Estadística multivariada: inferencia y métodos

Estadística multivariada: inferencia y métodos

Luis Guillermo Díaz Monroy

Departamento de Estadística
Facultad de Ciencias

Universidad Nacional de Colombia
Sede Bogotá

Estadística multivariada: inferencia y métodos

Departamento de Estadística
Facultad de Ciencias
Universidad Nacional de Colombia

Ignacio Mantilla, Decano
Eugenio Andrade, Vicedecano Académico

Jorge Ortiz Pinilla, Director de Publicaciones

Segunda edición, 2007
Bogotá, Colombia

ISBN 978-958-701-195-1

Impresión: Proceditor ltda.
proceditor@etb.net.co
Bogotá, Colombia

Diagramación: Margoth Hernández Quitián sobre original en LaTeX del autor
Diseño de carátula: Andrea Kratzer

Catalogación en la publicación Universidad Nacional de Colombia

Díaz Monroy, Luis Guillermo, 1958 –
Estadística multivariada : inferencia y métodos / Luis Guillermo Díaz Monroy. –
Bogotá : Universidad Nacional de Colombia. Facultad de Ciencias, 2007
xvii, 570 p.

ISBN 978-958-701-195-1

1. Análisis estadístico multivariable 2. Diseño experimental 3. Estadística matemática 4. Probabilidades

CDD-21 519.535 / 2007

A: María del Pilar,

María Camila,

Daniel Felipe y

Diego Alejandro

Mis componentes principales

Contenido

Índice de figuras

Índice de tablas

Introducción

La estadística multivariada, gracias a los avances de la computación, es un conjunto de técnicas ampliamente demandadas y usadas por estudiantes y profesionales de diversas áreas de la ciencia, las artes y la tecnología. La acogida que tuvo la primera edición de *Estadísistica multivariada: inferencia y métodos* me motiva y compromete a una nueva edición.

La segunda edición tiene en cuenta al lector para quien se pensó debe ir dirigido, es decir cualquier persona que posea concimientos básicos de matemáticas y estadística. Aunque se ha mantenido la estructura de la primera edición, ésta ha sido sometida a una revisión exhaustiva, cuyo resultado ha permitido la detección y corrección de algunas ambigüedades, la corrección de errores ortográficos y de edición, los cuales fueron advertidos al autor por juiciosos lectores. Algunos ejemplos fueron desarrollados con más detalle y sencillez. Para cada uno de los once capítulos y los dos anexos se ha incluido la sintaxis del paquete estadístico R, con el cual se desarrollan los cálculos, las tablas y las gráficas de algunos de los ejemplos contenidos en el respectivo capítulo.

Quiero expresar mis sentimientos de gratitud a mis alumnos y colegas del Departamento Estadística de la Universidad Nacional de Colombia y fuera de éste, quienes han colaborado con la corrección y orientación de la primera edición. Un reconocimiento especial para mi amigo y alumno el profesor Mario Alfonso Morales Rivera por la elaboración de los programas en R, al profesor Jorge Ortiz Pinilla, coordinador de Publicaciones de la Facultad de Ciencias y a Margoth Hernández Quitián, por su valiosa asistencia en el procesamiento del texto en LaTeX .

Este texto es una de las contribuciones académicas del grupo de investigación en Estadística Aplicada a la Investigación Experimental, Industria y Biotecnología.

Luis Guillermo Díaz Monroy

Parte I

Inferencia

Capítulo 1

Conceptos preliminares

1.1 Introducción

En este capítulo se mencionan algunos de los campos donde se usa y demanda la estadística multivariada, se hace una presentación descriptiva y exploratoria tanto de información multivariada como de algunas metodologías. También se presenta la caracterización probabilística de un vector aleatorio junto con los parámetros de localización, dispersión y asociación.

La información estadística proviene de respuestas o atributos, las cuales son observadas o medidas sobre un conjunto de individuos u objetos, referenciados generalmente en un espacio y un tiempo. Cada respuesta o atributo está asociado con una *variable*[1]; si tan sólo se registra un atributo por individuo, los datos resultantes son de tipo *univariado*, mientras que si más de una variable es registrada sobre cada objeto, los datos tienen una estructura *multivariada*. Aun más, pueden considerarse *grupos* de individuos, de los cuales se obtienen muestras de datos multivariados para comparar algunas de sus características o parámetros. En una forma más general, los datos multivariados pueden proceder de varios grupos o poblaciones de objetos; donde el interés se dirige a la exploración de las variables y la búsqueda de su interrelación dentro de los grupos y entre ellos.

Los valores que cualquier variable pueda tomar están, en su mayoría, en alguno de los niveles o escala de medición usuales; a saber: *nominal*, *ordinal*, *intervalo* o de *razón*. Una clasificación más útil es la de variables en escala *métrica* (*cuantitativa*) y la *no métrica* (*cualitativa o categórica*);

[1] La cual hace "visible" un concepto que se inscribe dentro de un marco teórico específico.

algunas técnicas multivariadas exigen más precisión respecto a la escala de medición de la variable. Al finalizar la sección se describen estas escalas de medición.

A riesgo de incurrir en omisión, a continuación se muestra un listado de casos sobre algunos campos del conocimiento, donde se requiere de técnicas multivariadas para el análisis o la exploración de datos.

Mercadeo

Se estudian seis características acerca de un producto percibidas por un grupo de consumidores, éstas son: calidad del producto, nivel de precio, velocidad de despacho o entrega, servicio, nivel de uso comparado con otros porductos sustitutos, nivel de satisfacción. Se quiere saber acerca de la incidencia, tanto individual como conjunta, de las variables anteriores en la decisión de compra del producto.

Geología

A lo largo de líneas transversales (en inglés "transects") toman varias muestras del suelo para estudiar los contenidos (en porcentaje) de arena, azufre, magnesio, arcilla, materia orgánica y pH. También se miden otras variables físicas tales como estructura, humedad, conductividad eléctrica y permeabilidad. El objetivo es determinar las características más relevantes del suelo y hacer una clasificación de éstos.

Psicología

A un grupo de jóvenes recién egresados de la educación media, se les registran las siguientes variables sicológicas: información, habilidad verbal, analogías verbales, intensidad del ego, ansiedad, memoria y autoestima. Se pretende encontrar unos pocos factores que den cuenta de estas variables.

Arqueología

Se realizan varias excavaciones en tres regiones donde se tiene la evidencia que habitaron comunidades indígeneas diferentes. Sobre los cráneos conseguidos se midió: la circunferencia, ancho máximo, altura máxima, altura nasal y longitud basialveolar. Esta información permitirá hacer comparaciones entre estas comunidades.

Medicina

Se considera el problema de distinguir entre "éxito" y "falla" de la efectividad de tratamientos aplicados sobre mujeres que padecen cáncer de mama, usando una variedad de indicadores de diagnóstico.

Antropología

Con base en algunas mediciones realizadas en algunos huesos pertenecientes a un cadáver, se quiere construir un modelo estadístico con el cual se pueda predecir el sexo, la edad, el grupo étnico, etc, de un individuo.

Biología

Con base en las medidas recogidas sobre varias plantas arbustivas, tales como: altura, área foliar, longitud de raíz, área basal, área radicular, biomasa, textura del tronco y textura de las hojas, se quiere hacer una clasificación de éstas.

Sociología

Se quiere establecer la relación entre diferentes tipos de crímenes y algunas variables socio-demográficas como: población, población económicamente activa, oferta de empleo, tipos de credos religiosos, credos políticos, índice de servicios públicos e índices de escolaridad.

▶ Escalas de medición

Se denomina *escalamiento* al desarrollo de reglas sistemáticas y de unidades significativas de medida para identificar o cuantificar las observaciones empíricas. La clasificación más común distingue cuatro conjuntos de reglas básicas que producen cuatro escalas de medida; éstas son:

- La escala de medida más simple implica una relación de identidad entre el sistema de números y el sistema empírico objeto de medida. La escala resultante se denomina *nominal*, porque los números empleados se consideran como "etiquetas" las cuales se asignan a los objetos con el propósito de clasificarlos, pero no poseen el significado numérico usual, aparte de la relación de igualdad; por tanto, tienen una naturaleza no-métrica. El género, la raza, la profesión, el credo religioso, son variables observadas en este tipo de escala.

- Una escala más compleja, implica además de la relación de igualdad como el caso anterior, una relación de orden que se preserva tanto en el sistema numérico como en el sistema empírico (medidas sobre los objetos). Éste tipo de escalas se denomina *ordinal* porque los números que se asignan a los atributos deben respetar (conservar) el orden de la característica que se mide. El tipo de datos que resulta tiene naturaleza no métrica. La valoración de la opinión en "de acuerdo", "indiferente" o "en desacuerdo", constituye un ejemplo de una variable típica de esta escala.

- El siguiente nivel de escalamiento implica, además de una relación de orden como la escala anterior, una relación de igualdad de diferencias entre pares de objetos respecto a una característica determinada. La escala resultante se denomina *de intervalo* porque las diferencias entre los números se corresponden con las diferencias entre la propiedad medida sobre los objetos, y por tanto tiene naturaleza métrica. La medición de la temperatura, la altura física, constituyen ejemplos de esta escala de medida. Una característica adicional de esta escala es la necesidad de precisar un origen o punto "cero" respecto al cual la medida tiene sentido, esto no necesariamente significa ausencia del atributo. En el ejemplo de la temperatura, el cero en la escala Celsius, es la temperatura de congelación del agua al nivel del mar; nótese que este cero no corresponde con el de la escala Farenheit.

- El nivel más complejo de escalamiento implica, además de una relación de igualdad de diferencias como en la escala anterior, un punto de origen fijo o natural, el cero absoluto. El resultado es la escala *de razón*, que tiene también naturaleza métrica. Ejemplos de este tipo de escala son el peso, la talla o la edad de los individuos.

1.2 Representación gráfica de datos multivariados

El objeto y materia prima del trabajo estadístico está contenido en los datos, los cuales suministran información referente a un objeto, en un tiempo determinado. Resultan entonces tres componentes del trabajo estadístico: de un lado están los *objetos* sobre los que se intenta desarrollar algún estudio, por otro las *características* o *atributos* inherentes a los primeros y finalmente el *momento* u *ocasión* en que están inscritos los dos primeros (objeto y variable). Una representación, meramente esquemática, de los objetos, las variables y el tiempo es un prisma cuyas aristas están sobre los ejes principales. (Figura 1.1).

Se puede concebir entonces una colección de información sobre un objeto $i = 1, \ldots, n$ con un atributo $j = 1, \ldots, p$ en un tiempo $t = 1, \ldots, s$. Un punto X_{ijt} del prisma corresponde al valor del atributo j-ésimo, para i-ésimo individuo, en el instante t.

Las diferentes técnicas estadísticas trabajan en alguna región de este prisma. Así por ejemplo, las regiones paralelas al plano OV son estudiadas por la mayoría de las técnicas del análisis multivariado; aveces se les llama estudios *transversales*, de las regiones paralelas a VT se ocupan los métodos de series cronológicas (*estudios longitudinales*). En general los procedimientos

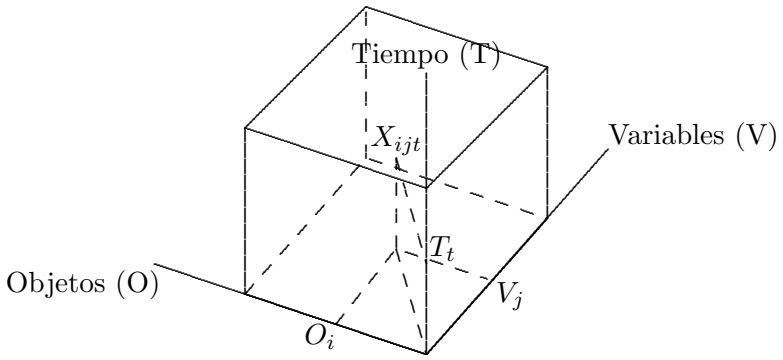


Figura 1.1. *Representación multivariada de datos.*

estadísticos consideran constantes o fijos algunos de los tres componentes señalados.

Algunos estudios consideran el sitio o espacio donde tienen lugar las mediciones observadas sobre los objetos. De este tipo de datos se ocupa la *estadística espacial* o la *geoestadística*. En ocasiones se considera que cada punto en el espacio define una población, con el esquema anterior corresponderían a varios prismas. Es preciso anotar que esta representación es más didáctica que formalmente matemática.

Cuando se dispone de dos variables su representación en un plano es relativamente sencilla. Para tres o más variables se han ideado algunas estrategias que permiten representar en el plano objetos definidos por dos o más atributos. Se debe tener presente, que el objetivo de estas representaciones es facilitar la lectura e interpretación acerca de la información contenida en los datos, de manera que las gráficas no resulten más complejas de leer que los mismos datos originales. A continuación se muestran algunas de estas herramientas gráficas.

Gráficos cartesianos. En estos gráficos se define un plano mediante la elección de dos variables, preferiblemente cuantitativas. Las variables restantes se pueden representar en este plano, con origen en el punto definido para las dos anteriores en cada objeto, y con orientación y trazado diferente para cada una. De esta manera, por ejemplo, cuatro individuos identificados por el vector de observaciones $(x_{i1}, x_{i2}, x_{i3}, x_{i4})$, $i = 1, 2, 3, 4$, se representan en un punto del plano $X_1 \times X_2$ cuyas coordenadas son las dos primeras; es decir, (x_{i1}, x_{i2}); las otras dos variables se ubican sobre sistemas coordenados construidos en cada uno de estos puntos (sistemas

"anidados"), con la orientación y escala decidida. Para más de cuatro variables, la representación de los sistemas "anidados" se construyen con ejes no perpendiculares (no ortogonales). En la figura 1.2 se representa el caso de cinco objetos A, B, C, D y E a los cuales se les registraron los atributos X_1, X_2, X_3 y X_4 (matriz $\mathbb{X}$).

$$\mathbb{X} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array} \\ \begin{pmatrix} 1.0 & 1.2 & 0.8 & 0.6 \\ 2.5 & 2.2 & 1.6 & 1.8 \\ 4.0 & 3.1 & 2.0 & 1.6 \\ 2.5 & 0.3 & 0.6 & 0.8 \\ 4.5 & 0.8 & 1.5 & 1.0 \end{pmatrix} \end{array}$$

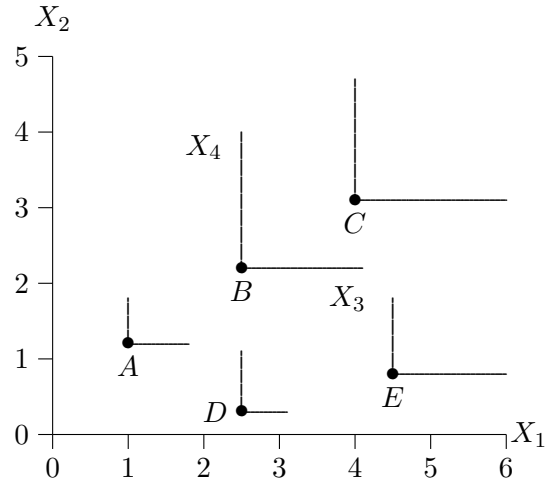


Figura 1.2 *Gráfico para cuatro dimensiones.*

Perfiles. Se representan a la manera de histogramas, donde cada barra corresponde a una variable y su altura al valor de la misma. A veces en lugar de barras se construye una línea poligonal. Cada diagrama corresponde a un objeto. La figura 1.3 muestra los perfiles para los datos de la matriz $\mathbb{X}$.

