle$, los estimadores de β_1 y β_2 han de ser necesariamente iguales. Estamos pues imponiendo una restricción lineal sobre la solución que permite una gran disminución de la varianza del estimador, a cambio de un pequeño sesgo.

8.2. Análisis de correspondencias

Esta disciplina relativamente moderna y de carácter eminentemente geométrico consiste en la reducción dimensional de una tabla de contingencias[4] mediante las técnicas del análisis de componentes principales. El objetivo final es la representación bidimensional de sus filas y columnas de manera que puedan identificarse claramente los distintos perfiles y la relación entre los mismos. Esta teoría se basa fundamentalmente en los teorema 7.7 y 7.8.

El punto de partida es pues la denominada tabla de contingencia, con el número de observaciones computadas en cada una de las $v \times w$ categorías posibles.

$(I \times J)$	W_1	$\dots$	W_I	Total
V_1	$O_{1,1}$	$\dots$	$O_{1,J}$	$O_{1\cdot}$
$\vdots$		$\ddots$		$\vdots$
V_I	$O_{I,1}$	$\dots$	$O_{I,J}$	$O_{I\cdot}$
Total	$O_{\cdot 1}$	$\dots$	$O_{\cdot J}$	$\mathtt{n}$

Se denotará

$$O = (O_{i,j}), \quad 1 \leq i \leq I,\ 1 \leq j \leq J.$$

La matriz de valores esperados en el caso de independencia se construye mediante

$$E_{ij} = \frac{O_{ij}}{O_{i\cdot}O_{\cdot j}}, \quad 1 \leq i \leq I,\ 1 \leq j \leq J.$$

En ese caso, se define la matriz de proporciones observadas mediante

$$P = \frac{1}{\mathtt{n}} \cdot O$$

y las sumas de sus filas y columnas son, respectivamente,

$$\mathtt{r} = P1_{\mathtt{J}}, \qquad \mathtt{c} = P'1_{\mathtt{I}}\ ^{5}.$$

Si se denota $\mathtt{r} = (\mathtt{r}_1, \dots, \mathtt{r}_I)'$ y $\mathtt{c} = (\mathtt{c}_1, \dots, \mathtt{c}_J)'$, se verifica

$$\sum_{i=1}^{I} \mathtt{r}_i = \sum_{j=1}^{J} \mathtt{c}_j = 1. \tag{8.14}$$

[4]Nos limitaremos aquí a estudiar el análisis de correspondecia simple. El caso múltiple podemos encontrarlo en Greenacre (1984).

[5]$1_{\mathtt{I}}$ y $1_{\mathtt{J}}$ denotan los vectores de $\mathbb{R}^I$ y $\mathbb{R}^J$, respectivamente, cuyas componentes son todas 1.

Asimismo, se denota

$$D_{\mathtt{r}} = \mathtt{diag}(\mathtt{r}), \qquad D_{\mathtt{c}} = \mathtt{diag}(\mathtt{c}).$$

Se supondrá por hipótesis que $\mathtt{r}_i > 0$, para todo $i = 1, \ldots, I$, y $\mathtt{c}_j > 0$, para todo $j = 1, \ldots, J$. En ese caso, se construyen las matrices

$$\mathtt{R} = D_{\mathtt{r}}^{-1} P,$$

cuyas filas, denominadas perfiles de fila, se denotan por $\mathtt{R}_i'$, $i = 1, \ldots, I$, y

$$\mathtt{C} = D_{\mathtt{c}}^{-1} P',$$

cuyas filas, denominadas perfiles de columna, se denotan por $\mathtt{C}_j'$, $j = 1, \ldots, J$. Precisamente y teniendo en cuenta las notación (7.10) y (7.11), se obtienen los siguientes centroides de $\mathtt{R}$ y $\mathtt{C}$

$$\overline{\mathtt{R}}^{D_{\mathtt{r}}} = 1_{\mathtt{I}} \mathtt{c}', \qquad \overline{\mathtt{C}}^{D_{\mathtt{c}}} = 1_{\mathtt{J}} \mathtt{r}'.$$

El primer centroide es la matriz de perfiles filas constante que cabría esperar de P en el caso de que V y W fueran independientes. Lo mismo puede decir del segundo: su traspuesta es la matriz de perfiles columna constante que cabría esperar de P en el caso de que V y W fueran independientes. Los perfiles observados $\mathtt{R}_i$'s y $\mathtt{C}_j$'s pueden considerarse puntos de $\mathbb{R}^J$ y $\mathbb{R}^I$ respectivamente. Se van a considerar las siguientes métricas, siguiendo la notación (7.17):

$$\mathtt{d}_{D_{\mathtt{r}}, D_{\mathtt{c}}^{-1}}\left(\mathtt{X}, \tilde{\mathtt{X}}\right) = \sum_{i=1}^{I} \mathtt{r}_i (\mathtt{X}_i - \tilde{\mathtt{X}}_i)' D_{\mathtt{c}}^{-1} (\mathtt{X}_i - \tilde{\mathtt{X}}_i)$$

$$\mathtt{d}_{D_{\mathtt{c}}, D_{\mathtt{r}}^{-1}}\left(\mathtt{Y}, \tilde{\mathtt{Y}}\right) = \sum_{j=1}^{J} \mathtt{c}_j (\mathtt{Y}_j - \tilde{\mathtt{Y}}_j)' D_{\mathtt{r}}^{-1} (\mathtt{Y}_j - \tilde{\mathtt{Y}}_j)$$

donde $\mathtt{X}$ y $\tilde{\mathtt{X}}$ son matrices $I \times J$ de filas $\mathtt{X}_i'$ y $\tilde{\mathtt{X}}_i'$, respectivamente. Igualmente, $\mathtt{Y}$ y $\tilde{\mathtt{Y}}$ denotan matrices $J \times I$ de filas $\mathtt{Y}_i'$ y $\tilde{\mathtt{Y}}_i'$, respectivamente. Utilizaremos dichas métricas para evaluar las distancias de $\mathtt{R}$ y $\mathtt{C}$ a sus respectivos centroides. De esta forma, obtenemos las siguientes sumas, denominadas inercias:

$$\mathtt{in}(I) = \mathtt{d}_{D_{\mathtt{r}}, D_{\mathtt{c}}^{-1}}\left(\mathtt{R}, \overline{\mathtt{R}}^{D_{\mathtt{r}}}\right) = \sum_{i=1}^{I} \mathtt{r}_i (\mathtt{R}_i - \mathtt{c})' D_{\mathtt{c}}^{-1} (\mathtt{R}_i - \mathtt{c}), \tag{8.15}$$

$$\mathtt{in}(J) = \mathtt{d}_{D_{\mathtt{c}}, D_{\mathtt{r}}^{-1}}\left(\mathtt{C}, \overline{\mathtt{C}}^{D_{\mathtt{c}}}\right) = \sum_{j=1}^{J} \mathtt{c}_j (\mathtt{C}_j - \mathtt{r})' D_{\mathtt{r}}^{-1} (\mathtt{C}_j - \mathtt{r}). \tag{8.16}$$

Desde luego, cuanto mayor es la inercia, más se diferencian los perfiles observados en filas y columnas de los que cabría esperar si V y W fueran independientes. La inercia 0 se corresponde con la independencia. Se trata, por lo tanto, de un concepto con reminiscencias de la Física que pretende ser una medida del grado de correlación existente entre las variables V y W. De hecho, puede demostrarse fácilmente (cuestión propuesta) que

$$\mathtt{in}(I) = \mathtt{in}(J) = \frac{1}{\mathtt{n}}\chi^2, \tag{8.17}$$

donde

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

el cual se utiliza como estadístico de contraste en el test de correlación. Este término puede entenderse en términos intuitivos como la distancia entre la tabla de contingencias obtenida y la que correspondería en el caso de independencia, multiplicada por el tamaño de la muestra. Su relación entre las inercias y el coeficiente de contingencia de Pearson es la siguiente

$$\mathtt{in}(I) = \mathtt{in}(J) = \frac{C^2}{1 - C^2}. \tag{8.18}$$

Nuestro objetivo es proyectar tanto los perfiles fila como columna en sendas subvariedades afines de baja dimensión de manera que se minimizen las distancias entre los perfiles originales y los proyectados. De esta forma tendremos una visión gráfica conjunta de todo el estudio. Para ello y siguiendo las directrices marcadas por el teorema 7.8, hay que considerar los ejes principales y las coordenadas de los perfiles respecto a los mismos. Consideremos primeramente los perfiles fila. Se trata de encontrar una nueva matriz cuyas I filas sean vectores de la menor dimensión posible, siempre y cuando la distancia $\mathtt{d}_{D_r, D_c{}^{-1}}$ a la matriz original sea suficientemente pequeña. Aplicando pues el teorema 7.8 se obtiene

$$\mathtt{R} = \bar{\mathtt{R}}^{D_r} + L \left(\begin{array}{c|c} D_\phi & 0 \\ \hline 0 & 0 \end{array} \right) M', \tag{8.19}$$

donde L y M son matrices verificando

$$L' D_r L = \mathtt{Id}_I, \qquad M' D_c{}^{-1} M = \mathtt{Id}_J.$$

D_Φ denota una matriz diagonal de elementos positivos y de orden $\mathtt{rg}P - 1$. En ese caso, M es la matriz de ejes principales para las filas y la matriz de coeficientes será

$$L \left(\begin{array}{c|c} D_\phi & 0 \\ \hline 0 & 0 \end{array} \right).$$

Si pretendemos reducir a dimensión $k < \mathtt{rg}(P)-1$, la distancia entre la matriz original y la proyectada coincidirá con la suma de los últimos $\mathtt{rg}(P)-(k+1)$ elementos de la diagonal de D_ϕ. La inercia correspondiente a la matriz proyectada será, por contra, la suma de los k primeros elementos de la diagonal. Se procede análogamente para las columnas, y se obtiene

$$\mathtt{C} = \overline{\mathtt{C}}^{D_c} + Y \left(\begin{array}{c|c} D_\gamma & 0 \\ \hline 0 & 0 \end{array} \right) N', \tag{8.20}$$

donde Y y N verifican

$$Y' D_c Y = \mathtt{Id}_J, \qquad N' {D_r}^{-1} N = \mathtt{Id}_I.$$

En ese caso, N es la matriz de ejes principales para las filas y la matriz de coeficientes será

$$Y \left(\begin{array}{c|c} D_\gamma & 0 \\ \hline 0 & 0 \end{array} \right).$$

Si pretendemos reducir a dimensión k, la distancia entre la matriz original y la proyectada coincidirá con la suma de los últimos $\mathtt{rg}(P)-(k+1)$ elementos de la diagonal de D_γ, y la inercia correspondiente a la matriz proyectada será la suma de los k primeros elementos de la diagonal. Ahora bien, multiplicando en (8.19) a la izquierda por D_r se obtiene

$$P - \mathtt{rc}' = (D_r L) \left(\begin{array}{c|c} D_\phi & 0 \\ \hline 0 & 0 \end{array} \right) M',$$

donde

$$(D_r L)' {D_r}^{-1} (D_r L) = \mathtt{Id}_I, \qquad M' {D_c}^{-1} M = \mathtt{Id}_J.$$

Así mismo, multiplicando en (8.20) a la izquierda por D_c y trasponiendo se obtiene

$$P - \mathtt{rc}' = N \left(\begin{array}{c|c} D_\gamma & 0 \\ \hline 0 & 0 \end{array} \right) (D_c Y)',$$

donde

$$(D_c Y)' {D_c}^{-1} (D_c Y) = \mathtt{Id}_J, \qquad N' {D_r}^{-1} N = \mathtt{Id}_I.$$

Entonces, dada la unicidad de la descomposición SVD [6], se verifica que $D_\phi = D_\gamma$ (denótese dicha matriz por $D = \mathtt{diag}(d_1, \ldots, d_{\mathtt{rg}(P)-1})$), y

$$P - \mathtt{rc}' = N \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array} \right) M', \tag{8.21}$$

[6] Tener en cuenta la unicidad en sentido débil de la misma, según se comentó en el capítulo anterior.

donde

$$N' {D_r}^{-1} N = \mathtt{Id}_I, \qquad M' {D_c}^{-1} M = \mathtt{Id}_J. \tag{8.22}$$

Es decir, los ejes principales sobre los que hemos de proyectar para obtener soluciones lo más próximas posibles a los perfiles originales se obtienen de la descomposición SVD de $P - \mathtt{rc}'$ para la métrica $\mathtt{d}_{\mathtt{D}_\mathtt{r}^{-1},\mathtt{D}_\mathtt{c}^{-1}}$. Las coordenadas de las filas y columnas respecto a dichos ejes se obtienen, respectivamente, mediante

$$F = {D_r}^{-1} N \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array} \right), \qquad G = {D_c}^{-1} M \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array} \right). \tag{8.23}$$

Podemos ofrecer una expresión alternativa de F y G. Según (8.19), se verifica que

$${D_r}^{-1} P - 1_\mathtt{I} \mathtt{c}' = F M'.$$

Luego, multiplicando a la derecha por ${D_c}^{-1} M$ y teniendo en cuenta (8.22), se verifica

$$F = ({D_r}^{-1} P - 1_\mathtt{I} \mathtt{c}') {D_c}^{-1} M. \tag{8.24}$$

Siguiendo un argumento análogo, se tiene que

$$G = ({D_c}^{-1} P' - 1_\mathtt{I} \mathtt{r}') {D_r}^{-1} N. \tag{8.25}$$

Multiplicando en (8.24) y (8.25) a la izquierda por $\mathtt{r}'$ y $\mathtt{c}'$, respectivamente, y teniendo en cuenta (8.14), se deduce que

$$\mathtt{r}' F = 0', \qquad \mathtt{c}' G = 0', \tag{8.26}$$

es decir, que las medias aritméticas de las filas de F y G, ponderadas por D_r y D_c, respectivamente, son nulas. En lo que sigue, $\overline{M}, \overline{N}, \overline{F}$ y $\overline{G}$ denotarán las $\mathtt{rg}(P) - 1$ primeras columnas de M, N, F y G, respectivamente, que, al fin y al cabo, son las únicas que interesan. En ese caso, consideramos (8.23) y tenemos en cuenta (8.26) se deduce fácilmente que

$$1_\mathtt{I}' \overline{M} = 0', \qquad 1_\mathtt{J}' \overline{N} = 0', \tag{8.27}$$

es decir, que las medias aritméticas de las filas de $\overline{M}$ y $\overline{N}$ son nulas. Según el teorema 7.8, las distancias entre las matrices originales y las proyecciones vale

$$\sum_{i=k+1}^{\mathtt{rg}(P)-1} d_i,$$

y la inercia de las proyecciones

$$\sum_{i=1}^{k} d_i,$$

tanto en el caso de los perfiles fila como en el de los perfiles columnas. Este último sumando se denomina parte de inercia explicada por los k primeros ejes principales, dado que la inercia total es, según el teorema 7.8,

$$\sum_{i=1}^{\mathtt{rg}(P)-1} d_i.$$

El término d_i, $i = 1, \ldots, \mathtt{rg}(P) - 1$, se denomina parte de inercia explicada por el i-ésimo eje principal. Teniendo en cuenta cuál es nuestro objetivo final, el número de ejes principales a seleccionar coincidirá con el número de elementos suficientemente grandes de la matriz diagonal D. Es así como se realiza la reducción de la dimensión. Una vez decidido el valor de k, podemos analizar gráficamente bien los perfiles de fila o bien los de columna, de manera que la proximidad de dos perfiles en la gráfica se entienda como un similar patrón de comportamiento de ambos niveles del carácter respecto a la otra cualidad estudiada. No obstante y en rigor, la métrica a considerar en este caso no es euclídea sino las métrica $d_{\mathtt{D_c}^{-1}}$ en el caso de los perfiles fila y $d_{\mathtt{D_r}^{-1}}$ en el caso de los perfiles columnas.

También pueden representarse conjuntamente los perfiles filas y columnas e interpretar el gráfico teniendo en cuenta lo siguiente. Si trasponemos (8.21) y multiplicamos a la izquierda por ${D_c}^{-1}$ se obtiene

$${D_c}^{-1}(P' - \mathtt{cr}') = {D_c}^{-1} M \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right) N'.$$

Lo cual, según (8.23), equivale a

$${D_c}^{-1} P' - 1_{\mathtt{J}} \mathtt{r}' = G N'.$$

Si multiplicamos a a la derecha por ${D_r}^{-1} N \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right)$ y tenemos en cuenta (8.22) y (8.23), se verifica

$${D_c}^{-1} P' F - 1'_{\mathtt{J}} \mathtt{r}' F = G \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right).$$

Según (8.26), el segundo sumando del primer término es nulo y, despejando, obtenemos por consiguiente

$$\overline{G} = \mathtt{C}\overline{F} D^{-1}. \tag{8.28}$$

Un razonamiento completamente análogo nos permite afirmar que

$$\overline{F} = \mathtt{R}\overline{G} D^{-1}. \tag{8.29}$$

Estas expresiones pueden resultar bastante útiles. De no ser por la presencia de la matriz D^{-1}, podríamos afirmar que, si dos niveles de los caracteres estudiados suelen presentarse simultáneamente, entonces tendrían unas coordenadas muy parecidas a la hora de proyectarlos sobres los correspondientes ejes principales, de manera que si se proyectaran los perfiles filas y columnas sobre los mismos ejes, podríamos entender la proximidad entre un perfil fila y otro columna como una asociación positiva entre ambos niveles. Por desgracia, los autovalores $d_1, \ldots, d_k$ *deforman* esta interpretación gráfica, aunque en una magnitud inversa al valor de los mismos y, por ende, a la importancia del eje correspondiente. Por ello, este tipo de representación gráfica, denominada *biplot*, puede ser de utilidad en una primera exploración acerca de la relación entre los caracteres cuando se presentan muchos niveles en los mismos, pero no es aconsejable en fases más rigurosas del estudio.

Veamos a continuación una forma más sencilla, desde un punto de vista operacional, de obtener los ejes principales y la proporción de inercia explicada por los mismos. Se trata de considerar la descomposición SVD tipo (7.21) de P

$$P = \tilde{N}\Delta\tilde{M}', \tag{8.30}$$

y eliminar las partes triviales. En efecto, se sigue de (8.21) que

$$P = (\mathtt{r}|\overline{N}) \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & D \end{array}\right) (\mathtt{c}|\overline{M})'. \tag{8.31}$$

Además, de (8.14) se sigue que $\mathtt{r}'{D_r}^{-1}\mathtt{r} = \mathtt{c}'{D_c}^{-1}\mathtt{c} = 1$, y de (8.27) deduce que $\mathtt{r}'{D_r}^{-1}\overline{N} = 0$ y $\mathtt{c}'{D_r}^{-1}\overline{M} = 0$. Es decir, que, por la unicidad de la descomposición (7.21), (8.30) y (8.31) coinciden. Por lo tanto, la matriz de autovalores D resulta de eliminar el término 1 de Δ, que coincide, según probaremos a continuación, con el mayor de los autovalores, y las matrices $\overline{M}$ y $\overline{N}$ resultan de eliminar en $\tilde{M}$ y $\tilde{N}$, respectivamente, las columnas asociadas al mismo, que son $\mathtt{c}$ y $\mathtt{r}$. Los matrices de coordenadas $\overline{F}$ y $\overline{G}$ se obtienen entonces automáticamente. Vamos a probar entonces que el resto de los autovalores han de ser iguales o inferiores a 1.

Lema 8.1.

Sean $Z = (z_{ik})_{i,k} \in \mathcal{M}_{p\times n}$ y $S = (s_{k,j})_{r,j} \in \mathcal{M}_{n\times p}$ cuyas componentes son todas no negativas y tales que

$$\sum_{k=1}^{n} z_{ik} = 1, \quad \forall i = 1, \ldots, p, \tag{8.32}$$

$$\sum_{j=1}^{n} s_{kj} = 1, \quad \forall k = 1, \ldots, n. \tag{8.33}$$

Entonces, para cada vector $x = (x_1, \ldots, x_p)' \in \mathbb{R}^p$, se verifica que todas las componentes del vector ZSx son menores o iguales que $\|x\|_{\texttt{max}} = \text{máx}\{|x_j| : \ j = 1, \ldots, p\}$.

Demostración.
Efectivamente, si $1 \geq i \geq p$, denótese por $(ZSz)_i$ a la i-ésima componente de ZSx. Entonces,

$$\begin{aligned}(ZSx)_i &= \sum_{k=1}^{n} z_{ik} \sum_{j=1}^{p} s_{kj} x_j \leq \sum_{k=1}^{n} z_{ik} \sum_{j=1}^{p} s_{kj} |x_j| \\ &\leq \sum_{k=1}^{n} z_{ik} \left(\sum_{j=1}^{p} s_{kj} \right) \|x\|_{\texttt{max}} = \left(\sum_{k=1}^{n} z_{ik} \right) \|x\|_{\texttt{max}} \\ &= \|x\|_{\texttt{max}}\end{aligned}$$

∎

Lema 8.2.

Sea $X = (x_i j)_{i,j}$ una matriz $n \times p$ de términos no negativos y definamos A y B como las matrices diagonales de términos $\{a_1, \ldots, a_n\}$ y $\{b_1, \ldots, b_p\}$, donde $a_1, \ldots, a_n$ y $b_1, \ldots, b_n$ son, respectivamente, las componentes de los vectores $a = X1_p$ y $b = X'1_n$, que supondremos todas estrictamente positivas. En ese caso, los autovalores de la matriz

$$B^{-1} X' A^{-1} X \tag{8.34}$$

se encuentran todos en el intervalo [0,1], siendo 1 el mayor de ellos, que se alcanza en el autovector 1_p.

Demostración.
En primer lugar, tener en cuenta que, trivialmente, los autovalores de la matriz (8.34) coinciden con los de la matriz simétrica semidefinida positiva

$$B^{-1/2} X' A^{-1/2} X B^{-1/2}, \tag{8.35}$$

y que $e \in \mathbb{R}^p$ es un autovector de la matriz (8.34) sii $B^{1/2}e$ lo es de (8.35). En consecuencia, los autovalores es son reales y no negativos, y también son reales las componentes de los correspondientes autovectores. Que 1_p sea autovector asociado al autovalor 1 se sigue directamente de las definiciones de A y B. Veamos que se trata del máximo autovalor, que se denota por λ. Por el teorema (13.4), sabemos que

$$\lambda = \max_{y \in \mathbb{R}^p \setminus \{0\}} \frac{y' B^{-1/2} X' A^{-1/2} X B^{-1/2} y}{y' y}.$$

Teniendo en cuenta que las componenetes de la matriz (8.35) son todas no negativas, se sigue trivialmente que el máximo anterior se alcanza en un vector y (autovector) cuyas componentes tengan todas el mismo signo. Podemos suponer, sin pérdida de generalidad, que todas las componentes de dicho vector son positivas y, por ende, lo son también las componentes del autovector de (8.34) asociado a λ, que se denotará por $\overline{e}$. Supongamos también que $\|\overline{e}\|_{\mathtt{max}}$ se alcanza en la componente i-ésima de $\overline{e}$, $\overline{e}_i$, para cierto $1 \leq i \leq p$. Dado que las matrices $A^{-1}X$ y $B^{-1}X'$ verifican las condiciones (8.32) y (8.33), respectivamente, se verifica por el lema anterior que $\lambda\overline{e}_i \leq \overline{e}_i$ y, en consecuencia, $\lambda \leq 1$.

∎

Teorema 8.3.
Las partes de la inercia explicadas por los distintos ejes principales son siempre menores o iguales que 1.

Demostración.
Se sigue de (8.31) que

$$D_c{}^{-1/2}P'D_r{}^{-1}PD_c{}^{-1/2} = D_c{}^{-1/2}(\mathtt{c}|\overline{M})\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & D^2 \end{array}\right)\left(D_c{}^{-1/2}(\mathtt{c}|\overline{M})\right)',$$

siendo $D_c{}^{-1/2}(\mathtt{c}|\overline{M})$ una matriz ortogonal, de lo cual se deduce que 1 junto con los elementos diagonales de D^2 son los autovalores de la matriz simétrica

$$D_c{}^{-1/2}P'D_r{}^{-1}PD_c{}^{-1/2},$$

que coinciden con los de la matriz $D_c{}^{-1}P'D_r{}^{-1}P$. Por el lema anterior, dichos autovalores están comprendidos entre 0 y 1 y, en consecuencia, también lo están

$$d_1, \ldots, d_{\mathtt{rg}(P)-1},$$

que son las partes de la inercia correspondientes a los distintos ejes principales.

∎

Este resultado nos permite, entre otras cosas, acotar los términos $\mathtt{in}(I)$ y $\mathtt{in}(J)$.

Corolario 8.4.
$\mathtt{in}(I) = \mathtt{in}(J) \leq \text{mín}\{I, J\} - 1$

Demostración.
Basta tener en cuenta que $\mathtt{in}(I) = \mathtt{in}(J) = \sum_{i=1}^{\mathtt{rg}(P)-1} d_i$. El teorema anterior concluye la prueba.

∎

Además, es fácil diseñar una tabla de contingencia de manera que dicha cota superior se alcance. Por ejemplo, considerar el caso $I = J$ y construir una tabla diagonal. Por último, teniendo en cuenta la cota anterior para las inercias junto con la igualdad (8.18), podemos obtener una cota superior para el coeficiente de contingencia de Pearson en función mín$\{I, J\}$, que puede llegar a alcanzarse por el mismo razonamiento. Concretamente

$$C \leq \sqrt{\frac{\text{mín}\{I, J\} - 1}{\text{mín}\{I, J\}}} \tag{8.36}$$

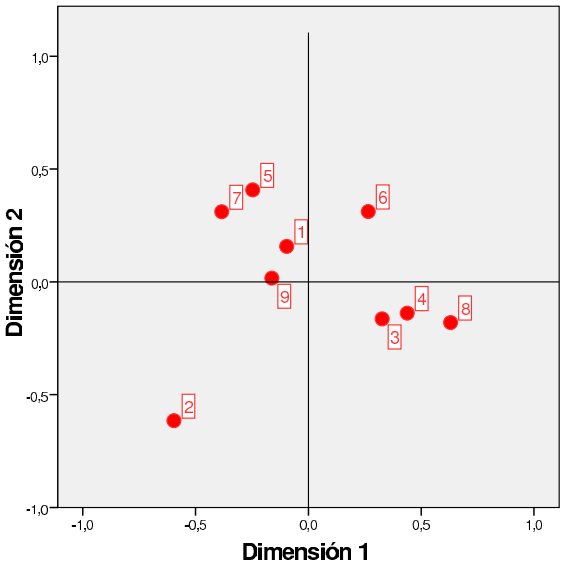
8.2.1. Ejemplo

En una cierta especie animal los espermatozoides se clasifican, en virtud de su motilidad, en cinco variedades diferentes. Nueve sementales de dicha especies aportaron entre todos una muestra de $\mathtt{n} = 176692$ espematozoides que fueron asignados a sus respectivas variedades, resultando una tabla de contingencia de dimensiones 9×5. El objetivo es determinar si existe correlación entre ambos caracteres, es decir, si los diferentes sementales presentan un mismo patrón de distribución (perfil) o, por el contrario, la distribución de las variedades de espermatozoide varían en función del semental. En ese caso, sería interesante dilucidar en qué radica exactamente dicha correlación. El problema puede plantearse de manera simétrica y equivalente desde el punto de vista de los tipos de espermatozoides, es decir, nos planteamos si todos los tipos de espermatozoides se distribuyen de igual manera entre los diferentes semantales o no.

En este caso se obtuvieron las siguientes medidas del grado de asociación, todas ellas relacionadas directamente en virtud de (8.17) y (8.18):

$$\mathtt{in} = 0{,}032 \qquad \chi^2 = 5723{,}6 \qquad C = 0{,}177$$

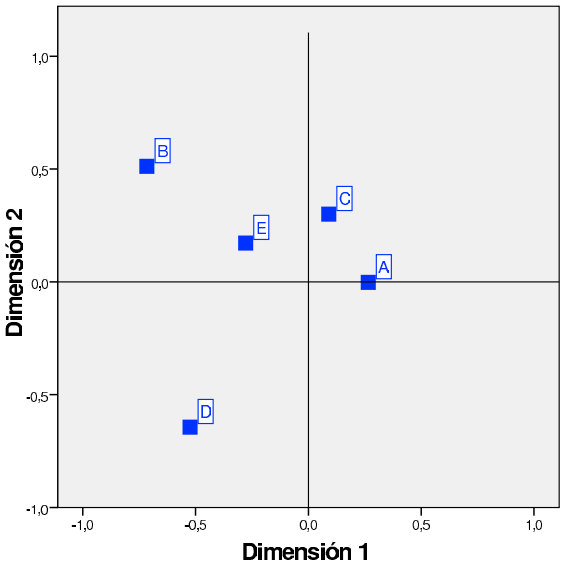
Dado que, según (8.36), el coeficiente dde contingencia C debe estar comprendido entre 0 y 0.89, podemos considerar que en la muestra se observa un grado de correlación medio-bajo que, dado el enorme tamaño de la misma, es, no obstante, claramente significativo. Para determinar las claves de esta sutil asociación sería interesante proyectar los perfiles *semental* y los perfiles *tipo de espermatozoide* sobre sendos espacios bidimensionales, de manera que quede explicada la mayor parte de la inercia. Concretamente, la proyección sobre el plano generado por los dos primeros ejes principales explica (en ambos caso, por supuesto) un 94.2 % de la inercia total, lo cual es muy satisfactorio. De esta forma obtenemos en primer lugar el siguiente gráfico para los perfiles *sementales*:



Observamos que, salvo el número 2, todos los sementales aparecen relativamente cercanos[7]. La proximidad entre dos sementales ha de interpretarse como una similar distribución respecto a las variedades de espermatozoides, mientras que la lejanía se corresponde con patrones de distribución opuestos. El semental responsable de la mayor parte de la inercia es sin duda el número 2. De ello puede deducirse que el perfil, es decir, el patrón de distribución de este semental, es claramente diferente al del resto. De no ser por él, la inercia y, por lo tanto, el grado de correlación, serían mucho menores.

Respecto a los perfiles *tipo de espermatozoide*, sucede algo similar a lo que ocurría con los sementales, pues todos los tipos se distribuyen en similares proporciones entre los diferentes sementales, excepto el tipo D, que aparece más distante, siendo responsable de la mayor parte de la inercia. Podemos apreciarlo en el siguiente gráfico:

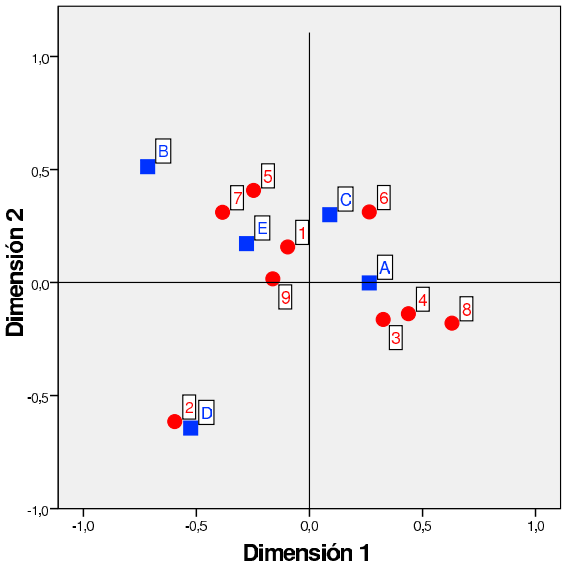
[7]Hemos de ser precavidos al realizar una afirmación de este tipo pues, según hemos visto, la distancia que rige en la interpretación del gráfico no es la euclídea, sino $d_{D_c^{-1}}$ en el caso de los perfiles fila y $d_{D_r^{-1}}$ en el caso de los perfiles columna.