Diagramas de tallo y hojas. Es un procedimiento seudo gráfico para representar datos cuantitativos. El procedimiento para construirlo es el siguiente:

1. Redondear convenientemente los datos en dos o tres cifras significativas.

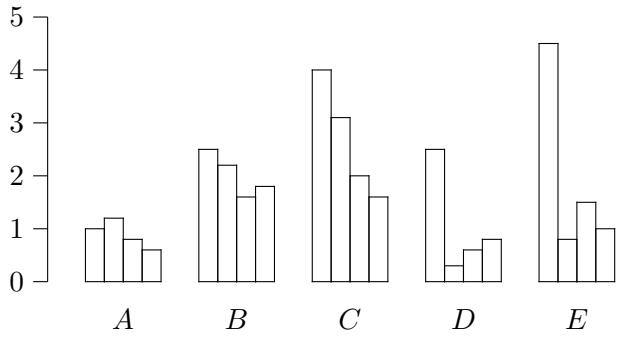


Figura 1.3 *Perfiles de la matriz de datos* $\mathbb{X}$.

2. Disponer los datos en una tabla con dos columnas como sigue:

   (a) Para datos con dos dígitos, escribir en la columna izquierda los dígitos de las decenas, éste es el tallo, y a la derecha, después de una línea o dos puntos, las unidades, que son las hojas. Así por ejemplo, 58 se escribe 5|8 o 5 : 8.

   (b) Para datos con tres dígitos el tallo estará formado por los dígitos de las centenas y decenas, los cuales se escriben en la columna izquierda, separados de las unidades (hojas). Por ejemplo, 236 se escribe 23|6 o 23 : 6.

3. Cada tallo define una clase, y se escribe una sola vez. El número de hojas representa la frecuencia de dicha clase.

La tabla 1.1 contiene el cociente de inteligencia (CI) de niños a los cuales se les registró el peso al nacer y la edad de la madre.

A continuación se muestra la representación de los datos de la tabla 1.1 mediante diagramas de tallo y hojas.

Diagramas de dispersión. Son gráficos en los cuales se representan los individuos u objetos por puntos asociados a cada par de coordenadas (valores de cada par variables).

En la figura 1.4 se han hecho los dispersogramas por pares de variables. Los dos dispersogramas que involucran el peso al nacer evidencian observaciones atípicas o "outliers" ("no usuales"). Además, en estas gráficas se puede advertir la posible asociación lineal entre pares de variables.

Tabla 1.1 Peso al nacer en 26 niños

Niño	CI	Peso	Edad	**Niño**	CI	Peso	Edad
1	125	2536	28	**14**	75	2350	23
2	86	2505	31	**15**	90	2536	24
3	119	2652	32	**16**	109	2577	22
4	113	2573	20	**17**	104	2464	35
5	101	2382	30	**18**	110	2571	24
6	143	2443	30	**19**	96	2550	24
7	132	2617	27	**20**	101	2437	23
8	106	2556	36	**21**	95	2472	36
9	121	2489	34	**22**	117	2580	21
10	109	2415	29	**23**	115	2436	39
11	88	2434	27	**24**	138	2200	41
12	116	2491	24	**25**	85	2851	17
13	102	2345	26				

Fuente: Everitt y Dunn (1991, pág. 27)

CI	*Peso*	*Edad*
7 : 5	22: 0	1: 7
8 : 568	23: 458	2: 012334444
9 : 056	24: 134446799	2: 67789
10: 1124699	25: 044567788	3: 00124
11: 035679	26: 25	3: 5669
12: 15	28: 51	4: 1
13: 28		
14: 3		

Diagramas de caja y "bigotes" (box-and-whisker plot). Un diagrama de estos consiste en una caja, y guiones o segmentos. Se dibuja una línea a través de la caja que representa la mediana. El extremo inferior de la caja es el primer cuartil (Q_1) y el superior el tercer cuartil (Q_3). Los segmentos o bigotes se extienden desde la parte superior de la caja a valores adyacentes; es decir, la observación más pequeña y la más alta que se encuentran dentro de la región definida por el límite inferior $Q_1 - 1.5 \cdot (Q_3 - Q_1)$ y el límite superior $Q_3 + 1.5 \cdot (Q_3 - Q_1)$. Las observaciones atípicas son puntos fuera de los límites inferior y superior, los cuales son señalados con estrellas ($\star$). Se pueden construir estos diagramas para varias variables conjuntamente. Este tipo de gráficas facilitan la lectura sobre localización, variabilidad, simetría, presencia de observaciones atípicas e incluso asociación entre variables, en un conjunto de datos.

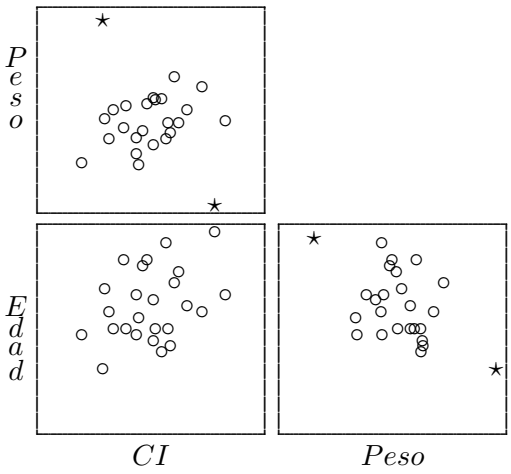


Figura 1.4 *Dispersograma para los datos de CI, peso y edad.*

En la figura 1.5 se muestran estos diagramas conjuntamente para los datos de las variables CI, peso y edad estandarizadas; se tuvo que estandarizar para eliminar el efecto de la escala de medición y posibilitar la comparación entre las variables. Se observa que la edad tiene más variabilidad que las otras dos variables, aunque es la de menor valor promedio. La variable peso es la de menor variabilidad o dispersión y tiene dos datos atípicos (uno en cada extremo).

Chernoff (1973), asocia a cada variable una característica del rostro; tal como longitud de la nariz, tamaño de los ojos, forma de los ojos, ancho de la boca, entre otras. La gráfica 1.6 presenta tres objetos mediante tres rostros. En el capítulo 7 se muestra el uso de estos gráficos en la construcción de conglomerados.

Andrews (1972), representa cada observación multidimensional como una función que toma una forma particular. A cada observación p dimensional $x' = (x_1, \ldots, x_p)$ se le asigna una función definida por:

$$x(t) = x_1/\sqrt{2} + x_2 \operatorname{sen}(t) + x_3 \cos(t) + x_4 \operatorname{sen}(2t) + x_5 \cos(2t) + \cdots$$

La función se grafica sobre el rango $-\pi \leq t \leq \pi$ para el número de p variables. La figura 1.7 contiene las curvas de Andrews para tres objetos hipotéticos. Estos y otros gráficos se presentan en el capítulo 7 para efectos de clasificación de objetos.

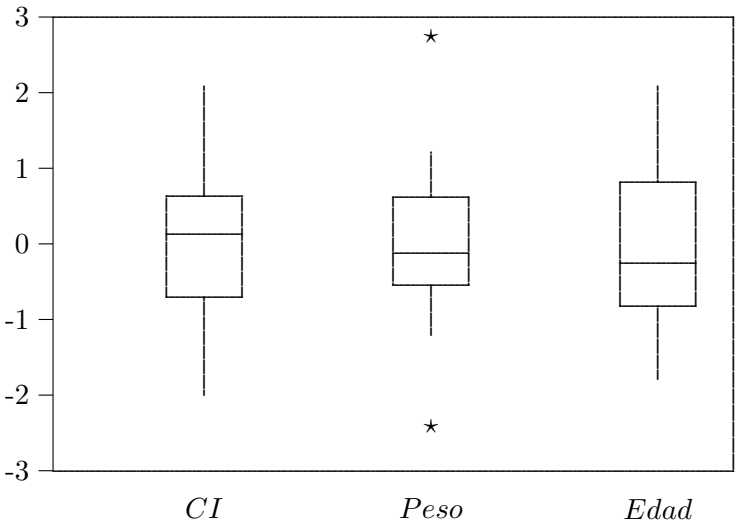


Figura 1.5 *Diagramas de cajas (box-plot) para los datos de la tabla 3.1.*

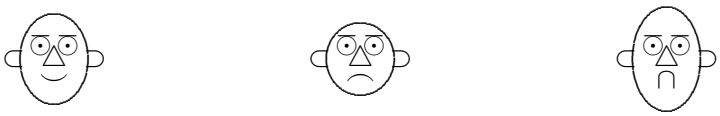


Figura 1.6 *Rostros de Chernoff*

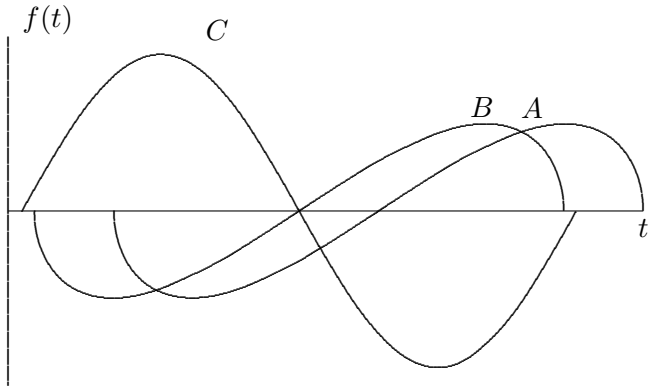


Figura 1.7 *Curvas de Andrews.*

Esta representación tiene, entre otras, la propiedad de preservar las medias de los datos y la distancia euclidiana entre las observaciones.

1.3 Técnicas multivariadas

Las técnicas del análisis multivariado (AM) tratan con datos asociados a conjuntos de medidas sobre un número de individuos u objetos. El conjunto de individuos junto con sus variables, pueden disponerse en un arreglo matricial $\mathbb{X}$, donde las filas corresponden a los individuos y las columnas a cada una de las variables. Las técnicas del AM se distinguen de acuerdo con el trabajo por filas (individuos) y/o columnas (variables).

$$\mathbb{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Algunos ejemplos de matrices de datos se presentan a continuación.

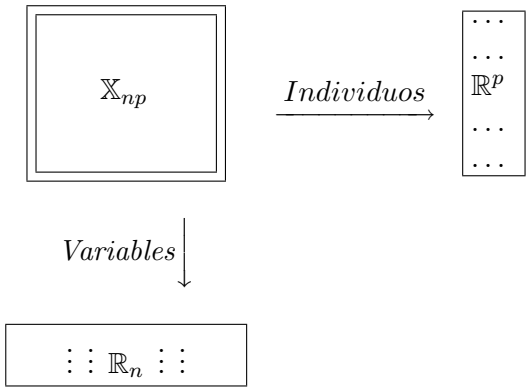
1. Se está interesado en el análisis de las notas de 6 áreas de conocimientos, registradas para un grupo de 200 estudiantes que ingresan a una carrera técnica; esta información se conforma en una matriz de tamaño (200×6).

2. La cantidad de azúcar y colesterol presente en la sangre, junto con la edad, presión arterial sistólica, el hábito de fumar y el género conforman la historia clínica de 120 pacientes que ingresaron a un centro de salud con dolencias renales; esta información está contenida en una matriz de datos 120×6. Con esta información se quiere encontrar las posibles asociaciones entre estas variables.

3. Sobre 65 ciudades diferentes de una región se emplean 7 indicadores de niveles de desarrollo; estos son: porcentaje de variación de la población 1995-2000, tasa de migración neta 1995-2000, ingreso per cápita a 1995, población económicamente activa a 1995, habitantes por médico en el año 2000, densidad de carreteras a 2000 (km por cien km^2) y líneas telefónicas por 1000 habitantes a 2000. Estos datos se consignan en una matriz de tamaño (65×7).

La mayoría de las técnicas multivariadas se dirigen a las filas, las columnas o las dos, de la matriz de datos. Así, trabajar sobre las filas de la matriz de

datos significa trabajar en el espacio de los individuos, es decir en $\mathbb{R}^p$. Análogamente, las técnicas estadísticas que trabajan sobre las columnas de la matriz de datos, están en el espacio de las variables $\mathbb{R}^n$. Esquemáticamente:



Diferentes son los enfoques y metodologías seguidos en el análisis multivariado. Algunos consideran éstas dos metodologías:

i) Los métodos *factoriales*, los cuales consideran a los individuos y/o variables ubicados en espacios referenciados por coordenadas (factores).

ii) De otro lado están las técnicas de *clasificación*, cuyo objetivo es la ubicación de individuos de manera espacial de acuerdo con las variables que los identifican; mediante estos métodos se consiguen mapas que ilustran el agrupamiento de los objetos.

Otro enfoque de las técnicas multivariadas considera que los objetivos del análisis y el tipo de datos obtenidos sugieren el tratamiento de la información. Dentro de esta visión se destacan las siguientes:

i) *Simplificación de la estructura de datos.* Tratan de encontrar una representación reducida del espacio de las variables en estudio mediante la transformación de algunas variables a un conjunto de menor dimensión.

ii) *Clasificación.* Análogo al primer enfoque, considera los individuos y las variables dispersos en un multiespacio; así, el objetivo es encontrar una ubicación espacial de éstos.

iii) *Interdependencia.* El propósito es estudiar la interdependencia entre las variables. Esta puede examinarse desde la independencia total de las variables hasta la dependencia de alguna con respecto a un subconjunto de variables (colinealidad).

iv) *Dependencia.* Interesa hallar la asociación entre dos conjuntos de variables, donde uno es considerado como la realización de mediciones dependientes de otro conjunto de variables.

v) *Formulación y pruebas de hipótesis.* Para un campo de estudio específico se postula un modelo estadístico, éste queda definido por unos parámetros que deben ser estimados y verificados de acuerdo con la información recopilada. Básicamente, se contemplan tres etapas: la *formulación*, la *estimación* y la *validación* del modelo.

Por considerar que los enfoques de dependencia y el de interdependencia cobijan la mayoría de metodologías multivariadas se esquematizan a continuación éstos dos. Existen otros enfoques del análisis multivariado tales como el bayesiano, el robusto, el no paramétrico, el no lineal y más recientemente el relacionado con la neurocomputación (Cherkassky y colaboradores, 1994); enfoques basados en el tipo de información utilizada y en los supuestos requeridos.