Todo esto nos induce a analizar la relación entre el semental 2 y el tipo D, pues puede ser la clave de la relación. Lo más aconsejable es acudir a la propia tabla de contingencia.

Según hemos comentado con anterioridad, puede resultar muy aventurado[8] combinar las proyecciones de filas y columnas en un único gráfico denominado *biplot*. No obstante, la práctica es bastante habitual. El biplot se interpreta así: la proximidad entre un semental y una variedad de espermatozoide se corresponde con la afinidad entre ambos, mientras que la lejanía se interpreta como la repulsión. En el gráfico que presentamos a continuación, queda bastante patente la afinidad existente entre el semental 2 y la variedad D, es decir, que en el semental 2 abundan más los espermatozoides del tipo D, en detrimento del resto de tipos, claro está, y en este hecho radica posiblemente la correlación observada entre ambos caracteres. Página 1

[8]Debido a que los autovalores d_1 y d_2 correspondientes a la descomposición SVD de $P - \mathtt{rc}'$ deforman, por así decirlo, la interpretación gráfica.



Cuestiones propuestas

1. Demostrar la validez de (8.4).
2. Demostrar la validez de (8.6).
3. Demostrar la validez de (8.1).
4. Demostrar la validez de (1.8).
5. Demostrar la validez de (8.17).

Página 1

Capítulo 9

Análisis discriminante I

En lo que se refiere a este término hemos encontrado diferentes acepciones según la bibliografía consultada. Entendemos que en sentido estricto se refiere al estudio de técnicas para distinguir o discriminar diferentes grupos de datos. Sin embargo, en muchas ocasiones se utiliza para hacer referencia al problema de decisión consistente en clasificar una observación p-dimensional en un determinado grupo entre r posibles. Es el caso, por ejemplo, del diagnóstico médico pues, basándonos en la observación en el individuo de determinadas variables sintomáticas (un análisis de sangre y orina, por ejemplo) debemos determinar la enfermedad que padece el mismo (considerar una enfermedad como una grupo humano que padece unos síntomas comunes). Esta ambigüedad del término no es casual, dado que para poder clasificar una observación respecto a r grupos debemos establecer primeramente criterios para diferenciarlos claramente. Queremos pues decir que el problema de clasificación consta de dos etapas, y hemos optado, siguiendo como en otras ocasiones el esquema de Rencher (1995), por dedicar un capítulo diferente a cada una de ellas. Además, dedicamos unas secciones adicionales a comentar otros métodos alternativos de clasificación que no guardan relación con las puntuaciones discriminantes.

Por lo tanto, en este primer capítulo se estudia la manera de discriminar r grupos. Este estudio guarda una estrecha relación con la comparación de r medias, denominado manova (de hecho, puede considerarse una continuación del mismo) y, por lo tanto, con el análisis de correlación canónica. Al igual que los capítulos precedentes se basa en el cálculos de ciertos valores teóricos que, en este caso, nos permitirán distinguir (si es que es posible) las distribuciones. Es aquí donde esperamos comprender definitivamente el significado de los autovalores que se obtienen en el manova. Se considerará como caso particular la discriminación de dos distribuciones y se analizará por último la influencia de cada una de las variables originales en la discrimi-

nación. Los métodos que estudiaremos se basan en los mismo supuestos del manova: p-normalidad e igualdad de matrices de covarianzas, lo cual no significa que hayan de ser descartados si estas hipótesis no se cumplen.

9.1. Ejes discriminantes

Consideremos, para cada $i = 1, \ldots, r$, sendas muestras aleatorias simples e independientes, $Y_{i1}, \ldots, Y_{i\mathtt{n}_i}$, de una distribución $N_p(\mu_i, \Sigma)$, y supongamos que se desea contrastar la hipótesis inicial

$$H_0 : \mu_1 = \ldots = \mu_r.$$

Nótese que son los supuestos del manova de una vía, es decir, se trata de un modelo lineal con un subespacio V de dimensión r y una hipótesis inicial W de dimensión 1. Sea $\mathtt{n} = \sum \mathtt{n}_i$ y considerense las matrices $n \times p$ siguientes

$$Y = \begin{pmatrix} Y'_{11} \\ \vdots \\ Y'_{1\mathtt{n}_1} \\ \vdots \\ Y'_{r1} \\ \vdots \\ Y'_{r\mathtt{n}_r} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu'_1 \\ \vdots \\ \mu'_1 \\ \vdots \\ \mu'_r \\ \vdots \\ \mu'_r \end{pmatrix}.$$

se verifica entonces

$$Y \sim N_{n,p}(\mu, \mathtt{Id}, \Sigma).$$

Consideremos las matrices $S_2 = Y' P_{V|\langle 1_{\mathtt{n}} \rangle} Y$ y $S_3 = Y' P_{V^\perp} Y$ (nos referimos, en este caso, a las matrices (4.16) y (4.19), respectivamente) y $t_1, \ldots, t_b$ los autovalores positivos de $S_3^{-1} S_2$, donde $b = \text{mín}\{p, r-1\}$, que constituye, como ya sabemos, un estadístico invariante maximal. En esta sección interpretaremos el significado exacto de cada uno de estos autovalores.

Si consideramos dos vectores $a, b \in \mathbb{R}^p$, tal que $\|a\| = 1$, el producto escalar $a'b \in \mathbb{R}$ es la longitud de la proyección del vector b sobre el eje que determina a. Para contrastar la hipótesis inicial

$$H_0^a : a'\mu_1 = \ldots = a'\mu_r,$$

podemos considerar, cada todo $i = 1, \ldots, r$, los conjuntos $a'Y_{i1}, \ldots, a'Y_{i\mathtt{n}_i}$, que constituyen sendas muestras aleatorias simples de tamaño $\mathtt{n}_i$ de cada distribución $N(a'\mu_i, a'\Sigma$

respectivamente. El contraste que proponemos se resuelve entonces mediante una anova de una vía. Por ejemplo, si $a = (1, 0, \ldots, 0)'$, estaremos contrastando la hipótesis inicial de igualdad de la primera componente de la media, a partir de la primera componente de los datos de la muestra (o lo que es lo mismo, el primer vector columna de la matriz de observaciones Y). El vector de datos (proyectados) es

$$Ya \sim N_n\big(\mu a, (a'\Sigma a) \cdot \mathtt{Id}\big).$$

El test F, que es UMP-invariante a nivel α y de razón de verosimilitudes, consiste, recordemos, en comparar el valor

$$\begin{aligned} F_a(Y) &= \frac{\mathtt{n}-r}{r-1} \cdot \frac{\|P_{V|W}Ya\|^2}{\|P_{V^\perp}Ya\|^2} \\ &= \frac{\mathtt{n}-r}{r-1} \cdot \frac{a'Y'P_{V|W}Ya}{a'Y'P_{V^\perp}Ya} \\ &= \frac{\mathtt{n}-r}{r-1} \cdot \frac{a'S_2a}{a'S_3a}, \end{aligned}$$

con el cuantil $F^{\alpha}_{r-1,\mathtt{n}-r}$. Dado que, si a y $\tilde{a}$ son vectores proporcionales, los estadísticos F_a y $F_{\tilde{a}}$ coinciden, debería denotarse más bien $F_{\langle a\rangle}$. Busquemos el máximo valor de $F_{\langle a\rangle}(Y)$ para todos los posibles ejes de proyección, es decir, buscamos

$$\max_{a\in\mathbb{R}^p\setminus\{0\}} F_{\langle a\rangle}(Y),$$

que es equivalente a buscar

$$\frac{\mathtt{n}-r}{r-1} \max_{a'S_3a=1} a'S_2a.$$

Como $\{a :\in \mathbb{R}^p : a'S_3a = 1\}$ es compacto, el máximo existe. Supongamos que se alcanza en un vector a_1 o, mejor dicho, en el eje $\langle a_1\rangle$. Situémonos en las condiciones del el teorema 13.3. Se trata de maximizar la función

$$\phi : a \mapsto a'S_2a$$

bajo la condición $f(a) = 0$, donde

$$f : a \mapsto a'S_3a - 1.$$

Entonces, existe necesariamente un único número real t tal que

$$\nabla(\phi - t \cdot f)(a_1) = 0.$$

Luego, considerando las derivadas parciales, debe verificarse $(S_2 - tS_3)a_1 = 0$, es decir,

$$S_3^{-1}S_2a_1 = ta_1. \tag{9.1}$$

Por lo tanto, t es un autovalor de $S_3^{-1}S_2$ y a_1 es un autovector asociado verificando que $a_1'S_3a_1 = 1$. Teniendo en cuenta (9.1), se sigue que el valor del máximo es $\frac{\mathtt{n}-r}{r-1}t$ y t es el primero de los autovalores, t_1, pues en el caso contrario incurriríamos, trivialmente, en una contradicción. Es decir,

$$\max_{a\in\mathbb{R}^p\setminus\{0\}} F_{\langle a\rangle}(Y) = \frac{\mathtt{n}-r}{r-1}t_1.$$

Hemos obtenido pues el eje sobre el cual debemos proyectar si queremos obtener una máxima discriminación entre los grupos. En consecuencia, el test F correspondiente al contraste univariante relativo a la proyección sobre dicho eje tiene como estadístico de contraste el primer autovalor t_1 [1]. El eje $\langle a_1\rangle$ se denomina primer eje discriminante, que tiene una íntima relación con los intervalos de confianza simultáneos obtenidos en la sección 2.5 y, por lo tanto, con el test de Roy (cuestión propuesta). El vector de $\mathbb{R}^{\mathtt{n}}$

$$w_1 = Ya_1,$$

que contiene las proyecciones sobre el eje $\langle a_1\rangle$ de los $\mathtt{n}$ vectores de $\mathbb{R}^p$ considerados, se denominará primer vector de puntuaciones discriminantes.

El proceso de obtención de ejes discriminantes puede proseguir. El paso siguiente es encontrar el eje que determine una proyección incorrelada con w_1 que maximice $F_{\langle a\rangle}$. Supongamos que $t_1 > 0$, cosa que sucede con probabilidad 1, y consideremos cualquier $a \in \mathbb{R}^p$. Se sigue de (9.1) que

$$a'S_3a_1 = \frac{1}{t_1}a'S_2a_1.$$

Luego,

$$a'S_3a_1 = 0 \Leftrightarrow a'S_2a_1 = 0. \tag{9.2}$$

Por otro lado, la covarianza muestral entre los vectores de datos Ya y w_1 puede descomponerse (cuestión propuesta) de la siguiente forma

$$s_{Ya,w_1} = \frac{1}{\mathtt{n}}\big[(\mathtt{n}-r)a'S_3a_1 + (r-1)a'S_2a_1\big]. \tag{9.3}$$

Entonces, se sigue que Ya y w_1 son incorrelados si, y sólo si, $a'S_3a_1 = 0$. Luego, estamos buscando

$$\max\big\{F_{\langle a\rangle}(Y)\colon a'S_3a_1 = 0\big\},$$

[1] Salvo la constante $(\mathtt{n}-r)/(r-1)$.

o, equivalentemente, buscamos

$$\text{máx}\left\{a'S_2a\colon a'S_3a=1\ \wedge\ a'S_3a_1=0\right\}.$$

Éste se alcanza en algún vector a_2. Aplicando nuevamente el teorema 13.3 con

$$f: a\longmapsto\begin{pmatrix} a'S_3a-1\\ a'S_3a_1\end{pmatrix},$$

se deduce que existe un único par $(t,\theta)\in\mathbb{R}^2$ tal que

$$S_2a_2-tS_3a_2-\theta S_3a_1=0.$$

Si multiplicamos a la izquierda por a_1 y tenemos en cuenta las restricciones impuestas a a_2 junto con la equivalencia (9.2), se deduce que $\theta=0$. Por tanto, razonando análogamente al primer paso se obtiene que el valor del máximo es t_2, que se alcanza en el autovector a_2 correspondiente. El eje $\langle a_2\rangle$ se denomina segundo eje de discriminación. El vector n-dimensional

$$w_2=Ya_2$$

se denomina segundo vector de puntuaciones discriminantes. Por lo tanto, se verifica que el test F correspondiente al contraste para la proyección sobre el segundo eje discriminante tiene como estadístico de contraste el autovalor t_2 (salvo cte.).

El proceso puede continuar hasta agotar todos los autovalores positivos $t_1,\ldots,t_b$, obteniéndose b ejes de discriminación $\langle a_1\rangle,\ldots,\langle a_b\rangle$ y b vectores de puntuaciones discriminantes $w_1,\ldots,w_b$. La matriz

$$w=(w_1\ldots w_b)\in\mathcal{M}_{\mathtt{n}\times b}$$

se denomina matriz de puntuaciones discriminantes. En conclusión, hemos probado el siguiente resultado:

Teorema 9.1.

En las condiciones anteriores se verifica

(a) $\frac{\mathtt{n}-r}{r-1}t_1$ es el valor máximo de $F_{\langle a\rangle}(Y)$, alcanzándose en el eje $\langle a_1\rangle$.

(b) Si $i=2,\ldots,b$, $\frac{\mathtt{n}-r}{r-1}t_i$ es el valor máximo de $F_{\langle a\rangle}(Y)$ si nos restringimos a los ejes tales que la proyección sobre los mismos es incorrelada con la proyección sobre los ejes $\langle a_1\rangle,\ldots,\langle a_{i-1}\rangle$. Dicho máximo se alcanza en el eje $\langle a_i\rangle$ [2].

[2]Relacionar con el teorema 13.7.

Por lo tanto, la capacidad en la discriminación viene dada por el correspondiente autovalor y va, en consecuencia, en orden decreciente: un autovalor t_i próximo a 0 indica que la proyección sobre el i-ésimo eje discriminante no evidencia diferencias entre los r grupos. En ese caso, dicho eje aportará poca información a la hora de asignar una observación a alguno de los grupos. Entonces, los ejes asociados a pequeños autovalores pueden ser despreciados en un problema de discriminación.

Respecto a la relación entre los distintos ejes discriminates, hemos de recordar que los autovalores de $S_3^{-1}S_2$ coinciden con los de $Z_2S_3Z_2'$, en los términos establecidos en el capítulo 2 (recordemos: $S_2 = Z_2'Z_2$). No obstante, se verifica que a es un autovector asociado a un autovalor t de la matriz $S_3^{-1}S_2$ si, y sólo si, Z_2a es un autovector asociado al mismo autovalor para la matriz simétrica $Z_2S_3Z_2'$. Entonces, si a_* y a_{**} son autovectores de $S_3^{-1}S_2$ asociados a dos autovalores distintos, debe verificarse que Z_2a_* y Z_2a_{**} son ortogonales, es decir,

$$a_*'S_2a_{**} = 0.$$

Luego, se sigue de (9.2) que

$$a_i'S_3a_j = 0, \qquad i \neq j.$$

Por lo tanto, no se trata, en principio, de ejes perpendiculares, a menos que $S_3 = \mathtt{Id}$. Por otro lado, recordemos que los autovalores $t_1, \ldots, t_b$ son parámetros muestrales que estiman, respectivamente, a $\theta_1, \ldots, \theta_b$, los autovalores de la matriz $\Sigma^{-1}\delta$, donde $\delta = \mu' P_{V|W}\mu$. Estos autovalores y sus respectivos autovectores gozan de una interesante interpretación, análoga por completo a la que hemos visto anteriormente (cuestión propuesta). Puede ser interesante realizar un test previo de significación de los mismos, es decir, contrastar una hipótesis inicial del tipo[3]

$$H_0 : \theta_{s+1} = \ldots = \theta_b = 0.$$

El contraste se resuelve comparando el estadístico de contraste

$$\lambda_1^d = n \sum_{j=s+1}^{b} \ln(1 + t_j) \tag{9.4}$$

con el cuantil

$$\chi^{2,\alpha}_{(p-s)(r-1-s)}{}^{4}.$$

[3] Expresiones de este tipo sólo tienen sentido si $r \geq 3$.

[4] Se trata del test de la razón de verosimilitudes. Recordemos que la hipótesis $\theta_1 = \ldots = \theta_b = 0$ se resuelve comparando $n \ln \lambda_1 = n \sum_{j=1}^{b} \ln(1 + t_j)$ con $\chi^{2,\alpha}_{p(r-1)}$.

Tiene validez asintótica y requiere de la hipótesis de normalidad. Si el resultado no es significativo, consideraremos únicamente las proyecciones sobre los s primeros ejes, es decir, las s primeras columnas de la matriz w. No obstante, en la mayoría de las ocasiones habremos de contentarnos con un análisis meramente descriptivo de los autovalores muestrales $t_1, \ldots, t_b$.

Nótese pues que, en definitiva y al igual que en el análisis de componentes principales, estamos hablando de un problema de reducción de la dimensión. No obstante, en componentes principales se eliminan ejes tales que las proyecciones sobre los mismos aportan poca variabilidad *intragrupo*; sin embargo, en el análisis discriminante, se eliminan ejes que aportan poca variabilidad *intergrupos*, en relación con la variabilidad intragrupo. Nótese también que si todos los autovalores salvo el primero son pequeños, la capacidad de discriminación intergrupos es prácticamente exclusividad del primer eje discriminante, lo cual nos aproxima a una situación univariante. En tales casos el test de Roy[5] se impone en la comparación de medias a los demás test considerados[6].

9.2. Análisis discriminate y correlación canónica

El análisis discriminate guarda una íntima relación con el análisis de correlación canónica. Recuérdese que un problema de comparación de r medias puede considerarse un problema de Regresión Lineal de Y respecto a un grupo de $r-1$ variables ficticias Z. En ese caso, tiene sentido hablar de los coeficientes de correlación canónica r_i^2, que se relacionan con los t_i's del manova de la siguiente forma:

$$r_i^2 = \frac{t_i}{1+t_i}, \quad i = 1, \ldots, b.$$

En ese caso, sabemos que se verifica

$$\begin{aligned} S_3 &= S_{yy} - S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y}, \\ S_2 &= S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y}, \end{aligned}$$

y que, por lo tanto, los autovectores de $S_{yy}^{-1} S_{y\mathbf{Z}} S_{\mathbf{ZZ}}^{-1} S_{\mathbf{Z}y}$ asociados a los autovalores $r_1^2, \ldots, r_b^2$ coinciden con los autovectores de $S_3^{-1} S_2$ asociados a los autovalores $t_1, \ldots, t_b$, respectivamente. Luego, los ejes discriminates $\langle a_1 \rangle, \ldots, \langle a_b \rangle$ coinciden con los que se obtienen en al análisis de correlación canónica con las variables ficticias como explicativas. Por ello, el primer eje discriminate es aquel sobre el cual hemos

[5] Recordemos: $\lambda_3 = t_1$.

[6] Wilks, Lawley-Hotelling y Pillay.

de proyectar para obtener una máxima correlación múltiple con las variables ficticias (que, recordemos, son las que determinan el grupo de pertenencia); el eje segundo aporta una máxima correlación múltiple entre todas aquellas proyecciones incorreladas con la primera, etc... En definitiva, los ejes discriminantes no son sino ejes canónicos y los autovalores se relacionan con los coeficientes de correlación canónica mediante la biyección creciente de $[0,1]$ en $[0,+\infty]$ definida mediante $f(\mathtt{x}) = (1-\mathtt{x})^{-1}\mathtt{x}$. Por lo tanto, todo lo dicho hasta ahora podría encuadrarse perfectamente en el análisis de correlación canónica.

Tener en cuenta que si $r_i^2 \simeq 1$, el i-ésimo eje tendrá un elevado poder de discriminación. De esta forma, r_i^2 expresa el poder de discriminación del i-ésimo eje principal, $i = 1, \ldots, b$. Por contra, una escasa correlación entre el eje i-ésimo y las variables ficticias supone una escasa capacidad discriminante de dicho eje. Por tanto, el coeficiente de correlación canónica i-ésimo puede considerarse una medida del poder de discriminación del i-ésimo eje discriminante, al igual que t_i, pero con la ventaja de que, mientras que t_i puede tomar en principio cualquier valor positivo, r_i^2 está acotado entre 0 y 1. De esta forma, pueden ser desechados aquellos ejes discriminantes que aporten un coeficiente de correlación canónica próximo a 0. Bajo la hipótesis de normalidad y con una muestra de gran tamaño, el estadístico de contraste (9.4) del test de la razón de verosimilitudes para contrastar una hipótesis inicial del tipo $H_0 : \theta_{s+1} = \ldots = \theta_b = 0$ puede expresarse mediante los coeficientes de correlación canónica según (6.5).

Esta idea de identificar la capacidad de discriminación con el grado de correlación con las variables ficticias será también de utilidad a la hora de eliminar variables observadas (lo veremos en la sección dedicada a la selección de variables).

9.3. Caso de dos grupos.

Aunque hemos obtenido un método general que engloba este caso, analicemos no obstante de manera aislada el caso $r = 2$, es decir, cuando se comparan dos grupos. Se tiene entonces que $b = 1$, es decir, existe un único autovalor y un único eje discriminante. Recordemos[7] que en tal caso el test consistía en contrastar el valor

$$T^2(Y) = \frac{\mathtt{n}_1\mathtt{n}_2}{\mathtt{n}_1+\mathtt{n}_2}(\overline{Y}_1 - \overline{Y}_2)' S_c^{-1} (\overline{Y}_1 - \overline{Y}_2)$$

[7] Cf. Aplicaciones del modelo lineal normal.

con el cuantil $T^{\alpha,2}_{p,\mathtt{n}-2}$. Si consideramos la proyección sobre un eje $\langle a\rangle$, el contraste de la igualdad entre las medias $\mu_1 a$ y $\mu_2 a$ se realiza[8] comparando el valor

$$T^2_{\langle a\rangle}(Y) = \frac{\mathtt{n}_1\mathtt{n}_2}{\mathtt{n}_1+\mathtt{n}_2}\frac{\left(a'\overline{Y}_1 - a'\overline{Y}_2\right)^2}{a'S_c a},$$

con el cuantil $F^{\alpha}_{1,\mathtt{n}-2}$. El estadístico $T^2_{\langle a\rangle}(Y)$ coincide, salvo una constante, con $F_{\langle a\rangle}(Y)$. Téngase en cuenta, además, que $F_{1,\mathtt{n}-2} = \left(t_{\mathtt{n}-1}\right)^2$. Es decir, el test equivale a comparar

$$t_{\langle a\rangle}(Y) = \frac{|a'\overline{Y}_1 - a'\overline{Y}_2|}{\sqrt{a'S_c a\left(\frac{1}{\mathtt{n}_1}+\frac{1}{\mathtt{n}_2}\right)}}$$

con el cuantil $t^{\alpha}_{\mathtt{n}-1}$ de la distribución t de Sudent. Existe, por compacidad, el máximo de $T^2_{\langle a\rangle}(Y)$ cuando a recorre $\mathbb{R}^p$. Puede comprobarse (cuestión propuesta) mediante el teorema 13.3 (o bien mediante el cálculo del autovector), que éste se alcanza en el eje

$$\langle S_c^{-1}(\overline{Y}_1 - \overline{Y}_2)\rangle.$$

Además,

$$\left[t_{\langle S_c^{-1}(\overline{Y}_1-\overline{Y}_2)\rangle}(Y)\right]^2 = T^2(Y).$$

Por lo tanto, al proyectar sobre el eje discriminante, se obtiene en el correspondiente estadístico t^2 (univariante) el mismo valor que corresponde al estadístico T^2 (multivariante). Es decir, que comparar las proyecciones sobre el eje discriminate es *similar* a comparar los datos en dimensión p.

9.4. Variables observadas y discriminación.

Proponemos a continuación distintas formas de evaluar la importancia de cada variable observada en la discriminación de los grupos. En ocasiones, puede suceder que la presencia de variables originales con escaso poder de discriminación entorpezca estudios posteriores, como puede ser la clasificación de observaciones, de ahí que sea interesante contar con criterios para optimizar nuestro modelo eliminando las variables originales inútiles. Esta selección de variables originales se une a la selección de ejes discriminantes efectuada en virtud de la magnitud de los coeficientes de correlación canónica asociados.

En primer lugar, podemos optar por la observación de las ponderaciones, es decir, de las componentes de los distintos vectores discriminantes. También contamos con

[8] Test de Student.

el análisis de la matriz de estructura, compuesta por las correlaciones entre los p vectores n-dimensionales observados (las columnas de la matriz de observaciones Y) y los b vectores de puntuaciones discriminantes (las columnas de la matriz de puntuaciones discriminantes w). Obviamente, tanto una ponderación como una correlación alta respecto a una variable observada indica una participación importante de dicha variable en la discriminación de los grupos. En ambos caso, el dato más importante es el que corresponde al primer eje discriminate, y la trascendencia de los demás ejes depende del tamaño de los respectivos coeficientes de correlación canónica.

No existe unanimidad acerca de cuál de estos dos métodos resulta más apropiado. No obstante, hemos de notar lo siguiente: supongamos que los ejes discriminates correspondientes a una matriz de datos Y son $\langle a_1\rangle, \ldots, \langle a_b\rangle$, asociados a los autovalores $t_1, \ldots, t_b$. Es decir, $S_3 = Y' P_{V^\perp} Y$ y $S_2 = Y' P_{V|W} Y$, $t_1, \ldots, t_b$ son lo autovalores de $S_3^{-1} S_2$ y $a_1, \ldots, a_b$ los respectivos autovectores. Consideremos un cambio de escala en las p componentes, es decir, una nueva matriz $\tilde{Y} = YD$, donde $D = \mathrm{diag}(d_1, \ldots, d_p)$, con $d_1, \ldots, d_p > 0$. Sean $\tilde{S}_2$ y $\tilde{S}_3$ las matrices $\tilde{Y} P_{V|W} \tilde{Y}$ e $\tilde{Y} P_{V^\perp} \tilde{Y}$, respectivamente, $\tilde{t}_1, \ldots, \tilde{t}_b$ los autovalores de $\tilde{S}_3^{-1} \tilde{S}_2$ y $\tilde{a}_1, \ldots, \tilde{a}_b$ los correspondientes autovectores. Se verifica entonces que

$$\tilde{S}_3 = DS_3D, \quad \tilde{S}_2 = DS_2D.$$

Por un argumento de invarianza se deduce

$$\tilde{t}_i = t_i, \quad \forall i = 1, \ldots b.$$

Por otra parte, la relación entre los autovectores es la siguiente:

$$\tilde{a}_i = D^{-1} a_i, \quad \forall i = 1, \ldots, b.$$

Efectivamente,

$$\tilde{S}_3^{-1} \tilde{S}_2 D^{-1} a_i = t_i D^{-1} a_i.$$

En consecuencia, no varían los autovalores pero sí lo ejes discriminantes. De esta forma, si se multiplicara por k la primera variable observada, la primera componente de cada eje discriminante quedaría dividida por k. Así, por ejemplo, si la primera variable es una medida expresada en metros, el hecho de pasarla a centímetros suponer dividir por 100 el peso de dicha variable en cada eje discriminante.

De todo ello se deduce la conveniencia de tipificar las variables si pretendemos ponderar el peso de cada una de ellas en la discriminación de los grupos. Es pues práctica habitual comenzar un análisis discriminate con la tipificación de todas las variables observadas (que no afecta a los autovalores $t_1, \ldots, t_b$). De esta forma, los

ejes discriminates tipificados $\tilde{a}_i, \ldots, \tilde{a}_b$ se relacionan con los no tipificados mediante

$$\tilde{a}_i = \begin{pmatrix} \sqrt{s_{11}} & & 0 \\ & \vdots & \\ 0 & & \sqrt{s_{pp}} \end{pmatrix} \cdot a_i, \quad i = 1, \ldots, b.$$

No obstante, ni la matriz de estructuras ni las puntuaciones discriminantes se ven afectada por la tipificación, pues las proyecciones de las vectores observados tipificados sobre los ejes discriminantes tipificados coinciden con las proyecciones de los vectores observados originales sobre los ejes discriminantes originales.

Sucede con frecuencia en la práctica que pueden deducirse implicaciones muy dispares respecto a las trascendencia de las variables originales en el problema de discriminación según analicemos la matriz de ponderaciones o la de cargas, de ahí que hayamos de ser muy prudentes a la hora de establecer conclusiones. No obstante, si nuestro propósito es suprimir variables que no intervengan de manera decisiva en la discriminación, es más frecuente, dado que el manova de una vía no es sino un problema de regresión multivariante respecto a veriables ficticias, hacer uso de algún algoritmo de selección de variables. Disponemos de varios, aunque sólo comentaremos el denominado método de la Lambda de Wilks.

Se trata de un método stepwise de selección de variables respuesta, considerando las r variables ficticias de asignación al grupo (recordemos que las variables ficticias toman valores 0 o 1 en función del grupo de pertenencia) como variables explicativas y las p variables observadas como respuesta. Se entiende pues que las variables originales más correlacionadas con las ficticias serán las que mejor discriminarán los grupos. Además, y con el fin de evitar multicolinealidad entre las variables observadas, podemos imponer la condición de que, si la entrada de una nueva variable en un paso determinado conlleva la presencia de un valor excesivamente bajo en la tolerancia de alguna de las variables en el nuevo modelo, la introducción de variables queda abortada. Puede demostrarse (cuestión propuesta) que la primera variable introducida por este método es la que maximiza el valor F de los p anovas posibles; que la segunda maximiza el valor λ_1 de Wilks de los $p-1$ manovas en dimensión 2 que puedan considerarse si se combina la variable ya introducida con cada una de las restantes, etcétera.

Cuestiones propuestas

1. Relaciona el primer eje discriminante con los intervalos de confianza obtenidos en la sección 2.5

2. Demostrar la igualdad (9.3).

3. Demostrar que

$$\begin{aligned} S_3 &= S_{yy} - S_{yz}S_{zz}^{-1}S_{zy}, \\ S_2 &= S_{yz}S_{zz}^{-1}S_{zy}, \end{aligned}$$

y que los autovectores de $S_{yy}^{-1}S_{yz}S_{xz}^{-1}S_{zy}$ asociados a los autovalores $r_1^2, \ldots, r_b^2$ coinciden con los autovectores de $S_3^{-1}S_2$ asociados a los autovalores $t_1, \ldots, t_b$, respectivamente.

4. Demostrar mediante el teorema 13.3 o mediante el cálculo de autovectores que el valor máximo de $T^2_{\langle a \rangle}(Y)$ cuando a recorre $\mathbb{R}^p$, se alcanza en el eje

$$\langle S_c^{-1}(\overline{Y}_1 - \overline{Y}_2) \rangle.$$

Además, dicho máximo vale $T^2(Y)$.