Se deja abierta la discusión sobre el "organigrama" de otros posibles enfoques y concepciones acerca del análisis estadístico multivariado.

1.3.1 Métodos de dependencia

1. Regresión múltiple

Se centra sobre la dependencia de una variable *respuesta* respecto a un conjunto de variables *regresoras* o *predictoras.* Mediante un modelo de regresión se mide el efecto de cada una de las variables regresoras sobre la respuesta. Uno de los objetivos es la estimación para la predicción del valor medio de la variable dependiente, con base en el conocimiento de las variables independientes o predictoras.

2. Análisis discriminante

Conocidas algunas características (variables) de un individuo y partiendo del hecho de que pertenece a uno de varios grupos (población) definidos de antemano, se debe asignar tal individuo en alguno de éstos, con base en la información que de él se dispone. La técnica del análisis discriminante suministra los requerimientos y criterios para tomar esta decisión.

3. Análisis de correlación canónica

Mediante este análisis se busca una relación lineal entre un conjunto de variables predictoras y un conjunto de criterios medidos u observados. Se inspeccionan dos combinaciones lineales, una para las variables predictoras

y otra para las variables criterio (dependientes). Cuando hay más de dos grupos se puede pensar en un análisis discriminante múltiple como un caso especial del análisis canónico.

4. Análisis logit

Es un caso especial del modelo de regresión, donde el criterio de respuesta es de tipo categórico o discreto. El interés se dirige a investigar los efectos de un conjunto de predictores sobre la respuesta, las variables predictoras pueden ser de tipo cuantitativo, categórico o de ambas.

5. Análisis de varianza multivariado

Cuando múltiples criterios son evaluados (tratamientos), y el propósito es determinar su efecto sobre una o más variables respuesta en un experimento, la técnica del análisis de varianza multivariado resulta apropiada. De otra manera, la técnica permite comparar los vectores de medias asociados a varias poblaciones multivariantes.

6. Análisis conjunto

Es una técnica que trata la evaluación de un producto o servicio, con base en las calidades que de éste requieren o esperan sus consumidores o usuarios. Consideradas las características o atributos que el producto o servicio debe tener, el problema se dirige a obtener la combinación *óptima* o adecuada de tales atributos. Ésta es una técnica que combina el diseño experimental, el análisis de varianza y las superficies de respuesta.

1.3.2 Métodos de interdependencia

Las técnicas de análisis de interdependencia buscan el *cómo* y el *por qué* se relacionan o asocian un conjunto de variables. En forma resumida las metodologías de este tipo son las siguientes:

1. Análisis de componentes principales

Técnica de reducción de datos, cuyo objetivo central es construir combinaciones lineales (componentes principales) de las variables originales que contengan una buena parte de la variabilidad total original. Las combinaciones lineales deben ser no correlacionadas (a veces se dice que están incorrelacionadas) entre sí, y cada una debe contener la máxima porción de variabilidad total respecto a las subsiguientes componentes.

2. Análisis de factores comunes

El análisis factorial describe cada variable en términos de una combinación lineal de un pequeño número de *factores comunes* no observables y un

factor único para cada variable. Los factores comunes reflejan la parte de la variabilidad que es compartida con las otras variables; mientras que el factor único expresa la variación que es exclusiva de esa variable. De esta manera, el objetivo es encontrar los factores comunes que recojan el máximo de información de las variables originales.

3. Análisis de correspondencias

En el caso más sencillo este método está dirigido al análisis de tablas de contingencia. Se intenta conseguir la mejor representación simultánea de los dos conjuntos de datos contenidos en la tabla (filas y columnas); de ahí el nombre de *correspondencias simples o binarias*. El análisis de *correspondencias múltiples* se desarrolla sobre varias variables categóricas, se considera una extensión de las correspondencias simples. Similar al análisis de componentes principales, se tiene una matriz de datos, donde las filas son los individuos y las columnas cada una de las modalidades o categorías de las variables.

4. Análisis de conglomerados

Es otra técnica de reducción de datos. Su objetivo es la identificación de un pequeño número de grupos, de tal manera que los elementos dentro de cada grupo sean similares (cercanos) respecto a sus variables y muy diferentes de los que están en otro grupo. El problema está en obtener una medida de distancia que garantice la cercanía o similitud entre los objetos.

5. Escalamiento multidimensional

Permite explorar e inferir criterios sobresalientes que la gente utiliza en la formación de percepciones acerca de la similitud y preferencia entre varios objetos. Con escalas métricas multidimensionales la similitud se obtiene sobre datos que tienen las propiedades de una métrica; de tal forma que la similitud entre dos objetos decrezca linealmente con la distancia.

Con el *escalamiento no-métrico* se transforman las similaridades percibidas entre un conjunto de objetos en distancias, para ubicar los objetos en algún espacio multidimensional. Se asume que los datos sólo tienen un rango ordenado, tal que las distancias son funciones monótonas de éstos. En resumen, el objetivo es la metrización de datos no métricos por transformación a un espacio métrico.

6. Modelos log-lineales

Con este tipo de modelos se puede investigar la interrelación entre variables categóricas que forman una tabla de contingencia o de clasificación cruzada. Los modelos log-lineales expresan las probabilidades de las celdas

en una tabla de contingencia múltiple en términos de efectos principales e interacción para las variables de la tabla.

7. Modelos estructurales

Aunque los modelos estructurales tienen aspectos de dependencia como de interdependencia, se considera como una técnica multivariada separada de éstas. Los objetivos de los modelos estructurales son tanto el modelamiento que permita descomponer las relaciones entre variables, a través de un sistema de ecuaciones lineales, como la prueba de las relaciones de causalidad involucradas en las variables observables (manifiestas) y en las variables no observables (latentes).

En el cuadro siguiente se resumen las principales técnicas multivariadas y se indica el tipo de medición requerida.

1.4 Variables aleatorias multidimensionales

En esta sección se presentan de manera muy resumida las definiciones, conceptos y propiedades básicas para el análisis estadístico multivariado. Como se puede apreciar en algunos casos, éstas son una extensión del caso univariado.

1.4.1 Distribuciones conjuntas

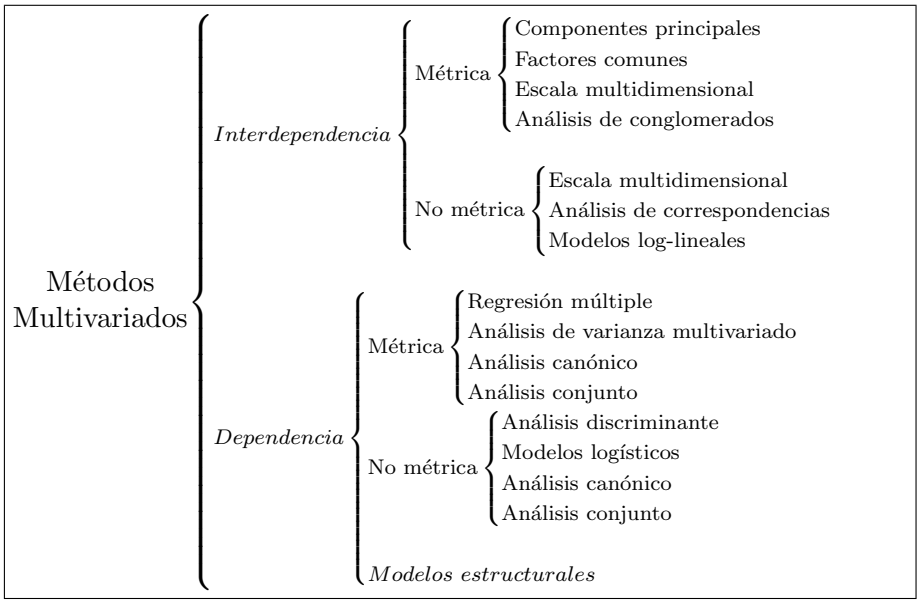
Una *variable aleatoria* p-dimensional, es un vector en el que cada una de sus componentes es una variable aleatoria. Así,

$$X' = (X_1, \ldots, X_p), \tag{1.1}$$

es un vector aleatorio, con X_i variable aleatoria para cada $i = 1, \ldots, p$.

Por la definición anterior los vectores aleatorios pueden estar conformados por variables aleatorias de tipo discreto, continuo o ambos. Los análisis y métodos multivariados señalan en cada caso los tipos de variables a los cuales se les puede aplicar adecuadamente tales procedimientos.

Los vectores aleatorios pueden considerarse como el objeto central del trabajo en el análisis y métodos de la estadística multivariada. Las filas de la matriz de datos, presentada al iniciar este capítulo, está conformada por vectores aleatorios.



A continuación se presentan algunos casos de aplicación práctica:

1. A una persona se le registra la estatura (X_1), el peso (X_2), su edad (X_3), años de escolaridad (X_4) y sus ingresos (X_5). De esta forma un individuo queda definido, para el estudio a desarrollar, por los valores que tome el vector $(X_1, X_2, X_3, X_4, X_5)'$.

2. En un estudio sobre el consumo de un producto en hogares de una ciudad, se consultó acerca de su frecuencia mensual de compra (X_1), número de miembros del hogar (X_2), producto sustituto (X_3) e ingresos (X_4). Los valores del vector $(X_1, X_2, X_3, X_4)'$ definen estos hogares.

3. Con el objeto de conocer la situación en el sector lechero en una región, se recogió la siguiente información en algunas fincas: superficie total de la finca (X_1), número total de vacas (X_2), promedio semanal de leche producida por vaca (X_3), índice de tecnificación (X_4), índice sanitario (X_5) e índice de instalaciones (X_6). La información para cada finca queda determinada por los valores que asuma el vector $(X_1, X_2, X_3, X_4, X_5, X_6)'$.

Como en el caso univariado, se define la *función de distribución conjunta* para el vector X mediante:

$$F(x_1, \ldots, x_p) = P(X_1 \leq x_1, \ldots, X_p \leq x_p). \tag{1.2}$$

Corresponde a la probabilidad de que cada una de las componentes del vector aleatorio X asuma valores menores o iguales que el respectivo componente de $(x_1, \ldots, x_p)$.

1.4.2 Algunos parámetros y estadísticas asociadas

Dado un vector aleatorio X, como el definido en (1.1), el *valor esperado* de X, notado $\mathcal{E}(X)$, es el vector de valores esperados de cada una de las variables aleatorias, así:

$$\boldsymbol{\mu} = \mathcal{E}(X) = \begin{pmatrix} \mathcal{E}(X_1) \\ \vdots \\ \mathcal{E}(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}. \tag{1.3}$$

La *matriz de varianzas y covarianzas* de X, la cual notaremos por $\boldsymbol{\Sigma}$, está dada por:

$$\boldsymbol{\Sigma} = Cov(X) = \mathcal{E}\left\{(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})'\right\} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}. \tag{1.4}$$

Donde σ_{ij} denota la covarianza entre la variable X_i y la variable X_j, la cual se define como:

$$\sigma_{ij} = \mathcal{E}[(X_i - \mu_i)(X_j - \mu_j)].$$

Al desarrollar el producto y aplicar las propiedades del valor esperado, se obtiene una expresión alterna para la matriz de varianzas y covarianzas; ésta es

$$\boldsymbol{\Sigma} = Cov(X) = \mathcal{E}(XX') - \boldsymbol{\mu}\boldsymbol{\mu}'. \tag{1.5}$$

Los elementos de la diagonal de la matriz (1.4) corresponden a las varianzas de cada una de las variables, los elementos fuera de la diagonal son las covarianzas entre las variables correspondientes de la fila y la columna.

Gran número de las metodologías señaladas en la primera parte de este capítulo se basan en la estructura y propiedades de $\boldsymbol{\Sigma}$; se destacan entre otras las siguientes propiedades:

1. La matriz $\boldsymbol{\Sigma}$ es simétrica; es decir, $\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}$, puesto que $\sigma_{ij} = \sigma_{ji}$.

2. Los elementos de la diagonal de $\boldsymbol{\Sigma}$ corresponden a la varianza de las respectivas variables ($\sigma_{ii} = \sigma_i^2$).

3. Toda matriz de varianzas y covarianzas es *definida no negativa* ($|\boldsymbol{\Sigma}| \geq 0$). Y es definida positiva, cuando el vector aleatorio es continuo.

4. Si $\mathcal{E}(X) = \boldsymbol{\mu}$ y $Cov(X) = \boldsymbol{\Sigma}$, entonces:

$$\mathcal{E}(AX + b) = A\boldsymbol{\mu} + b \text{ y } Cov(AX + b) = A\boldsymbol{\Sigma}A',$$

con A matriz de constantes de tamaño $(q \times p)$ y $\mathbf{b}$ vector $(q \times 1)$ también de constantes.

En adelante se hablará de la matriz de varianzas y covarianzas o de la matriz de covarianzas en forma indistinta.

A continuación se desarrollan algunas estadísticas descriptivas ligadas a los parámetros anteriores.

Se dice que un conjunto de datos es una *muestra aleatoria* multivariada si ésta tiene la misma probabilidad de extrarse que cualquier otra del mismo tamaño. A cada individuo (objeto) seleccionado de manera aleatoria de la población de individuos, se le registran una serie de atributos u observaciones (valores de las variables aleatorias). Sea x_{ij} la observación de la j-ésima variable en el i-ésimo individuo, se define la *matriz de datos multivariados* como el arreglo

$$\mathbb{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}. \tag{1.6}$$

Observaciones:

- La matriz $\mathbb{X}$ puede definirse como el arreglo de vectores fila o vectores columna. El i-ésimo vector fila se nota por $\boldsymbol{X}_{(i)}$ y el j-ésimo vector columna se nota por $\boldsymbol{X}^{(j)}$. Así cada uno denota el i-ésimo individuo o la j-ésima variable respectivamente.

- El vector formado por las p-medias muestrales, es el vector de promedios o de medias (centroide de los datos)

$$\bar{\boldsymbol{X}}' = \frac{1}{n}\mathbf{1}'\mathbb{X} = (\bar{x}_1, \ldots, \bar{x}_p), \tag{1.7}$$

donde $\mathbf{1}'$ es el vector columna de unos.

Se define la *media muestral* de la j-ésima variable por

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \text{ con } j = 1, \ldots, p.$$

La matriz constituida por las covarianzas s_{ij}, es la *matriz de varianzas y covarianzas muestral*, ésta es:

$$\boldsymbol{S} = \frac{1}{n}\mathbb{X}'(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}'\mathbf{1})\mathbb{X} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}, \tag{1.8}$$

donde:

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k); \quad j,k = 1,\ldots,p.$$

es la covarianza muestral entre la variable columna j y la variable columna k. Nótese que si $j = k$, se obtiene la varianza muestral asociada a la variable j-ésima. La matriz $\boldsymbol{S}$ es simétrica, es decir, $s_{jk} = s_{kj}$, para todas las entradas $j,k = 1,2,\ldots,p$.