5. Clarifica la importancia de la tipificación a la hora de evaluar el peso de cada variable observada en la discriminación de los grupos.

6. ¿Qué ventaja presentan los coeficientes de correlación canónica a la hora de determinar el poder discriminante de los ejes?

7. ¿Han de ser perpendiculares los ejes discriminantes? Si no fuera así, qué condición habría que imponer a la matriz S_3 para que ello sucediese? ¿Cómo interpretar dicha condición en términos de las p variables observadas?

8. Demostrar que en el método de la Lambda de Wilks, la primera variable introducida es la que maximiza el valor F de los p anovas posibles; que la segunda maximiza el valor λ_1 de Wilks de los $p-1$ manovas en dimensión 2 que puedan considerarse si se combina la variable ya introducida con cada una de las restantes, etcétera.

9. Recordemos que, en el contraste de la media para el modelo lineal normal multivariante, $\theta_1, \ldots, \theta_b$ denotan los b primeros autovalores ordenados de la matriz simétrica $\nu_2\Sigma^{-1}\nu_2'$, que coinciden con los b primeros autovalores ordenados de $\Sigma^{-1}\delta$. La distribución del estadístico $(t_1, \ldots, t_b)$ depende de μ y Σ únicamente a través de $\theta_1, \ldots, \theta_b$, y la hipótesis inicial W se corresponde con el caso $\theta_1 = \ldots = \theta_b = 0$. Interpretar de la manera más clara posible los autovalores y autovectores de la matriz $\Sigma^{-1}\delta$ en relación con el problema de discriminación de r grupos.

10. Continuando con la cuestión anterior, interpreta de la forma más clara posible el significado del único autovalor poblacional θ que se obtiene en la comparación de dos medias.

11. ¿Qué trasformación lineal debemos aplicara los datos para que os ejes discriminantes sean perpendiculares?

Capítulo 10

Análisis discriminante II

La segunda parta del análisis discriminante está dedicada a la clasificación de observaciones. Consideremos p caracteres cuantitativos observados en r poblaciones distintas (es decir, r distribuciones p-dimensionales). Nuestro objetivo es asignar un individuo cualquiera (observación) a alguna de las r poblaciones (distribuciones) mediante el análisis del valor que toman los p caracteres cuantitativos en dicho individuo. Se trata pues de un problema de decisión que resolveremos mediante una estrategia no aleatoria. Estas estrategias pueden justificarse en virtud de distintos criterios. En la primera sección del capítulo demostramos que, bajo débiles condiciones de regularidad, distintos criterios coinciden en seleccionar la estrategia que asigna a cada observación el modelo probabilístico que la haga más probable o verosímil. Es decir, que en última instancia, se decide siguiendo el principio de máxima verosimilitud.

En este capítulo se estudia exclusivamente la clasificación partiendo inicialmente de modelos probabilísticos p-normales. Bajo este supuesto distinguiremos dos casos: el de matrices de covarianzas distintas, que dará lugar a regiones de clasificación cuadráticas, y el de matrices de covarianzas iguales, que dará lugar a una estrategia de tipo lineal atribuida al propio Fisher (se denomina método de clasificación de Fisher). Éstos estudios (en especial el último) se relacionan muy estrechamente con los ejes discriminantes estudiados en el capítulo anterior. Por supuesto que la aplicación de la técnica no debe supeditarse al estricto cumplimiento del supuesto de normalidad. Más bien hemos de intentar garantizar que el riesgo asociado a la misma es bajo, lo cual puede ser analizado a posteriori una vez diseñada. De hecho, en muchas ocasiones, el fracaso de la estrategia no se debe tanto a la violación de los supuestos como a la escasa diferencia entre las distribuciones consideradas. No obstante, si entendemos que el fracaso de la estrategia puede atribuirse al incumplimiento de la hipótesis de normalidad, contamos con otros métodos de clasificación alternativos,

que abordamos en las últimas secciones. Empezaremos estudiando el caso $r = 2$. Posteriormente generalizaremos al caso general.

10.1. Dos grupos: planteamiento general.

Consideremos dos probabilidades P_1 y P_2 sobre $(\mathbb{R}^p, \mathcal{R}^p)$ dominadas por la medida de Lebesgue, siendo p_1 y p_2 sus respectivas densidades. Nuestro problema consistirá en decidir por uno de los dos modelos probabilísticos posibles a partir de una observación $x \in \mathbb{R}^p$. Así pues, se trata del siguiente problema de decisión:

$$\Big\{\big(\mathbb{R}^p, \mathcal{R}^p, \{P_\theta : \theta \in \Theta\}\big), \Delta\Big\}, \qquad \Theta = \Delta = \{1, 2\}.$$

Para solucionar este problema nos restringiremos a estrategias no aleatorias, es decir, funciones medibles de $(\mathbb{R}^p, \mathcal{R}^p)$ en Δ. Éstas se identifican con los elementos de $\mathcal{R}^p$ de la siguiente forma: dada una estrategia no aleatoria S, se denota

$$\mathcal{S} = S^{-1}(\{1\}),$$

es decir, la región medible de $\mathbb{R}^p$ que determina la aceptación de P_1. En ocasiones, nos permitiremos el abuso de notación consistente en hablar de $\mathcal{S}$ como una estrategia no aleatoria. Consideraremos una función de pérdida

$$W : \Theta \times \Delta \longrightarrow [0, +\infty[$$

verificando

$$W(1|1) = W(2|2) = 0, \quad W(1|2), W(2|1) > 0.$$

Recordemos que, dada una estrategia S y una función de pérdida W, se define la función de riesgo asociada a S de la siguiente forma:

$$\begin{aligned} R_S(1) &= W(1|2)P_1(S = 2) \\ &= W(1|2)P_1(\mathcal{S}^{\mathsf{C}}) \\ &= W(1|2)\int_{\mathcal{S}^{\mathsf{C}}} p_1(x)\, dx. \\ R_S(2) &= W(2|1)P_2(S = 1) \\ &= W(2|1)P_2(\mathcal{S}) \\ &= W(2|1)\int_{\mathcal{S}} p_2(x)\, dx. \end{aligned}$$

En el problema de clasificación de una observación vamos a adoptar un tipo de estrategia que veremos a continuación, íntimamente ligado al principio de máxima

verosimilitud. El propósito de esta sección es justificar tales estrategias. Para ello, debemos considerar los tres criterios a seguir a la hora de seleccionar una estrategia:

(a) Se dice que $S_1 \succeq S_2$ cuando $R_{S_1}(\theta) \leq R_{S_2}(\theta)$ para $\theta = 1, 2$. La relación $\succeq$ es un preorden. Se dice que $S_1 \succ S_2$ cuando $S_1 \succeq S_2$ y alguna de las anteriores desigualdades es estricta. En ese sentido, se dice que una estrategia S es óptima cuando es maximal para el preorden $\succeq$. Este requisito peca, en general, de ambicioso. Se dice que S es admisible cuando no existe ninguna otra estrategia S^* tal que $S^* \succ S$. Una familia $\mathcal{F}$ de estrategias se dice completa cuando, para toda estrategia S^* fuera de $\mathcal{F}$, existe una estrategia $S \in \mathcal{F}$ tal que $S \succ S^*$. Así mismo, $\mathcal{F}$ se dice completa minimal cuando es completa y no posee ninguna subclase propia completa.

(b) Criterio minimax: se dice que una estrategia S_m es minimax cuando minimiza el máximo riesgo, es decir, verifica que

$$\text{máx}\{R_{S_m}(1), R_{S_m}(2)\} \leq \text{máx}\{R_S(1), R_S(2)\}, \qquad \forall S.$$

(c) Criterio bayesiano: cualquier distribución a priori Q sobre el espacio de parámetros
$\big(\{1,2\}, \mathcal{P}(\{1,2\})\big)$ se identifica con un número $q \in [0,1]$ de la forma $Q(\{1\}) = q$ (y, por tanto, $Q(\{2\}) = 1 - q$). Entonces, si suponemos que el espacio de parámetros está, efectivamente, dotado de una distribución a priori Q (es decir, $q \in [0,1]$), se define el riesgo de Bayes de la forma

$$\begin{aligned} R_S^q &= \int_\Theta R_S(\theta)\, dQ(\theta) = qR_S(1) + (1-q)R_S(2) \\ &= qW(1|2)\int_{\mathcal{S}^{\mathsf{c}}} p_1(x)\, dx + (1-q)W(2|1)\int_{\mathcal{S}} p_2(x)\, dx. \end{aligned}$$

Se dice que S es la estrategia q-Bayes cuando R_S^q es mínimo. Veamos cuál es la estrategia q-Bayes: dada una estrategia S, se tiene

$$\begin{aligned} R_S^q &= W(1|2)q\int_{\mathbb{R}^p} p_1(x)\, dx + \int_{\mathcal{S}} \big[W(2|1)(1-q)p_2(x) - W(1|2)qp_1(x)\big]\, dx = \\ &= W(1|2)q - \int_{\mathcal{S}} \big[W(1|2)qp_1(x) - W(2|1)(1-q)p_2(x)\big]\, dx. \end{aligned}$$

Entonces, R_S^q es mínimo y, por tanto, S es q-Bayes si, y sólo si, se verifica

$$\begin{aligned} \big\{x \in \mathbb{R}^p :\ W(1|2)qp_1(x) - W(2|1)(1-q)p_2(x) > 0\big\} &\subset \mathcal{S} \\ \big\{x \in \mathbb{R}^p :\ W(1|2)qp_1(x) - W(2|1)(1-q)p_2(x) < 0\big\} &\subset \mathcal{S}^{\mathsf{c}} \end{aligned}$$

De esta forma, una estrategia q-Bayes viene dada por la región

$$\mathcal{S}_q = \big\{x \in \mathbb{R}^p :\ W(1|2)qp_1(x) - W(2|1)(1-q)p_2(x) \geq 0\big\}.$$

Si se verifica la hipótesis

$$P_\theta\Big(W(1|2)qp_1(x) = W(2|1)(1-q)p_2(x)\Big) = 0, \quad \theta = 1, 2, \tag{10.1}$$

entonces la estrategia q-Bayes es esencialmente única, es decir, dos estrategias q-Bayes se diferenciarán en un suceso nulo. Además, la estrategia puede expresarse de esta forma, más intuitiva:

$$\mathcal{S}_q = \left\{x \in \mathbb{R}^p :\ \frac{p_1(x)}{p_2(x)} \geq \frac{1-q}{q} \cdot \frac{W(2|1)}{W(1|2)}\right\}.$$

Nótese que dicha estrategia concuerda con la idea de máxima verosimilitud, pues clasifica la observación x en el modelo 1 cuando $p_1(x)$ es grande en relación con $p_2(x)$. Este tipo de estrategias será el que utilicemos para resolver nuestro problema. Para justificar esta elección, demostraremos que, bajo ciertas condiciones de regularidad, la familia de las estrategias q-Bayes, donde q recorre $[0, 1]$, es completa minimal, y que toda estrategia minimax es de Bayes.

Teorema 10.1.
Consideremos una distribución a priori $q \in [0, 1]$ y S_q una estrategia q-Bayes. Supongamos que se verifica la hipótesis

$$P_1\big(\{p_2 = 0\}\big) = P_2\big(\{p_1 = 0\}\big). \tag{10.2}$$

Entonces S_q es admisible.

Demostración.
Consideremos una estrategia S tal que $S \succeq S_q$. Tenemos que demostrar que la desigualdad no puede ser estricta. Supongamos primeramente que $q \in (0, 1)$. Como S_q es q-Bayes, se verifica que $R^q_{S_q} \leq R^q_S$, es decir,

$$q[R_{S_q}(1) - R_S(1)] \leq (1-q)[R_S(2) - R_{S_q}(2)].$$

Luego, si $R_S(2) < R_{S_q}(2)$, tendríamos que $R_{S_q}(1) < R_S(1)$, lo cual es contradictorio. Por un razonamiento simétrico concluiríamos.
Si $q = 0$ y S_q es q-Bayes, se verifica necesariamente que

$$\mathcal{S}_q \subset \big\{x \in \mathbb{R}^p :\ W(1|2)qp_1(x) - W(2|1)(1-q)p_2(x) \geq 0\big\} = \big\{p_2 = 0\big\}.$$

Luego,

$$R_{S_q}(2) = \int_{\mathcal{S}_q} p_2(x)\, dx = 0.$$

Entonces, por hipótesis,

$$0 = R_S(2) = \int_{\mathcal{S}} p_2(x)\, dx,$$

lo cual equivale a afirmar que $\mathcal{S} \subset \{p_2 = 0\}$, es decir, $\{p_2 > 0\} \subset \mathcal{S}^{\mathsf{c}}$ que, como $P_1\big(\{p_2 = 0\}\big) = 0$ (es decir, $P_1\big(\{p_2 > 0\}\big) = 1$), implica $P_1\big(\mathcal{S}^C\big) = 1$, es decir,

$$\int_{\mathcal{S}^{\mathsf{c}}} p_1(x)\, dx = 1,$$

que será, desde luego, mayor que $\int_{\mathcal{S}_q^{\mathsf{c}}} p_1(x)\, dx$. Por tanto, $R_S(1) \geq R_{S_1}(1)$. Razonando simétricamente en el caso $q = 1$ se concluye.

∎

Una hipótesis más fuerte que (10.1) y (10.2) es la siguiente[1]

$$P_\theta\left(\frac{p_1(x)}{p_2(x)} = k\right) = 0, \qquad \forall k \in [0, +\infty], \quad \forall \theta = 1, 2. \tag{10.3}$$

Bajo esta hipótesis, si F_θ denota la función de distribución de la variable aleatoria p_1/p_2 respecto a la probabilidad P_θ, donde $\theta = 1, 2$, se verifica que F_θ es una biyección de $[0, +\infty]$ en $[0, 1]$ continua y estrictamente creciente.

Teorema 10.2.
Si se verifica (10.3) y S es una estrategia cualquiera, existe $q \in [0, 1]$ tal que la estrategia q-Bayes S_q verifica $S_q \succeq S$.

Demostración.
Dado que $P_1\left(\mathcal{S}^{\mathsf{c}}\right) \in [0, 1]$, existe un único número $k \in [0, +\infty]$ tal que

$$P_1\left(\mathcal{S}^{\mathsf{c}}\right) = F_1\left(k \frac{W(2|1)}{W(1|2)}\right).$$

Entonces, existe un único $q \in [0, 1]$ tal que $q = \frac{1}{1+k}$, es decir, $\frac{q-1}{q} = k$. Consideremos entonces S_q la estrategia q-Bayes (única, pues se verifica (10.1), y admisible), es decir,

$$\mathcal{S}_q = \left\{x \in \mathbb{R}^p : \ \frac{p_1(x)}{p_2(x)} \geq \frac{1-q}{q} \cdot \frac{W(2|1)}{W(1|2)}\right\} = \left\{x \in \mathbb{R}^p : \ \frac{p_1(x)}{p_2(x)} \geq k \frac{W(2|1)}{W(1|2)}\right\}.$$

[1] Se considera la siguiente aritmética: si $a > 0$, $\frac{a}{0} = +\infty$, y $\frac{0}{0} = 1$ No obstante, ambos casos son despreciables, pues sólo se verifican en un suceso nulo.

Entonces se verifica que

$$\begin{aligned} R_{S_q}(1) &= P_1\left(\mathcal{S}_q^{\mathsf{c}}\right) \\ &= P_1\left(\frac{p_1(x)}{p_2(x)} < k\frac{W(2|1)}{W(1|2)}\right) \\ &= F_1\left(k\frac{W(2|1)}{W(1|2)}\right) = R_S(1) \end{aligned}$$

Al ser S_q admisible, debe verificarse que $R_{S_q}(2) \leq R_S(2)$. Por tanto, $S_q \succeq S$.

∎

Corolario 10.3.

Bajo la hipótesis (10.3), se verifica que toda estrategia admisible es q-Bayes para algún $q \in [0,1]$.

Demostración.

Si S es admisible y S_q es la estrategia de Bayes en las condiciones del teorema anterior, se verifica que $R^q_{S_q} = R^q_S$. De (10.1) se sigue la unicidad de la estrategia Bayes y se concluye.

∎

Corolario 10.4.

Bajo la hipótesis (10.3), la familia $\mathcal{B} = \{S_q \colon q \in [0,1]\}$ de las estrategias de Bayes es completa minimal.

Demostración.

En virtud del teorema anterior, se tiene que si $S \notin \mathcal{B}$, entonces no es admisible. Por tanto, existe otra estrictamente mejor que ella que, según el teorema anterior, puede ser seleccionada entre las estrategias de Bayes. Por tanto, $\mathcal{B}$ es completa. Además, si excluimos cualquier estrategia S_q q-Bayes, con $q \in [0,1]$, al ser ésta admisible, no podremos encontrar ninguna estrictamente mejor en $\mathcal{B}\backslash\{S_q\}$. Luego, $\mathcal{B}$ es minimal.

∎

Por tanto, hemos demostrado que existe una fuerte concordancia entre los criterios $\succeq$ y Bayes. Relacionemos, por último, el criterio Bayes con el criterio minimax bajo la hipótesis (10.3). Consideremos las siguientes aplicaciones definidas sobre $[0,1]$

$$h_1(q) = F_1\left(\frac{1-q}{q} \cdot \frac{W(2|1)}{W(1|2)}\right)$$

$$h_2(q) = 1 - F_2\left(\frac{1-q}{q} \cdot \frac{W(2|1)}{W(1|2)}\right)$$

Se tiene pues que h_1 y h_2 son biyecciones continuas de $[0,1]$ en $[0,1]$, decreciente la primera y creciente la segunda. Por tanto, existe un único valor $q_m \in [0,1]$ tal que $h_1(q_m) = h_2(q_m)$.

Teorema 10.5.
Bajo la hipótesis (10.3), se verifica que la estrategia minimax es única y es la estrategia q_m-Bayes, es decir, aquella tal que $R_{S_{q_m}}(1) = R_{S_{q_m}}(2)$.

Demostración.
Se verifica, en general, que

$$\begin{aligned} h_1(q) &= P_1\left(\frac{p_1(x)}{p_2(x)} \geq \frac{1-q}{q}\cdot\frac{W(2|1)}{W(1|2)}\right) = R_{S_q}(1) \\ h_2(q) &= P_2\left(\frac{p_1(x)}{p_2(x)} < \frac{1-q}{q}\cdot\frac{W(2|1)}{W(1|2)}\right) = R_{S_q}(2), \end{aligned}$$

donde S_q es la estrategia q-Bayes. Luego, S_{q_m} verifica que $R_{S_{q_m}}(1) = R_{S_{q_m}}(2)$. Si S es otra estrategia tal que $\max\{R_S(1), R_S(2)\} < R_{S_{q_m}}(1)$, entonces S_{q_m} no sería admisible, lo cual es contradictorio, pues es de Bayes. Por lo tanto, S_{q_m} es una estrategia minimax.

Por otro lado, si S es minimax distinta de S_{q_m}, se verificará necesariamente

$$R_S(1) \neq R_{S_{q_m}}(1) \ \vee \ R_S(2) \neq R_{S_{q_m}}(2)$$

pues, de lo contrario, S sería q_m-Bayes y, por unicidad, $S = S_{q_m}$. Por tanto, si se verifica que $\max\{R_S(1), R_S(2)\} \leq R_{S_{q_m}}(1)$, entonces $\min\{R_S(1), R_S(2)\} < R_{S_{q_m}}(1)$ y S_{q_m} no sería admisible. Luego llegamos a una contradicción.

∎

En consecuencia, dada la consistencia con los demás criterios de selección, buscaremos estrategias de Bayes a la hora de solucionar nuestro problema de decisión: si suponemos conocida la probabilidad a priori q en el espacio de parámetros, consideraremos la estrategia S_q, que será admisible. Si no la suponemos conocida, buscaremos entre las estrategias de Bayes aquella que sea minimax. Nótese que las estrategias de Bayes son las de la forma

$$\mathcal{S} = \left\{ x :\in \mathbb{R}^p : \frac{p_1(x)}{p_2(x)} \geq k \right\}, \qquad k \in [0, +\infty],$$

relacionándose estrechamente con el principio de máxima verosimilitud.

10.2. Dos normales con covarianzas común

Supongamos que $P_1 = N_p(\mu_1, \Sigma)$ y $P_2 = N_p(\mu_2, \Sigma)$. En ese caso,

$$p_\theta(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu_\theta)'\Sigma^{-1}(x-\mu_\theta)\right\}, \qquad \theta = 1, 2.$$

Por tanto,

$$\frac{p_1(x)}{p_2(x)} = \exp\left\{-\frac{1}{2}\left[(x-\mu_1)'\Sigma^{-1}(x-\mu_1)' - (x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right]\right\}.$$

Si $k \in [0, +\infty]$, se verifica

$$\frac{p_1(x)}{p_2(x)} \geq k \Leftrightarrow \log \frac{p_1(x)}{p_2(x)} \geq \log k.$$

Dado que $\log k$ recorre $[-\infty, +\infty]$ cuando k recorre $[0, \infty]$, se deduce que las estrategias de Bayes serán de la forma

$$\mathcal{S}^a = \left\{x \in \mathbb{R}^p : \ -\frac{1}{2}\left[(x-\mu_1)'\Sigma^{-1}(x-\mu_1)' - (x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right] \geq a\right\},$$

donde $a \in [-\infty, +\infty]$. Nótese que una estrategia de este tipo consiste en compara las distancias de Mahalanobis

$$(x-\mu_\theta)'\Sigma^{-1}(x-\mu_\theta), \qquad \theta = 1, 2,$$

asignando la observación x al grupo tal que la distancia correspondiente sea menor (salvo la cte. a). Operando en la expresión de $\mathcal{S}^a$ se obtiene

$$\mathcal{S}^a = \left\{x \in \mathbb{R}^p : \ x'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) \geq a\right\}.$$

Hemos de tener en cuenta que, dado $k \in [0, +\infty]$, se verifica que

$$\frac{p_1(x)}{p_2(x)} = k \Leftrightarrow x'\Sigma^{-1}(\mu_1-\mu_2) = \frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) + \log k,$$

lo cual sucede, en todo caso, con probabilidad 0 [2]. Por tanto, se verifica la hipótesis (10.3). Luego, las estrategias de Bayes (que son únicas) se identifican con las admisibles y constituyen una familia completa minimal, y la estrategia minimax es la de

[2] Se trata de una subvariedad afín de dimensión $p-1$, cuya medida de Lebesgue p-dimensional es 0. Recordemos que P_1 y P_2 están dominadas por la medida de Lebesgue.

Bayes con ambos riesgos idénticos. En estos términos, S^a es la estrategia q-Bayes, donde

$$q = \frac{1}{1 + e^a \frac{W(1|2)}{W(2|1)}} \in [0, 1].$$

Recíprocamente, dado $q \in [0, 1]$, $S_q = S^a$, donde

$$a = \log \left\{ \frac{W(2|1)}{W(1|2)} \cdot \frac{1-q}{q} \right\} \in [-\infty, +\infty] \tag{10.4}$$

Nótese que, en el caso $W(1|2) = W(2|1) = 1$ y si la probabilidad a priori considerada es la uniforme, es decir, $Q(\{1\}) = Q(\{2\}) = 0{,}5$, entonces la estrategia de Bayes queda determinada por la región

$$\mathcal{S}_{0,5} = \left\{ x \in \mathbb{R}^p : \; x'\Sigma^{-1}(\mu_1 - \mu_2) \geq \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \right\},$$

lo cual equivale a asignar la observación al grupo cuya media minimice la distancia de Mahalanobis, es decir, al grupo que la haga más verosímil. Si no suponemos conocida ninguna probabilidad a priori, hemos de buscar la estrategia minimax. Recordemos que ésta es la estrategia S_q de Bayes donde tal que $R_{S_q}(1) = R_{S_q}(2)$ o, equivalentemente, $R_{S^a}(1) = R_{S^a}(2)$, donde a y q guardan la relación anteriormente expresada. Previamente debemos conocer las distribuciones de p_1/p_2 (o, equivalentemente, de $\log(p_1/p_2)$) respecto a P_1 y P_2. Consideremos pues la variable

$$\begin{aligned} U &= \log \frac{p_1}{p_2} \\ &= X'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2), \end{aligned}$$

que sigue un modelo de distribución 1-normal, y denótese

$$\Delta^2 = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2).$$

Se trata de una distancia del tipo Mahalanobis, que expresa el grado de proximidad entre los dos modelos probabilísticos. Se verifica entonces, trivialmente,

$$\mathtt{E}_1[U] = \frac{1}{2}\Delta^2.$$

Por otro lado, se verifica (cuestión propuesta)

$$\begin{aligned} \mathtt{var}_1[U] &= \mathtt{E}_1\left[(\mu_1 - \mu_2)'\Sigma^{-1}(X - \mu_1)(X - \mu_1)'\Sigma^{-1}(\mu_1 - \mu_2)\right] \\ &= \Delta^2 \end{aligned}$$

Luego,

$$P_1^U = N\left(\frac{1}{2}\Delta^2, \Delta^2\right).$$

Análogamente,

$$P_2^U = N\left(-\frac{1}{2}\Delta^2, \Delta^2\right).$$

Volviendo a la búsqueda de la solución minimax, debemos encontrar $a \in \mathbb{R}$ tal que $R_{S^a}(1) = R_{S^a}(2)$, donde

$$\mathcal{S}^a = \{U \geq a\}.$$

En ese caso, $R_{S^a}(1) = W(1|2)P_1(U < a)$ y $R_{S^a}(2) = W(2|1)P_2(U \geq a)$. Debe pues verificarse

$$W(1|2)\int_{-\infty}^{a} \frac{1}{\Delta\sqrt{2\pi}}\, \mathsf{e}^{-\frac{1}{2}\frac{\left(z-\frac{1}{2}\Delta^2\right)^2}{\Delta^2}}\, dz = W(2|1)\int_{a}^{+\infty} \frac{1}{\Delta\sqrt{2\pi}}\, \mathsf{e}^{-\frac{1}{2}\frac{\left(z+\frac{1}{2}\Delta^2\right)^2}{\Delta^2}}\, dz.$$

Mediante un cambio de variables (cuestión propuesta), si $f(y)$ denota la función de densidad de $N(0,1)$, se obtiene

$$W(1|2)\int_{-\infty}^{\frac{a-\frac{1}{2}\Delta^2}{\Delta}} f(y)\, dy = W(2|1)\int_{\frac{a+\frac{1}{2}\Delta^2}{\Delta}}^{+\infty} f(y)\, dy. \tag{10.5}$$

Si $W(1|2) = W(2|1)$, la solución es, por la simetría de f, $a = 0$, con lo cual se obtiene como estrategia minimax

$$S_m = \left\{x \in \mathbb{R}^p :\ x'\Sigma^{-1}(\mu_1 - \mu_2) \geq \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\right\}.$$

Coincide con la estrategia Bayes en el caso $q = 0{,}5$. En ese caso,

$$R_{S_m}(1) = R_{S_m}(2) = \int_{\frac{1}{2}\Delta}^{+\infty} f(y)\, dy. \tag{10.6}$$

Luego, obviamente,

$$\lim_{\Delta^2 \longrightarrow \infty} R_{S_m}(1) = \lim_{\Delta^2 \longrightarrow \infty} R_{S_m}(2) = 0.$$

Es decir, que cuanto más grande es la distancia entre los dos modelos probabilísticos, menos riesgo implica la clasificación respecto a ellos.

Todo lo expuesto hasta el momento en esta sección tiene un valor puramente teórico, pues las estrategias de clasificación propuestas requieren del conocimiento de

las medias y covarianza poblaciones. No obstante, marca el camino a seguir cuando los parámetros han de ser estimados.

Consideremos $X_{i1}, \ldots, X_{i\mathtt{n}_i}$ $i = 1, 2$, sendas muestras aleatorias simples e independientes de $N_p(\mu_i, \Sigma)$, respectivamente. Sea Y la variable observada, que sigue un modelo de distribución $N_p(\mu_1, \Sigma)$ o $N_p(\mu_2, \Sigma)$ (nuestro propósito es decidir al respecto) y es independiente de las muestras anteriores. Recordemos que, en el caso de parámetros conocidos, consideramos, a la hora de tomar la decisión, las estrategias de Bayes, que eran de la forma

$$S = \left\{ x \in \mathbb{R}^p : \ \frac{p_1(x)}{p_2(x)} > k \right\}.$$

Recuérdese que este tipo de estrategias concuerdan con el principio de máxima verosimilitud. En el caso de que los parámetros poblacionales sean desconocidos, podemos determinar la estrategia siguiendo el principio de sustitución, es decir, reemplazando los valores poblaciones desconocidos por estimadores de los mismos. Veremos, sin embargo, que el principio de máxima verosimilitud conduce a la misma estrategia que el de sustitución:

Si $i = 1, 2$, consideremos los siguientes parámetros muestrales.

$$\begin{aligned} \overline{X}_i &= \frac{1}{\mathtt{n}_i} \sum_{j=1}^{\mathtt{n}_i} X_{ij}, \\ S_i &= \frac{1}{\mathtt{n}_i} \sum_{j=1}^{\mathtt{n}_i} (X_{ij} - \overline{X}_i)(X_{ij} - \overline{X}_i)', \\ S_c &= \frac{\mathtt{n}_1 S_1 + \mathtt{n}_2 S_2}{\mathtt{n}_1 + \mathtt{n}_2}. \end{aligned}$$

Si consideramos ambas muestras por separado, tenemos sendos modelos lineales normales con dimensión $r = 1$. $\overline{X}_i$ y S_i son los EMV de μ_i y Σ, respectivamente, para $i = 1, 2$. Si consideramos conjuntamente ambas muestras tendremos un modelo lineal normal con dimensión $r = 2$. En ese caso, S_c es el EMV de Σ.

Supongamos que la observación Y corresponde al modelo de distribución $N_p(\mu_1, \Sigma)$ En ese caso, podemos agregarla a la muestra 1. No obstante, si consideramos conjuntamente ambas muestras obtenemos un nuevo modelo lineal normal de dimensión $r = 2$ con $\mathtt{n}_1 + \mathtt{n}_2 + 1$ datos y función de densidad $p^1_{\mu_1,\mu_2,\Sigma}$ [3]. Los nuevos EMV para

[3] Función de $\mathcal{M}_{(\mathtt{n}_1+1+\mathtt{n}_2)\times p}$ en $\mathbb{R}^+$.