La escritura de $\boldsymbol{S} = \frac{1}{n}\mathbb{X}'(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}'\mathbf{1})\mathbb{X}$, para el caso de una matriz de datos con n observaciones y tres variables, por ejemplo, corresponde a la siguiente expresión de la respectiva matriz de varianzas y covarianzas es:

$$\begin{aligned}
\boldsymbol{S} &= \frac{1}{n}\mathbb{X}'(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}'\mathbf{1})\mathbb{X} \\
&= \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \\ x_{13} & \cdots & x_{n3} \end{pmatrix} \left[\begin{pmatrix} 1 & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} - \frac{1}{n} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \right] \begin{pmatrix} x_{11} & x_{13} \\ x_{21} & x_{23} \\ \vdots & \vdots \\ x_{n1} & x_{n3} \end{pmatrix} \\
&= \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \\ x_{13} & \cdots & x_{n3} \end{pmatrix} \left[\begin{pmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n} \end{pmatrix} \right] \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \\
&= \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{12} & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{pmatrix}.
\end{aligned}$$

La matriz $\boldsymbol{S}$ expresa tanto la dispersión de los datos en torno a la media (elementos de la diagonal), como la asociación lineal entre las variables (elementos fuera de la diagonal). En algunas circunstancias se necesita disponer de un solo número que señale la dispersión de los datos; la *varianza generalizada* y la *variabilidad total* son dos de tales parámetros. La varianza generalizada se define como el determinante de la matriz $\boldsymbol{S}$, y se nota $|\boldsymbol{S}|$; es decir,

$$VG = |\boldsymbol{S}|. \tag{1.9}$$

La varianza total se define como la traza de la matriz $\boldsymbol{S}$; téngase presente que los elementos de la diagonal de $\boldsymbol{S}$ son las varianzas de cada una de las variables:

$$VT = tra(\boldsymbol{S}) = \sum_{j=1}^{p} s_j^2. \tag{1.10}$$

Aunque a mayor variabilidad, los valores de VG y de VT aumentan, se debe tener cuidado por la influencia de valores extremos en la varianza. Su raíz cuadrada se denomina la *desviación típica generalizada*. Nótese que si $p = 1$; $VG = VT = s^2$.

Estas varianzas se emplean en métodos de análisis de varianza multivariado, en la construcción de componentes principales, en el análisis de factores comunes y únicos, en el análisis de correspondencias, entre otros.

También a partir de la matriz $\boldsymbol{S}$ se puede obtener la matriz de correlación $\boldsymbol{R}$, cuyos elementos son los coeficientes de correlación entre cada par de variables. Cada elemento r_{jk} de $\boldsymbol{R}$ es de la forma:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}},$$

donde r_{jk} es el *coeficiente de correlación lineal* entre la variable j y la variable k.

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}, \tag{1.11}$$

donde $\boldsymbol{D}^{-\frac{1}{2}}$ es la matriz diagonal con los inversos de las desviaciones estándar sobre la diagonal; es decir, $\boldsymbol{D}^{-\frac{1}{2}} = \text{Diag}(1/s_j)$.

El coeficiente de correlación muestral r_{jk} está relacionado con el *coseno* del ángulo entre los vectores $X^{(j)'} = (x_{1j}, \ldots, x_{nj})$ y $X^{(k)'} = (x_{1k}, \ldots, x_{nk})$, los cuales están centrados en sus repectivas medias; es decir, $X^{(j)} - \bar{X}_j\mathbf{1}$ y $X^{(k)} - \bar{X}_k\mathbf{1}$, con $\mathbf{1}$ vector de unos de tamaño $(n \times 1)$. El coseno del ángulo θ formado entre estas variables es (A1.10)

$$\begin{aligned} \cos\theta &= \frac{(X^{(j)} - \bar{X}_j\mathbf{1})'(X^{(k)} - \bar{X}_k\mathbf{1})}{\sqrt{[(X^{(j)} - \bar{X}_j\mathbf{1})'(X^{(j)} - \bar{X}_j\mathbf{1})][(X^{(k)} - \bar{X}_k\mathbf{1})'(X^{(k)} - \bar{X}_k\mathbf{1})]}} \\ &= \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}} = r_{jk}. \end{aligned}$$

De esta forma, si el ángulo θ, entre los dos vectores centrados, es pequeño, tanto su coseno como el coeficiente de correlación r_{jk} son cercanos a 1. Si los dos vectores son perpendiculares, $\cos\theta$ y r_{jk} son iguales a cero. Si los dos vectores tienen, aproximadamente, direcciones opuestas, $\cos\theta$ y r_{jk} tendrán un valor cercano a

-1. Ésta es una manera de expresar la proximidad entre variables, propiedad sobre la cual se apoyan los métodos factoriales.

Como toda matriz de covarianzas es definida positiva, su determinante es positivo; además, la varianza generalizada está asociada con el área (para $p = 2$) o volumen (para $p \geq 3$) ocupado por el conjunto de datos. Para ilustrar estas afirmaciones considérese el caso $p = 2$. La matriz de covarianzas puede escribirse como:

$$\boldsymbol{S} = \begin{pmatrix} s_1^2 & rs_1s_2 \\ rs_1s_2 & s_2^2 \end{pmatrix}.$$

La varianza generalizada es

$$\begin{aligned} VG = |\boldsymbol{S}| &= s_1^2 s_2^2 - r^2 s_1^2 s_2^2 \\ &= s_1^2 s_2^2 (1 - r^2) \\ &= s_1^2 s_2^2 (1 - \cos^2 \theta) \\ &= (s_1 s_2 \operatorname{sen} \theta)^2, \end{aligned}$$

y la desviación típica generalizada es: $|\boldsymbol{S}|^{\frac{1}{2}} = s_1 s_2 \sqrt{1 - r^2}$.

La figura 1.8 representa las variables x_1 y x_2 como vectores en el espacio de observaciones (fila). Los vectores han sido escalados dividiéndolos por $\sqrt{n-1}$, y θ es el ángulo formado entre ellos, el cual puede ser obtenido desde el coeficiente de correlación, pues anteriormente se mostró que es igual al coseno del ángulo formado entre los vectores. Se observa, en esta figura, que si x_1 tiene una relación lineal perfecta con x_2 entonces los vectores x_1 y x_2 son colineales, y por tanto, el área del paralelogramo es igual a cero. Correlación perfecta entre variables implica *redundancia* en los datos; es decir, que las dos variables miden lo mismo. De lo contrario, si la correlación es cero los vectores son ortogonales, esto sugiere que no hay redundancia en los datos.

De la figura 1.8 es claro que el área es mínima (cero) para vectores colineales y máxima para vectores ortogonales. Así, el área del paralelogramo se relaciona con la cantidad de redundancia en la información contenida en el conjunto de datos. El área al cuadrado del paralelogramo es usada como una medida de la varianza generalizada; o equivalentemente, la desviación típica generalizada está asociada con el área del paralelogramo.

En la figura 1.9 se muestra también la relación entre la desviación típica generalizada y el área determinada por un conjunto de datos.

Si las variables son independientes, la mayoría de las observaciones están máximo a 3 desviaciones estándar de la media; es decir, dentro de un rectángulo de lados $6s_1$ y $6s_2$. Por la desigualdad de Tchebychev, se espera que al menos el 90% de los datos esté entre la media y 3 desviaciones típicas a cada lado; esto se muestra en la figura 1.9a. Así, el área ocupada por las variables es directamente proporcional con el producto de las desviaciones típicas.

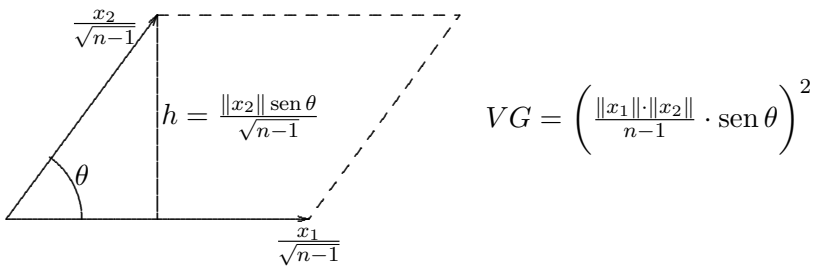


Figura 1.8 *Varianza generalizada.*

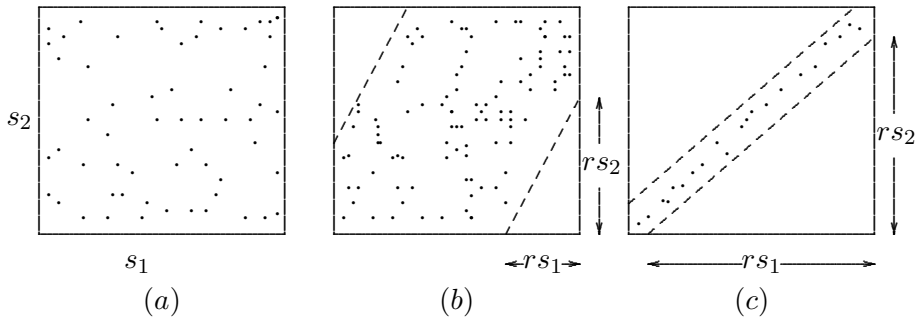


Figura 1.9 *Desviación típica generalizada.*

Si las variables tienen una asociación lineal, el coeficiente de correlación r será diferente de cero. Asúmase, sin pérdida de generalidad que r es positivo. De esta manera los puntos se ubicarán dentro de una franja como se indica en la figura 1.9b. Esta área tenderá a reducirse en tanto que r sea grande. En el caso de $r \approx 1$, los puntos se dispondrán cerca de una línea recta como se muestra en la figura 1.9c, y el área será próxima a cero. Para $p \geq 3$, la varianza generalizada, o la desviación típica generalizada, tendrá una relación inversa con el volumen del sólido (o hipersólido) que contiene los datos.

Ejemplo 1.3 Los siguientes datos se refieren a la altura de una planta X_1 (en m.), su longitud radicular X_2 (en cm), su área foliar X_3 (en cm^2) y su peso en pulpa X_4 (en gm.), de una variedad de manzano. Los datos (matriz $\mathbb{X}$) se presentan en la tabla 1.2.

Tabla 1.2 Medidas sobre manzanos

Obs.	X_1	X_2	X_3	X_4
1	1.38	51	4.8	115
2	1.40	60	5.6	130
3	1.42	69	5.8	138
4	1.54	73	6.5	148
5	1.30	56	5.3	122
6	1.55	75	7.0	152
7	1.50	80	8.1	160
8	1.60	76	7.8	155
9	1.41	58	5.9	135
10	1.34	70	6.1	140

La media para la variable altura de planta X_1 se calcula de las siguientes dos formas alternativas:

$$\begin{aligned}\bar{x}_1 &= \frac{1.38 + 1.40 + \cdots + 1.34}{10} \\ &= \frac{1}{10}\mathbf{1}(1.38, 1.40, \ldots, 1.34)' \\ &= \frac{1}{10}(1, 1, \ldots, 1)(1.38, 1.40, \ldots, 1.34)' \\ &= 1.44.\end{aligned}$$

Con un cálculo similar para las demás medias se obtiene el vector de medias muestrales, éste es:

$$\bar{X} = (1.44,\ 66.80,\ 6.29,\ 139.50)'.$$

La varianza muestral para la variable altura de planta X_1 se calcula como sigue:

$$\begin{aligned}s_{11} = s_1^2 &= \frac{1}{9}\sum_{i=1}^{10}(x_{i1} - \bar{x}_1)^2 \\ &= \frac{1}{9}\{(1.38 - 1.44)^2 + (1.40 - 1.44)^2 + \cdots + (1.34 - 1.44)^2\} \\ &= 0.0096.\end{aligned}$$

La covarianza muestral entre la variable altura de planta X_1 y la variable longitud radicular X_2 se calcula así:

$$\begin{aligned}s_{12} &= \frac{1}{9}\sum_{i=1}^{10}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \\ &= \frac{1}{9}\{(1.38 - 1.44)(51 - 66.80) + \cdots + (1.34 - 1.44)(70 - 66.80)\} \\ &= 0.7131.\end{aligned}$$

Las demás se calculan en forma análoga.

El coeficiente de correlación entre las dos variables anteriores es el siguiente:

$$\begin{aligned} r_{12} &= \frac{s_{12}}{\sqrt{s_{11}s_{22}}} \\ &= \frac{0.7131}{\sqrt{(0.0096)(96.6222)}} \\ &= 0.7369. \end{aligned}$$

Mediante cálculos como los anteriores (considerando su extensión) se obtienen las demás entradas de la matriz de covarianzas $\boldsymbol{S}$ y la matriz de correlación $\boldsymbol{R}$; está son, respectivamente:

$$\boldsymbol{S} = \begin{pmatrix} 0.010 & 0.713 & 0.083 & 1.150 \\ 0.713 & 96.622 & 9.509 & 138.556 \\ 0.083 & 9.509 & 1.134 & 14.883 \\ 1.150 & 138.556 & 14.883 & 212.056 \end{pmatrix},$$

y

$\boldsymbol{R} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}$

$$= \begin{pmatrix} 0.010 & 0 & 0 & 0 \\ 0 & 96.622 & 0 & 0 \\ 0 & 0 & 1.134 & 0 \\ 0 & 0 & 0 & 212.056 \end{pmatrix}^{-\frac{1}{2}} \boldsymbol{S} \begin{pmatrix} 0.010 & 0 & 0 & 0 \\ 0 & 96.622 & 0 & 0 \\ 0 & 0 & 1.134 & 0 \\ 0 & 0 & 0 & 212.056 \end{pmatrix}^{-\frac{1}{2}}$$

$$= \begin{pmatrix} 1.000 & 0.737 & 0.790 & 0.802 \\ 0.737 & 1.000 & 0.908 & 0.968 \\ 0.790 & 0.908 & 1.000 & 0.960 \\ 0.802 & 0.968 & 0.960 & 1.000 \end{pmatrix}.$$

Al comparar las respectivas entradas de las dos matrices se observa un cambio en su orden por magnitud dentro de cada matriz. Por ejemplo $s_{13} = s_{31}$ es el valor más bajo en $\boldsymbol{S}$, mientras que $r_{13} = r_{31}$ no lo es en $\boldsymbol{R}$.

Se nota la alta relación lineal que tiene el peso en pulpa con el área foliar y la longitud radicular, éstos son los elementos responsables en la fisiología de la planta.

La varianza total y la varianza generalizada son, respectivamente:

$$VT = tra(\boldsymbol{S}) = \sum_{j=1}^{4} s_j^2 = (0.0096 + 96.6222 + 1.1343 + 212.0555) = 309.8216$$

$$VG = |\boldsymbol{S}| = 0.330259.$$

Nótese que la variable que más participa de la varianza total es la variable peso en pulpa X_4, pues esta corresponde a $(212.0555/309.8216) \times 100 = 68.4\%$ de la variabilidad total, de manera análoga y decreciente, las participaciones de las otras variables son: 31.20% para la longitud radicular X_2, 0.37% para el área foliar X_3, y, 0.003% para la altura de planta X_1. ☑

1.4.3 Distancia

El concepto de distancia es uno de los más importantes y sobre el cual se han elaborado muchos conceptos matemáticos, como la convergencia y los espacios métricos. La estadística no ha sido ajena a su uso, aun más, para el desarrollo de algunas técnicas ha tenido que "inventar" o definir y adaptar algunas de tales distancias. En esta parte se hace referencia al concepto de distancia dentro de un contexto estadístico sin prentender hacer una presentación rigurosa del tema.

Uno de los problemas al que más esfuerzos ha dedicado la estadística es el estudio de la variabilidad, *¿de qué se ocuparían los estadísticos si no existiera variabilidad en los datos*? Para esto ha sido necesario crear formas de medir, emplear y modelar la heterogeneidad de la información contenida en los datos u observaciones.

Para un investigador puede ser importante determinar si dos individuos, con determinadas características (variables), se deben considerar cercanos o no. El interés puede consistir en la ubicación de los individuos en alguna de varias poblaciones con base en su proximidad a ellas. Otra situación consiste en decidir si se rechaza o no una hipótesis estadística de acuerdo con su discrepancia con datos observados (muestra). Una de las formas de estimar los parámetros asociados a un modelo de regresión es a través de la minimización de la distancia, en dirección de la variable respuesta, entre los puntos observados y la línea, curva o superficie de regresión propuesta; metodología que se conoce con el nombre de mínimos cuadrados. La bondad de un estimador se juzga, aveces, por su distancia al parámetro; distancia que se traduce muy comúnmente en sesgo, error de estimación, varianza, o consistencia, entre otros (Apéndice B).

A continuación se presentan los tipos de distancia de gran utilidad en la mayoría de las técnicas de la estadística multivariada.

1. Distancia euclidiana

Dados dos puntos (objetos) de $\mathbb{R}^n$, $X_h = (X_{h1}, \ldots, X_{hp})$ y $X_i = (X_{i1}, \ldots, X_{ip})$, se define su *distancia euclidiana* como el número

$$d_{hi} = \left(\sum_{j=1}^{p} (X_{hj} - X_{ij})^2 \right)^{1/2}. \tag{1.12}$$

Dada una muestra aleatoria $X_1, \ldots, X_n$, se puede escribir la varianza muestral $\widehat{\sigma}^2$ como

$$\sqrt{n}\widehat{\sigma} = \left(\sum_{j=1}^{p} {X_j}^{*2} \right)^{1/2} = \|X'\|.$$

La desviación tipíca $\widehat{\sigma}$ se toma como la distancia euclidiana promedio entre los datos y su constante más próxima, la media aritmética.

El error cuadrático medio (B.23) es la distancia cuadrática promedio entre un estimador $\widehat{\theta}$ y el respectivo parámetro θ.

2. Distancia de Mahalanobis

Las variables empleadas en un estudio suelen estar en escalas de medición diferente y correlacionadas. Así, por ejemplo, la altura y el peso de las personas, son cantidades con distintas unidades (metros y kilogramos), de manera que el número que representa la distancia entre dos individuos no solo cambiará de acuerdo con las unidades de medida empleadas sino por el grado de asociación que hay entre estas variables; de esta forma, si dos variables están muy relacionadas y en dos objetos o individuos toman valores bastante diferentes, éstos deben considerarse más separados que si los mismos valores se hubieran observado en variables independientes. La distancia de *Mahalanobis* entre los objetos $X_h = (X_{h1}, \ldots, X_{hp})$ y $X_i = (X_{i1}, \ldots, X_{ip})$ se define mediante la siguiente forma cuadrática

$$D^2_{hi} = (X_h - X_i)' S^{-1} (X_h - X_i), \text{ con } h, i = 1, \ldots, n, \tag{1.13}$$

la cual considera tanto el efecto de las unidades de medición como la correlación entre las variables.

Para el caso bidimensional, la distancia de Mahalanobis entre las observaciones h e i está dada por la siguiente expresión

$$D^2_{hi} = \frac{1}{1-r^2}\left[\frac{(X_{h1}-X_{i1})^2}{s_1^2} + \frac{(X_{h2}-X_{i2})^2}{s_2^2} - 2r\frac{(X_{h1}-X_{i1})(X_{h2}-X_{i2})}{s_1 s_2}\right]. \tag{1.13a}$$

En esta expresión s_1^2 y s_2^2 son las varianzas para las variables X_1 y X_2, respectivamente, y r es el coeficiente de correlación entre las dos variables. Se observa que si las variables no se correlacionan ($r = 0$) se tiene la llamada *"distancia estadística"* entre las dos variables, y si además, las variables tienen varianza igual a 1 esta distancia se reduce a la distancia euclidiana al cuadrado. Es decir, la distancia estadística y euclidiana son casos especiales de la distancia de Mahalanobis. Nótese además que el tercer término de (1.13a), que incluye el coeficiente de correlación r, influye sobre la distancia entre dos objetos.

La distancia de Mahalanobis es usada frecuentemente para medir la distancia entre una observación multivariada (individuo) y el centro de la población de donde procede la observación. Si $x_i = (x_{i1}, \ldots, x_{ip})'$ representa un individuo particular, seleccionado aleatoriamente de una población con centro $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$, y matriz de covarianzas $\boldsymbol{\Sigma}$, entonces

$$D_i^2 = (x_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (x_i - \boldsymbol{\mu}), \tag{1.14}$$

se considera como una medida de la distancia entre el individuo x_i y el centroide $\boldsymbol{\mu}$ de la población.

El valor D_i^2 puede considerarse como un *residual* multivariado para la observación x_i, donde residual significa la distancia entre una observación y el "centro de gravedad" de todos los datos. Si la población puede asumirse como normal multivariada (capítulo 2), entonces los valores de D_i^2 se distribuyen *ji-cuadrado* con p grados de libertad; el cual es un instrumento útil para la detección de valores atípicos.

La distribución *ji-cuadrado* se presenta asociada con la distancia de Mahalanobis. Si se considera un vector aleatorio conformado por p variables aleatorias normales e independientes; es decir, $X' = (X_1, \ldots, X_p)$, con $X_j \sim n(\mu_j, \sigma_j^2)$ para $j = 1, \ldots, p$, entonces la distancia estandarizada entre el vector X y el vector de medias $\boldsymbol{\mu}$ está dado por

$$\sum_{j=1}^{p} \left(\frac{x_j - \boldsymbol{\mu}_j}{\sigma_j} \right)^2 = (X - \boldsymbol{\mu})' \boldsymbol{D}^{-1} (X - \boldsymbol{\mu}) = \sum_{j=1}^{p} z_j^2 = \chi^2_{(p)},$$

donde $z_j \sim n(0, 1)$ y $\boldsymbol{\Sigma} = \boldsymbol{D} = \text{Diag}(\sigma_j^2)$. Así, la distribución χ^2 se interpreta como la distancia estandarizada entre un vector de variables normales independientes X y su vector de medias, o también, como la longitud (norma) de un vector de variables aleatorias $n(0, 1)$ e independientes.

La distancia euclidiana es un caso particular de distancia de Mahalanobis, basta hacer $\boldsymbol{\Sigma} = \boldsymbol{I}_p$.

3. Otras Distancias

Finalmente se resumen algunas otras distancias que pueden emplearse en el trabajo estadístico; con estas no se agota el tema (en el capítulo 10, tabla 10.1, se consideran otras distancias).

La distancia de *Minkowski* entre el par de observaciones identificadas como los vectores fila $X_h = (X_{h1}, \ldots, X_{hp})$ y $X_i = (X_{i1}, \ldots, X_{ip})$, se define por:

$$d_{hi} = \left(\sum_{j=1}^{p} |X_{hj} - X_{ij}|^r \right)^{\frac{1}{r}}, \tag{1.15}$$

donde d_{hi} denota la distancia entre el objeto h y el objeto i. La distancia euclidiana se obtiene de esta última haciendo $r = 2$.

Otra distancia, es la denominada de *ciudad* dada por

$$d_{hi} = \sum_{j=1}^{p} |X_{hj} - X_{ij}|, \tag{1.16}$$

que resulta de hacer $r = 1$ en la distancia de Minkowski. El calificativo de ciudad es porque la distancia entre dos puntos de ésta es igual al número de cuadras (calles o carreras) que se deben recorrer para ir de un punto a otro.

Ejemplo 1.4 Con relación a los datos del ejemplo 1.3 (tabla 1.2) se calculan la distancia euclidiana y de Mahalanobis entre cada observación y el centroide de los datos.

Para la primera observación $X_1 = (1.38,\ 51,\ 4.8,\ 115)$, la distancia euclidiana respecto al vector de medias muestral $\bar{X} = (1.44,\ 66.80,\ 6.29,\ 139.50)'$ se calcula como sigue:

$$
\begin{aligned}
d_1 &= \sqrt{(X_1 - \bar{X})(X_1 - \bar{X})'} \\
&= \sqrt{(1.38 - 1.44)^2 + (51 - 66.80)^2 + (4.8 - 6.29)^2 + (115 - 139.50))^2} \\
&= 29.19.
\end{aligned}
$$

También, la distancia de Mahalanobis entre la primera observación y el centroide de los datos es:

$$
\begin{aligned}
D_1^2 &= (X_1 - \bar{X})' \boldsymbol{S}^{-1} (X_1 - \bar{X}) \\
&= (-0.06, -15.80, -1.49, -24.50) \begin{pmatrix} 0.010 & 0.713 & 0.083 & 1.150 \\ 0.713 & 96.622 & 9.509 & 138.556 \\ 0.083 & 9.509 & 1.134 & 14.883 \\ 1.150 & 138.556 & 14.883 & 212.056 \end{pmatrix}^{-1} \begin{pmatrix} -0.06 \\ -15.80 \\ -1.49 \\ -24.50 \end{pmatrix} \\
&= (-0.06, -15.80, -1.49, -24.50) \begin{pmatrix} 311.608 & 1.858 & -2.854 & -2.703 \\ 1.850 & 0.191 & 0.415 & -0.164 \\ -2.854 & 0.415 & 12.208 & -1.112 \\ -2.703 & -0.164 & -1.112 & 0.204 \end{pmatrix} \begin{pmatrix} -0.06 \\ -15.80 \\ -1.49 \\ -24.50 \end{pmatrix} \\
&= 4.9626427.
\end{aligned}
$$

En la tabla 1.3 se muestran la distancias euclidiana y de Mahalanobis entre cada una de las observaciones y el centroide de los datos. De acuerdo con los resultados

Tabla 1.3 Distancias de manzanos respecto a la media

Obs.	Distancia Euclidiana	Distancia de Mahalanobis
1	29.190995 (10)	4.962643 (9)
2	11.703334 (5)	0.512610 (1)
3	2.707522 (1)	2.586287 (3)
4	10.523465 (4)	3.043581 (5)
5	20.588609 (8)	3.041331 (4)
6	14.966808 (6)	1.570419 (2)
7	24.449320 (9)	4.716541 (8)
8	18.088517 (7)	4.339042 (7)
9	9.891575 (3)	7.298924 (10)
10	3.246062 (2)	3.928625 (6)

contenidos en la tabla 1.3, se observa que las magnitudes de las distancias son notoriamente diferentes; cosa natural, pues mientras la distancia euclidiana se hace sobre las medidas originales, la distancia de Mahalanobis "corrige" por el inverso de la varianza y de acuerdo con la covarianza entre las variables. No hay concordancia en las distancias, es decir, el orden de separación de cada observación (indicado dentro de los paréntesis) respecto al centroide de los datos resulta diferente. ☑

1.4.4 Datos faltantes

Frecuentemente ocurre que un número de entradas en la matriz de datos son vacios o faltantes, lo que produce observaciones o registros incompletos. Por ejemplo:

- En datos sobre pacientes, puede darse que algunos no asistan el día que se registra parte de su información.
- En un laboratorio puede ocurrir un accidente el cual produce información incompleta.
- Ante una encuesta una persona puede negarse a dar cierta información.
- En el proceso de captura por medio magnético de la información se pueden cometer errores de omisión.

Aunque algunas técnicas multivariadas pueden sufrir modificaciones leves ante la presencia de observaciones incompletas, otras sólo trabajan con información completa. Una salida ante esta situación (seguida por varios paquetes estadísticos) es la exclusión de observaciones incompletas. Esta solución puede resultar complicada cuando se tenga un número determinado de observaciones con uno o más valores faltantes, pues el tamaño de muestra se reduciría notablemente. Una alternativa más conveniente es la estimación de las observaciones faltantes ("llenar huecos"); este proceso se le llama *imputación*.

La distribución de los valores faltantes en los datos es importante. Valores faltantes dispuestos aleatoriamente en las variables de una matriz de datos representa menos problema que cuando la información faltante tiene un patrón que depende, para algún rango, de los valores de las variables.

• Varias han sido las técnicas de imputación propuestas en los últimos años. La más vieja y simple es la de reemplazar un valor faltante por el promedio de los valores presentes en la variable correspondiente. Reemplazar una observación por su media reduce la varianza y la covarianza en valor absoluto. En consecuencia, la matriz de covarianzas muestral $\boldsymbol{S}$ calculada desde la matriz de datos $\mathbb{X}$ con medias imputadas para valores faltantes es sesgada; aunque, definida positiva.

• Un segundo método de estimación consta de una serie de regresiones múltiples en la cual cada variable que tenga valores faltantes se trata como la variable dependiente y las demás como variables regresoras o explicativas. El procedimiento se desarrolla así:

- La matriz de datos se particiona en dos, una parte contiene todas las filas u observaciones que tienen entradas *faltantes* y la otra contiene las observaciones que están *completas*. Supóngase que x_{ij}, que corresponde al dato del individuo i en la variable j, es un dato faltante. Entonces, empleando la matriz de observaciones completas, la variable x_j es regresada sobre las otras variables para obtener el siguiente modelo de predicción: $\hat{x}_j = b_0 + b_1x_1 + \ldots + b_{j-1}x_{j-1} + b_{j+1}x_{j+1} + \ldots + b_px_p$. Las entradas no faltantes de la i ésima fila son reemplazadas en el miembro izquierdo de esta ecuación para obtener el valor de predicción $\hat{x}_{ij}$.

Este procedimiento se puede desarrollar en forma iterativa de la siguiente manera: estimar todos los datos faltantes desde la respectiva ecuación de regresión. Después de "tapar todos los huecos" usar la matriz de datos que se completó para estimar nuevas ecuaciones de predicción. Con estas ecuaciones de predicción calcular nuevamente los valores $\hat{x}_{ij}$ para las entradas faltantes. Usar nuevamente la matriz de datos completada en la segunda etapa para predecir los nuevos valores $\hat{x}_{ij}$ correspondientes a los datos faltantes. Continuar este proceso hasta que se observe una convergencia o estabilización de los valores estimados.

Ejemplo 1.5 Para los datos del ejemplo 1.3, asúmase que las observaciones 1 y 2 tienen información faltante (notadas por ϕ) como se ilustra en la tabla 1.4.