μ_1, μ_2 y Σ son:

$$\begin{aligned}
\hat{\mu}_1^1 &= \frac{\mathtt{n}_1\overline{X}_1 + Y}{\mathtt{n}_1+1},\\
\hat{\mu}_2^1 &= \overline{X}_2,\\
\hat{\Sigma}^1 &= \frac{1}{\mathtt{n}_1+\mathtt{n}_2+1}[\sum_{j=1}^{\mathtt{n}_1}(X_{1j}-\hat{\mu}_1^1)(X_{1j}-\hat{\mu}_1^1)' + (Y-\hat{\mu}_1^1)(Y-\hat{\mu}_1^1)'\\
&+ \sum_{j=1}^{\mathtt{n}_2}(X_{2j}-\hat{\mu}_2^1)(X_{2j}-\hat{\mu}_2^1)']\\
&= \frac{\mathtt{n}_1+\mathtt{n}_2}{\mathtt{n}_1+\mathtt{n}_2+1}S_c + \frac{\mathtt{n}_1}{(\mathtt{n}_1+1)(\mathtt{n}_1+\mathtt{n}_2+1)}(Y-\overline{X}_1)(Y-\overline{X}_1)'.
\end{aligned}$$

De manera simétrica se obtienen la función de densidad $p^2_{\mu_1,\mu_2,\Sigma}$ y los los EMV $\hat{\mu}_1^2$, $\hat{\mu}_2^2$ y $\hat{\Sigma}^2$, en el caso de que Y corresponda al modelo de distribución $N_p(\mu_2,\Sigma)$. El cociente entre las máximas verosimilitudes es el siguiente:

$$RV = \frac{\sup_{\mu_1,\mu_2,\Sigma} p^1_{\mu_1,\mu_2,\Sigma}\big(x_{11},\ldots,x_{2\mathtt{n}_2},y\big)}{\sup_{\mu_1,\mu_2,\Sigma} p^2_{\mu_1,\mu_2,\Sigma}\big(x_{11},\ldots,x_{2\mathtt{n}_2},y\big)}.$$

Un valor grande del mismo nos invita a aceptar el primer modelo, y un valor pequeño, el segundo. Los máximos se alcanzan[4] con $\left(\hat{\mu}_1^1,\hat{\mu}_2^1,\hat{\Sigma}^1\right)$ y $\left(\hat{\mu}_1^2,\hat{\mu}_2^2,\hat{\Sigma}^2\right)$, respectivamente, siendo entonces

$$\begin{aligned}
RV &= \left(\frac{\left|\hat{\Sigma}^2\right|}{\left|\hat{\Sigma}^1\right|}\right)^{\frac{\mathtt{n}_1+\mathtt{n}_2+1}{2}} = \left(\frac{\left|(\mathtt{n}_1+\mathtt{n}_2)S_c + \frac{\mathtt{n}_2}{\mathtt{n}_2+1}(Y-\overline{X}_2)(Y-\overline{X}_2)'\right|}{\left|(\mathtt{n}_1+\mathtt{n}_2)S_c + \frac{\mathtt{n}_1}{\mathtt{n}_1+1}(Y-\overline{X}_1)(Y-\overline{X}_1)'\right|}\right)^{\frac{\mathtt{n}_1+\mathtt{n}_2+1}{2}}\\
&= \left(\frac{1+\frac{\mathtt{n}_2}{(\mathtt{n}_2+1)(\mathtt{n}_1+\mathtt{n}_2)}(Y-\overline{X}_2)'S_c^{-1}(Y-\overline{X}_2)}{1+\frac{\mathtt{n}_1}{(\mathtt{n}_1+1)(\mathtt{n}_1+\mathtt{n}_2)}(Y-\overline{X}_1)'S_c^{-1}(Y-\overline{X}_1)}\right)^{\frac{\mathtt{n}_1+\mathtt{n}_2+1}{2}}
\end{aligned}$$

La última igualdad se deduce fácilmente del lema 2.6. Dado que una función del tipo $g(x) = (1+bx)^m$ es una biyección creciente, se trata pues de aceptar el modelo 1 cuando el cociente

$$\frac{(Y-\overline{X}_2)'S_c^{-1}(Y-\overline{X}_2)}{(Y-\overline{X}_1)'S_c^{-1}(Y-\overline{X}_1)}$$

toma un valor grande, lo cual equivale a afirmar que que

$$Y'S_c^{-1}(\overline{X}_1-\overline{X}_2) - \frac{1}{2}(\overline{X}_1+\overline{X}_2)'S_c^{-1}(\overline{X}_1-\overline{X}_2)$$

[4]Cf. EMV y test de razón de verosimilitudes en el modelo lineal normal.

sea grande. Así pues, el principio de máxima verosimilitud nos conduce a considerar estrategias del tipo

$$\mathcal{S}^a = \big\{x \in \mathbb{R}^p :\ Y' S_c^{-1}(\overline{X}_1 - \overline{X}_2) - \frac{1}{2}(\overline{X}_1 + \overline{X}_2)' S_c^{-1}(\overline{X}_1 - \overline{X}_2) > a\big\}, \qquad a \in \mathbb{R}.$$

Nótese la similitud con el caso de parámetros conocidos. En esta ocasión, hemos sustituido dichos parámetros por sus respectivas estimaciones. Así pues, los principios de máxima verosimilitud y de sustitución conducen al mismo tipo de estrategia. El valor del parámetro a dependerá de la función de pérdida y de la probabilidad a priori considerada mediante la ecuación (10.4). Así, en el caso $q = 0{,}5$ y $W(1|2) = W(2|1)$, a vale 0. Esta estrategia equivale a asignar la observación a la distribución que minimice (salvo la constante a) la distancia de Mahalanobis con la observación

$$(Y - \overline{X}_i)' S_c^{-1}(Y - \overline{X}_i), \qquad i = 1, 2.$$

Por analogía con la sección anterior y con el objeto de calcular los riesgos asociados, podemos definir la variable

$$V_{\mathtt{n}_1,\mathtt{n}_2} = Y' S_c^{-1}(\overline{X}_1 - \overline{X}_2) - \frac{1}{2}(\overline{X}_1 + \overline{X}_2)' S_c^{-1}(\overline{X}_1 - \overline{X}_2).$$

Se verifica[5] entonces la siguiente convergencia en probabilidad, con P_1 y P_2,

$$\lim_{\mathtt{n}_1 \to \infty, \mathtt{n}_2 \to \infty} (V_{\mathtt{n}_1,\mathtt{n}_2} - U) = 0.$$

Por tanto, se da también la convergencia en distribución, con P_1 y P_2. En consecuencia, en el caso $W(1|2) = W(2|1)$ la estrategia

$$\mathcal{S} = \left\{x \in \mathbb{R}^p :\ Y' S_c^{-1}(\overline{X}_1 - \overline{X}_2) > \frac{1}{2}(\overline{X}_1 + \overline{X}_2)' S_c^{-1}(\overline{X}_1 - \overline{X}_2)\right\}$$

tiene los mismos riesgos asintóticos que las correspondientes estrategias de Bayes para el caso de parámetros conocidos. Concrétamente, en el caso $a = 0$, los riesgos asintóticos son los siguientes

$$R_S(1) = R_S(2) = \int_{\frac{1}{2}\Delta}^{+\infty} f(y)\, dy.$$

[5] Se está aplicando la ley débil de los grandes números para μ_1, μ_2 y Σ. Tener en cuenta que el determinante es una función continua y la expresión de la inversa de una matriz por adjuntos para garantizar la convergencia en probabilidad de S^{-1} a Σ^{-1}. Tener en cuenta también que, para todo $\varepsilon > 0$, existe $M > 0$ tal que $P_\theta(\|Y\| > 0) < \varepsilon$, $\theta = 1, 2$.

Desde el punto de vista Bayesiano, si consideramos que existe una probabilidad a priori en el espacio de parámetros q y que la probabilidad de que los datos de las dos muestras se obtienen conforme a dicha distribución, podemos estimar q mediante

$$\hat{q} = \frac{\mathtt{n}_1}{\mathtt{n}_1 + \mathtt{n}_2}$$

De esa forma y teniendo en cuenta $W(1|2)$ y $W(2|1)$, podemos calcular el valor a correspondiente a la estrategia.

Por último, supongamos que hemos obtenido ya los valores de las dos muestras. Nótese entonces que estas estrategias consisten básicamente en, dado un valor observado Y, calcular el número real

$$\left(Y - \frac{1}{2}(\overline{X}_1 + \overline{X}_2)\right)' S_c^{-1}(\overline{X}_1 - \overline{X}_2)$$

y ver si es grande o pequeño. En el caso q=0.5 con pérdidas iguales se comparará con 0. Si se considera una probabilidad a priori o pérdidas alternativas, se comparará con otro número. En definitiva, estamos comparando las proyecciones de Y y $\frac{1}{2}(\overline{X}_1 + \overline{X}_2)$ sobre el eje discriminate[6]

$$\langle S_c^{-1}(\overline{X}_1 - \overline{X}_2)\rangle.$$

De esta forma, la estrategia considerada asigna la observación Y al modelo 1 cuando la distancia de la proyección de ésta sobre el eje discriminate con la proyección de $\overline{X}_1$ sobre dicho eje es menor que la distancia de la proyección de Y con la proyección de $\overline{X}_2$.

10.3. Caso general: r distribuciones p-normales

En este caso, debemos clasificar una observación Y respecto a r distribuciones de probabilidad. Tenemos pues r posibles decisiones. Mediante un desarrollo análogos al caso de dos probabilidades, aunque lógicamente más laborioso, y si se consideran iguales todos los costes de clasificación errónea, se obtiene[7] que las estrategias de Bayes son aquéllas que asignan la observación x al modelo P_i cuando

$$\frac{p_i(x)}{p_j(x)} > \frac{q_j}{q_i}, \quad \forall j \neq i,$$

[6] Es único en este caso, pues $b = 1$.

[7] No vamos a desarrollar aquí el mismo proceso. El lector interesado puede encontrarlo en Anderson (1958), sec. 6.6.

lo cual concuerda nuevamente con el principio de máxima verosimilitud, especialmente si se suponen iguales todas las probabilidades a priori, en cuyo caso estaremos hablando de la estrategia minimax. Se prueba, igualmente, que las estrategias Bayes constituyen una familia completa minimal.

Caso de que los modelos probabilísticos considerados sean p-normales con matriz de covarianzas común, las estrategias a considerar consistirán en asignar la observación X al modelo i-ésimo cuando se verifique

$$(X-\mu_i)'\Sigma^{-1}(X-\mu_i) < (X-\mu_j)'\Sigma^{-1}(X-\mu_j) + a_{ij}, \qquad \forall j \neq i,$$

donde a_{ij} dependerá de las probabilidades a priori y función de pérdida consideradas. Si son todas idénticas se tiene $a_{ij} = 0$, lo cual encaja perfectamente con el principio de máxima verosimilitud. Tener en cuenta que los términos anteriores son las distancias de Mahalanobis entre la observación y la media de cada distribución. De esta forma, la estrategia consiste en asignar la observación a la distribución más *próxima*. La anterior expresión es equivalente a la siguiente:

$$X'\Sigma^{-1}(\mu_i-\mu_j) - \frac{1}{2}(\mu_i+\mu_j)'\Sigma^{-1}(\mu_i-\mu_j) > a_{ij}, \quad \forall j \neq i.$$

Pueden calcularse los riesgos implícitos a estas estrategias de manera totalmente análoga al caso $r = 2$. No obstante, en la práctica los parámetros $\mu_1, \ldots, \mu_r$ y $(n-r)\Sigma$ serán desconocidos y habremos de estimarlos mediante $\overline{X}_1, \ldots, \overline{X}_r$ y S_3, respectivamente. Análogamente al caso $r = 2$, consideraremos estrategias consistentes en asignar la observación Y al grupo i-ésimo cuando

$$Y'S_3^{-1}(\overline{X}_i - \overline{X}_j) - \frac{1}{2}(\overline{X}_i + \overline{X}_j)S_3^{-1}(\overline{X}_i - \overline{X}_j) > a_{ij}.$$

Si suponemos iguales probabilidades a priori, tendremos $a_{ij} = 0$. Como en el caso $r = 2$, pueden calcularse los riesgos asintóticos, que coinciden con los que se obtienen cuando los parámetros son conocidos. Además, por un razonamiento inverso al que realizamos anteriormente, la expresión anterior equivale (suponemos $a_{ij} = 0$) a la siguiente

$$(Y-\overline{X}_i)'S_3^{-1}(Y-\overline{X}_i) < (Y-\overline{X}_j)'S_3^{-1}(Y-\overline{X}_j), \qquad \forall j \neq i,$$

es decir, se asigna la observación al grupo del que dista[8] menos.

[8] Estamos hablando, de nuevo, de una distancia del tipo Mahalanobis (elíptica) igual a la anterior, salvo que esta se basa en los parámetros muestrales.

10.4. Relación con los ejes discriminantes.

En este sección se establece claramente la relación entre los dos capítulos dedicados al análisis discriminante. En el caso $r = 2$ quedó patente la relación entre el problema de clasificación y la proyección sobre el eje discriminante (único). Lo mismo sucederá en el caso general, donde tendremos b[9] ejes discriminantes, $\langle a_1\rangle, \ldots, \langle a_b\rangle$, asociados a $t_1, \ldots, t_b$, resp., que son los autovalores positivos de $S_3^{-1}S_2$. Si $p > b$, el proceso puede completarse con los $p - b$ autovalores nulos de $S_3^{-1}S_2$, obteniendo así nuevos ejes $\langle a_{b+1}\rangle, \ldots, \langle a_p\rangle$, de manera que

$$a_i' S_3 a_i = 1, \qquad a_i' S_2 a_i = t_i = 0, \qquad i = b+1, \ldots, p.$$

Consideremos entonces la matriz $p \times p$ invertible $A = (a_1, \ldots, a_p)$, que verifica, según la construcción de los ejes discriminante,

$$A' S_3 A = \mathtt{Id}, \qquad A' S_2 A = \left(\begin{array}{ccc|ccc} t_1 & & 0 & 0 & & 0 \\ & \ddots & & & \ddots & \\ 0 & & t_b & 0 & & 0 \\ \hline 0 & & 0 & 0 & & 0 \\ & \ddots & & & \ddots & \\ 0 & & 0 & 0 & & 0 \end{array}\right).$$

En el estudio de clasificación de una observación $Y \in \mathbb{R}^p$ se parte de r muestras aleatoria simples independientes $X_{j1}, \ldots, X_{j\mathtt{n}_j}$, $j = 1, \ldots, r$, siendo $\overline{X}_j$ la media aritmética correspondiente a la muestra j-ésima. Consideremos el vectores W y W_{jk}, donde $j = 1, \ldots, r$ y $k = 1, \ldots, \mathtt{n}_j$, definidos mediante

$$\begin{aligned} W &= A'Y \\ W_{jk} &= A'X_{jk} \end{aligned}$$

Estamos pues proyectando los datos originales sobre los p ejes discriminantes, lo cual no es sino un cambio de la base, es decir, $W, W_{11}, \ldots, W_{r\mathtt{n}_r}$ son, respectivamente, las coordenadas de los vectores $Y, X_{11}, \ldots, X_{r\mathtt{n}_r}$ respecto a la base $(A^{-1})'$ [10] (si $S_3 = \mathtt{Id}$, estaremos hablando de una base ortonormal A). De la misma forma pueden proyectarse tanto las medias aritméticas de los distintos grupos, $\overline{X}_j$, $j =, \ldots, r$, como

[9] $b = \text{mín}\{p, r-1\}$.

[10] Nótese que los vectores fila W_{jk}' configuran la matriz de puntuaciones discriminantes que estudiáramos en el capítulo anterior.

la media aritmética de todos los datos, $\overline{X}$, obteniéndose respectivamente

$$\begin{aligned}\overline{W}_j &= A'\overline{X}_j = \frac{1}{\mathtt{n}_j}\sum_{k=1}^{\mathtt{n}_j} W_{jk},\\ \overline{W} &= A'\overline{X} = \frac{1}{\sum_{j=1}^{r}\mathtt{n}_j}\sum_{j=1}^{r}\sum_{k=1}^{\mathtt{n}_j} W_{jk}.\end{aligned}$$

Descompongamos los vectores resultantes mediante

$$W = \begin{pmatrix} W^1 \\ \vdots \\ W^p \end{pmatrix}, \quad \overline{W}_j = \begin{pmatrix} \overline{W}_j^1 \\ \vdots \\ \overline{W}_j^p \end{pmatrix} \quad \overline{W} = \begin{pmatrix} \overline{W}^1 \\ \vdots \\ \overline{W}^p \end{pmatrix}.$$

Entonces, para todo $i = 1, \ldots, p$, se verifica (cuestión propuesta)

$$\sum_{j=1}^{r} \mathtt{n}_j \left(\overline{W}_j^i - \overline{W}^i\right)^2 = t_i.$$

Por tanto, si $i > b$, y $j, h \in \{1, \ldots, r\}$, se verifica que

$$\overline{W}_j^i = \overline{W}_h^i. \tag{10.7}$$

Además, si $i \leq b$ pero t_i es pequeño, entonces

$$\overline{W}_j^i \simeq \overline{W}_h^i. \tag{10.8}$$

Recordemos que la estrategia a seguir en el problema de clasificación se reduce a seleccionar el grupo que minimice (salvo una constante) las distancias de Mahalanobis

$$(Y - \overline{X}_j)' S_3^{-1} (Y - \overline{X}_j), \quad j = 1, \ldots, r. \tag{10.9}$$

Dado que $Y = (A')^{-1}W$ y $\overline{X}_j = (A')^{-1}\overline{W}_j$, la expresión anterior equivale a la siguiente

$$\|W - \overline{W}_j\|^2, \quad j = 1, \ldots, r. \tag{10.10}$$

Por tanto, la estrategia consiste en asignar la observación Y al grupo que minimice la distancia euclídea anterior. Tener en cuenta que éstas se obtiene de la forma

$$\|W - \overline{W}_j\|^2 = \left(W^1 - \overline{W}_j^1\right)^2 + \ldots + \left(W^p - \overline{W}_j^p\right)^2, \quad j = 1, \ldots, r.$$

En definitiva, proyectando sobre los ejes discriminantes hemos realizamos un cambio de base que transforma las distancias de Mahalanobis (10.9) en las distancias euclídeas

(10.10). En estos términos, el problema de clasificación se reduce a minimizar estas últimas, es decir, se trata de buscar el grupo cuya media $\overline{W}_j$ esté más próxima (en el sentido usual) a la observación W. Las $p-b$ últimas coordenadas no influyen a la hora de buscar el mínimo, por lo que son eliminadas, lo cual puede suponer una primera reducción en la dimensión[11]. Además, teniendo en cuenta (10.8) y con el objeto de reducir aún más la dimensión, se puede prescindir de las coordenadas correspondientes a ejes asociados a autovalores pequeños[12] (los últimos). En la práctica, es raro encontrar más de dos autovalores *grandes*, con lo cual la clasificación queda fundamentalmente reducida a proyectar la observación en el plano que determinan los dos primeros ejes discriminantes y determinar el grupo cuyo centroide (media aritmética) proyectado diste menos. Estas proyecciones pueden ser representadas mediante el gráfico denominado mapa territorial.

Así pues, el fin último del análisis discriminante I (construcción de los ejes discriminantes) es conseguir una reducción de la dimensión en el problema de clasificación de una observación, de manera que, a ser posible, podamos tener una visión gráfica del mismo.

Veamos qué sucede si aplicamos una análisis dicriminante a los datos del archivo `irisdata` de Fisher. Se supone que nuestro propósito es discriminar o distinguir las tres especies de flores en función de las cuatro variables medidas (longitud y anchura de sépalos y pétalos). En este caso, $b = \text{mín}\{4, 3-1\} = 2$, es decir, el manova presenta 2 autovalores, asociados a sendos coeficientes de correlación canónica. Concretamente $t_1 = 32{,}192$ con $r_1 = 0{,}985$ y $t_2 = 0{,}285$ con $r_2 = 0{,}471$. Así pues, la representación de los datos en el plano discriminante no supone en este caso pérdida alguna de información en lo referente a la seperación entre grupos.

El resultado del manova es, por supuesto, significativo, es decir, el primer autovalor es significativo. A pesar de la desproporción existente entre los dos autovalores, el segundo también resulta ser significativo, es decir, el segundo eje discriminante también posee cierta capacidad de discriminación, aunque despreciable respecto a la del primero. Las matrices de estructura y ponderaciones son, respectivamente, las siguientes:

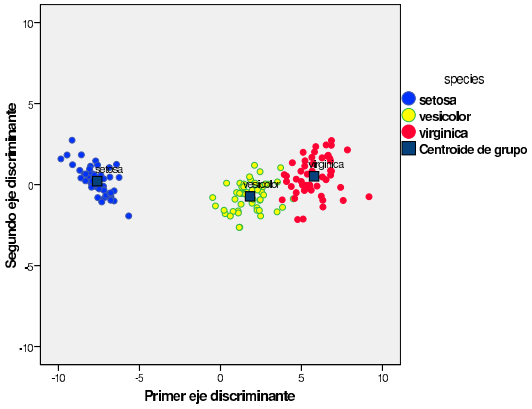
	w_1	w_2
long-sep	0.223	0.311
anch-sep	-0.119	0.864
long-pet	0.706	0.168
anch-pet	0.633	0.737

	$\langle a_1 \rangle$	$\langle a_2 \rangle$
long-sep	-0.427	0.012
anch-sep	-0.521	0.735
long-pet	0.947	-0.401
anch-pet	0.575	0.581

[11] Sólo valida bajo la hipótesis de igualdad de las matrices de covarianzas.

[12] Ya hemos estudiado, en el anterior capítulo, un test de significación de autovalores.

En ambos casos podemos despreciar la segunda columna pues el 99 % de la capacidad de discriminación recae sobre el primer eje discriminante. Del análisis de la primera columna en la matriz de estructura se deduce que los parámetros del pétalo son los que más correlacionan con la primera puntuación discriminante, lo que nos podría llevar a pensar que la diferencia entre las especies radica únicamente en el pétalo, pero esta conclusión es muy aventurada. De hecho, en la matriz de ponderaciones se otorga un peso similar a las variables del sépalo en el primer eje. La proyección de los datos sobre el plano discriminante es la siguiente:



Se observa claramente una perfecta separación de las especies, muy especialmente de setosa, lo cual favorece sin duda una buena clasificación. La estrategia que proponemos, denomina de Fisher, consiste pues en lo siguiente: dada una flor de especie desconocida, se procederá a medir las cuatro variables y proyectar el punto de $\mathbb{R}^4$ resultante sobre el plano anterior, de manera que se le asignará la especie cuyo centroide quede más próximo según la distancia euclíidea.

10.5. Caso de matriz de covarianzas distintas

Hasta ahora hemos supuesto en todo momento la igualdad de las matrices de covarianzas. Dicha hipótesis puede contrastarse, bajo el supuesto de p-normalidad, mediante el test M de Box. Consideremos el caso más general de clasificación respecto a las distribuciones de probabilidad $N_p(\mu_j, \Sigma_j)$, $j = 1, \ldots, r$. Según el criterio de

Bayes, una observación x se asignará al grupo i-ésimo cuando se verifique

$$\frac{1}{2}(x-\mu_j)'\Sigma_j^{-1}(x-\mu_j)-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)+a_{ij}>0,\quad \forall j\neq i.$$

donde la cte a_{ij} depende de las probabilidades a priori y de los costes considerados, siendo cero cuando las unas y los otros son iguales. Desarrollando la anterior expresión, se tiene

$$\frac{1}{2}x'\Sigma_j^{-1}x-\frac{1}{2}x'\Sigma_i^{-1}x-x'\left(\Sigma_j^{-1}\mu_j-\Sigma_i^{-1}\mu_i\right)+b_{ij}>0,\quad \forall j\neq i.$$

Por tanto, las regiones de clasificación quedarán determinadas por formas cuadráticas. Cuando las matrices de covarianzas sean idénticas, los dos primeros términos se anulan y tendremos (como ya sabemos) regiones de tipo lineal (delimitadas por semirrectas).

Cuando los parámetros poblacionales sean desconocidos se procederá, análogamente a los casos anteriores, a sustituirlos por sus respectivos estimadores, de tal forma que la expresión anterior se transforma en la siguiente:

$$\frac{1}{2}Y'S_j^{-1}Y-\frac{1}{2}Y'S_i^{-1}Y-Y'\left(S_j^{-1}\overline{X}_j-S_i^{-1}\overline{X}_i\right)+b_{ij}>0,,\quad \forall j\neq i,$$

donde S_k, $k=1,\ldots,r$ es, respectivamente, el EIMV de Σ_k, es decir,

$$S_k=\frac{1}{\mathtt{n}_k-1}\sum_{j=1}^{\mathtt{n}_k}(X_{kj}-\overline{X}_k)(X_{kj}-\overline{X}_k)',\quad k=1,\ldots,r.$$

De esta forma, invirtiendo el proceso se deduce que las estrategias consideradas consisten en asignar la observación Y al grupo i que minimice (salvo cte.) la distancia de Mahalanobis

$$(Y-\overline{X}_i)'S_i^{-1}(Y-\overline{X}_i),\quad i=1,\ldots,r.$$

Aunque en rigor no procede en este caso, el program SPSS resuelve el problema mediante las puntuaciones discriminantes. A menos que p sea menor que r, ello supondrá con toda seguridad un error (cuya cuantía dependerá de la diferencia entre las matrices de covarianzas), dado que (10.7) no se verifica y puede que (10.8) tampoco. De hecho, así queda advertido en la propia salida del programa.

10.6. Validez de la estrategia.

En principio, el método de clasificación lineal es únicamente aplicable bajo las condiciones de normalidad e igualdad de las matrices de covarianzas. Estos supuestos difícilmente se dan en la realidad. Sin embargo, el método lineal de clasificación

de Fisher manifiesta ser bastante robusto frente a violaciones moderadas de estas hipótesis. Por ello, en el caso (frecuente) de que no se verifiquen las mismas, no debemos descartar la estrategia lineal sino que hemos de ejecutarla y evaluar su validez a posteriori. Si los resultados no son satisfactorios, optaremos por otro tipo de clasificación, como la cuadrática. Igualmente, la estrategia cuadrática tiene plena validez bajo la hipótesis de normalidad (aunque las matrices de covarianzas no sean iguales), lo cual no quiere decir que deba ser desechado cuando ésta no se verifica, sino que debe evaluarse también a posteriori antes de buscar estrategias alternativas. Aunque el método cuadrático (distintas matrices de covarianzas) exige menos condiciones de validez que el lineal (misma matriz de covarianzas), merece la pena comparar en todo caso la validez de ambos y elegir el más satisfactorio.

Así pues, hemos de resolver dos cuestiones: primero, ¿cómo podemos evaluar a posteriori la validez de la estrategia de clasificación? Y segundo, ¿con qué métodos alternativos al lineal y cuadrático contamos? La respuesta a la segunda cuestión intentaremos darla en las próximas secciones. Respecto a la primera, la forma de evaluar la validez de una estrategia de clasificación es estimando la probabilidad de clasificación correcta. La manera más simple de hacerlo es considerar la muestra de datos que se ha utilizado para construir la propia estrategia, cuya ubicación conocemos de antemano, y reclasificarlos en función de la misma. La proporción de datos reubicados en su grupo original es una estimación de la probabilidad de clasificación correcta o, equivalentemente si multiplicamos por la función de pérdida, de los riesgos asociados a la estrategia.

No obstante, dado que los datos ya han sido utilizados a la hora de determinar las regiones de clasificación, esta estimación tiende a sobrevalorar la probabilidad de aciertos. Para solucionar este problema puede dividirse la muestra en dos partes. Con la primera se construye las regiones de clasificación y con la segunda (se supone que los datos son independientes) se estima la probabilidad de acierto Este método se denomina con frecuencia jacknife. Otro método, denominado *Holdout* o método de validaciones cruzadas, consiste en construir las regiones de clasificación sin tener en cuenta el primer dato y clasificarlo en función de dicha estrategia; el proceso se repite con todos y cada uno y al final se calcula la proporción de aciertos.

La cuestión que se plantea a continuación a es la siguiente: ¿cuál es el mínimo valor permitido para la probabilidad de aciertos? Veamos un criterio convencional al respecto, válido únicamente si el grupo de pertenencia de cada dato de la muestra es aleatorio, es decir, si los datos de la muestra han sido escogidos sin considerar a qué grupo pertenecían. En ese caso, podemos estimar las probabilidades a priori calculando las proporciones de datos correspondientes a cada grupo, q_i. Una estrategia

ficticia sería clasificar cada observación tras un sorteo donde al grupo i-ésimo le corresponde una probabilidad q_i. El riesgo de Bayes respecto a esta distribución a priori sería

$$\sum_{i=1}^{r} q_i(1-q_i) = 1 - \sum_{i=1}^{r} q_i^2.$$

Desde luego, es de desear que la estrategia que consideremos, si bien no es óptima (por no verificarse las hipótesis requeridas) sea al menos considerablemente mejor que ésta, o equivalentemente, que la probabilidad de acierto sea considerablemente mayor que $\sum_{i=1}^{r} q_i^2$. Se conviene una diferencia mínima de 25 %. Si los resultados no son positivos conviene buscar métodos alternativos de clasificación. No obstante, el hecho de obtener una probabilidad de clasificación correcta baja no debe achacarse únicamente a la violación de los supuestos. Según queda de manifiesto en (10.6), aunque éstos se verifiquen, si las distribuciones de referencia son muy próximas, las probabilidades de error son, lógicamente, altas. En el caso extremo de que todas las distribuciones sean idénticas, la estrategia a seguir sería precisamente la estrategia ficticia mencionada anteriormente.

10.7. Estimación de densidades

Las técnicas estudiadas hasta ahora requieren del supuesto de normalidad multivariante. No obstante y en no pocas ocasiones, el fracaso, si es que se da, de este tipo de estrategias, debe achacarse a una excesiva *proximidad* entre las distribuciones de referencia antes que a una violación del supuesto de la normalidad. De todas formas, si consideramos que la no normalidad de las distribuciones pude ser responsable de la ineficacia de ambos métodos (lineal y cuadrático), es conveniente analizar procedimientos alternativos, más robustos, que puedan proporcionar una estrategia mejor. En este capítulo mostramos tres, siendo la primera de ellas la estimación de densidades.

Las estrategias lineal o cuadrática de Fisher se justifican mediante la aplicación del principio de máxima verosimilitud, pues consisten simplemente en asignar la observación a la distribución que la haga más verosímil, es decir, aquélla que maximiza el valor de la función de densidad. Pero siempre suponiendo la particularidad de que las densidades correspondan a distribuciones p-normales, tengan o no la misma matriz de covarianzas. Hemos de tener claro que, al margen de la relación que pueda existir con los ejes discriminantes (y, en consecuencia, con el resultado del manova), lo único que necesitamos para construir la estrategia es concocer (o, en su defecto, suponer) las densidades de las distribuciones. Ello indujo a Fix y Hodges, allá por el año 1951,

a iniciar el estudio de estimación de dendidades (que se encuadra en el ámbito de la estimación funcional), estudio que a la postre vino a revolucionar la estadística no paramétrica. En esta sección nos limitaremos a una descripción heurística del denominado método del núcleo y a algunos comentarios adicionales. Para un estudio más detallado remitimos al lector a Silverman (1986).