Tabla 1.4 Medidas sobre manzanos con datos faltantes (ϕ)

Obs.	X_1	X_2	X_3	X_4
1	ϕ	51	4.8	115
2	1.40	60	ϕ	130
3	1.42	69	5.8	138
4	1.54	73	6.5	148
5	1.30	56	5.3	122
6	1.55	75	7.0	152
7	1.50	80	8.1	160
8	1.60	76	7.8	155
9	1.41	58	5.9	135
10	1.34	70	6.1	140

Esta tabla o matriz se particiona en dos: una que contiene las observaciones faltantes (1 y 2); y la otra que contiene las observaciones con entradas completas (3 a 10).

- Para encontrar un valor que "tape el hueco" de la primera observación se estima la ecuación de regresión de la variable dependiente X_1 sobre las variables X_2, X_3 y X_4, mediante la matriz de observaciones completas; la ecuación estimada es igual a:

$$\hat{X_1} = 0.05406 - 0.00770X_2 - 0.03661X_3 + 0.01517X_4.$$

A partir de esta ecuación se estima el valor de la variable X_1 para la primera observación, es decir para: $X_2 = 51$, $X_3 = 4.8$ y $X_4 = 115$; este valor es $\hat{X_1} = 1.2302$. De manera similar se estima el dato faltante en la segunda observación; esto se logra regresando la variable X_3 sobre las variables X_1, X_2 y X_4. Con la porción de datos completos la ecuación estimada es igual a:

$$\hat{X_3} = -4.94374 - 1.21246X_1 - 0.04414X_2 + 0.11371X_4.$$

La estimación para el dato faltante en la segunda observación se obtiene mediante la predicción en los valores $X_1 = 1.40$, $X_2 = 60$ y $X_4 = 130$, esta es $\hat{X_3} = 5.4927$.

- Hasta aquí, se han "llenado los huecos" en una primera etapa; se dispone de una matriz de 10 datos *completada*. El procedimiento que sigue es la estimación de la regresión de X_1 sobre las variables X_2, X_3 y X_4 con los datos "completados". El modelo estimado es

$$\hat{X_1} = 0.04685 - 0.00833X_2 - 0.04309X_3 + 0.01584X_4.$$

El valor estimado de X_1 en $X_2 = 51$, $= 4.8$ y $X_4 = 115$ es $\hat{X_1} = 1.2368$

Con los mismos datos, la estimación para la segunda observación viene dada por:

$$\hat{X_3} = -4.05205 - 1.55239X_1 - 0.04491X_2 + 0.11147X_4.$$

De donde se tiene que en $X_1 = 1.40$, $X_2 = 60$ y $X_4 = 130$, la estimación de la observación faltante es ahora $\hat{X_3} = 5.5711$.

- Por un proceso similar, en dos etapas más, se obtienen los valores

$$\{\hat{X_1} = 1.2450,\ \hat{X_3} = 5.514726\} \text{ y } \{\hat{X_1} = 1.243778,\ \hat{X_3} = 5.499036\},$$

respectivamente. De manera iterativa se puede observar que estos valores tienden a estabilizarse entorno a $\{\hat{X_1} = 1.25,\ \hat{X_3} = 5.60\}$, los cuales corresponden a una estimación de esta información faltante.

De otra parte la inputación a través de la media de los datos produce la estimación $\{\hat{X_1} = 1.45,\ \hat{X_3} = 6.37\}$, valores bastante diferentes a los conseguidos mediante regresión. El juicio sobre la conveniencia de cada uno de estos métodos, en general, es dado por las caracteísticas que se requieran acerca de las técnicas en donde estos datos sean empleados: por ejemplo: sesgo y varianza de los estimadores, calidad de la predicción, etc. No obstante el juez más apropiado, como ocurre con la mayoría de las metodologías estadísticas, es la calidad que muestren los modelos estadísticos que incorporen este tipo de datos para explicar, controlar y predecir algún fenómeno conceptuado y observado .

Cabe aclarar que se trata de un procedimiento con bastantes limitaciones, toda vez que se han construido modelos de regresión sin indagar sobre la validez de los supuestos requeridos para su misma estimación. No obstante, es una herramienta útil para estos casos. ☑

Una mezcla de los procedimientos anteriores, propuesta por Buck (1960), consiste en la imputación de medias en una primera etapa y las regresiones en una segunda. Una discusión más completa del tratamiento estadístico para observaciones faltantes se puede consultar en Little y Rubin (1987).

1.4.5 Visión Geométrica

Tal como se expuso en la sección (1.1), la matriz de datos multivariados se puede abordar, fundamentalmente, de dos formas: desde el conjunto de individuos o desde las variables. En el primer caso, se denomina el espacio de los individuos (espacio fila), que corresponde a un conjunto de n-individuos en un espacio definido por p-variables, los individuos quedan representados por puntos de p-coordenadas (p-variables), cada eje es una variable. En el segundo caso se denomina el espacio de las variables (columnas), las cuales quedan representadas por los valores que toman en ellas cada uno de los n-individuos. Así, se puede pensar en un espacio de n dimensiones, en el cual cada uno de los individuos está representado por un eje en este espacio. En resumen, el espacio fila o de individuos tiene dimensión p y el espacio columna o de variables tiene dimensión n. Como se afirmó anteriormente, las diferentes técnicas multivariadas se dirigen sobre alguno de estos dos espacios o sobre ambos simultáneamente. Por ejemplo, el análisis discriminante o el análisis por conglomerados, clasifican individuos en función de sus atributos o variables; es decir, se comparan vectores fila. Al comparar vectores columna, se obtiene información de la relación entre los atributos estudiados en términos de los individuos. Técnicas tales como las componentes principales, el análisis de correlación canónica y de regresión múltiple, se concentran sobre el espacio fila para el desarrollo de estas metodologías.

Para facilitar, admítase que se tienen n-individuos sobre los que se han medido las variables X_1 y X_2; es decir, se dispone de una muestra de n-puntos en $\mathbb{R}^2$. El vector $\bar{\boldsymbol{X}}$, se llama el *centroide* de los datos; y se define así

$$\bar{\boldsymbol{X}} = \frac{1}{n}\mathbf{1}'\mathbb{X} = (\bar{x}_1, \bar{x}_2),$$

donde $\mathbf{1}$ es el vector de unos de tamaño $(n \times 1)$ y $\mathbb{X}$ es la matriz de datos de tamaño $(n \times 2)$.

Llamando $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$, con $i = 1, \ldots, n$ y $j = 1, 2$, se tiene que

$$\sqrt{n}\sigma_{X_j} = \left[\sum_{i=1}^{n}(\tilde{x}_{ij})^2\right]^{\frac{1}{2}} = \|\tilde{X}_j\|, \text{ con } j = 1, 2. \tag{1.17}$$

La última expresión relaciona la desviación estándar de un conjunto de datos con la longitud del vector corregido por la media (norma).

La distancia de cada punto (x_{i1}, x_{i2}) al centroide $(\bar{x}_1, \bar{x}_2)$ se estandariza dividiendo por la respectiva norma. El vector resultante, de dividir cada componente por su norma, es unitario. El vector centrado y unitario se nota por X_j^*, $j = 1, 2$; es decir,

$$X_1^* = \frac{\tilde{X}_1}{\sigma_{X_1}} \qquad X_2^* = \frac{\tilde{X}_2}{\sigma_{X_2}}.$$

La matriz de datos originales $\mathbb{X}$, la matriz de datos centrados en la media $\tilde{\mathbb{X}}$ y la matriz de datos estandarizados (reescalados) $\mathbb{X}^*$, respectivamente, se presentan a continuación,

$$\mathbb{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{i1} & x_{i2} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}, \quad \tilde{\mathbb{X}} = \begin{pmatrix} \tilde{x}_{11} & \tilde{x}_{12} \\ \tilde{x}_{21} & \tilde{x}_{22} \\ \vdots & \vdots \\ \tilde{x}_{i1} & \tilde{x}_{i2} \\ \vdots & \vdots \\ \tilde{x}_{n1} & \tilde{x}_{n2} \end{pmatrix}, \quad \mathbb{X}^* = \begin{pmatrix} x^*_{11} & x^*_{12} \\ x^*_{21} & x^*_{22} \\ \vdots & \vdots \\ x^*_{i1} & x^*_{i2} \\ \vdots & \vdots \\ x^*_{n1} & x^*_{n2} \end{pmatrix}.$$

La figura 1.10 muestra los datos originales, los datos corregidos por la media y los datos estandarizados. Nótese que se han realizado dos transformaciones sobre los datos: con la primera transformación, cambio de origen, se obtiene una traslación al origen $(0, 0)$ de los datos, mediante la resta del vector de medias a cada una de las observaciones; mientras que con la segunda se consigue un reescalamiento. Un tercera transformación correspondería a una rotación rígida de los ejes coordenados; este tipo de transformaciones se tratan en el capítulo 5.

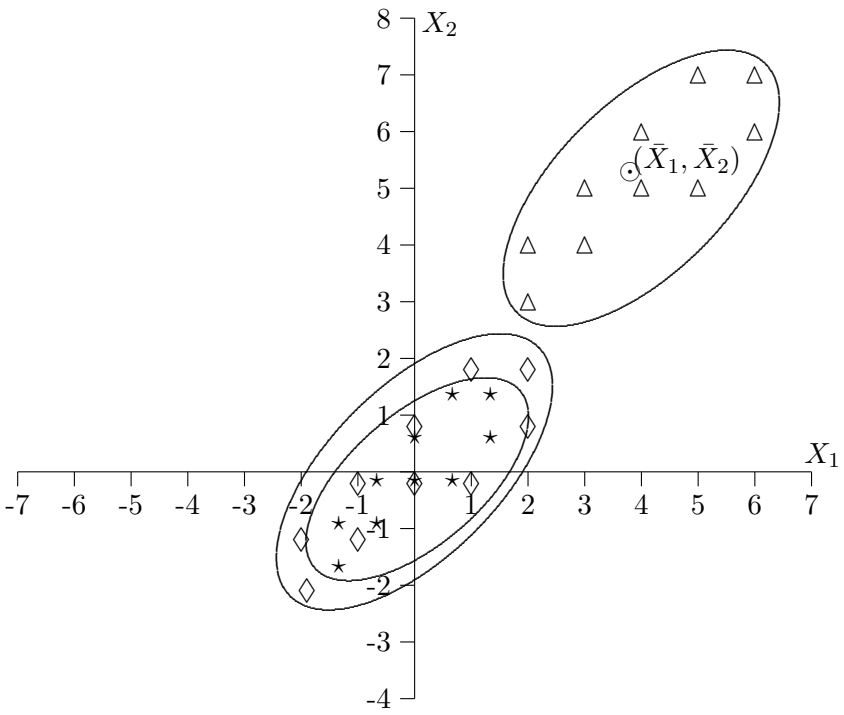


Figura 1.10 *Datos: (△) originales, (◊) corregidos por la media y ⋆ estandarizados.*

1.5 Procesamiento de los datos con SAS/IML

Las siguientes instrucciones permiten calcular el vector de medias, la matriz de covarianzas, la matriz de correlación, para un conjunto de datos. El programa se hace mediante el procedimiento IML (Interactive Matrix Language). Al frente de cada instrucción se explica su propósito dentro de los símbolos /∗ y ∗/. La sintaxis se escribe en mayúsculas fijas, esto no es necesario, simplemente se hace para resaltar los comandos SAS.

```
TITLE 'Procedimiento IML para manipulación de matrices';
OPTIONS NOCENTER PS=60 LS=80;
DATA EJER\_1;  /*  Archivo de datos Ejer\_1 */
INPUT X1 X2 X3 X4 X5 \@\@'; /* Ingreso de las variables  */
CARDS; /*  Para ingresar datos */
insertar aquí  la matriz de datos
;
PROC IML;
USE EJER\_1; /* invoca el archivo Ejer\_1   */
READ ALL INTO X; /* Pone los datos del archivo */
                 /* Ejer\_1 en la matriz X */
N=NROW(X); /* N es el número de observaciones */ UNOS=J(N,1,1); /*
Genera un vector de tamaño (Nx1) de unos */ GL=N-1;
MEDIA=((UNOS)‘*X)/N; /*  Calcula el vector de medias */
XC=X-(UNOS*MEDIA); /* Calcula la matriz XC de datos */
                   /* centrados en la media */
S=(XC)‘*(XC)/GL; /* Calcula la matriz de covarianzas S   */
D=DIAG(S); /*  Diagonaliza la matriz S dejando los elememtos */
           /* de la diagonal */
XS=XC*SQRT(INV(D)); /* Calcula la matriz XS de datos estandarizados
*/ R=(XS)‘*(XS)/GL ; /* Calcula la matriz de correlaciones   */
VG=DET(S); /*  Calcula el determinante de S; es decir, */
          /*  la varianza generalizada */
VT=TRACE(S); /* Calcula la traza de S; es decir, */
             /* la varianza total */
PRINT MEDIA XC S D R VG VT; /*  Imprime cada una de éstos   */
```

1.6 Procesamiento de datos con R

El siguiente código de `R` lee los datos de la tabla 1.1 y con ellos realiza los diagramas de tallos y hojas que están inmediatamente debajo, el dispersograma de la figura 1.4, el box plot de la figura 1.5, los perfiles de la figura 1.3 y por último un gráfico de estrellas y los rostros de Chernoff.

```
# Lectura de los datos de la tabla 1.1
datos<-scan()
125  2536  28   86   2505 31   119 2652 32   113 2573 20
101  2382  30   143  2443 30   132 2617 27   106 2556 36
121  2489  34   109  2415 29   88  2434 27   116 2491
```

```
24   102   2345 26   75   2350 23  90   2536 24  109  2577
22   104   2464 35   110  2571 24  96   2550 24  101  2437
23   95    2472 36   117  2580 21  115  2436 39  138
2200 41    85   2851 17

datos2<-matrix(datos,ncol=3,byrow=TRUE)
tabla1_1<-data.frame(datos2)
colnames(tabla1_1)<-c("CI","Peso","Edad")
# termina la lectura de datos
```

Los gráficos que se crean con este código corresponden a los de la sección 1.2

```
# Diagramas de tallos y hojas
stem(tabla1_1$CI)
stem(tabla1_1$Peso)
stem(tabla1_1$Edad)
# Dispersograma figura 1.4
pairs(tabla1_1)
# En la siguiente linea de código, la función scale()
# estandariza los datos. La función stack()
# convierte la tabla1_1 a un data frame con dos columnas
# la primera contiene los valores y la segunda es un
# factor que identifica a que variable corresponde el valor
datos3<-stack(data.frame(scale(tabla1_1)) )
# Boxplot, figura 1.5  con los datos de la tabla 1.1
plot(datos3$ind,datos3$values)
# El siguiente código crea los perfiles de la matriz de datos
# (figura 1.3) pero adicionalmente dibuja sobre cada
# histograma la gráfica de la densidad normal. Requiere
#  la librería lattice.
 library(lattice)
 histogram(~values|ind,data=datos3, layout = c(3,1),
 type = "density",   panel = function(x, ...)
 {panel.histogram(x, ...)
 panel.mathdensity(dmath = dnorm, col = "black",
 args = list(mean=mean(x),sd=sd(x)))
 } )
 # Gráfico de estrellas
 stars(tabla1_1)
 # Rostros de Chernoff,  requiere la librería  aplpack
 library(aplpack)
 faces(tabla1_1)
```

Calculo de las estadísticas de resumen

```
# Estadísticas de resumen
summary(tabla1_1)
# vector de medias
mean(tabla1_1)
# Matriz de varianzas y covarianzas
```

```
cov(tabla1_1)
# Matriz de correlaciones
cor(tabla1_1)
# VT (traza de la matriz de covarianzas)
sum(diag(cov(tabla1_1)))
# VG (determinante de la matriz de covarianzas)
det(cov(tabla1_1))
```

El mismo procedimiento anterior ahora con los datos de la tabla 1.2.

```
datos<-scan()
1.38    51  4.8     115
1.40    60  5.6     130
1.42    69  5.8     138
1.54    73  6.5     148
1.30    56  5.3     122
1.55    75  7.0     152
1.50    80  8.1     160
1.60    76  7.8     155
1.41    58  5.9     135
1.34    70  6.1     140

datos2<-matrix(datos,ncol=4,byrow=TRUE)
tabla1_2<-data.frame(datos2)
colnames(tabla1_2)<-c("X1","X2","X3","X4")
# termina la lectura de datos
# Diagramas de tallos y hojas
stem(tabla1_2$X1)
stem(tabla1_2$X2)
stem(tabla1_2$X3)
stem(tabla1_2$X3)
# Dispersograma figura 1.4 (con los datos de la tabla 1.2)
pairs(tabla1_2)
# En el siguiente código, la función scale() estandariza
# La función stack() convierte la tabla1_2 a un data frame
# con dos columnas  la primera contiene los valores de
# la variable,  y la segunda es un factor que identifica
# a  qué variable corresponde el valor
datos3<-stack(data.frame(scale(tabla1_2)) )
# Boxplot, figura 1.5  con los datos de la tabla 1.2
plot(datos3$ind,datos3$values)
# El siguiente código crea los perfiles de la matriz de datos
# (figura 1.3) pero adicionalmente dibuja sobre cada
# histograma la gráfica de la densidad normal. Requiere
# la librería lattice.
 library(lattice)
 histogram(~values|ind,data=datos3, layout = c(3,1),
 type = "density",   panel = function(x, ...)
 {panel.histogram(x, ...)
 panel.mathdensity(dmath = dnorm, col = "black",
 args = list(mean=mean(x),sd=sd(x)))
```

```
 } )
 # Gráfico de estrellas
 stars(tabla1_2)
 # Rostros de Chernoff,  requiere la librería  aplpack
  library(aplpack)
  faces(tabla1_2)
# Estadísticas de resumen
summary(tabla1_2)
# vector de medias
round(mean(tabla1_2),2)
# Matriz de varianzas y covarianzas
round(cov(tabla1_2),3)
# Matriz de correlaciones
round(cor(tabla1_2),3)
# VT  (traza de la matriz de covarianzas)
sum(diag(cov(tabla1_2)))
# VG  (determinante de la matriz de covarianzas)
det(cov(tabla1_2))
# Cuadrado de la distancia de mahalanobis entre cada
# observación y el vector de medias
S<-cov(tabla1_2)
mahalanobis(tabla1_2,center=mean(tabla1_2),cov=S)
# Distancia de mahalanobis entre cada
# observación y el vector de medias
 sqrt( mahalanobis(tabla1_2,center=mean(tabla1_2),cov=S) )
# la distancia euclidiana se obtiene tomando la matriz de
# covarianza igual a la  identidad la cual se obtiene con
# el comando  diag(1,n)  (identidad de orden n)
 I4<-diag(1,4)
 sqrt(mahalanobis(tabla1_2,center=mean(tabla1_2),cov=I4) )
```

Capítulo 2

Distribuciones multivariantes

2.1 Introducción

Los valores de la mayoría de las medidas asociadas con objetos se aglomeran simétricamente en torno a un valor central específico. La mayoría de estas medidas se ubican dentro de alguna distancia determinada respecto a un valor central, a la izquierda o a la derecha, las demás se presentan de manera cada vez más escasa, en tanto que la distancia al valor central es grande. Lo anterior corresponde a una descripción intuitiva de la variable cuyos valores se distribuyen conforme a una distribución *normal*. El nombre de "normal" procede del uso en algunas disciplinas, las cuales asumen como *normales* a los individuos cuyos atributos se ubican dentro de cierto intervalo centrado en un valor específico[1].

Un número amplio de los métodos de inferencia estadística, para el caso univariado, se apoya sobre el supuesto de distribución normal e independencia entre las observaciones. En casos de no normalidad, existen algunas alternativas, como las que se nombran a continuación, para conseguirla o enfrentarla: (i) mediante teoremas límites, ii) a través de transformación de los datos, (iii) el empleo de técnicas de libre distribución o no paramétricas, o (iv) técnicas robustas a la normalidad.

De manera análoga, muchas de las metodologías del análisis multivariado se apoyan sobre la distribución *normal multivariante*, aunque muchos de los procedimientos son útiles aún sin la normalidad de los datos. Las siguientes son algunas de las justificaciones para el empleo de la distribución normal multivariante:

- Es una fácil extensión de la distribución normal univariante; tanto en su definición como en su aplicación.
- Queda completamente definida por los dos primeros momentos. El número de parámetros asociado es $(1/2)p(p+3)$, con lo cual se facilita la estimación.
- Bajo normalidad, variables aleatorias con covarianza cero son independientes dos a dos y en conjunto, además, recíprocamente, la no correlación implica independencia. Esto no siempre se tiene bajo otras distribuciones.
- La combinación lineal de variables aleatorias con distribución normal tiene distribución normal.

[1] Aunque el aforismo estadístico dice que "lo más anormal es la normalidad"... en un conjunto de datos.

- Cuando los datos no tienen distribución multinormal, se recurre a teoremas límites que garantizan normalidad en muestras de tamaño grande.

Se desarrollan en este capítulo los conceptos y características ligadas a la distribución normal multivariante, en la forma clásica a través de la función de densidad de probabilidad. También se trata con algunas distribuciones básicas conectadas a la distribución normal multivariada, tales como la distribución ji-cuadrado no central, t-Student no central, la F no central y la distribución de Wishart; distribuciones que justifican algunas propiedades y métodos de la inferencia estadística (capítulos 3 y 4). Algunas herramientas para inspeccionar si un conjunto de datos se ajusta a una normal multivariante son tratadas junto con transformaciones que les permiten acondicionarse a una distribución normal multidimensional. Finalmente se aborda, con un enfoque geométrico, la distribución normal multivariante y así se hacen más asequibles tales conceptos mediante la distribución normal bivariada.

2.2 La distribución normal multivariante

Aunque existen varias formas de presentar la distribución normal multivariada, se expone a continuación, casi que por construcción, la distribución normal multivariante. El camino a seguir es la identificación de su distribución mediante la función generadora de momentos. Con esta definición resulta sencillo construir un algoritmo computacional para simular datos procedentes de una determinada distribución normal multivariada.

Sea $Z' = (Z_1, \ldots, Z_p)$ un vector con p variables aleatorias independientes y cada una con distribución normal estándar; es decir, $Z_i \sim n(0, 1)$. Entonces

$$\mathcal{E}(Z) = 0, \quad \text{Cov}(Z) = \boldsymbol{I}, \quad M_Z(t) = \prod_{i=1}^{p} \exp\left\{\frac{t_i^2}{2}\right\} = \exp\frac{t't}{2}.$$

Considérese el vector $\boldsymbol{\mu}$ y la matriz A de tamaño $(p \times p)$. El vector $X = AZ + \boldsymbol{\mu}$ es tal que

$$\mathcal{E}(X) = \boldsymbol{\mu}, \quad \text{Cov}(X) = AA'.$$

La función generadora de momentos de X es dada por

$$\begin{aligned} M_X(t) &= \exp\left\{\boldsymbol{\mu}'t\right\} M_Z(A't) \\ &= \exp\left\{\boldsymbol{\mu}'t + \frac{t'(A'A)t}{2}\right\} \\ &= \exp\left\{\boldsymbol{\mu}'t + \frac{t'\boldsymbol{\Sigma}t}{2}\right\}, \end{aligned}$$

con $\boldsymbol{\Sigma} = AA'$.

En consecuencia, se puede afirmar que un vector p-dimensional X, tiene distribución *normal p-variante*, con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$, si y sólo si, la función generadora de momentos de X es:

$$M_X(t) = \exp\left(\boldsymbol{\mu}'t + \frac{t'\boldsymbol{\Sigma}t}{2}\right)$$

Se nota $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Ahora se encuentra la función de densidad para X. Del resultado anterior se afirma que

$$Z \sim N_p(0, I), \quad \textit{con } Z' = (Z_1, \ldots, Z_p).$$

Por la independencia entre los Z_i, su densidad conjunta es, de acuerdo con (B.39),

$$f_Z(z) = \prod_{i=1}^{p} \frac{1}{(2\pi)^{1/2}} \exp\left\{-\frac{1}{2} z_i^2\right\} = \frac{1}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2} z'z\right\}.$$

Sea $X = \boldsymbol{\Sigma}^{1/2} Z + \boldsymbol{\mu}$, entonces por el resultado anterior $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. El vector Z se puede expresar como $Z = \boldsymbol{\Sigma}^{-1/2}(X - \boldsymbol{\mu})$, expresión que es una transformación invertible. El jacobiano de la transformación (sección (B.4)) es $J = \mid \boldsymbol{\Sigma} \mid^{-1/2}$. Por tanto la función de densidad conjunta de X es

$$f_X(x) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\Big\{-\frac{1}{2}(x - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(x - \boldsymbol{\mu})\Big\}, \tag{2.1}$$

donde $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ y $\boldsymbol{\Sigma}$ es una matriz simétrica definida positiva de tamaño $p \times p$.

Observación:

> Otra definición alterna de *distribución normal multivariante* es la siguiente: un vector X de tamaño $(p \times 1)$ tiene distribución normal p-variante, si para todo $\boldsymbol{a} \in \mathbb{R}^p$ la distribución de $\boldsymbol{a}'X$ es normal univariada. Muirhead (1982, pág. 5) .

Las propiedades que se muestran a continuación pueden derivarse desde cualquiera de las dos definiciones anteriormente dadas.

2.2.1 Propiedades de la distribución normal multivariada

A continuación se hace una caracterización muy sucinta sobre la distribución normal p variante. Los interesados en seguir este desarrollo en una forma más detallada pueden consultar a Anderson (1984) o Rencher (1998).

Observación: generación de datos normales multivariados (simulación)

> El procedimiento seguido para obtener la función de densidad conjunta dada en la ecuación (2.1), se puede emplear para generar vectores aleatorios con distribución normal multivariante, a través de simulación en un computador. Si se decide generar una matriz de datos $\mathbb{X}$ desde $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con los valores de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ conocidos, se puede usar $X = \boldsymbol{\Sigma}^{1/2} Z + \boldsymbol{\mu}$, donde Z es $N_p(0, I_p)$. Alternativamente, se puede factorizar a $\boldsymbol{\Sigma}$ como $\boldsymbol{\Sigma} = AA'$, usando la descomposición de Cholesky (A2.35) y definir $X = AZ + \boldsymbol{\mu}$, donde $A = \boldsymbol{\Sigma}^{1/2}$. El vector Z está conformado por p variables normales estándar independientes, las cuales se pueden obtener fácilmente en el computador. Al final del capítulo se presenta un programa con el procedimieto IML del SAS para simular distribuciones normales p-variantes.

Propiedad 2.2.1

Determinación. Si un vector aleatorio $X_{p\times 1}$, tiene distribución normal multivariante, entonces su media es $\boldsymbol{\mu}$ y su matriz de varianzas y covarianzas es $\boldsymbol{\Sigma}$.

En adelante se indica que un vector aleatorio X tiene distribución normal p-variante con vector de medias $\mathcal{E}(X) = \boldsymbol{\mu}$ y matriz de covarianzas $\mathrm{Cov}(X) = \boldsymbol{\Sigma}$, escribiendo:

$$X \sim N_p\ (\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Esto significa que la distribución normal queda completamente determinada a través del vector $\boldsymbol{\mu}$ y la matriz $\boldsymbol{\Sigma}$.

Propiedad 2.2.2

Linealidad. Si X es un vector aleatorio p-dimensional distribuido normalmente, con vector de medias $\boldsymbol{\mu}$ y matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$, entonces el vector $Y = AX+b$, con A una matriz de tamaño $(q \times p)$ y b un vector de tamaño $(q \times 1)$, tiene distribución normal q-variante, con vector de medias $A\boldsymbol{\mu} + b$ y matriz de varianzas y covarianzas $A\boldsymbol{\Sigma}A'$.

En símbolos, si $X \sim N_p\ (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ entonces

$$Y = (AX + b) \sim N_q(A\boldsymbol{\mu} + b; A\boldsymbol{\Sigma}A').$$

Propiedad 2.2.3

Marginales. Considérese el vector X particionado como $X = (X_{(1)}, X_{(2)})$, con $X_{(1)} = (X_1, \ldots, X_{p_1})$, $X_{(2)} = (X_{p_1+1}, \ldots, X_p)$, y sea $\boldsymbol{\mu}$ particionado similarmente como $\boldsymbol{\mu}' = (\boldsymbol{\mu}_{(1)}, \boldsymbol{\mu}_{(2)})$ y además $\boldsymbol{\Sigma}$ particionada:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

donde $\boldsymbol{\Sigma}_{11}$ es la submatriz superior izquierda de $\boldsymbol{\Sigma}$ de tamaño $p_1 \times p_1$. Si X tiene distribución normal con media $\boldsymbol{\mu}$ y matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ (definida positiva) y $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21} = \mathbf{0}$, entonces los vectores $X_{(1)}$ y $X_{(2)}$ son independientes y normalmente distribuidos con vectores de medias $\boldsymbol{\mu}_{(1)}$, $\boldsymbol{\mu}_{(2)}$ y matrices de varianzas y covarianzas $\boldsymbol{\Sigma}_{11}$ y $\boldsymbol{\Sigma}_{22}$ respectivamente. De otra manera, cualquier subvector de un vector con distribución normal p variante tiene distribución normal, con subvector de medias y submatriz de covarianzas los asociados a las componentes de éste[2].