En el caso univariante, el método del núcleo funciona de la forma siguiente. Supongamos que contamos con una muestra aleatoria $x_1, \ldots, x_{\mathtt{n}}$, correspondiente a una determinada distribución continua con función de densidad p y queremos estimar el valor de p en $\mathtt{x}$, $\hat{p}(\mathtt{x})$. Para ello escogemos un número $\delta > 0$, que denominaremos ancho de banda, y consideramos el intervalo $[\mathtt{x} - \delta, \mathtt{x} + \delta]$, de amplitud 2δ. Sea $N(\mathtt{x})$ la cantidad de datos de la muestra en el anterior intervalo. Entonces, si $\mathtt{n}$ es grande se verifica[13]

$$P\big([\mathtt{x} - \delta, \mathtt{x} + \delta]\big) \simeq \frac{N(\mathtt{x})}{\mathtt{n}}.$$

Por otra parte, si δ pequeño se verifica[14]

$$P\big([\mathtt{x} - \delta, \mathtt{x} + \delta]\big) \simeq p(\mathtt{x}) \cdot 2\delta,$$

lo cual nos induce a considerar la estimación

$$\hat{p}(\mathtt{x}) = \frac{N(\mathtt{x})}{2\mathtt{n}\delta}.$$

Si queremos expresar $\hat{p}$ en función de los datos de la muestra, hemos de tener en cuenta que un dato x_i pertence al intervalo anterior si, y sólo si, $\delta^{-1}|x_i - \mathtt{x}| \leq 1$. Definimos entonces la función (denominada núcleo)

$$K(u) = \begin{cases} \frac{1}{2} \text{ si } |u| \leq 1 \\ 0 \text{ si } |u| > 1 \end{cases}, \quad u \in \mathbb{R}.$$

De esta forma,

$$\hat{p}(\mathtt{x}) = \frac{1}{\mathtt{n}\delta} \sum_{i=1}^{\mathtt{n}} K\left(\frac{\mathtt{x} - x_i}{\delta}\right), \qquad \mathtt{x} \in \mathbb{R}.$$

En el caso multivariante (dimensión p) no consideraremos intervalos de amplitud 2δ centrados en x sino *cubos* de volumen $2^p\delta^p$, y el núcleo K^p asigna el valor 2^{-p} a un punto u cuando $\|u\|_\infty \leq 1$. De esta forma, la función de densidad se estima como sigue:

$$\hat{p}(\mathtt{x}) = \frac{1}{\mathtt{n}\delta^p} \sum_{i=1}^{\mathtt{n}} K^p\left(\frac{\mathtt{x} - x_i}{\delta}\right), \qquad \mathtt{x} \in \mathbb{R}^p.$$

[13] Ley débil de los grandes números.

[14] Teorema fundamental del cálculo integral.

No obstante, la función de densidad estimada será de tipo escalonado. Un procedimiento comúnmente utilizado para suavizarla es considerar, en vez del núcleo anterior, el siguiente:

$$\tilde{K}(u) = \frac{1}{(2\pi S)^{p/2}} \exp\left\{-\frac{1}{2}u'S^{-1}u\right\}, \qquad u \in \mathbb{R}^p,$$

donde S es la matriz de covarianzas muestral correspondiente al grupo considerado. Así, la función de densidad se estima mediante

$$\hat{p}(\mathtt{x}) = \frac{1}{\mathtt{n}\delta^p\,(2\pi S)^{p/2}} \sum_{i=1}^{\mathtt{n}} \exp\left\{-\frac{1}{2\delta^2}(\mathtt{x}-x_i)'S^{-1}(\mathtt{x}-x_i)\right\}.$$

Podemos comprobar que la función anterior se trata, efectivamente, de una densidad. Una vez estimadas las densidades de los distintos grupos procederemos a establecer las regiones de clasificación según el criterio de máxima verosimilitud. Otros núcleos que aparecen en la literatura, amén de éste denominado gaussiano, son el triangular, el del coseno, de Epanechnikov, etc.

Hay que tener en cuenta que la estimación de las densidades, y por ende la estrategia de clasificación, depende de la elección del núcleo K y del ancho de banda δ. Diversos trabajos vienen a convencernos de que la elección del núcleo no es demasiado determinante. No se puede decir lo mismo de la selección del ancho de banda. No podemos hablar, desde luego, de un ancho de banda universal, sino que debe depender del problema considerado. En la teoría denominada L_2, el ancho de banda se escoge de tal forma que minimize el denominado error cuadrático integrado medio[15]. La selección de un ancho de banda excesivamente grande tenderá a estimar la densidad demasiado plana, sobresuavizada, mientras que uno excesivamente pequeño la estimará más abrupta de lo que realmente es. Por desgracia, y como cabía esperar, para obtener la expresión del ancho de banda óptimo necesitamos poseer cierta información de la distribución considerada, en concreto su curvatura (precisamente). De esta forma, entramos uno de esos círculos viciosos tan frecuentes en Estadística.

Otro inconveniente a tener en cuenta es la denominada maldición de la dimensión, que consiste en que el número de datos requerido para lograr una estimación satisfactoria de la densidad crece exponencialmente en relación con la dimensión considerada[16]. Por lo tanto, cuando tengamos un amplio número de variables precisaremos de una cantidad ingente de datos para obtener una estimación fiable de la densidad.

Otro importante método alternativo propuesto es la regresión logística, al que dedicamos la sección siguiente.

[15] Ver Silverman (1986)

[16] Ver Silverman (1986), tabla 2.2.

10.8. Regresión logística

Este método de regresión aporta en muchos casos una estrategia bastante satisfactoria cuando las distribuciones consideradas no son normales, e incluso cuando algunas de las variables son discretas o cualitativas. En esta sección nos limitaremos a una breve descripción del método. Para un estudio más detallado, remitimos al lector al capítulo 8 del volumen dedicado a los Modelos Lineales.

Efectivamente, un problema de clasificación puede entenderse como un análisis de regresión en el cual, las p variables consideradas actúan como explicativas, mientras que la variable cualitativa de asignación al grupo, que es la que se pretende predecir, desempeña por lo tanto la función de variable respuesta. Por razones didácticas, consideraremos el problema de clasificación respecto a dos distribuciones, a las que asignamos los valores 0 y 1, que se resolverá mediante la regresión logística binaria. Posteriormente, daremos la clave para pasar al caso general, que se resuelve mediante la regresión logística multinomial.

Pues bien, en el caso binario tenemos una variable dependiente dicotómica. Resultaría pues en todo punto descabellado intentar ajustar los datos observados mediante una regresión lineal, dado que en ese caso el rango de la variable dependiente sería toda la recta real. Debemos pues considerar funciones cuya imagen sea el conjunto $\{0,1\}$, o mejor el intervalo $[0,1]$, de tal forma que si el valor resultante es inferior a 0.5 tomamos el cero y si es superior el 1. De entre las funciones continuas de $\overline{\mathbb{R}}$ en $[0,1]$ destacamos la función de distribución del modelo probabilístico $N(0,1)$ y la función logística

$$\mathtt{L}(z) = \frac{e^z}{1+\mathtt{e}^z} \tag{10.11}$$

El uso de la primera da lugar a los modelos *probit* y el de la segunda a los modelos *logit* o de regresión logística. Son estos últimos los más utilizados y los que nos ocupan en estos momentos. En el capítulo 8 del volumen dedicado a los Modelos Lineales se razona la utilidad de esta función cuando se aborda un problema de discriminación.

Para ser breves diremos que, si Z denota el vector aleatorio p-dimensional de variables observadas e Y la variable (dicotómica en este caso) de asignación a grupo, que se supone también aleatoria, y se verifica[17]

$$Z|Y = i \sim N_p(\mu_i, \Sigma), \quad i = 0, 1,$$

se sigue entonces de la regla de Bayes (cuestión propuesta) que, si $X = -(1|Z)$ y

[17] Téngase en cuenta que en el modelo condicional consideramos los supuestos del manova y, por lo tanto, de la estrategia lineal.

$\beta = (\beta_0, \beta_1')'$,

$$P(Y = 1|Z) = \mathtt{L}(X\beta) \tag{10.12}$$

donde

$$\begin{aligned} \beta_0 &= \log \frac{1-p}{p} + \mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0, \\ \beta_1 &= \Sigma^{-1}(\mu_0 - \mu_1). \end{aligned}$$

La estimación de los parámetros del modelo no se realizará por el método de mínimos cuadrados sino por el de máxima verosimilitud, es decir, se buscarán los valores de β_0 y β_1 que maximicen el logaritmo de la función de verosimilitud de la distribución condicional $\prod_{j=1}^{\mathtt{n}} P^{Y|Z=z_j}$, donde $\mathtt{n}$ es el tamaño de muestra y z_j el vector p-dimensional de observaciones correspondientes al dato j-ésimo de la misma. Los detalles referentes a los problemas de estimación y contraste de hipótesis podemos encontralos en el capítulo 8 del primer volumen.

Hay que recordar que el uso de la función logística se ha argumentado partiendo de los supuestos de normalidad e igualdad de las matrices de covarianzas, los mismos de la estrategia lineal. No obstante, esta técnica tiene como principal ventaja su robustez, siendo su ámbito de aplicación bastante amplio.

Cuando se trabaja con una variable discreta tenemos la opción de calcular la media o media ponderada de los datos que coinciden en esa variable. Si se trabaja con variables cualitativas, conviene especificarlas como tales para que el programa les asigne las correspondientes variables ficticias.

En el caso de que la clasificación se realice respecto a r grupos, la técnica a utilizar es similar. Si suponemos ahora que la variable de asignación al grupo Y toma valores en $\{0, 1, \ldots, r-1\}$ y sigue, no un modelo de distribución binomial, como en el caso anterior, sino un modelo multinomial, se sigue por razonamientos en todo análogos a los anteriores que existen $\alpha_1, \ldots, \alpha_{r-1} \in \mathbb{R}$ y $\beta_1, \ldots, \beta_{r-1} \in \mathbb{R}^p$, tales que

$$P^{Y=i|Z=z} = \frac{\mathsf{e}^{\alpha_i + \beta_i z}}{1 + \sum_{j=1}^{r-1} \mathsf{e}^{\alpha_j + \beta_j z}}, \quad i = 1, \ldots, r-1. \tag{10.13}$$

Lo cual se relaciona nuevamente con la función logística. La técnica a desarrollar, denominada análisis de regresión logística multinomial, es análoga a la anterior aunque, lógicamente, más compleja.

10.9. k-proximidad

Este método podría considerarse una variante de las estrategia lineal. Recordemos que ésta última consiste en asignar la observación x a la distribución que la haga más

verosímil, lo cual equivale a adjudicarlo al centroide más próximo según la distancia de Mahalanobis

$$d^2(x, \overline{x}_j) = (x - \overline{x}_j)' S_3^{-1} (x - \overline{x}_j), \quad j = 1, \ldots, r.$$

El método de la k- proximidad consiste en seleccionar un entero k y considerar, para la observación x, las distancias

$$d^2(x, x_i)$$

donde x_i recorre todos los valores de las r muestras. Se considera entonces los k puntos más cercanos a x según esta distancia, de manera que tendremos k_j puntos correspondientes al grupo j-ésimo, de tamaño $\mathtt{n}_j$, $j = 1, \ldots, r$. Se verifica, por tanto,

$$\sum_{j=1}^{r} k_j = k, \quad \sum_{j=1}^{r} \mathtt{n}_j = \mathtt{n}.$$

Entonces, se asigna la observación al grupo que maximice el cociente

$$\frac{k_j}{\mathtt{n}_j}, \quad j = 1, \ldots, r.$$

Al igual que sucediera en el caso de la estimación de densidades, la estrategia se ve sensiblemente afectada por la elección de k, y no se dispone de un criterio universal para determinarlo. Se aconseja en todo caso probar con distintos valores y escoger el que aporte el mejor resultado en la validación posterior.

Cuestiones propuestas

1. Probar que $\mathtt{var}_1[U] = \Delta^2$.
2. Probar (10.5).
3. Demostrar que $\sum_{j=1}^{r} \mathtt{n}_j \left(\overline{W}_j^i - \overline{W}_{\cdot\cdot}^i \right)^2 = t_i$.
4. Consideremos un problema de clasificación de una observación respecto a dos distribuciones con idénticas probabilidades a priori, donde la función de pérdida verifica

   $$W(1|1) = W(2|2) = 0, \quad W(1|2) = W(2|1) = 1.$$

   Si las distribuciones se ajustan satisfactoriamente a sendos modelos normales con matrices de covarianzas idénticas y se utiliza la estrategia de clasificación lineal, la probabilidad de cometer error en la clasificación o, equivalentemente, el riesgo asociado a dicha estrategia, debe ser muy pequeña. ¿Puedes matizar esta afirmación en virtud de resultados ya conocidos?

5. Razona cómo y cuando se justifica la reducción a dimensión 2 (mapa territorial) en el problema de clasificación de una observación respecto a r modelos de distribución p-normales con matriz de covarianza común.

6. Relacionar la perspectiva de éxito en un problema de clasificación con el resultado del manova.

7. Deducir detalladamente las expresiones de α y β correspondientes a la regresión logística binaria. Deducir asimismo los coeficientes que se obtendrían en el aso multinomial.

8. Obtener (10.12).

9. Obtener las probabilidades $P^{Y|Z=z}(\{i\})$, $i = 0, 1, \ldots, r-1$, en el caso de la regresión logística multinomial.

Capítulo 11

Análisis factorial

El análisis factorial tiene como objetivo la representación de las p variables del vector aleatorio como puntos de un espacio de la menor dimensión posible, de manera que las variables con una fuerte correlación lineal directa queden próximas en el gráfico obtenido. Se trata pues, al igual que el análisis cluster que veremos en el próximo capítulo, de un método de formación de conglomerados, con la diferencias de que, mientras en este último se agrupan datos por un criterio de similitud a determinar, aquí agruparemos variables en virtud de un criterio de correlación lineal. Agrupar las variables partiendo de la mera observación de la matriz R es francamente complicado cuando su número es grande, de ahí que precisemos, como paso previo, de una técnica de simplificación. Dicha técnica está estrechamente ligada al análisis de componentes principales. Tanto es así que con frecuencia se confunden ambas disciplinas, sin bien el objeto final de las mismas es francamente dispar. No obstante, ambas se basan, fundamentalmente, en la diagonalización de la matriz de covarianzas o correlaciones. En lo que sigue, trabajaremos con las variables previamente tipificadas, como ya lo hicimos en el análisis de componentes principales, lo cual equivale a centrarnos en el estudio de la matriz de correlaciones R en lugar de la matriz de covarianzas S.

El análisis factorial es una técnica multivariante que no está exenta de polémica. Se debe a que el modelo estadístico de partida supone la existencia de unas variables no observadas o *latentes* denominadas *factores*, a partir de las cuales se obtienen mediante una ecuación lineal las variables realmente observadas, salvo errores incorrelados entre sí. Además, estos factores no están unívocamente determinados, sino que cualquier rotación permite obtener factores igualmente válidos. Desde luego, no parece éste un punto de partida que inspire confianza. No obstante, intentaremos dejar claro hasta qué punto es estrictamente necesaria la asunción de este modelo a la hora de establecer los conglomerados de variables, que es nuestro objetivo final.

Puede consultarse Rencher (1995) y Uriel et al. (2005) para ampliar la información.

11.1. Planteamiento del problema

En esta sección intentaremos explicar qué ventajas comporta la representación de la matriz de correlaciones R a partir de otra matriz Λ de menores dimensiones.

Un ejemplo

Empezaremos con un ejemplo extremo para dejar claros nuestro objetivos. Consideremos una muestra aleatoria simple de tamaño `n` de 5 variables que aporta la siguiente matriz de correlaciones.

$$R = \begin{pmatrix} 1 & 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ -1 & 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 1 \end{pmatrix}$$

Esta matriz simétrica de dimensiones 5×5 puede expresarse a partir de una matriz Λ de dimensiones 5×2 mediante

$$R = \Lambda\Lambda'$$

Podemos considerar, por ejemplo, la matriz

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{11.1}$$

No es ésta la única matriz 5×2 que nos permite reconstruir R. Podemos optar también, por ejemplo, por la matriz

$$\Lambda_* = \begin{pmatrix} -1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{11.2}$$

Decimos que se trata de un caso extremo porque dos variables cualesquiera son incorreladas o presentan una correlacion lineal perfecta, ya sea positiva o negativa. No hay posibilidad de término medio alguno.

Representación de R en un espacio k-dimensional

En general, nuestro problema consiste en reproducir de la manera más aproximada posible una matriz de correlación a través de otra matriz $\Lambda \in M_{p\times k}$, con k sensiblemente menor que p. Es decir, encontrar $\Lambda \in M_{p\times k}$ tal que

$$R \simeq \Lambda\Lambda' \tag{11.3}$$

Veamos qué beneficios se derivarían de semejante reproducción: denótese por $\lambda_1, \ldots, \lambda_p$ los vectores de $\mathbb{R}^k$ que constituyen, traspuestos y por ese orden, las filas de Λ. Es claro que el vector λ_i está asociado a la variable aleatoria i-ésima, pues la matriz R se expresaría, aproximadamente, como sigue:

$$R \simeq \begin{pmatrix} \|\lambda_1\|^2 & \ldots & \langle\lambda_1, \lambda_p\rangle \\ \vdots & \ddots & \vdots \\ \langle\lambda_p, \lambda_1\rangle & \ldots & \|\lambda_p\|^2 \end{pmatrix} \tag{11.4}$$

Denótese por Ψ la matriz diferencia $R - \Lambda\Lambda' \in \mathcal{M}_{p\times p}$, es decir,

$$r_{ij} = \langle\lambda_i, \lambda_j\rangle + \psi_{ij}, \qquad 1 \leq i, j \leq p \tag{11.5}$$

Si la aproximación (11.4) es satisfactoria, podemos identificar, en lo que respecta al problema de correlación lineal, cada variable con su correspondiente punto $\lambda_i \in \mathbb{R}^k$. ¿En qué sentido? En primer lugar, se siguen de (11.5) y de la desigualdad de Cauchy-Schwarz las siguientes desigualdades

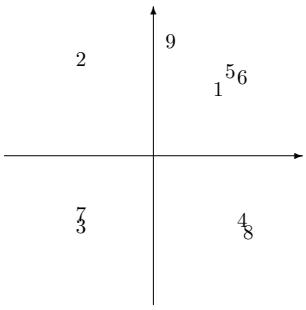
$$1 - r_{ij} \leq |\psi_{ii} + \psi_{jj}| + \sqrt{1 + \psi_{ii}} \cdot \|\lambda_i - \lambda_j\| \tag{11.6}$$

$$1 + r_{ij} \leq |\psi_{ii}| + \sqrt{1 + \psi_{ii}} \cdot \|\lambda_i + \lambda_j\| \tag{11.7}$$

Por lo tanto, si λ_i y λ_j son puntos próximos según la métrica euclídea, es decir, si $\|\lambda_i - \lambda_j\| \simeq 0$, se sigue de (11.6) que la correlación lineal entre las variables i-ésima y j-ésima será próxima a uno, tanto más cuanto mejor sea la aproximación (11.4) y cuanto mayor sea la cercanía entre los puntos. Análogamente, se sigue de (11.7) que, si λ_j se aproxima al opuesto a λ_i, es decir, si $\|\lambda_i + \lambda_j\| \simeq 0$, la correlación entre ambas variables será próxima a -1. Por último, si los puntos λ_i y λ_j se sitúan en direcciones aproximadamente perpendiculares, es decir, si $\langle\lambda_i, \lambda_j\rangle \simeq 0$, se sigue directamente de (11.5) que las variables correspondientes son prácticamente incorreladas.

Por lo tanto, mediante la observación de p puntos en un espacio k-dimensional podemos llegar a determinar, en el mejor de los casos, grupos o conglomerados de variables, de manera que las variables de un mismo grupo presenten fuertes correlaciones lineales, ya sean positivas o negativas, mientras que las correlaciones con las variables de otros grupos será muy débil. Dado que esta clasificación se realiza atendendiendo a criterios de proximidad y perpendicularidad, el número de conglomerados no puede exceder en ningún caso la dimensión del espacio en cuestiones, es decir, k.

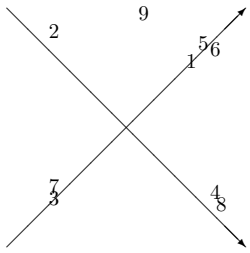
Veamos un ejemplo gráfico con ocho variables. Supongamos que la matriz de correlaciones R puede reproducirse salvo pequeños errores mediante un matriz Λ de dimensiones 8×2. Las fila se corresponden con puntos de $\mathbb{R}^2$ que se representan como sigue:



En este caso, la interpretación es clara: las variables 1,5 y 6 correlacionan fuerte y directamente entre sí; lo mismo sucede con 3 y 7 que, su vez, correlacionan fuerte pero inversamente con las primeras; por otra parte las variables 4 y 8 correlacionan fuerte y directamente entre sí e inversamente con 2; estas tres últimas variables son prácticamente incorreladas con las anteriores. Por último, la variable 9 presenta una correlación moderada con todas las demás. En consecuencia, pueden distinguirse claramente dos conglomerados de variables y otra variable más que queda en una situación intermedia.

Rotación de la solución

Se ha dado pues la circunstancia de que todos los puntos salvo uno, el 9, quedan recogidos en dos direcciones perpendiculares. Una rotación en $\mathbb{R}^2$ que convirtiera esas direcciones en ejes de coordenadas podría facilitar la interpretación del resultado. Esta podría ser la solución:



Como vemos, tenemos dos claros conglomerados de variables se identifica con uno de los ejes de coordenadas. Hay que tener en cuenta que la aplicación de una rotación $\Gamma \in \mathcal{O}_{k\times k}$ a los valores $\lambda_1, \ldots, \lambda_p$ no afecta a la ecuaciones (11.5). Efectivamente, si se verifica que $R = \Lambda\Lambda' + \Psi$, también se verificará $R = \Lambda_*\Lambda_*' + \Psi$ para $\Lambda_* = \Lambda\Gamma'$, es decir, cualquier rotación que apliquemos a una solución Λ conduce a un idéntica reproducción de la matriz de correlaciones. Una solución que identifique los conglomerados de variables con los ejes de coordenadas es más comprensible, en especial cuando $k > 2$, y ese será, por lo tanto, nuestro objetivo. En definitiva, buscamos una matriz ortogonal Γ tal que las componentes de los vectores $\Gamma\lambda_i$ sean próximas a ± 1 ó a 0, de manera que queden claramente asociados a un eje. Existen diversas técnicas para intentar conseguirlo. Destacamos dos de ellas:

- **Rotación varimax:** busca la máxima varianza entre las columnas de $\Lambda\Gamma$. Por lo tanto, pretende asociar a cada eje el menor número posible de variables.
- **Rotación cuartimax:** busca la máxima varianza entre las filas de $\Lambda\Gamma$. Por lo tanto, pretende asociar a cada variable el menor número posible de ejes.

Existen otros métodos, como el equamax, así como otro tipo de rotación denominada oblicua asociadas a matrices no ortogonales que, por lo tanto no son rotaciones en el sentido estricto de la palabra. Podemos encontrar más información al respecto en Rencher (1995) y Uriel et al. (2005).

Fases en la resolución del problema

Teniendo esto en cuenta, distinguiremos cuatro fases en la resolución del problemas

1. Analizar las condiciones.

2. Buscar una matriz Λ que, con el menor número posible de columnas reproduzca lo más presisamente posible la matriz de correlaciones R.

3. Una vez escogida Λ, comprobar que, efectivamente, proporciona una reproducción satisfactoria de la matriz de correlaciones.

4. Buscar la rotación que acerque lo más posibles los puntos fila de Λ a los ejes de coordenadas.

¿A qué condiciones nos referimos en el primer apartado? Pues depende de qué pretendemos conseguir exactamente, y la respuesta es simplificar el problema de correlación agrupando las variables en pocos conglomerados. Es decir, querríamos pocos conglomerados y muchas variables en cada uno de ellos. La situación que más se aleja de esta simplificación es la incorrelación entre todas las variables, que se correspondería con una matriz de correlaciones igual a la identidad. Esta hipótesis podría contrastarse mediante el test de Barlett, de manera que un resultado no significativo abortaría el análisis factorial pues no permitiría ninguna reducción de la dimensión original.

También hemos de tener en cuenta que, con el esquema deseado, los coeficientes de correlación al cuadrado deben ser muy altos o muy bajos, mientras que las correlaciones parciales al cuadrado entre dos variables cualesquiera dadas las demás deben ser muy bajas. ¿Por qué? Si las variables pertenecen a distintos conglomerados parece relativamente claro; si pertenecen a un mismo conglomerado y éste es suficientemente numeroso, el resto de las variables que lo configuran se encargarán de reducir la correlación parcial. Partiendo de esa idea se propone el coeficiente KMO[1]:

$$\mathsf{KMO} := \frac{\sum_{i\neq j} r_{ij}^2}{\sum_{i\neq j} r_{ij}^2 + \sum_{i\neq j} a_{ij}^2},$$

donde los r'_{ij}s denotan las correlaciones y los a'_{ij}s, las correlaciones parciales dadas las demás variables. Un valor alto de KMO invita al optimismo en lo que respecta a la simplificación del problema. En la práctica, se considera que con $\mathsf{KMO} < 0{,}60$ el análisis factorial no aportará una simplificación satisfactoria.

El tercer apartado lo hemos estudiado ya, aunque brevemente. El cuarto apartado está muy claro: una vez escogida la matriz Λ se calcula $MCR = \Lambda\Lambda'$ (matriz de correlaciones reproducidas) y se evalúa su diferencia con R, la cual se denomina matriz residual, que ha sido denotada anteriormente por Ψ. Se trata, obviamente, de que sus componentes sean próximas a 0. Respecto al apartado 2, falta todavía por concretar cómo se escogen k y Λ. Es lo que veremos en las siguientes secciones.

[1] Kaiser-Meyer-Olkin.

En el primer ejemplo, se obtiene una matriz Λ (11.1) que proporciona dos conglomerados perfectos de variables: el primero compuesto por las variables 1,3 y 5 (3 correlaciona inversamente); el segundo está compuesto por 2 y 4. Se trata, como dijimos, de un caso extremo, pues las correlaciones dentro de los conglomerados son perfectas, al igual que las incorrelaciones entre conglomerados distintos. En este caso, $\mathsf{KMO} = 1$ y la matriz Ψ es nula (pues R es de rango 2). Además, esta solución identifica el primer conglomerado con el eje OX y el segundo con OY. Por lo tanto, no precisa ser rotada. No obstante, la solución (11.2) es una de las infinitas soluciones que pueden obtenerse a partir de la anterior mediante una rotación. La configuración de los conglomerados no depende, como vemos, de la solución escogida.

11.2. Método de componentes principales

En esta sección expondremos la manera más natural de descomponer la matriz de correlaciones en un producto del tipo $\Lambda\Lambda'$, con $\Lambda \in \mathcal{M}_{p\times k}$. Se trata pues de una técnica para afrontar la segunda fase del estudio, según vimos anteriormente.

Ya hemos comentado que, con frecuencia, se confunden los análisis factorial y de componentes principales, y habría que pensar hasta que punto se trata de una confusión puesto que la aplicación del análisis e componentes principales es la forma más natural reducir la dimensión original del problema y, por lo tanto, de configurar los conglomerados de variables.

Diagonalización de R

La idea es muy simple. Sabemos que el análisis de componentes principales se basa en la descomposición canónica de una matriz simétrica, que en este caso va a ser la propia matriz de correlaciones

$$R = GDG'$$

donde D es la matriz diagonal de los autovalores ordenados de mayor a menor y G una matriz ortogonal cuyas columnas son los respectivos autovectores. Razonando en términos heurísticos, podemos distinguir entre autovalores grandes, los k primeros, y pequeños los $p-k$ restantes. El número k se determina de tal forma que $p^{-1}\sum_{j=1}^{k} d_j$ supere cierta cota a convenir. Tener en cuenta que, al tratarse de variables tipificadas, p es la varianza total del vector aleatorio y, por lo tanto, la fracción anterior se denomina proporción de varianza total explicada por las k primeras componentes principales. Incidiremos un poco más adelante en esta cuestión.

Si los k primeros autovalores y autovectores se agrupan, respectivamente, en las matrices D_1 y G_1 y hacemos lo mismo con los $p-k$ restantes la descomposición anterior queda como sigue:

$$R = (G_1|G_2)\left(\begin{array}{c|c} D_1 & 0 \\ \hline 0 & D_2 \end{array}\right)\left(\begin{array}{c} G_1' \\ \hline G_2' \end{array}\right) = G_1 D_1 G_1' + G_2 D_2 G_2' \tag{11.8}$$

Denótese

$$\Lambda = G_1 D_1^{1/2} \in \mathcal{M}_{p\times k}, \qquad \Psi = G_2 D_2 G_2' \in \mathcal{M}_{p\times p}$$

En ese caso, teniendo en cuenta que los autovalores de D_2 se han escogido de manera que sean próximos a 0 y que ninguna componente de G_2 puede ser superior a 1 pues sus columnas las constituyen un sistema ortonormal de vectores, se tiene que

$$R = \Lambda\Lambda' + \Psi, \qquad \Psi \simeq 0 \tag{11.9}$$

Así pues, ya tenemos lo que queríamos.

Regresión respecto a las componentes principales

Analicemos el resultado en términos de las componentes principales. La matriz $\mathtt{Z}$ de datos que corresponde a la observación de las p variables tipificadas en los $\mathtt{n}$ individuos estudiados y la matriz $\mathtt{U} \in \mathcal{M}_{\mathtt{n}\times p}$ correspondiente a las componentes principales se relacionan mediante $\mathtt{Z} = \mathtt{U}G'$. Podemos expresar la matriz $\mathtt{U}$ por columnas de la forma $\mathtt{U} = (\mathtt{U}_1, \mathtt{U}_2)$ de acuerdo con la descomposición anterior de D y G. Entonces, si se denota $\mathtt{F} = \mathtt{U}D_1^{-1/2}$ y $\mathtt{E} = \mathtt{U}_2 G_2'$, obtenemos la igualdad

$$\mathtt{Z} = \mathtt{F}\Lambda' + \mathtt{E}, \tag{11.10}$$

Con $\mathtt{E} \simeq 0$. La matriz $\mathtt{F}$, de media 0 y matriz de varianzas-covarianzas identidad, recoge las k primeras componentes principales, salvo una homotecia. Nos permitiremos el abuso de denominarlas igualmente componentes principales. La matriz $\mathtt{E}$ tiene también media 0 y es incorrelada con $\mathtt{F}$.

Esto es en definitiva lo que ya sabemos por el corolario 7.6, es decir, que los datos originales pueden reconstruirse de manera muy precisa partiendo de las k primeras componentes principales. Concretamente, si $\mathtt{Z}_i'$, $\mathtt{F}_i'$ y $\mathtt{E}_i'$ denotan las i-esimas filas de $\mathtt{Z}$, $\mathtt{F}$ y $\mathtt{E}$, donde $i = 1, \ldots, \mathtt{n}$, la ecuación (11.10) equivale a

$$\mathtt{Z}_i = \Lambda \mathtt{F}_i + \mathtt{E}_i, \quad 1 \leq i \leq \mathtt{n} \tag{11.11}$$

Es decir, salvo los errores $\mathtt{E}_i$, tenemos una ecuación lineal que relaciona las observaciones $\mathtt{Z}_i$ con las componentes principales $\mathtt{F}_i$, y al componer todos los datos obtenemos

el modelo de regresión lineal multivariante (11.10). Los coeficientes de la ecuación son las componentes de Λ. Así, el vector $\lambda_j \in \mathbb{R}^k$ cuya traspuesta es la j-ésima fila de Λ contiene los coeficientes que han de multiplicarse por los distintas componentes principales (k en total) para obtener (aproximadamente) los $\mathtt{n}$ valores correspondientes a la j-ésima variable. Según se ha obtenido Λ y teniendo en cuenta (7.27), queda claro que Λ es la matriz de cargas compuesta por las correlaciones entre las variables originales y las componentes principales seleccionadas.

Podemos pues interpretar la descomposición de R en los términos de una regresión lineal de $\mathtt{Z}$ respecto a una matriz explicativa $\mathtt{F}$. Primeramente, que $\lambda_i \simeq \lambda_j$ significa que las variables i-ésima y j-ésima pueden expresarse, salvo un pequeño error, como una casi idéntica combinación de las componentes principales y, por lo tanto, son casi iguales (lo cual equivale a $r_{ij} \simeq 1$, pues están tipificadas); que $\lambda_i \simeq -\lambda_j$ equivale, por el mismo razonamiento, a que las variables sean casi opuestas (lo cual equivale a $r_{ij} \simeq -1$). Por otra parte, podemos aplicar cualquier rotación $\Gamma \in \mathcal{O}_{k\times k}$ a los vectores fila de Λ, obteniendo otra matriz $\Lambda_* = \Lambda\Gamma'$. Si aplicamos a $\mathtt{F}$ la rotación inversa $\mathtt{F}_* = \mathtt{F}\Gamma$, se conserva la ecuación

$$\mathtt{Z} = \mathtt{F}_*\Lambda'_* + \Psi \tag{11.12}$$

siendo también la matriz de varianzas y covarianzas de $\mathtt{F}_*$ la identidad. Se trata pues de una rotación (inversa) de los ejes de coordenadas que podemos aplicar a nuestro antojo, transformando las componentes principales en otras variables también incorreladas cuyas varianzas suman $\sum_{j=1}^{k} d_j$. En consecuencia, si λ_i y λ_j son casi perpendiculares, existirá una rotación que convierta las variables i-ésima y j-ésima en la expresión casi exacta de sendas variables incorreladas y, por lo tanto, son casi incorreladas.

Por último, una variable como la novena en el ejemplo anterior podría interpretarse como una combinación lineal de las dos primera componentes principales (o de sus rotaciones), de ahí que correlacione moderadamente con los dos conglomerados.

Parte explicada y parte no explicada

De (11.10) se puede deducir la descomposición (11.9) de la matriz de correlaciones R, evidenciando la correspondencia de los sumandos $\Lambda\Lambda'$ y Ψ con $\mathtt{F}$ y $\mathtt{E}$, respectivamente. De hecho, podemos entender $\Lambda\Lambda'$ como la parte de R explicada por $\mathtt{F}$ y, en definitiva, por las componentes principales seleccionadas. La varianza total de $\mathtt{F}\Lambda'$ se

obtiene mediante

$$\begin{aligned}\mathtt{var}_T[\mathtt{F}\Lambda'] &= \mathtt{tr}\left[\mathtt{n}^{-1}\Lambda\mathtt{F}'\mathtt{F}\Lambda'\right] = \mathtt{tr}[\Lambda\Lambda'] \\ &= \mathtt{tr}[G_1 D_1 G_1'] = \mathtt{tr}[D_1] = \sum_{j=1}^{k} d_j\end{aligned}$$

En ese sentido dijimos anteriormente que $p^{-1}\sum_{j=1}^{k} d_j$ se entiende como la proporción de varianza total explicada por las k primeras componentes principales. Además, trabajando directamente con la matriz Λ se tiene también

$$\sum_{j=1}^{k} d_j = \sum_{j=1}^{k} \|\lambda_j\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{k} \lambda_{ij}^2 \tag{11.13}$$

Por lo tanto y en particular, cada sumando $\sum_{i=1}^{p} \lambda_{ij}^2$, $j = 1, \ldots, k$ se interpreta como la parte de variabilidad total explicada por la j-ésima componente principal.

Si, para $i = 1, \ldots, p$, se denota $h_i^2 = \|\lambda_i\|^2$ (téngase en cuenta que este número permanece invariante si a Λ se le aplica una rotación), la parte diagonal de la ecuación matricial (11.9) queda como sigue

$$1 = h_i^2 + \psi_i, \qquad i = 1, \ldots, p. \tag{11.14}$$

El término h_i^2 expresa la parte de la varianza de la i-ésima variable reproducida por Λ o, equivalentemente, explicada por la matriz explicativa $\mathtt{F}$, mientras que ψ_i expresa el error cometido al intentar reproducir la varianza i-ésima Λ o, equivalentemente, la parte de a varianza no explicada por la matriz $\mathtt{F}$. De (11.13) se sigue que la proporción de varianza explicada por la k primeras componentes principales es igual a la media aritmética de los términos h_i^2, que se denotará por $\overline{h}$.

Este término puede servir para acotar las diferencias entre los términos de R y los de la matriz de de correlaciones reproducidas $\Lambda\Lambda'$. Los elementos de la diagonal, ψ_{ii}, $i = 1, \ldots, p$, están acotados por $1 - \overline{h}$. Para acotar los elementos de fuera de la diagonal, ψ_{ij}, $i \neq j$, es decir, las diferencias entre las correlaciones reales y las reproducidas por Λ, hemos de tener en cuenta la descomposición (11.8) junto con la desigualdad de Cauchy-Schwarz. De esta forma se tiene que

$$\psi_{ij} \leq \sqrt{p(1-\overline{h})}, \qquad i \neq j \tag{11.15}$$

Esta desigualdad es, no obstante, bastante conservadora. Podemos apurar más si sustituimos $\overline{h}$ por $p^{-1}d_{k+1}$.

Otro ejemplo

Veamos otro ejemplo. Presentamos la matriz de correlaciones muestral R correspondiente a la medición de 4 variables psicológicas sobre 5 individuos:

$$R = \begin{pmatrix} 1{,}000 & & & \\ 0{,}296 & 1{,}000 & & \\ 0{,}881 & -0{,}022 & 1{,}000 & \\ 0{,}995 & 0{,}326 & 0{,}867 & 1{,}000 \end{pmatrix}.$$

Se observa una fuerte correlación entre las variables 1, 3 y 4 que, por otro lado, correlacionan débilmente con 2, lo cual nos induce a probar con un análisis de dos factores. Es un caso claro de aplicación del análisis factorial, y además extraordinariamente sencillo, pues una simple ojeada a la matriz R determina los grupos a formar.

Mediante el método de componentes principales representamos la matriz R mediante $R \simeq \Lambda\Lambda'$ donde

$$\Lambda = \begin{pmatrix} 0{,}992 & -0{,}007 \\ 0{,}315 & 0{,}945 \\ 0{,}911 & -0{,}347 \\ 0{,}991 & 0{,}026 \end{pmatrix}.$$

En este caso, obtenemos los siguientes h_i^2's:

$$h_1^2 = 0{,}984, \quad h_2^2 = 0{,}992, \quad h_3^2 = 0{,}950, \quad , h_4^2 = 0{,}983$$

Podemos apreciar una excelente aproximación a las distintas varianzas. De hecho, la proporción de varianza total explicada, que puede calcularse como la media aritmética de los cuatro términos anteriores, es

$$\overline{h} = 0{,}977$$

Así pues, podemos acotar

$$|r_{ij} - \lambda_i'\lambda_j| \leq \sqrt{4(1 - 0{,}977)} = 0{,}303, \quad i \neq j.$$

Si conociéramos el tercer autovalor de R podríamos obtener una cota más baja. Observando la matriz Λ podemos hacernos una idea bastante clara de los conglomerados de variables existentes. De todas formas, una rotación tipo varimax aporta una solución más clara aún:

$$\Lambda_* = \begin{pmatrix} 0{,}969 & 0{,}216 \\ 0{,}094 & 0{,}992 \\ 0{,}965 & -0{,}133 \\ 0{,}960 & 0{,}248 \end{pmatrix}$$

Puede observarse claramente que las variables 1,3 y 4 se asocian al eje OX (correlacionan entre sí y positivamente), mientras que 2 se asocia al eje OY, es decir, correlaciona débilmente con las demás.

11.3. Modelo basado en el concepto de factor

Con lo visto hasta ahora ya estamos en condiciones de resolver íntegramente nuestro problema sin necesidad de añadir nada más. Hemos de tener en cuenta que, hasta el momento, no ha sido necesaria la imposición de ningún supuesto estadístico. Por contra, nada nos garantiza a priori una reducción satisfactoria de la dimensión ni una configuración clara de las variables como solución final. En aras de obtener una solución lo más satisfactoria posible, puede resultar apropiado, aunque controvertido, establecer un modelo estadístico de partida para nuestro estudio, lo cual supondrá la aceptación de una serie de supuestos, cosa a la que estamos más que habituados. Pero decimos controvertido porque el modelo se basará en la existencia de ciertas variables latentes denominadas factores. Con el término latentes queremos decir que estas variables no se observan en la realidad. ¿A qué nos referimos pues?

Hemos visto en la sección anterior un par de ejemplos. En el segundo de ellos teníamos un par de conglomerados de variables muy bien configurados: el primero compuesto por las variables 1, 3, 5, 6 y 7, mientras que el segundo lo constituyen 2, 4 y 8. Hay que añadir una novena que queda entre ambos conglomerados. Dos variables que pertenecen a un mismo conglomerado pueden expresarse linealmente una a partir de la otra con bastante precisión. Si las suponemos tipificadas, como es nuestro caso, ello se traduce en que las variables son prácticamente iguales u opuestas. En definitiva, todas las variables de un mismo conglomerado poseen *algo* en común, y ese algo no guarda relación lineal con las variables de un conglomerado distinto. Ese algo que, estrictamente hablando no es sino una clase de equivalencia, es lo que denominamos factor, y puede tomar como representante a cualquiera cualquiera de las variables del conglomerado. Efectivamente, es claro que conocida una de ellas obtendremos casi con exactitud el resto. En nuestro teoría, los factores serán unas variables incorreladas no especificadas.

El modelo del análisis factorial consiste en llevar un poco más lejos la ecuación (11.11). Partiremos de una muestra aleatoria simple de tamaño `n` de un vector aleatorio p-dimensional de componentes $X_1, \ldots, X_p$, con media $\mu = (\mu_1, \ldots, \mu_p)'$ y matriz de varianzas-covarianzas Σ, y supondremos que existen un vector aleatorio k-dimensional f de media 0 y componentes $f_1, \ldots, f_k$ incorreladas con varianza 1, denominadas factores, y otro vector p-dimensional $\mathcal{E}$ de componentes $\mathcal{E}_1, \ldots, \mathcal{E}_p$ e

incorrelado con f, tales que X puede expresarse mediante el sistema de ecuaciones lineales

$$\begin{array}{rcccccccc} X_1-\mu_1 & = & \underline{\lambda}_{11}\mathtt{f}_1 & + & \ldots & + & \underline{\lambda}_{1k}\mathtt{f}_k & + & \mathcal{E}_1 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ X_p-\mu_p & = & \underline{\lambda}_{p1}\mathtt{f}_1 & + & \ldots & + & \underline{\lambda}_{pk}\mathtt{f}_k & + & \mathcal{E}_p \end{array} \tag{11.16}$$

Los coeficientes $\underline{\lambda}_{ij}$ componen una matriz $\underline{\Lambda} \in \mathcal{M}_{p\times k}$ que permite expresar matricialmente (11.16) mediante

$$X-\mu = \underline{\Lambda} f + \mathcal{E} \tag{11.17}$$

donde

(a) $\mathtt{E}[f]=0, \quad \mathtt{Cov}[f]=\mathtt{Id}_{k\times k}$

(b) $\mathtt{Cov}[f,\mathcal{E}]=0$

De todo ello se sigue inmediatamente que $\mathtt{E}[\mathcal{E}]=0$, que $\mathtt{Cov}[X,f]=\underline{\Lambda}$ y que

$$\Sigma = \underline{\Lambda}\underline{\Lambda}' + \mathtt{Cov}[\mathcal{E}] \tag{11.18}$$

Sabemos, por lo visto en el capítulo dedicado al análisis de componentes principales, que un modelo de este tipo se verifica automáticamente sin más que escoger como factores las k primeras componentes principales de X divididas por sus desviaciones típicas y, como $\mathcal{E}$, las $p-k$ últimas, multiplicadas matricialmente por sus correspondientes autovectores de Σ.

Nuestro objetivo será estimar la matriz $\underline{\Lambda}$ porque, aplicando los mismos razonamientos de las secciones anteriores, nos permitirá identificar nuestras variables con puntos de $\mathbb{R}^k$ para así agruparlas en conglomerados en función del grado de correlación lineal. Igualmente, podemos aplicar una rotación $\Gamma \in \mathcal{O}_{k\times j}$ al vector f de factores obteniendo un nuevo vector $f_* = \Gamma f$ de tal manera que las ecuaciones (11.17) y (11.18) se siguen verificando con $\underline{\Lambda}_* = \underline{\Lambda}\Gamma'$. En cosecuencia, cualquier rotación de los factores conduce a k nuevos factores igualmente válidos, es decir, que el vector f no está unívocamente determinado. Se buscará la versión f_* que permita una interpretación más clara de la matriz $\underline{\Lambda}_*$ correspondiente. Normalmente, como ya sabemos, se procurará representar los conglomerados sobre los distintos ejes, de manera que se identificarán ejes con factores.

En lo que sigue, impondremos una condición adicional a nuestro modelo. En términos heurísticos podríamos formularla así: la variabilidad de cada componente X_i descompone en una parte común a todas las componentes más otra puramente específica. En términos formales, el enunciado sería éste:

(c) $\mathtt{cov}[\mathcal{E}_i, \mathcal{E}_j] = 0$, para todo $i \neq j$.

En consecuencia, la matriz $\underline{\Psi} = \mathtt{Cov}[\mathcal{E}]$ será diagonal y la ecuación (11.18) podrá expresarse mediante

$$\Sigma = \underline{\Lambda}\underline{\Lambda}' + \underline{\Psi}, \tag{11.19}$$

de manera que, si para $i = 1, \ldots, p$, se denota

$$\overline{h}_i^2 = \sum_{j=1}^{k} \underline{\lambda}_{ij}^2$$

se sigue de (11.19) que

$$\begin{aligned} \mathtt{var}[X_i] &= \underline{h}_i^2 + \underline{\psi}_{ii} \\ \mathtt{cov}[X_i, X_l] &= \sum_{j=1}^{k} \underline{\lambda}_{ik}\underline{\lambda}_{lk}, \quad i \neq l \end{aligned}$$

Los términos $\underline{h}_i^2$, $1 \leq i \leq p$, se denominarán **comunalidades** pues expresan la parte de cada varianza explicada por los factores comunes. En contraposición tenemos los términos $\underline{\psi}_{ii}$, $1 \leq i \leq p$, que denominaremos **varianzas específicas**, pues expresan la parte de cada varianza no explicada por los factores comunes sino por un error $\mathcal{E}_i$ específico de cada variable, en el sentido de que, según nuestro modelo, los p errores son incorrelados. Por lo tanto, se verifica que, mediante la matriz $\underline{\Lambda}\underline{\Lambda}'$, podemos reproducir de manera aproximada las varianzas de las componentes X y de manera exacta las covarianzas.

Esta suposición es muy controvertida y en absoluto contrastable, dado que nuestro modelo se construye a partir de un cierto número de variables no observadas. Será, no obstante, de utilidad a la hora de estimar la matriz $\underline{\Lambda}$.

Lo dicho hasta ahora se enmarca en un contexto puramente probabilístico. Debemos traducirlo pues al lenguaje estadístico. Recordemos que partimos de una muestra aleatoria simple de tamaño $\mathtt{n}$ del vector aleatorio $(X_1, \ldots, X_p, f_1, \ldots, f_k)$ en las condiciones supuestas. Nótese que las k últimas componentes de cada unidad experimental no son observables. Cada unidad experimental satisfará la ecuación (11.17) junto con las propiedades (a), (b) y (c). Dado que nuestro propósito es la estimación de la matriz $\underline{\Lambda}$, procederemos a descomponer, de la manera más aproximada posible, la matriz de covarianzas muestral S de forma análoga a la descomposición (11.18) de Σ, es decir, de la forma

$$S \simeq \Lambda\Lambda' + \Psi$$

con $\Lambda \in \mathcal{M}_{p\times k}$ y $\Psi \in \mathcal{M}_{p\times p}$ diagonal y semidefinida positiva. De esta forma, Λ y Ψ constituirán sendos estimadores de $\underline{\Lambda}$ y $\underline{\Psi}$, respectivamente. Ya hemos comentado al principio del capítulo que es habitual trabajar con las variables tipificadas, cuya matriz de covarianzas es la matriz de correlaciones $\mathtt{P}$ de las variables originales. Así pues, en lo que sigue nuestro objetivo será obtener Λ y $\Psi \geq 0$ diagonal tales que

$$R \simeq \Lambda\Lambda' + \Psi \tag{11.20}$$

Las comunalidades $\underline{h}_i^2$ se estimarán, lógicamente, mediante $h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2$; por su parte, las varianzas específicas se estimarán mediante $\psi_i = 1 - h_i^2$, de manera que (11.20) sea una ecuación exacta, al menos para los elementos de la diagonal. Como podemos ver, estamos en las mismas condiciones de la sección anterior con la salvedad de que la matriz Ψ ha de ser diagonal. De hecho, el método de componentes principales que estudiamos en dicha sección sigue siendo perfectamente válido en este contexto: se desecha la parte de R asociada a los autovalores más pequeños obteniéndose $R \simeq \Lambda\Lambda'$. La matriz $\underline{\Psi}$ se estima entonces como la parte diagonal de $R - \Lambda\Lambda'$. Posteriormente se procede a la rotación de la solución.

Está claro que la polémica propiedad (c) no se tiene en cuenta a la hora de aplicar el método de componentes principales. Sin embargo, existen otras técnicas de estimación de $\underline{\Lambda}$ que sí se basan en dicho supuesto, entre las que destacamos brevemente dos: el método de los ejes principales y el de máxima verosimilitud.

Método de los ejes principales

Recordemos que el método de componentes principales realiza una estimación de Λ, de la cual se deriva la estimación de Ψ como la diagonal del residuo. En este caso, procederemos al contrario: realizaremos una estimación inicial $(R - \Psi)_0$ de $\mathtt{P} - \underline{\Psi}$ y, entonces, determinaremos una primera estimación Λ_1 de $\underline{\Lambda}$ eliminado la parte de la primera correspondiente a sus $p - k$ últimos autovalores, con lo cual se verificará

$$(R - \Psi)_0 \simeq \Lambda_1\Lambda_1'.$$

Sabemos que la matriz $\mathtt{P} - \Psi$ debería expresarse de la siguiente forma:

$$\begin{pmatrix} \underline{h}_1^2 & r_{12} & \dots & r_{1p} \\ r_{21} & \underline{h}_2^2 & \dots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \dots & \underline{h}_p^2 \end{pmatrix}.$$

Los elementos fuera de la diagonal son ya conocidos. Necesitamos conocer estimaciones de las comunalidades. Teniendo en cuenta el significado de estos términos según nuestro modelo[2], suele estimarse cada comunalidad $\underline{h}_i^2$ mediante

$$h_i^2(0) := R_i^2$$

siendo R_i^2 el coeficiente de correlación múltiple de la i-ésima variable respecto al resto. Dicho coeficiente puede expresarse (cuestión propuesta) de la forma

$$R_i^2 = 1 - \frac{1}{r^{ii}}, \tag{11.21}$$

donde r^{ii} denota el i-ésimo elemento del eje[3] o diagonal de R^{-1}. Es preciso que R no sea singular para poder estimar de esta forma las comunalidades. Si lo fuera, se estimaría cada comunalidad $\underline{h}_i^2$ mediante

$$h_i^2(0) := \text{máx}\{r_{ij}^2 : j = 1, \ldots, p\}.$$

Una vez obtenido obtenido $(R - \Psi)_0$, se determina una primera estimación de $\underline{\Lambda}$, $\Lambda(1)$, tal que

$$(R - \Psi)_0 \simeq \Lambda(1)\Lambda(1)',$$

eliminando la parte de $(R - \Psi)_0$ correspondiente a sus $p - k$ últimos autovalores. De esta forma, la proporción de varianza explicada por los k factores se estima mediante

$$\frac{\sum_{j=1}^k d_j}{\text{tr}(R - \Psi)_0},$$

siendo $d_1, \ldots, d_k$ los k primeros autovalores de $(R - \Psi)_0$. Tener en cuenta que la matriz $(R - \Psi)_0$ no es necesariamente semidefinida positiva, y puede llegar a tener autovalores negativos. Por tanto, el anterior cociente puede ser mayor que 1. Esta situación problemática se denomina caso *Heywood*. Una vez estimadas las cargas factoriales, podemos mejorar la estimación inicial de las comunalidades mediante

$$h_i^2(1) = \sum_{j=1}^k \lambda_{ij}(1)^2, \quad i = 1, \ldots, p.$$

De esta forma, podemos construir una nueva estimación $(R - \Psi)_1$, sustituyendo las comunalidades iniciales por las nuevas, y realizar una nueva estimación de la matriz de $\underline{\Lambda}$, $\Lambda(2)$, de manera que

$$(R - \Psi)_1 \simeq \Lambda(2)\Lambda(2)'$$

[2] Entiéndase como la parte de la varianza explicada linealmente por los factores *comunes* a todas las variables.

[3] De ahí su nombre.

Este proceso puede continuar de forma iterativa hasta que las comunalidades se estabilizan. Si se presenta un *caso Heywood*, el ordenador puede optar por finalizar el proceso.

En la práctica, los dos métodos estudiados hasta el momento presentan resultados muy similares cuando existe una gran cantidad de variables o cuando las correlaciones entre éstas son fuertes.

Método de máxima verosimilitud

Este método es bastante complicado desde el punto de vista operacional. Además requiere como hipótesis que el vector X siga un modelo de distribución p-normal. Se trata de buscar los estimadores de máxima verosimilitud de $\underline{\Lambda}$ y $\underline{\Psi}$, que se obtienen igualando a 0 las distintas derivadas parciales y resolviendo las consiguientes ecuaciones lineales. La solución no está unívocamente determinada, por lo cual es necesario añadir a las condiciones del modelo la restricción de que la matriz $\underline{\Lambda}'\underline{\Psi}^{-1}\underline{\Lambda}$ sea diagonal. Entonces se busca primero la estimación de $\underline{\Lambda}$ y, posteriormente, la de $\underline{\Psi}$. Las ecuaciones se resuelven por un método iterativo.

No obstante su complicación, este método tiene dos importantes ventajas: en primer lugar, nos proporciona un test para contrastar una hipótesis inicial que puede indentificarse parcialmente con las condiciones del modelo para k factores. Se trata del test de razón de verosimilitudes, cuyo estadístico de contrate es el siguiente[4]:

$$\left(\mathtt{n} - \frac{2p - 4k + 11}{6}\right) \ln\left(\frac{|\Lambda\Lambda' + \Psi|}{|S|}\right),$$

que se compara con el correspondiente cuantil de la distribución

$$\chi^2_{\frac{1}{2}[(p-k)^2-p-k]}$$

Si se rechaza la hipótesis alternativa hay que considerar la posibilidad de introducir más factores. No obstante, este método es bastante exigente en ese sentido, es decir, si lo aplicamos rigurosamente acabaremos introduciendo demasiados factores.

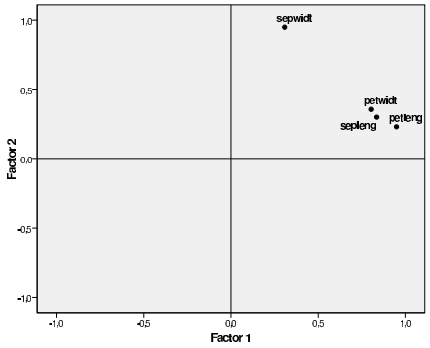
La otra ventaja de este método estriba en que el multiplicar la variable i-ésima por una cte c sólo conlleva multiplicar los coeficentes de Λ relativos a dicha variable por $\sqrt{c}$. Por tanto, las matrices R y S aportan filas proporcionales en la matriz Λ. Para más detalles, consultar Rencher (1995) y Uriel et al. (2005).

[4]En este caso, no deberíamos trabajar con datos tipificados si queremos asumir la normalidad

11.4. Ejemplo

Para terminar vamos a ver cómo quedarían configuradas las variables tipificadas del archivo `irisdata` de Fisher. En primer lugar, hemos de tener en cuenta que la especie setosa no debe mezclarse con versicolor y virginica pues su matriz de correlaciones es netamente diferente a las del resto de especies. Así pues efectuaremos dos análisis por separado. En todo caso, los factores se extraerán mediante el método de componentes principales y se aplicará una rotación varimax.

En el caso de virginica y vesicolor juntas, sabemos que las dos primeras componentes principales explican el 87.82 % de la varianza total. De esta forma, la representación de las variables rotadas respecto a un modelo con dos factores es la siguiente:



Las comunalidades nos dan una idea de la fiabilidad e esta gráfica. Concretamente, obtenemos las siguientes:

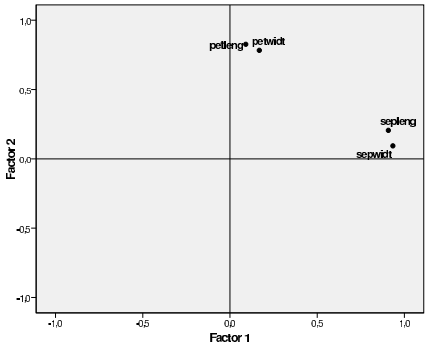
	comunalidad
long-sep	0.788
anch-sep	0.997
long-pet	0.954
anch-pet	0.774

En el caso de setosa, se obtiene las siguientes comunalidades para dos factores:

	comunalidad
long-sep	0.867
anch-sep	0.881
long-pet	0.692
anch-pet	0.641

Página 1

Podemos apreciar que las variables de setosa quedan peor representadas mediante dos factores. Concretamente, basta calcular la media aritmética de las cuatro comunalidades para concluir que se explica únicamente el 77.01 % de la varianza total. La representación de las variables rotadas es la siguiente:



Aunque el gráfico ofrece menos garantías que el anterior, muestra de forma más clara dos conglomerados de variables: uno que se indentifica con el factor sépalo y otro con el pétalo.

Cuestiones propuestas

1. Obtener (11.6) y (11.7).
2. Demostrar que, si se verifican las cuatro hipótesis del modelo de análisis factorial, entonces $\mathtt{Cov}[X, \mathtt{f}] = \Lambda$. Página 1
3. Probar que una rotación en los factores no altera las condiciones del modelo ed análisis factorial ni el valor de las comunalidades.
4. Demostrar que $\Sigma = \Lambda\Lambda' + \Psi$.
5. Demostrar que, en el método de componentes principales, la media aritmética de las comunalidades coincide con la proporción de varianza total explicada por la k primeras componentes principales.
6. Obtener la desigualdad (11.15).
7. Demostrar que $\sum_{i=1}^{p} \lambda_{ij}^2 = d_j, \quad j = 1, \ldots, k.$

8. Probar la igualdad (11.21). Indicación: utilizar el lema 13.6.

9. Crees conveniente realizar un análisis factorial cuando todos los autovalores de la matriz de covarianzas S son similares. ¿Puedes contrastar esta hipótesis mediante algún test conocido?

10. Con los datos del archivo iris.sav, comparar el diagrama de dispersión de las dos primeras componentes principales con el de los dos primeros factores estimados.

Capítulo 12

Análisis cluster

El análisis cluster o de conglomerados es una técnica multivariante de agrupación de datos teniendo en cuenta su afinidad respecto a un vector Y de p variables observadas. Se trata de un método mayormente computacional, donde no es del todo fácil la aplicación de las técnicas de inferencia estadística. De hecho, existen varios caminos a seguir a la hora de resolver el problema de aglomeración sin que exista unanimidad alguna al respecto. Ante esta disyuntiva, lo más aconsejable es realizar el análisis mediante diferentes métodos y contrastar los resultados obtenidos.

Primeramente, hemos de tener en cuenta que un cambio de escala en alguna variable puede afectar sensiblemente a la formación de conglomerados, de ahí que, en una primera encrucijada, hemos de decidir si tipificamos o no las variables. Además, si existen varias variables fuertemente correlacionadas, puede que estemos sobrevalorando un factor latente común, que tendrá más peso del debido en la formación de los conglomerados. Por ello es necesario en ciertas ocasiones realizar un análisis de componentes principales y considerar la posibilidad de reducir la dimensión del vector observado. También tendrán un gran impacto en la formación de conglomerados la presencia de valores atípicos. Éstos pueden detectarse a priori (mediante un método univariante, como la representación del diagrama de cajas, o multivariante, como el cálculo de la distancia de Mahalanobis), o bien a posteriori, dado que los datos atípicos constituirán conglomerados unitarios. Una vez detectados, estos datos deben ser analizados desde el punto de vista experimental para decidir si se mantienen o eliminan del estudio.

El capítulo se divide en tres secciones: una primera donde se establecen las distintas formas de medir el grado de afinidad entre datos; en la segunda sección se expone cómo formar los conglomerados una vez seleccionada la medida de afinidad. En este aspecto, hemos de distinguir dos métodos de aglomeración: jerárquico y de k-medias.

Por último, se realiza la valoración de los conglomerados obtenidos.

12.1. Medidas de afinidad

Presentamos a continuación las medidas de afinidad más utilizadas a la hora de formar conglomerados. Todas ellas, excepto la de Mahalanobis, están presentes en SPSS.

- **Distancia euclídea:** es la utilizada por defecto. La distancia euclídea entre dos datos $x_1 = (x_{11}, \ldots, x_{1p})'$ y $x_2 = (x_{21}, \ldots, x_{2p})'$ se define mediante

$$d_e(x_1, x_2) = \|x_1 - x_2\|_2 = \sqrt{\sum_{j=1}^{p} (x_{1j} - x_{2j})^2}.$$

  SPSS ofrece la posibilidad de personalizar el exponente pues da opción a considerar cualquier medida del tipo

$$\|x_1 - x_2\|_p$$

- **Correlación de Pearson:** es la alternativa más popular a la distancia euclídea. Se trata de permutar el papel que juegan los datos y las variables. De esta forma, dados dos datos x_1 y x_2, se calculan

$$\overline{x}_i = \frac{1}{p} \sum_{j=1}^{p} x_{ij}, \quad s_{ik} = \frac{1}{p} \sum_{j=1}^{p} (x_{ij} - \overline{x}_i)(x_{kj} - \overline{x}_k), \quad i, k = 1, 2.$$

  Entonces, se considera el coeficiente

$$r_{x_1, x_2} = \frac{s_{12}}{\sqrt{s_{11} s_{22}}}.$$

  Nótese que un alto valor de este coeficiente se corresponde con un patrón de comportamiento análogo de los datos x_1 y x_2 en relación con las componentes del vector Y. Una medida similar al coeficiente de correlación es el coseno del ángulo entre los vectores x_1 y x_2.

- **Distancia absoluta:** se define mediante

$$d_a(x_1, x_2) = \sum_{j=1}^{p} |x_{1j} - x_{2j}|.$$

- **Distancia de Mahalanobis:** se define mediante

$$d_m^2(x_1, x_2) = (x_1 - x_2)' S^{-1} (x_1 - x_2),$$

donde S denota la matriz de covarianza correspondientes a las p variables. Esta medida es de gran importancia en el análisis multivariante, como hemos podido comprobar. En este apartado, presenta la ventaja de que, al ser una mediada invariante ante cambios de escala, su utilización permite eludir el dilema de la tipificación de los datos. Además, los datos atípicos en sentido multivariante quedarán identificados como conglomerados unitarios.

12.2. Formación de conglomerados

Una vez establecida una medida de afinidad o proximidad entre datos, el siguiente paso es la formación de conglomerados en función de la misma. En este sentido, hemos de distinguir dos métodos: el jerárquico y el de k-medias. A su vez, cada uno de ellos puede llevarse a cabo mediante distintos procedimientos, como veremos a continuación.

Método jerárquico

Inicialmente se considera cada dato como un conglomerado unitario. Partiendo de esa situación, cada paso que se dé consistirá en unir los dos conglomerados más *próximos* entre sí para formar un único conglomerado más grande. El procedimiento se repite, en principio, hasta que quede un único conglomerado constituido por todos los datos. El proceso de formación de los conglomerados queda registrado y nosotros podemos analizar el estado más interesante, que será aquél en el que queden patente grandes diferencias interconglomerados y pequeñas diferencias intraconglomerados. Es decir, que en todos los pasos anteriores se unieron conglomerados próximos, pero en el inmediatamente posterior se unen dos conglomerados muy distantes. Esto puede verse gráficamente de forma muy sencilla mediante un gráfico, disponible en SPSS, denominado dendrograma. El diagrama de témpanos aporta una información similar. Mediante el análisis de los mismos debemos pues determinar el número de

conglomerados en la solución final. No obstante, SPSS ofrece la opción de detener el proceso cuando se haya llegado a cierto número de conglomerados que podemos especificar previamente.

Hemos dicho anteriormente que cada paso consistirá en la fusión de los dos conglomerados más *próximos* entre sí. Obviamente, la proximidad se determinará en virtud de la medida de afinidad que hayamos escogido. No obstante, ésta se aplica a cada par de puntos, mientras que los conglomerados son conjuntos (unitarios o no). Por ello, queda aún pendiente determinar una medida de proximidad entre conjuntos partiendo de la medida d de proximidad entre puntos seleccionada. En ese sentido, contamos con las opciones siguientes:

- **Vinculación intergrupos:** si d es la distancia entre puntos determinada, se define la distancia $\tilde{d}$ entre dos conglomerados A y B como la media aritmética de las distancias $d(a,b)$ donde $a \in A$ y $b \in B$, es decir

$$\tilde{d}(A,B) = \frac{1}{\mathsf{card}(A \times B)} \sum_{a \in A,\ b \in B} d(a,b).$$

- **Vinculación intragrupos:**

$$\tilde{d}(A,B) = \frac{1}{\mathsf{card}\big((A \cup B) \times (A \cup B)\big)} \sum_{x,y \in A \cup B} d(x,y).$$

- **Vecino más próximo:**

$$\tilde{d}(A,B) = \text{mín}\,\big\{d(a,b)\colon a \in A,\ b \in B\big\}.$$

- **Vecino más lejano:**

$$\tilde{d}(A,B) = \text{máx}\,\big\{d(a,b)\colon a \in A,\ b \in B\big\}.$$

- **Agrupación de centroides:** la distancia entre A y B es la distancia entre sus respectivos centroides (medias aritméticas).

Método de k-medias

Se utiliza fundamentalmente cuando, por las razones que sean, el número k de conglomerados finales está determinado *a priori*. Se trata de aglomerar todos los datos en torno a k puntos (que se denominan semillas) en función de la proximidad a éstos. En muchos caso, estas semillas son establecidas de antemano en función

de conocimientos previos. En ese caso, el método consiste en asignar cada dato a la semilla más próxima. Al final, habremos formado tantos conglomerados como semillas introducidas. No obstante y una vez acabado el proceso, las semillas iniciales pueden reemplazarse por los centroides (medias) de los conglomerados finales y volver a repetir el procedimiento de aglomeración en torno a las nuevas semillas *mejoradas*. Se trata pues de un proceso iterativo que finalizará cuando se haya alcanzado una cierta estabilidad en las semillas, o bien cuando se hayan realizado un número de iteraciones que podemos determinar previamente.

Si queremos formar k conglomerados pero no contamos con semillas, puede procederse de la siguiente forma: se seleccionan k datos, bien aleatoriamente o bien los k primeros, que serán las semillas iniciales. Los datos restantes se irán aglomerando en torno a ellos. No obstante, si la menor semilla más cercana a un datos dista del mismo más que que la semilla más cercana a ésta, dicho dato la reemplaza como semilla y conquista, por así decirlo, su conglomerado. Al final del proceso, se reconstruyen las semillas como centroides de los conglomerados finales y el procedimiento se repite sucesivamente hasta conseguir cierta estabilidad en los centroides finales, o bien cuando se hayan realizado un determinado número de iteraciones.

La ventaja del método de k-medias respecto al jerárquico radica en que su algoritmo es mucho más rápido (especialmente con muestras de gran tamaño) y se ve menos afectado ante la presencia de valores atípicos. Su desventaja estriba en la elección de las semillas iniciales, que puede tener una gran trascendencia en el resultado final y, sin embargo, son frecuentemente designadas según criterios bastante arbitrarios. No obstante, puede ser muy interesante la combinación de los dos métodos: nos referimos a utilizar el método jerárquico y analizar el dendrograma para determinar el número de conglomerados a formar, y utilizar como semillas los centroides de los mismos en un posterior análisis de k medias. También puede invertirse el orden: clasificamos primeramente respecto a un número elevado m de semillas y obtenemos m centroides finales, que se someterán a un análisis jerárquico, de manera que los grupos correspondientes a centroides próximos se unirán dando lugar a un número menor de conglomerados homogéneos.

12.3. Interpretación de los conglomerados

Una vez configurados los conglomerados definitivos, conviene caracterizarlos mediante un patrón de comportamiento respecto a las variables observadas. El método más usual de caracterización consiste en representar los perfiles de las medias aritméticas por variables de los distintos centroides. La interpretación de los mismos

excede el estudio meramente estadístico y es tarea que corresponde al investigador experimental.

El análisis cluster puede unirse a otras técnicas multivariantes: por ejemplo, puede aplicarse un manova para ratificar la correcta separación de los conglomerados finales. También puede realizarse un análisis cluster como paso previo a un análisis discriminate. En el primer análisis se procedería a la configuración de grupos y en el segundo, a la clasificación de observaciones respecto a los grupos constituidos.

Cuestiones propuestas

1. Cuando se ha propuesto el coeficiente de correlación como medida de proximidad entre datos, hemos advertido una permutación en el papel que juegan variables y datos. ¿Puedes relacionar estos con el análisis factorial? Realiza una agrupación de variables con los datos de `psicologic.sav` mediante técnicas de análisis cluster, y compara los resultados con los que obtendríamos con los procedimientos del capítulo anterior.

2. Aplica un análisis cluster a los datos de `iris.sav` y valora los resultados obtenidos en relación con las especies existentes. Plantea una estrategia de clasificación respecto a los conglomerados obtenidos y analiza la validez de la misma.

Capítulo 13

Apéndice

Este capítulo se divide en dos partes. La primera presenta una selección de definiciones y resultados correspondientes, fundamentalmente, al Álgebra de matrices, que serán de utilidad en el desarrollo de nuestra teoría. La mayor parte de las demostraciones pueden encontrarse en el Apéndice del volumen dedicado a los Modelos Lineales. En general, dicho Apéndice es bastante más completo que éste pues pretende abarcar los conceptos necesarios para abordar el estudio de los Modelos Lineales, el cual podría considerarse en buena medida preliminar del Análisis Multivariante. De ahí que hayamos optado por no incluirlos aquí para evitar excesivas redundancias.

La sección B es de carácter eminentemente técnico y está dedicada íntegramente a la demostración del teorema 6.1 . Ha sido postergada a este Apéndice para aligerar la lectura del capítulo 6.

(A) Álgebra de matrices.

En primer lugar, en lo que sigue y dado $n \in \mathbb{N}$, se denotarán mediante 1_n y $1_{n\times n}$ el vector y las matrices constantes 1 de $\mathbb{N}^n$ y $\mathcal{M}_{n\times n}$, respectivamente. Por otra parte, dada una matriz $A \in \mathcal{M}_{n\times n}$ (entendemos que sus coeficientes son reales), $\delta \in \mathbb{C}$ se dice autovalor de A cuando es raíz del polinomio de grado n $p(x) = |A - x\mathtt{Id}|$, lo cual significa que existe un vector $e \in \mathbb{C}^n$ tal que $Ae = \delta e$. De un vector en tales condiciones decimos que es un autovector de A asociado al autovalor δ, cosa que sucede para toda la recta $\langle e \rangle$.

Consideremos $y = (y_1, \ldots, y_n)'$ y $x = (x_1, \ldots, x_n)'$ dos vectores (los vectores se consideran, por convenio, columnas) cualesquiera de $\mathbb{R}^n$. El producto interior definido mediante

$$\langle x, y \rangle = x'y = \sum_{i=1}^{n} x_i y_i,$$

dota a $\mathbb{R}^n$ de una estructura de espacio de Hilbert. Se deice que x e y son ortogonales cuando $\langle x, y\rangle = 0$, lo cual se denota mediante $x \perp y$. La norma asociada al mismo,

$$\|x\| = \sqrt{\langle x, x\rangle},$$

se denomina euclídea, e induce una distancia, del mismo nombre, entre los puntos de $\mathbb{R}^n$

$$\|x - y\| = \sqrt{\langle x - y, x - y\rangle} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

Por lo tanto, la región del espacio formada por los puntos cuya distancia respecto a x sea igual a un cierto número positivo k es un esfera. El cuadrado de la distancia puede expresarse de esta forma

$$\|y - x\|^2 = (y - x)'\mathtt{Id}(y - x).$$

Si sustituimos la matriz identidad por cualquier matriz simétrica definida positiva A, la región anterior será un elipsoide, cuyas características dependerán de los autovectores y autovalores de A (ver teorema de diagonalización). Una expresión de este tipo pueden encontrarse en la densidad de la distribución normal multivariante.

Una sistema de vectores e de $\mathbb{R}^n$ se dice ortonormal cuando los vectores son de norma euclídea 1 y ortogonales entre sí. Una matriz $\Gamma \in \mathcal{M}_{n\times n}$ se dice ortogonal cuando Γ' es su inversa, lo cual equivale a afirmar que sus columnas constituyen una base ortonormal de $\mathbb{R}^n$. En ocasiones las denominaremos *rotaciones*, ya veremos el porqué. El conjunto de todas las matrices ortogonales de orden n se denotará por $\mathcal{O}_{n\times n}$. Dado un subespacio vectorial $V \subset \mathbb{R}^n$, $V^\perp$ denota el subespacio vectorial de dimensión $n - \mathtt{dim}V$ constituidos por todos los vectores ortogonales a V. Asimismo, si $W \subset V$, $V|W$ denotará el subespacio $V \cap W^\perp$, de dimensión $V|W$.

Una matriz $A \in \mathcal{M}_{n\times n}$ se dice semidefinida positiva cuando es simétrica[1] y verifica que $e'Ae \geq 0$, para todo $e \in \mathbb{R}^n$, en cuyo caso se dice definida positiva, denotándose por $A \geq 0$. Esta definición permite establecer un preorden en $\mathcal{M}_{n\times n}$. Concretamente,

$$A \geq B \text{ cuando } x'Ax \geq x'Bx, \text{ para todo } x \in \mathbb{R}^n. \tag{13.1}$$

Decimos que A es definida positiva cuando verifica $e'Ae > 0$, para todo $e \in \mathbb{R}^n\backslash\{0\}$, en cuyo caso se denota $A > 0$.

Lema 13.1.

Todos los autovalores de una matriz simétrica son reales.

[1]En rigor, no es necesario que la matriz sea simétrica para que sea definida positiva, pero en nuestra teoría lo supondremos siempre.

En consecuencia, dado que sólo consideraremos autovalores de matrices reales simétricas, tanto éstos como las componentes de sus autovectores serán reales. El resultado siguiente precede al más importante de este capítulo.

Lema 13.2.
Si $A \in \mathcal{M}_{n\times n}$ simétrica y $\Gamma \in \mathcal{M}_{n_\times n}$ ortogonal, los autovalores de A coinciden con los de $\Gamma' A \Gamma$.

El siguiente resultado, denominado Teorema de Diagonalización, permite expresar de forma natural cualquier matriz simétrica. Para la demostración del la segunda parte del mismo se precisa del Teorema de los Multiplicadores Finitos de Lagrange, abreviadamente TMFL, que será a su vez necesario para probar diversos resultados de nuestra teoría. El TMFL se divide en dos parte: la primera establece condiciones necesarias que debe verificar un extremos relativo condicionado; la segunda establece condiciones suficientes. Su enunciado es el siguiente.

Teorema 13.3.
Sean n y m números naturales tales que $n < m$ y $\mathcal{U} \subset \mathbb{R}^m$ abierto. Consideremos las aplicaciones $\phi : \mathcal{U} \longrightarrow \mathbb{R}$ y $f : \mathcal{U} \longrightarrow \mathbb{R}^n$, ambas con derivadas parciales segunda continuas. Sean $M = \{x \in \mathcal{U} \colon f(x) = 0\}$ y $c \in M$. Supongamos que el rango de la matriz $\left(\frac{\partial f_i}{\partial x_k}(c)\right)$ es n, y que existe un vector $\lambda \in \mathbb{R}^n$ tal que $\bigtriangledown(\phi - \lambda' f)(c) = 0$. Entonces, para que $\phi_{|M}$ tenga un máximo (mínimo) relativo en c, es condición suficiente que $D^2 L_\lambda(c)(h,h) < 0$ (respectivamente > 0) cada vez que $h \in \mathbb{R}^m \setminus \{0\}$ verifique que $Df_i(c)(h) = 0,\ i = 1, \ldots, n$, donde $L_\lambda = \phi - \lambda' f$.

Obsérvese la analogía que guarda con las condiciones necesaria y suficiente para máximos y mínimos no condicionados. La primera parte (necesariedad) se obtiene como aplicación del teorema de la función implícita, mientras que la segundo (suficiencia) se deduce del teorema de Taylor. Para más detalles, consultar Fdez. Viñas II, pag. 126. Dicho esto, vamos a enunciar el teorema fundamental al que hacía alusión anteriormente.

Teorema 13.4 (Diagonalización)**.**
Si $A \in \mathcal{M}_{n\times n}$ simétrica, existe una matriz $n \times n$ ortogonal Γ y una matriz $n \times n$ diagonal Δ=diag$(\delta_1, \ldots, \delta_n)$, con $\delta_1 \geq \ldots \geq \delta_n$, tales que

$$A = \Gamma \Delta \Gamma'.$$

En ese caso, los δ_i's son los autovalores de A y las columnas γ_i's de Γ constituyen una base ortonormal de autovectores asociados, siendo igualmente válida cualquier otra base

ortonormal de autovectores asociados. Se verifica, además, que

$$\delta_1 = \sup_{\alpha \in \mathbb{R}^n \setminus \{0\}} \frac{\alpha' A \alpha}{\|\alpha\|^2},$$

alcanzándose en $\alpha = \gamma_1$, y que, para cada $i = 2, \ldots, n$,

$$\delta_i = \sup_{\alpha \in \langle \gamma_1, \ldots, \gamma_{i-1} \rangle^{\perp}} \frac{\alpha' A \alpha}{\|\alpha\|^2},$$

alcanzándose el máximo en $\alpha = \gamma_i$.

Obsérvese que, si los autovalores de la matriz son distintos, la descomposición es única salvo reflexiones de los autovectores. En caso contrario, será única salvo reflexiones y rotaciones de éstos. El siguiente corolario es inmediato:

Corolario 13.5. (i) Dos autovectores asociados a distintos autovalores de una matriz simétrica son ortogonales.

(ii) Si A es simétrica, su rango coincide con el número de autovalores no nulos.

(iii) Si $A \geq 0$, sus autovalores son todos no negativos. Si $A > 0$, son todos estrictamente positivos.

(iv) Si $A \geq 0$, existe[2] una matriz simétrica $A^{1/2}$ tal que $A = A^{1/2} A^{1/2}$. Si $A > 0$, existe también una matriz simétrica $A^{-1/2}$ tal que $A^{-1} = A^{-1/2} A^{-1/2}$.

(v) Si $A \geq 0$, existe una matriz X con las mismas dimensiones tal que $A = X'X$.

(vi) Dada $A \in \mathcal{M}_{n \times n}$ semidefinida positiva de rango r, existe $X \in \mathcal{M}_{n \times r}$ de rango r tal que $A = XX'$.

(vii) La traza de una matriz simétrica es la suma de sus autovalores y el determinante, el producto de los mismos.

Exponemos a continuación un resultado fundamental, relacionado con la matriz de covarianzas parciales.

Lema 13.6.

Consideremos una matriz cuadrada

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

[2] En Arnold(1981) se prueba además la unicidad.

(i) Si S_{22} es invertible, entonces $|S| = |S_{22}| \cdot |S_{11} - S_{12}S_{22}^{-1}S_{21}|$.

(ii) Si $S > 0$, entonces $S_{22} > 0$. Además, si la inversa de S es

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix},$$

se verifica que $V_{11}^{-1} = S_{11} - S_{12}S_{22}^{-1}S_{21}$.

El siguiente resultado se utilizará para justificar los test para la media en el modelo lineal normal multivariante.

Teorema 13.7.
Sean S y U matrices $p \times p$ simétricas, definida positiva y semidefinida positiva, respectivamente, y sea el polinomio en t $p(t) = |U - tS|$. Entonces, $p(t)$ tiene todas sus raíces reales y no negativas, $t_1 \geq \ldots \geq t_p$, verificándose que

$$t_1 = \max_{x \in \mathbb{R}^p \setminus \{0\}} \frac{x'Ux}{x'Sx}.$$

Además, existe una matriz $A \in \mathcal{M}_{p\times p}$ tal que

$$ASA' = \mathtt{Id}_p, \quad AUA' = \begin{pmatrix} t_1 & & 0 \\ & \ddots & \\ 0 & & t_p \end{pmatrix}.$$

A continuación un resultado que será de interés en el estudio de la distribución de Wishart.

Teorema 13.8.
Para toda $S \in \mathcal{M}_{p\times p}$ semidefinida positiva existe una matriz $C \in \mathcal{M}_{p\times p}$ triangular superior tal que $S = CC'$.

El resultado siguiente se utiliza en una de las reducciones por invarianza a la hora de obtener el test de contraste para la media.

Teorema 13.9.
Sean $X, Y \in \mathcal{M}_{p\times k}$. Se verifica entonces que $X'X = Y'Y$ sii existe una matriz $\Gamma \in \mathcal{M}_{p\times p}$ ortogonal tal que $Y = \Gamma X$.

Nótese que, si $k = 1$, estamos afirmando que $\|X\| = \|Y\|$ sii existe una matriz $\Gamma \in \mathcal{M}_{p\times p}$ ortogonal tal que $Y = \Gamma X$. Por ello se identifican las matrices ortogonales con las rotaciones y la norma euclídea constituye un invariante maximal para el grupo de las rotaciones. El siguiente resultado se utiliza también en las reducciones por invarianza para el contraste de la media.

Teorema 13.10.
Sean $X, Y \in \mathcal{M}_{p\times k}$ y $S, T \in \mathcal{M}_{p\times p}$ definidas positivas. Si $X'S^{-1}X = Y'T^{-1}Y$, existe una matriz $A \in \mathcal{M}_{p\times p}$ invertible tal que $Y = AX$ y $T = ASA'$.

El siguiente resultado es de utilidad a la hora de encontrar el estimador de máxima verosimilitud en el modelo lineal normal multivariante.

Teorema 13.11.
Sean A una matriz $p \times p$ definida positiva y f la función que asigna a cada matriz U del mismo tipo el número $f(U) = \frac{1}{|U|^{n/2}} \exp\left\{-\frac{1}{2}\mathtt{tr}(U^{-1}A)\right\}$. Entonces, dicha función alcanza el máximo en $U = \frac{1}{n}A$.

A continuación, repasaremos brevemente el concepto de proyección ortogonal, de gran importancia en el Modelo Lineal. Aunque se define sobre $\mathbb{R}^n$, el concepto tiene sentido en cualquier espacio dotado de un producto interior, por ejemplo L_2. Dado un subespacio lineal $V \subset \mathbb{R}^n$ de dimensión k, se define la proyección ortogonal sobre V como la aplicación P_V que asigna a cada vector $u \in \mathbb{R}^n$ el único vector $v \in V$ tal que $u - v \in V^\perp$. Puede probarse que se trata del vector de V más próximo a u según la distancia euclídea. Dicha aplicación es lineal y sobreyectiva. Por lo tanto, se identifica con una matriz $n \times n$ de rango k, que se denotará igualmente por P_V. Se verifica además que, si $X \in \mathcal{M}_{n\times k}$ es una base de V,

$$P_V = X(X'X)^{-1}X'. \tag{13.2}$$

La anterior expresión tiene sentido, pues $\mathtt{rg}(X) = \mathtt{rg}(X'X) = k$, es decir, $X'X$ es invertible. Así pues, dado $u \in \mathbb{R}^p$, se tiene que $X(X'X)^{-1}X'u \in V$. Además, dado cualquier $y \in \mathbb{R}^k$, se tiene que

$$\left\langle u - X(X'X)^{-1}X'u, Xy\right\rangle = u'Xy - u'X(X'X)^{-1}X'Xy = 0,$$

es decir, que $u - X(X'X)^{-1}X'u \in V^\perp$. Además, $X(X'X)^{-1}X'u$ es el único vector de V que lo verifica pues, si existiesen dos vectores $v_1, v_2 \in V$ tales que $u - v_1, u - v_2 \in V^\perp$, entonces se tendría que $v_1 - v_2 \in V \cap V^\perp = 0$. Además, dado que

$$\mathtt{rg}\left(X(X'X)^{-1}X'\right) = \mathtt{rg}(X) = k,$$

la aplicación es sobreyectiva. Por lo tanto, la proyección ortogonal está bien definida y es, efectivamente, una aplicación lineal sobreyectiva cuya matriz es (13.2). Nótese que, si X es una base ortonormal de V, entonces $P_V = XX'$.

La matriz P_V es idempotente, es decir, es simétrica y verifica que $P_V^2 = P_V$. Puede demostrarse, recíprocamente (ver, por ejemplo, Arnold (1981)), que toda matriz $n \times n$

idempotente de rango k es la matriz de la proyección ortogonal sobre el subespacio k-dimensional de $\mathbb{R}^n$ generado por sus vectores columna. Veamos algunas propiedades elementales de la proyección ortogonal.

Proposición 13.12.
Sean $V, W \subset \mathbb{R}^n$, con $W \subset V$. Se verifica:

(i) $P_V = P_{V|W} + P_W$.

(ii) Para todo $y \in \mathbb{R}^n$, $\|P_V y\|^2 = \|P_W y\|^2 + \|P_{V|W} y\|^2$. En particular, $\|y\|^2 = \|P_V y\|^2 + \|P_{V^\perp} y\|^2$.

(iii) $P_V y = y$ sii $y \in V$.

(iv) $P_W \cdot P_V = P_W$.

(v) $\mathtt{tr} P_V = \mathtt{dim} V$.

(vi) $P_{V^\perp} = \mathtt{Id} - P_V$.

El concepto de proyección ortogonal puede extenderse a subvariedades afines de la siguiente forma: dada una subvariedad afín k-dimensional H (esto es, un subconjunto de $\mathbb{R}^n$ de la forma $x + V$, siendo x un vector cualquiera de $\mathbb{R}^n$ y V un subespacio k-dimensional del mismo), definiremos la proyección ortogonal sobre H (que se denotará P_H) como la aplicación que asigna a cada vector u el vector de H que minimiza la distancia euclídea a u, en cuyo caso se tiene

$$P_{x+V} u = x + P_V(u - x).$$

Hemos de tener en cuenta que, para cada $v \in V$, se verifica

$$P_{x+V} = P_{(x+v)+V}. \tag{13.3}$$

(B) Demostración del teorema 6.1.

En esta sección nos proponemos probar el teorema 6.1. La demostración utilizará técnicas básicas del álgebra lineal junto con el Teorema de los Multiplicadores Finitos de Lagrange.

Con las notaciones introducidas en el capítulo 6, consideremos, para cada $\lambda \in \mathbb{R}$, la siguiente matriz

$$A_\lambda = \begin{pmatrix} -\lambda \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & -\lambda \Sigma_{zz} \end{pmatrix}$$

Vamos a analizarla detenidamente. Consideremos el polinomio en λ

$$P(\lambda) = |A_\lambda|$$

Dado que las matrices Σ_{zz} y Σ_{yy} no son singulares, el grado de $P(\lambda)$ es $p+q$. Si $\lambda \neq 0$, se sigue del lema 13.6

$$\begin{aligned} P(\lambda) &= |-\lambda\Sigma_{zz}| \cdot |\lambda\Sigma_{yy} + \lambda^{-1}\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}| \\ &= \lambda^{q-p}|\Sigma_{yy}||\Sigma_{zz}| \cdot |\Sigma_{yy}^{-1}\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy} - \lambda^2 \mathtt{Id}| \\ &= k\lambda^{q-p}Q(\lambda^2), \end{aligned}$$

donde

$$k = (-1)^q|\Sigma_{yy}||\Sigma_{zz}|$$

y Q es el polinomio característico de la matriz

$$C_1 = \Sigma_{yy}^{-1}\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}, \tag{13.4}$$

de grado p. Siendo la igualdad válida para todo λ distinto de 0, también lo es para 0. Puede comprobarse fácilmente que los autovalores de C_1 coinciden (aunque con distintos autovectores asociados) con los de la matriz

$$\Sigma_{yy}^{-1/2}\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}\Sigma_{yy}^{-1/2}, \tag{13.5}$$

que es simétrica y semidefinida positiva. Luego, son todos reales y no negativos y, por tanto, las raíces de $P(\lambda)$ son todas reales. Si $\lambda_1, \ldots, \lambda_{p+q}$ son tales raíces, ordenadas de mayor a menor, ha de verificarse

$$\begin{array}{ccccccc} \lambda_1 & \geq & \ldots & \geq & \lambda_p & \geq & 0 \\ \lambda_{p+1} & = & \ldots & = & \lambda_q & = & 0 \end{array}$$

$$\begin{aligned} \lambda_{p+q} &= -\lambda_1 \\ &\vdots \\ \lambda_{q+1} &= -\lambda_p \end{aligned}$$

También se verifica que los autovalores positivos de C_1 coinciden (con distintos autovectores asociados) con los autovalores positivos de la matriz

$$C_2 = \Sigma_{zz}^{-1}\Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz}. \tag{13.6}$$

El lema siguiente es inmediato:

Lema 13.13.
Para cada $i = 1, \ldots, p$, se verifica que $(\alpha', \beta')' \in \mathtt{ker} A_{\lambda_i}$ si, y sólo si, α y β son autovectores de las matrices C_1 y C_2, asociados ambos al mismo autovalor λ_i^2. Además, si $\lambda_i \neq 0$, se verifica

$$\begin{aligned} \beta &= \lambda_i^{-1}\Sigma_{zz}^{-1}\Sigma_{zy}\alpha \\ \alpha &= \lambda_i^{-1}\Sigma_{yy}^{-1}\Sigma_{yz}\beta \end{aligned}$$

Lema 13.14.
Dada una raíz λ_i de $P(\lambda)$, donde $i = 1, \ldots, p+q$, se verifica que

$$\mathtt{dim}(\mathtt{ker} A_{\lambda_i}) = \mathtt{mult}_P(\lambda_i).$$

Además, si $\lambda_i \neq 0$,

$$\mathtt{ker} A_{\lambda_i} \cap \mathtt{ker} A_{-\lambda_i} = 0.$$

Demostración.
Si $\lambda_i \neq 0$ (lo cual implica $i \notin \{p+1, \ldots, q\}$), se sigue del lema anterior que

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \in \mathtt{ker} A_{\lambda_i} \Leftrightarrow \left[\lambda_i^2\alpha_i = C_1\alpha_i\right] \wedge \left[\beta_i = \frac{1}{\lambda_i}\Sigma_{zz}^{-1}\Sigma_{zy}\alpha_i\right]$$

Por tanto, $\mathtt{dim}(\mathtt{ker} A_{\lambda_i}) = \mathtt{mult}_Q(\lambda_i^2) = \mathtt{mult}_P(\lambda_i)$. Por otro lado,

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \in \mathtt{ker} A_{\lambda_i} \Leftrightarrow \begin{pmatrix} -\alpha_i \\ \beta_i \end{pmatrix} \in \mathtt{ker} A_{-\lambda_i},$$

de lo cual se deduce que

$$\mathtt{ker} A_{\lambda_i} \cap \mathtt{ker} A_{-\lambda_i} = 0, \qquad \mathtt{dim}(\mathtt{ker} A_{-\lambda_i}) = \mathtt{mult}_P(\lambda_i).$$

Veamos qué sucede con las raíces nulas: supongamos que $\lambda_1 \geq \ldots \geq \lambda_k > 0,\ \lambda_{k+1} = 0$ (ello implica $k \leq p$) En ese caso, C consta de k autovalores positivos, luego $k = \mathtt{rg}(C)$ o, equivalentemente,

$$k = \mathtt{rg}\left(\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}\right) = \mathtt{rg}\left(\Sigma_{zz}^{-1/2}\Sigma_{zy}\right) = \mathtt{rg}(\Sigma_{yz}).$$

Es decir,

$$\mathtt{rg}\begin{pmatrix} 0 & \Sigma_{yz} \\ \Sigma_{zy} & 0 \end{pmatrix} = 2k.$$

Por lo tanto, $\mathtt{dim}\left(\mathtt{ker} A_0\right) = p + q - 2k = \mathtt{mult}_P(0)$. ∎

Denótese por Υ la siguiente matriz invertible

$$\Upsilon = \begin{pmatrix} \Sigma_{zz}^{1/2} & 0 \\ 0 & \Sigma_{yy}^{1/2} \end{pmatrix}.$$

Dados dos vectores $a = (a_1', a_2')'$ y $b = (b_1', b_2')'$ en $\mathbb{R}^{p+q}$, decimos que $a \perp_\Upsilon b$ cuando $a_1' \Sigma_{yy} b_1 = a_2' \Sigma_{zz} b_2 = 0$, es decir,

$$a \perp_\Upsilon b \Leftrightarrow \Upsilon a \perp \Upsilon b.$$

Si A y B son subespacios vectoriales de $\mathbb{R}^{p+q}$, se denota $A \perp_\Upsilon B$ cuando $a \perp_\Upsilon b$ para todo $a \in A$ y $b \in B$. Obviamente, $A \perp_\Upsilon B$ implica $A \cap B = 0$. Por último, $A \oplus_\Upsilon B$ denotará la suma directa de ambos subespacios siempre y cuando se verifique además que $A \perp_\Upsilon B$. En estas condiciones, podemos añadir al lema anterior obtenemos el resultado siguiente.

Lema 13.15.
Si $\lambda_i^2 \neq \lambda_j^2$, entonces $\mathtt{ker} A_{\lambda_i} \perp_\Upsilon \mathtt{ker} A_{\lambda_j}$.

Demostración.
Supongamos primeramente que $\lambda_i, \lambda_j \neq 0$. Si $(\alpha_i', \beta_i')' \in \mathtt{ker} A_{\lambda_i}$ y $(\alpha_j', \beta_j')' \in \mathtt{ker} A_{\lambda_j}$, entonces se verifica

$$\begin{aligned} \beta_i &= \frac{1}{\lambda_i} \Sigma_{zz}^{-1} \Sigma_{zy} \alpha_i \\ \lambda_i \Sigma_{yy} \alpha_i &= \Sigma_{yz} \beta_i. \end{aligned}$$

Multiplicando a la izquierda por α_j' en la segunda ecuación y sustituyendo el valor de β_i según la primera, se tiene que

$$\lambda_i^2 \alpha_j' \Sigma_{yy} \alpha_i = \alpha_j' \Sigma_{yz} \Sigma_{zz}^{-1} \Sigma_{zy} \alpha_i.$$

Análogamente,

$$\lambda_j^2 \alpha_i' \Sigma_{yy} \alpha_j = \alpha_i' \Sigma_{yz} \Sigma_{zz}^{-1} \Sigma_{zy} \alpha_j.$$

De ambas ecuaciones se deduce que

$$(\lambda_i^2 - \lambda_j^2) \alpha_j' \Sigma_{yy} \alpha_i = 0,$$

luego $\alpha_j' \Sigma_{yy} \alpha_i = 0$. Por una razonamiento simétrico, se obtiene también que $\beta_j' \Sigma_{zz} \beta_i = 0$.

Supongamos ahora que $\lambda_j = 0$. Si $(\alpha_0', \beta_0')' \in \mathtt{ker}(A_0)$, entonces $\Sigma_{zy}\alpha_0 = 0$. Aplicando la primera parte del razonamiento anterior se tiene que

$$\lambda_i^2 \alpha_0' \Sigma_{yy} \alpha_i = \alpha_0' \Sigma_{yz} \Sigma_{zz}^{-1} \Sigma_{zy} \alpha_i = 0$$

y, por tanto, $\alpha_0' \Sigma_{yy} \alpha_i = 0$. Por un razonamiento simétrico, se verifica $\beta_0' \Sigma_{zz} \beta_0 = 0$. ∎

Supongamos que existen k raíces de $P(\lambda)$ estrictamente positivas ($k \leq p$) y que, de esas k raíces, existen r distintas ($r \leq k$), $\lambda_{i_1} > \ldots > \lambda_{i_r}$, con multiplicidades $d_1, \ldots, d_r$, respectivamente ($d_1 + \ldots + d_r = k$). Si $d_0 = \mathtt{mult}_P(0)$. Se tiene pues que $p + q = d_0 + 2\sum_{j=1}^{r} d_j$. En esas condiciones, de los lemas anteriores se deduce el siguiente resultado:

Teorema 13.16.
$\mathbb{R}^{p+q}$ descompone de la siguiente forma:

$$\mathbb{R}^{p+q} = \left[(\mathtt{ker} A_{\lambda_{i_1}}) \oplus (\mathtt{ker} A_{-\lambda_{i_1}})\right] \oplus_{\Upsilon} \ldots \oplus_{\Upsilon} \left[(\mathtt{ker} A_{\lambda_{i_r}}) \oplus (\mathtt{ker} A_{-\lambda_{i_r}})\right] \oplus_{\Upsilon} \left[\mathtt{ker} A_0\right].$$

Hecho esto, vamos a construir las variables canónicas. Descompondremos este proceso en tres partes:

1. Nuestro primer propósito es, según hemos comentado anteriormente, encontrar dos vectores $\alpha \in \mathbb{R}^p$ y $\beta \in \mathbb{R}^q$, tales que el coeficiente de correlación lineal simple $\rho_{\alpha' Y, \beta' Z}$ alcance un valor máximo. Dado que el coeficiente ρ es invariante ante homotecias en ambas variables, bastará con encontrar

$$\text{máx}\left\{\rho_{\alpha' Y, \beta' Z} \colon \mathtt{var}[\alpha' Y] = \mathtt{var}[\beta' Z] = 1\right\},$$

es decir,

$$\text{máx}\left\{\alpha' \Sigma_{yz} \beta \colon \alpha' \Sigma_{yy} \alpha = \beta' \Sigma_{zz} \beta = 1\right\}.$$

Dado que se trata de maximizar una función continua sobre un compacto, el máximo existe. Supongamos que se alcanza en $(\alpha_1', \beta_1')' \in \mathbb{R}^{p+q}$, y consideremos las funciones siguientes:

$$\phi \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \alpha' \Sigma_{yz} \beta, \qquad f \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \alpha' \Sigma_{yy} \alpha - 1 \\ \beta' \Sigma_{zz} \beta - 1 \end{pmatrix}.$$

Se trata pues de maximizar ϕ bajo la condición $f \equiv 0$. Dado que el máximo se alcanza en $(\alpha_1', \beta_1')'$, se verifica, por el teorema 13.3, que existen $\rho_1, \overline{\rho}_1 \in \mathbb{R}$ (únicos), tales que

$$\bigtriangledown \left(\phi - (\rho_1, \overline{\rho}_1) f\right) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0,$$

o equivalentemente, que existen $\rho_1, \overline{\rho}_1 \in \mathbb{R}$ (únicos), tales que

$$\left\{ \begin{array}{rcl} \Sigma_{yz}\beta_1 - \rho_1\Sigma_{yy}\alpha_1 & = & 0 \\ \Sigma_{zy}\alpha_1 - \overline{\rho}_1\Sigma_{zz}\beta_1 & = & 0 \end{array} \right. , \tag{13.7}$$

lo cual equivale a afirmar

$$\left\{ \begin{array}{rcl} \Sigma_{yz}\beta_1 & = & \rho_1\Sigma_{yy}\alpha_1 \\ \Sigma_{zy}\alpha_1 & = & \overline{\rho}_1\Sigma_{zz}\beta_1 \end{array} \right. \tag{13.8}$$

Multiplicando a la izquierda por α_1' y β_1' en (13.7), se obtiene

$$\left\{ \begin{array}{rcl} \alpha_1'\Sigma_{yz}\beta_1 - \rho_1\alpha_1'\Sigma_{yy}\alpha_1 & = & 0 \\ \beta_1'\Sigma_{zy}\alpha_1 - \overline{\rho}_1\beta_1'\Sigma_{zz}\beta_1 & = & 0 \end{array} \right. \Leftrightarrow \quad \left\{ \begin{array}{rcl} \alpha_1'\Sigma_{yz}\beta_1 & = & \rho_1 \\ \beta_1'\Sigma_{zy}\alpha_1 & = & \overline{\rho}_1 \end{array} \right. ,$$

es decir,

$$\rho_1 = \overline{\rho}_1 = \alpha_1'\Sigma_{yz}\beta_1 = \rho_{\alpha_1'Y,\beta_1'Z}.$$

Por tanto, (13.7) equivale a

$$\left(\begin{array}{cc} \rho_1\Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \rho_1\Sigma_{zz} \end{array} \right) \left(\begin{array}{c} \alpha_1 \\ \beta_1 \end{array} \right) = \left(\begin{array}{c} 0 \\ 0 \end{array} \right),$$

es decir,

$$A_{\rho_1} \left(\begin{array}{c} \alpha_1 \\ \beta_1 \end{array} \right) = \left(\begin{array}{c} 0 \\ 0 \end{array} \right).$$

Por lo tanto, $(\alpha_1', \beta_1')' \in \mathtt{ker} A_{\rho_1}$. Luego, $|A_{\rho_1}| = 0$. Entonces,

$$\rho_1 \in \{\lambda_1, \ldots, \lambda_{p+q}\}.$$

Si se denota $U_1 = \alpha_1'Y$ y $V_1 = \beta_1'Z$, se tiene que $\rho_{U,V}$ es máxima.

2. Consideremos todos los vectores $\alpha \in \mathbb{R}^p$ y $\beta \in \mathbb{R}^q$ tales que

$$\mathtt{cov}[\alpha'Y, U_1] = \mathtt{cov}[\beta'Z, V_1] = 0,$$

lo cual, por (13.8), equivale a afirmar lo siguiente:

$$\mathtt{Cov}\left[\left(\begin{array}{c} \alpha'Y \\ \beta'Z \end{array} \right), \left(\begin{array}{c} U_1 \\ V_1 \end{array} \right) \right] = 0.$$

Nuestro siguiente objetivo es encontrar la máxima correlación dentro de esta clase, es decir,

$$\text{máx}\left\{ \alpha'\Sigma_{yz}\beta \colon \alpha'\Sigma_{yy}\alpha = \beta'\Sigma_{zz}\beta = 1 \ \wedge \ \alpha'\Sigma_{yy}\alpha_1 = \beta'\Sigma_{zz}\beta_1 = 0 \right\}.$$

Supongamos que el máximo se alcanza en $(\alpha_2', \beta_2')'$ y consideremos ϕ como antes y f definida mediante

$$f\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}\alpha'\Sigma_{yy}\alpha_1\\ \beta'\Sigma_{zz}\beta_1\\ \alpha'\Sigma_{yy}\alpha - 1\\ \beta'\Sigma_{zz}\beta - 1\end{pmatrix}.$$

Por el teorema 13.3 se verifica que existen $\rho_2, \overline{\rho}_2, \theta, \gamma \in \mathbb{R}$ (únicos) tales que

$$\begin{cases} \Sigma_{yz}\beta_2 - \rho_2\Sigma_{yy}\alpha_2 - \theta\Sigma_{11}\alpha_1 &= 0\\ \Sigma_{zy}\alpha_2 - \overline{\rho}_2\Sigma_{zz}\beta_2 - \gamma\Sigma_{22}\beta_1 &= 0\end{cases}. \tag{13.9}$$

Multiplicando a la izquierda por α_1' en la primera ecuación, se tiene

$$\alpha_1'\Sigma yZ\beta_2 - \rho_2\alpha_1'\Sigma_{yy}\alpha_2 - \theta\alpha_1'\Sigma_{11}\alpha_1 = 0,$$

es decir, $\theta = 0$. Igualmente, multiplicando por β_1' la segunda se deduce que $\gamma = 0$. Luego, (13.9) implica

$$\begin{cases} \Sigma_{yz}\beta_2 - \rho_2\Sigma_{yy}\alpha_2 &= 0\\ \Sigma_{zy}\alpha_2 - \overline{\rho}_2\Sigma_{zz}\beta_2 &= 0\end{cases}.$$

Por un razonamiento análogo al del primer paso, se concluye que $\rho_2 = \overline{\rho}_2$, coincidiendo éstos con $\rho_{\alpha_2'Y,\beta_2'Z}$, que es la correlación máxima bajo las condiciones anteriores. Además, $(\alpha_2', \beta_2')' \in \mathtt{ker}\big(A_{\rho_2}\big)$ y $\rho_2 \in \{\lambda_1, \ldots, \lambda_{p+q}\}$. Por supuesto, $\rho_2 \leq \rho_1$.

3. Supongamos que el proceso se ha completado m veces, es decir, hemos obtenido dos matrices, $A = (\alpha_1, \ldots, \alpha_m) \in \mathcal{M}_{p\times m}$ y $B = (\beta_1, \ldots, \beta_m) \in \mathcal{M}_{q\times m}$, y m números no negativos $\rho_1 \geq \ldots \geq \rho_m$, tales que, si consideramos la matriz $m \times 2$ aleatoria

$$\begin{pmatrix} U_1 & \ldots & U_m\\ V_1 & \ldots & V_m\end{pmatrix} = \begin{pmatrix} A'Y\\ B'Z\end{pmatrix},$$

se verifica entonces

$$\mathtt{var}[U_i] = \mathtt{var}[V_i] = 1, \quad (\alpha_i', \beta_i')' \in \mathtt{ker}A_{\rho_i}, \quad i = 1, \ldots, m.$$

Además,

$$\begin{aligned}\rho_1 &= \text{máx}\{\rho_{\alpha'Y,\beta'Z}\}\\ \rho_i &= \text{máx}\left\{\rho_{\alpha'Y,\beta'Z} \colon \mathtt{Cov}\left[\begin{pmatrix}\alpha'Y\\ \beta'Z\end{pmatrix}, \begin{pmatrix}U_j\\ Vj\end{pmatrix}\right] = 0,\ \forall j < i\right\},\ i = 2, \ldots, m.\end{aligned}$$

Nótese que se afirma de manera implícita que $\rho_{U_i,V_i} = \rho_i$, $i = 1, \ldots, m$. Se verifica también que

$$m = \mathtt{rg}\,(\mathtt{Id}_{m\times m}) = \mathtt{rg}(A'\Sigma_{yy}A) \leq \mathtt{rg}(\Sigma_{yy}) = p,$$

es decir, $m \leq p$. Vamos a demostrar que, si $m < p$, el proceso puede completarse hasta p. Para ello consideremos los vectores

$$u_i = \Sigma_{yy}^{-1/2}\alpha_i, \quad v_i = \Sigma_{zz}^{-1/2}\beta_i, \quad i = 1, \ldots, m.$$

En esas condiciones, $\{u_1, \ldots, u_m\}$ y $\{v_1, \ldots, v_m\}$ constituyen sendos sistema ortonormales de $\mathbb{R}^p$ y $\mathbb{R}^q$, resp. Si $m < p$, pueden añadirse otros dos vectores, u_{m+1} y v_{m+1} a los sistemas respectivos. Considerando entonces los vectores

$$\alpha = \Sigma_{yy}^{-1/2}u_{m+1}, \qquad \beta = \Sigma_{zz}^{-1/2}v_{m+1},$$

se verifica que

$$\mathtt{Cov}\left[\begin{pmatrix} \alpha'Y \\ \beta'Z \end{pmatrix}, \begin{pmatrix} U_i \\ Vi \end{pmatrix}\right] = 0, \qquad \forall j = 1, \ldots, m. \tag{13.10}$$

Luego,el proceso puede continuar, buscando la máxima correlación para los vectores α y β que verifiquen (13.10). Es decir, que culmina en p, obteniéndose las parejas de variables aleatorias $(U_1, V_1), \ldots, (U_p, V_p)$, con correlaciones $\rho_1 \geq \ldots \geq \rho_p \geq 0$, respectivamente. Más concretamente, se obtiene la siguiente matriz de covarianzas:

$$\mathtt{Cov}\left[\begin{pmatrix} U_1 & \ldots & U_p \\ V_1 & \ldots & V_p \end{pmatrix}\right] = \begin{pmatrix} 1 & & 0 & \rho_1 & & 0 \\ & \ddots & & & \ddots & \\ 0 & & 1 & 0 & & \rho_p \\ \rho_1 & & 0 & 1 & & 0 \\ & \ddots & & & \ddots & \\ 0 & & \rho_p & 0 & & 1 \end{pmatrix}.$$

A continuación, vamos a analizar la relación existente entre los coeficientes $\rho_1, \ldots, \rho_p$ así obtenidos y las raíces $\lambda_1, \ldots, \lambda_{p+q}$ del polinomio $P(\lambda)$. Ya sabemos que

$$\{\rho_1, \ldots, \rho_p\} \subset \{\lambda_1, \ldots, \lambda_{p+q}\}.$$

En primer lugar, $\lambda_1 = \rho_1$ pues, en caso contrario, consideraríamos $(\alpha', \beta')' \in \mathtt{ker}A_{\lambda_1}$. Entonces se verificaría

$$\begin{cases} \alpha'\Sigma_{yz}\beta &= \lambda_1\alpha'\Sigma_{yy}\alpha \\ \beta'\Sigma_{zy}\alpha &= \lambda_1\beta'\Sigma_{zz}\beta \end{cases},$$

es decir,

$$\lambda_1 \mathtt{var}[\alpha' Y] = \lambda_1 \mathtt{var}[\beta' Z] = \mathtt{cov}[\alpha' Y, \beta' Z].$$

Luego, ello implicaría

$$\rho_{\alpha' Y, \beta' Z} = \lambda_1 > \rho_1,$$

lo cual es contradictorio.

Si $\mathtt{mult}_P(\lambda_1) = 1$, entonces $\lambda_2 = \rho_2$, pues, en caso contrario, consideraríamos $(\alpha', \beta')' \in \mathtt{ker} A_{\lambda_2}$. Recordemos que, en ese caso, por el teorema 13.16,

$$\mathtt{ker} A_{\lambda_1} \perp_{\Upsilon} \mathtt{ker} A_{\lambda_2}.$$

Entonces, razonaríamos como en el caso anterior.

Si $\mathtt{mult}_P(\lambda_1) = d_1 > 1$, entonces $\rho_1 = \ldots = \rho_{d_1} = \lambda_1$ y $\rho_{d_1+1} = \lambda_{d_1+1}$: en efecto, sabemos por el lema 13.14 que $\mathtt{dim}(\mathtt{ker} A_{\lambda_1}) = d_1$. Consideremos entonces una base ortonormal $\{e_i \colon i = 1, \ldots, d_1\}$ del subespacio imagen por Υ de $\mathtt{ker} A_{\lambda_1}$, y un sistema de vectores de $\mathtt{ker} A_{\lambda_1}$, $\{(\alpha_i', \beta_i')' \colon i = 1, \ldots, d_1\}$, tales que

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} = \Upsilon^{-1} e_i, \quad i = 1, \ldots, d_1,$$

que constituye una base de $\mathtt{ker} A_{\lambda_1}$ y verifica , además,

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \perp_{\Upsilon} \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix}, \quad \forall i \neq j.$$

Se verifica pues que

$$\rho_{\alpha_i' Y, \beta_i' Z} = \lambda_1, \quad i = 1, \ldots, d_1.$$

Por lo tanto, $\rho_1 = \ldots = \rho_{d_1} = \lambda_1$. Además, ρ_{d+1} ha de ser distinto de λ_1 pues, en caso contrario, $\mathtt{dim}(\mathtt{ker} A_{\lambda_1}) > d_1$. Por un razonamiento idéntico al primero, se concluye que $\rho_{d+1} = \lambda_{d+1}$.

El proceso se repite de manera análoga hasta el final. Se verifica además que, si $(\alpha_i', \beta_i')'$ está asociado a ρ_i, entonces $(-\alpha_i', \beta_i')'$ lo está a $-\rho_i$. En conclusión, hemos obtenido el resultado siguiente:

Proposición 13.17.
Para cada $i = 1, \ldots, p$, se verifica

$$\rho_i = \lambda_i, \quad -\rho_i = \lambda_{p+q-i}.$$

Hasta ahora, hemos supuesto $p \leq q$. En el caso $p > q$, procederíamos, teniendo en cuenta que los vectores Z e Y desempeñan papeles completamente simétricos en lo que al problema de correlación lineal se refiere, a considerar los q autovalores de la matriz (13.6), que coincidirían con los q primeros autovalores de la matriz (13.4) (el resto son nulos). Por lo tanto, en general, si se denota $b = \text{mín}\{p, q\}$, hemos de calcular los b primeros autovalores de la matriz (13.4) con sus respectivos autovectores, además de los autovectores asociados a esos mismos autovalores respecto a la matriz (13.6). Luego el proceso de obtención de variables canónicas culminará en el paso b-ésimo. Recapitulando, hemos demostrado lo siguiente:

Teorema: Sean $\rho_1^2 \ldots, \rho_b^2$ son los b primeros autovalores de la matriz $\Sigma_{yy}^{-1}\Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy}$, contados con su multiplicidad, y consideremos, para cada $i = 1, \ldots, b$, las variables reales $U_i = \alpha_i' Y$ y $V_i = \beta_i' Z$, donde α_i es el autovector asociado al autovalor ρ_i^2 para la matriz anterior tal que $\texttt{var}(U_i) = 1$, mientras que β_i es el autovector asociado al mismo autovalor para la matriz $\Sigma_{zz}^{-1}\Sigma_{zy}\Sigma_{yy}^{-1}\Sigma_{yz}$, y tal que $\texttt{var}(V_i) = 1$. Se verifica entonces:

(i) $\texttt{var}[U_i] = \texttt{var}[V_i] = 1,\ i = 1, \ldots, b,$

(ii) $\texttt{Cov}\left[\begin{pmatrix} U_i \\ V_i \end{pmatrix}, \begin{pmatrix} U_j \\ Vj \end{pmatrix}\right] = 0,\ \forall i \neq j.$

(iii) $\rho_{U_i,V_i} = \rho_i,\ i = 1, \ldots, b.$

(iv) ρ_1 es la máxima correlación entre una variable del tipo $\alpha' Y$ y otra del tipo $\beta' Z$.

(v) Si $1 < i \leq b$, ρ_i es la máxima correlación entre entre una variable del tipo $\alpha' Y$ y otra del tipo $\beta' Z$, si imponemos la condición

$$\texttt{Cov}\left[\begin{pmatrix} \alpha' Y \\ \beta' Z \end{pmatrix}, \begin{pmatrix} U_j \\ Vj \end{pmatrix}\right] = 0,\ \forall j < i.$$

Nótese que, si la multiplicidad de ρ_i es 1, el par (U_i, V_i) está bien determinado salvo el signo. En caso contrario, la indeterminación es más complicada.

Bibliografía

- **S.F. Arnold.** *The Theory of Linear Models and Multivariate Analysis.* Wiley (1981).
- **T. W. Anderson.** *An Introduction to Multivariate Statistical Analysis.* Wiley (1958).
- **R.B. Ash.** *Real Analysis and Probability.* Academic Press (1972).
- **M. Bilodeau & D.Brenner.** *Theory of Multivariate Statistics.* Springer (1999)
- **W.R. Dillon & M. Goldstein.** *Multivariate Analysis. Methods and Aplications.* Wiley (1984).
- **T.S. Ferguson.** *A Course in Large Sample Theory.* Chapman & Hall (1996).
- **B. Flury.** *A First Course in Multivariate Statistics.* Springer (1997).
- **M.J. Greenacre.** *Theory and Applications of Correspondence Analysis.* Academic Press (1984).
- **J.F. Hair, R.E. Anderson R.L. Tatham & C.B. Black.** *Análisis Multivariante.* Prentice Hall (1999).
- **A.C. Lehmann.** *Testing Statistical Hypotheses.* Wiley (1986)
- **A.C. Lehmann.** *Elements of Large Sample Theory.* Springer (1998)
- **Mardia, Kent & Bibby.** *Multivariate Analysis.* Academic Press (1979).
- **A.G. Nogales.** *Estadística Matemática.* Uex (1998).
- **E.S. Pearson & H.O. Heartley.** *Biometrika Tables for Staticians, Volumen 2.* Cambridge University Press.

- **D. Peña & S. Rivera.** *Estadística. Modelos y métodos (partes I y II).* Alianza Editorial (1986).
- **J.O. Rawlings, S.G. Pantula & D.A. Dickey.** *Applied Regression Analysis.* Springer (1998).
- **A.C. Rencher.** *Methods of Multivariate Analysis.* John Wiley & Sons (1995).
- **B.W. Silverman.** *Density Estimation for Statistics and Data Analysis.* Chapman & Hall (1986).
- **E. Uriel, J. Aldás.** *Análisis Multivariante Aplicado.* Thomson (2005).

Índice alfabético

índices de condicionamiento, 179

análisis cluster o de conglomerados, 259
análisis de correspondencias, 184
análisis de la covarianza, 135
análisis de perfiles, 116
análisis de redundancia, 147
análisis discriminante, 112, 144
análisis factorial, 239
anova, 105
autovalor, 265
autovector, 265
autucia de Cramer-Wold, 82

biplot, 190, 195

cargas canónicas, 146
cargas canónicas cruzadas, 147
centroides, 263
Cf. volumen dedicado a los Modelos Lineales., 244
coeficiente de contingencia de Pearson, 186
coeficiente de corrección de Barlett, 95, 99
coeficiente de correlación canónica, 135
coeficiente de correlación múltiple, 138
coeficiente de correlación simple, 137
coeficiente de determinación, 138
coeficientes de correlación canónica, 97, 129, 140, 203
colinealidad, 175
comparaciones múltiples, 115
completitud, 57
componentes principales, 155
comunalidades, 252
condición de Huber, 90, 107, 110, 114, 134
consistencia, 82, 85
contraste de dos medias, 108
contraste de hipótesis generalizado, 92
contraste para una media, 106
contraste parcial para variables respuesta, 131
contrastes parciales, 130
convergencia en distribución, 79
convergencia en probabilidad, 79

dendrograma, 261
densidad de una normal multivariante, 22
desigualdad de Bonferroni, 158
desigualdad de Cauchy-Schwarz, 77
desigualdad de Chebichev, 157
diagonalización de una matriz simétrica, 267
distancia χ^2, 186
distancia de Mahalanobis, 24, 120, 134, 218, 219, 227, 230
distancia euclídea, 227, 266
distribución χ^2, 25
distribución F-Snedecor, 27
distribución t-Student, 28
distribución T^2 de Hotelling, 48
distribución a priori, 213
distribución Beta, 27
distribución binomial, 236

distribución condicional de la normal matricial, 37
distribución condicional de una normal multivariante, 22
distribución de Poisson, 26
distribución de Wishart, 43
distribución normal degenerada, 22
distribución normal matricial, 33
distribución normal multivariante, 20
distribución normal multivariante esférica, 24

EIMV, 106, 125, 230
ejes discriminantes, 144, 200, 227
ejes principales, 187
EMV, 59, 138, 139, 141
estadístico completo, 57
estadístico de contraste, 73
estadístico de la razón de verosimilitudes, 69
estadístico suficicente, 57
estimador, 57
estimador de máxima verosimilitud, 59
estimador insesgado, 59
estimador insesgado de mínima varianza, 59
estimador sesgado, 182
estrategia admisible, 213
estrategia cuadrática, 230
estrategia de bayes, 214
estrategia lineal, 230
estrategia minimax, 213
estrategia no aleatoria, 212
experimento estadístico con razón de verosimilitudes monótona, 68
exponencial, 58

factor, 250
factor de inflación de la varianza, 176
familia de estrategias completa, 213
familia de estrategias completa minimal, 213
familia de intervalos de confianza simultán 76
familia de Scheffé, 76, 115
familia exponencial, 58
FIV, 176, 178
función característica de una matriz aleatoria, 31
función de densidad de la normal matricial, 37
función de distribución empírica, 29
función de pérdida o costes, 212
función de riesgo, 212
función logística, 235

grado de libertad, 26

Huber, 133

IC, 179
incorrelación, 21
independencia, 21
independencia condicional, 45, 132
inercia, 185
información, 62
invariante maximal, 63
invarianza, 25

jacobiano, 36

KMO, 244
kurtosis, 96

lema fundamental de Neyman-Pearson, 68
ley débil de los grandes números, 85

m.a.s., 106

método backward, 132
método de clasificación de Fisher, 211
método de componentes principales, 245
método de k medias o quick cluster, 262
método de los ejes principales, 253
método de sustitución, 73, 223
método de unión-intersección, 75
método de validación cruzada, 231
método del núcleo, 233
método Delta, 159
método forward, 133
método jacknife, 231
método jerárquico, 261
método Lambda de Wilks, 207
método stepwise, 133
maldición de la dimensión, 234
mancova, 134
manova, 112, 198
mapa territorial, 228
matriz $\mathbf{X}$ de regresión, 122
matriz $\mathbf{Z}$ de regresión, 122
matriz aeatoria, 30
matriz de covarianzas muestral, 60
matriz de cargas principales o componentes, 171
matriz de componentes o cargas principales, 170
matriz de correlaciones reproducida, 244
matriz de covarianzas de una matriz aleatoria, 32
matriz de covarianzas parciales, 22
matriz de estructura, 206, 228
matriz de ponderaciones, 170
matriz de una proyección ortogonal, 270
matriz definida positiva, 266
matriz idempotente, 270
matriz ortogonal, 266
matriz semidefinida positiva, 266
media de una matriz aleatoria, 32
modelo asintótico, 80
modelo completo, 130
modelo de correlación lineal, 123
modelo exacto, 80
modelo lineal normal multivariante, 56
modelo reducido, 130
modelos de crecimiento, 116
multicolinealidad, 175

norma euclídea, 266

orden de lectura de una matriz, 32

P-P plot, 29
perfiles de columna, 185
perfiles de fila, 185
ponderaciones canónicas, 145
ponderaciones discriminante, 205
principio de invarianza, 62
principio de máxima verosimilitud, 211
problema de clasificación, 211
producto de Kronecker, 32
producto interior de dos matrices, 31
producto interior de dos vectores, 265
proporciones de la varianza, 180
proyección ortogonal, 270
proyección ortogonal sobre una subvariedad afín, 163, 271
puntuaciones discriminantes, 200

Q-Q plot, 29

reducción por suficiencia, 62
región de confianza, 107, 136
regiones de confianza, 93
regresión lineal múltiple, 121
regresión lineal multivariante, 121
regresión lineal simple, 121

regresión logística, 235
riesgo de Bayes, 213
rotación, 269
rotación cuartimax, 243
rotación varimax, 243
RV, 69

selección de variables, 132, 207
selector de ancho de banda, 234
semilla de conglomerado, 262
sesgo, 182
sistema ortonormal, 266
subespacio V/W, 266
suficiencia, 57
SVD, 165

tabla de contingencia, 184
teorema central del límite, 82, 107
teorema de descomposición en valores singulares, 165
teorema de diagonalización, 152
teorema de Glivenko-Cantelli, 29
teorema de Lehmann-Scheffé, 59
teorema de los multiplicadores finitos de Lagrange, 267
teorema del cambio de variables, 36
test F, 68
test M de Box, 99, 110, 229
test de Barlett, 99
test de correlación, 96, 128
test de esfericidad de Barlett, 101
test de Hotelling, 107, 109, 118, 130
test de la razón de verosimilitudes, 69, 107, 110
test de Lawley-Hotelling, 74
test de Pillay, 74, 130, 144
test de Roy, 75, 115
test de Snedecor, 104
test de Student, 110
test de Welch, 110
test de Wilks, 71, 130, 132
test insesgado, 68, 95
test UMP, 68
test UMP-invariante, 62, 68, 107, 110, 114, 117, 118
tests de normalidad multivariante, 28
tipificación, 146, 170, 178, 206, 250, 259
TMFL, 267
transformaciones de Box-Cox, 42
traza de una matriz cuadrada, 30
TRV, 69

valores teóricos, 137
variable explicativa, 121
variable latente, 250
variable repuesta, 121
variables canónicas, 140
variables ficticias, 134
varianza específica, 252
varianza generalizada, 52
varianza parcial, 138
varianza total, 52, 158, 160
vecino más lejano, 262
vecino más próximo, 262
vectores ortogonales, 266
versión coordenada del modelo lineal normal multivariante, 56
versión coordenada del modelo lineal normal multivariante, 122
vinculación intergrupos, 262
vinculación intragrupos, 262