Observaciones:

- Se enfatiza en que la independencia debida a la incorrelación se garantiza por el supuesto de normalidad; de lo contrario, no siempre es válida la proposición.
- Para el caso bivariado la partición es:

  $$\begin{array}{lll} X_{(1)} = X_1, & & X_{(2)} = X_2; \\ \boldsymbol{\mu}_{(1)} = \boldsymbol{\mu}_1, & & \boldsymbol{\mu}_{(2)} = \boldsymbol{\mu}_2; \\ \boldsymbol{\Sigma}_{11} = \sigma_1^2, & \boldsymbol{\Sigma}_{22} = \sigma_2^2 \ y & \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = \sigma_1\sigma_2\rho_{12}; \end{array}$$

  con ρ_{12} el coeficiente de correlación lineal entre X_1 y X_2. Así, las variables aleatorias X_1 y X_2 con distribución normal conjunta, son independientes si y sólo si son incorrelacionadas. Si son incorrelacionadas la distribución marginal de X_i es normal con media $\boldsymbol{\mu}_i$ y varianza σ_i^2 (para $i = 1, 2$).
- La partición anterior se hace para ilustrar cómodamente, pero bien pueden escogerse las p_1 y p_2 variables de cualquier manera dentro del vector X (con $p_1+p_2 = p$).

Propiedad 2.2.4

[2]Se asume que un subvector (submatriz) es un arreglo reordenado de algunas componentes de un vector (matriz).

Independencia. La matriz de varianzas y covarianzas de un vector aleatorio $X_{p\times 1}$, con distribución normal p variante es diagonal si y sólo si los componentes de X son variables aleatorias normales e independientes.

De esta propiedad se puede expresar (2.1) como el producto de las funciones de densidad asociadas con cada una de las componentes del vector aleatorio X, así:

$$f_X(x) = f_1(x_1)\cdots f_p(x_p) = \prod_{i=1}^{p} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_i} \exp\Big\{-\frac{1}{2}\Big(\frac{x_i-\mu_i}{\sigma_i}\Big)^2\Big\}.$$

La siguiente propiedad es un caso particular de la transformación dada en la propiedad (2.2.2), la cual es el equivalente a la estandarización para el caso univariado.

Propiedad 2.2.5

"Estandarización". Sea X un vector aleatorio p-dimensional distribuido normalmente con vector de medias $\boldsymbol{\mu}$ y matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$. Si $\boldsymbol{\Sigma}$ es una matriz no singular entonces:

$$Z = \boldsymbol{\Sigma}^{-1/2}(X-\boldsymbol{\mu})$$

tiene distribución normal p variante con vector de medias cero y matriz de varianzas y covarianzas la identidad $\mathbf{I}_p$, donde $\boldsymbol{\Sigma}^{-1/2} = (\boldsymbol{\Sigma}^{-1})^{1/2}$ tal como se define en (A2.34). En símbolos, si

$$X \sim N_p\ (\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \text{entonces},\ Z = \boldsymbol{\Sigma}^{-1/2}(X-\boldsymbol{\mu}) \sim N_p\ (0, I).$$

Nótese que es equivalente al caso univariado ($p = 1$), pues si

$$x \sim n(\mu, \sigma^2),\ \text{entonces},\ z = \frac{x-\mu}{\sigma} = (\sigma^2)^{-\frac{1}{2}}(x-\mu) \sim n(0,1).$$

Propiedad 2.2.6

Distribución condicional. Considérese la misma partición efectuada para la propiedad (2.2.3), con $X_{(1)}$ y $X_{(2)}$ de tamaños $(p_1 \times 1)$ y $(p_2 \times 1)$, respectivamente $(p_1 + p_2 = p)$. La función de densidad condicional de $X_{(1)}$ dado $X_{(2)} = x_{(2)}$, de acuerdo con la sección (B.4), se obtiene de

$$g(x_{(1)} \mid x_{(2)}) = \frac{f(x_{(1)}, x_{(2)})}{h(x_{(2)})},$$

donde h es la función de densidad marginal para $X_{(2)}$, es decir

$$h(x_{(2)}) = \frac{1}{(2\pi)^{p_2/2} \mid \boldsymbol{\Sigma}_{22} \mid^{\frac{1}{2}}} \exp\Big\{-\frac{1}{2}(x_{(2)}-\boldsymbol{\mu}_{(2)})'\boldsymbol{\Sigma_{22}}^{-1}(x_{(2)}-\boldsymbol{\mu}_{(2)})\Big\}.$$

La función de densidad conjunta $f(x_{(1)}, x_{(2)})$ es la normal multivariante expresada en (2.1). Nótese que la forma cuadrática del exponente contiene la inversa de la matriz $\boldsymbol{\Sigma}$, la cual está particionada en bloques, y se obtiene a través de (A2.40). Ésta es

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} (\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\Sigma}_{12})')^{-1} & -(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\Sigma}_{12})')^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\Sigma}_{12})'(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\Sigma}_{12})')^{-1} & \boldsymbol{\Sigma}_{22}^{-1}+\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\Sigma}_{12})'(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\Sigma}_{12})')^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix}.$$

El determinante de la matriz particionada $\boldsymbol{\Sigma}$, de acuerdo con (A2.41), es

$$\mid \boldsymbol{\Sigma} \mid = \mid \boldsymbol{\Sigma}_{22} \mid \mid \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}' \mid .$$

Al reemplazar las dos últimas expresiones en el numerador de $g(x_{(1)} \mid x_{(2)})$, después de hacer las operaciones y las simplificaciones pertinentes, se llega al siguiente resultado:

$$g(x_{(1)} \mid x_{(2)}) = \frac{1}{(2\pi)^{p_1/2} \mid \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}' \mid^{1/2}}$$

$$\exp\Big\{-\frac{1}{2}\big[x_{(1)} - (\boldsymbol{\mu}_{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x_{(2)} - \boldsymbol{\mu}_{(2)}))\big]'$$

$$\big(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}'\big)^{-1}\big[x_{(1)} - (\boldsymbol{\mu}_{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x_{(2)} - \boldsymbol{\mu}_{(2)}))\big]\Big\}.$$

La función $g(x_{(1)} \mid x_{(2)})$ es la función de densidad normal p_1 variante con vector de medias

$$\boldsymbol{\mu}_{X_{(1)}|X_{(2)}} = \boldsymbol{\mu}_{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x_{(2)} - \boldsymbol{\mu}_{(2)}), \tag{2.2a}$$

y matriz de covarianzas

$$\boldsymbol{\Sigma}_{X_{(1)}|X_{(2)}} = \boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}'. \tag{2.2b}$$

Se nota que la media de $X_{(1)}$, dado $X_{(2)}$ es simplemente una función lineal de $X_{(2)}$ y que la matriz de varianzas y covarianzas de $X_{(1)}$, dado $X_{(2)}$, no depende del todo de $X_{(2)}$.

La matriz $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ es la matriz de coeficientes de la *regresión* de $X_{(1)}$ sobre $X_{(2)}$.

El vector $\boldsymbol{\mu}_{X_{(1)}|X_{(2)}} = \boldsymbol{\mu}_{(1)} + \boldsymbol{\beta}(x_{(2)} - \boldsymbol{\mu}_{(2)})$ se llama frecuentemente *función de regresión* de $X_{(1)}$ sobre $X_{(2)}$.

Nótese que en el caso unidimensional ($p_1 = p_2 = 1$), se trata de una regresión lineal simple; es decir, se espera que bajo normalidad el dispersograma de $X_{(2)}$ frente a $X_{(1)}$ se aproxime a una línea recta de la forma $X_{(1)} = \beta_0 + \beta_1 X_{(1)}$, con

$\beta_1 = \sigma_{12}/\sigma_2^2$ y $\beta_0 = \mu_{(1)} - [\sigma_{12}/\sigma_2^2]\mu_{(2)}$.

Esta propiedad es útil para el diagnóstico de multinormalidad en un conjunto de datos como se aduce en la sección (2.5).

Propiedad 2.2.7

Los subvectores $X_{(2)}$ y $X_{(1)}^* = X_{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(X_{(2)} - \boldsymbol{\mu}_{(2)})$ son independientes y normalmente distribuidos con medias

$$\boldsymbol{\mu}_{(2)} \text{ y } \boldsymbol{\mu}_{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_{(2)},$$

y matrices de varianzas y covarianzas (definidas positivas)

$$\boldsymbol{\Sigma}_{22} \; y \; \boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21},$$

respectivamente.

La independencia entre los subvectores $X_{(2)}$ y $X_{(1)}^*$ se garantiza demostrando que: $\boldsymbol{\Sigma}_{21}^* = \mathcal{E}([X_{(2)} - \boldsymbol{\mu}_{(2)}]X_{(1)}^*) = 0$.

Para terminar el paralelo con la regresión lineal, el vector

$$\begin{aligned}\mathbf{E}_{(1\cdot 2)} = X_{(1)} - \boldsymbol{\mu}_{X_{(1)}|X_{(2)}} &= X_{(1)} - \big[\boldsymbol{\mu}_{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x_{(2)} - \boldsymbol{\mu}_{(2)})\big] \\ &= X_{(1)} - \big[\boldsymbol{\mu}_{(1)} + \boldsymbol{\beta}(x_{(2)} - \boldsymbol{\mu}_{(2)})\big],\end{aligned}$$

es el vector de residuales entre $X_{(1)}$ y los valores predichos por la regresión de $X_{(1)}$ sobre $X_{(2)}$. De lo anterior se establece que bajo el supuesto de normalidad, los residuales y las variables regresoras (fijas) son independientes.

Propiedad 2.2.8

Combinación lineal de multinormales. Sean $X_1, \ldots, X_n$ vectores aleatorios independientes de tamaño $(p \times 1)$ con distribución $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Entonces, la combinación lineal $L_1 = c_1X_1 + \cdots + c_nX_n$ se distribuye $N_p\Big(\sum_{i=1}^n c_i\boldsymbol{\mu}_i, \big(\sum_{i=1}^n c_i^2\big)\boldsymbol{\Sigma}\Big)$. Además, L_1 y $L_2 = d_1X_1 + \cdots + d_nX_n$ tienen distribución normal conjunta, con vector de medias

$$\begin{pmatrix} \sum_{i=1}^n c_i\boldsymbol{\mu}_i \\ \sum_{i=1}^n d_i\boldsymbol{\mu}_i \end{pmatrix},$$

y matriz de covarianzas

$$\begin{bmatrix} \big(\sum_{i=1}^n c_i^2\big)\boldsymbol{\Sigma} & (d'c)\boldsymbol{\Sigma} \\ (d'c)\boldsymbol{\Sigma} & \big(\sum_{i=1}^n d_i^2\big)\boldsymbol{\Sigma} \end{bmatrix}.$$

Las dos combinaciones lineales L_1 y L_2 son independientes si $d'c = \sum_{i=1}^n c_i d_i = 0$.

Ejemplo 2.1 Sea X un vector aleatorio de tamaño (4×1) con distribución $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde

$$\boldsymbol{\mu} = \begin{pmatrix} 2 \\ -1 \\ 3 \\ 1 \end{pmatrix} \text{ y } \boldsymbol{\Sigma} = \begin{pmatrix} 7 & 3 & -3 & 2 \\ 3 & 6 & 0 & 4 \\ -3 & 0 & 5 & -2 \\ 2 & 4 & -2 & 4 \end{pmatrix}.$$

La matriz $\boldsymbol{\Sigma}$ es una auténtica matriz de covarianzas, pues $det(\boldsymbol{\Sigma}) = 16 > 0$. A través de la matriz $A = \begin{pmatrix} 1 & -2 & 0 & 0 \\ 0 & 1 & -1 & 3 \end{pmatrix}$ y el vector $b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ se hace la transformación $Y = AX + b$, la cual corresponde a un vector de tamaño (2×1)

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 - 2X_2 + 1 \\ X_2 - X_3 + 3X_4 + 2 \end{pmatrix}.$$

Por la propiedad (2.2.2) el vector Y tiene distribución normal bivariada con vector de medias $\boldsymbol{\mu}_Y = A\boldsymbol{\mu} + b = \left(\begin{smallmatrix} 5 \\ 1 \end{smallmatrix}\right)$ y matriz de covarianzas $\boldsymbol{\Sigma}_Y = A\boldsymbol{\Sigma}A' = \left(\begin{smallmatrix} 19 & -24 \\ -24 & 83 \end{smallmatrix}\right)$.

La distribución de cada una de las componentes del vector X es normal univariada; en particular, la variable X_3 tiene distribución normal con media 3 y varianza 5; ésta se obtiene de $Y = X_3 = \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix} X$, la cual tiene media varianza $\begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix} \boldsymbol{\mu} = \mu_3 = 3$ y varianza $\begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix} \boldsymbol{\Sigma} \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}' = 5$. En general, cada componente se obtiene al hacer la multiplicación entre el respectivo vector canónico y el vector X.

La distribución del subvector $X_{(1)} = (X_1, X_4)$ se obtiene de la transformación

$$Y = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} X,$$

este subvector tiene distribución normal con media $\boldsymbol{\mu}'_{(1)} = (2, 1)$ y matriz de covarianzas $\boldsymbol{\Sigma}_{(1)} = \left(\begin{smallmatrix} 7 & 2 \\ 2 & 4 \end{smallmatrix}\right)$. Nótese que el vector de medias y la matriz de covarianzas del subvector $X_{(1)}$ se obtienen tomando los elementos correspondientes de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, respectivamente.

Los subvectores $X_{(1)} = \binom{X_2}{X_4}$ y $X_{(2)} = \binom{X_1}{X_3}$ corresponden a un "reordenamiento" de X, $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ de la siguiente manera

$$Y = \begin{pmatrix} X_2 \\ X_4 \\ \cdots \\ X_1 \\ X_3 \end{pmatrix}, \ \boldsymbol{\mu}_Y = \begin{pmatrix} \mu_2 \\ \mu_4 \\ \cdots \\ \mu_1 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ \cdots \\ 2 \\ 3 \end{pmatrix} \text{ y}$$

$$\boldsymbol{\Sigma}_Y = \begin{pmatrix} \sigma_{22} & \sigma_{24} & : & \sigma_{21} & \sigma_{23} \\ \sigma_{42} & \sigma_{44} & : & \sigma_{41} & \sigma_{43} \\ \cdots & \cdots & \cdot & \cdots & \cdots \\ \sigma_{12} & \sigma_{14} & : & \sigma_{11} & \sigma_{13} \\ \sigma_{32} & \sigma_{34} & : & \sigma_{31} & \sigma_{33} \end{pmatrix} = \begin{pmatrix} 6 & 4 & : & 3 & 0 \\ 4 & 4 & : & 2 & -2 \\ \cdots & \cdots & \cdot & \cdots & \cdots \\ 3 & 2 & : & 7 & -3 \\ 0 & -2 & : & -3 & 5 \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & : & \Sigma_{12} \\ \cdots & \cdot & \cdots \\ \Sigma_{21} & : & \Sigma_{22} \end{pmatrix}.$$

La partición anterior se deriva de la trasformación

$$\begin{aligned} Y &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \\ &= \begin{pmatrix} X_2 \\ X_4 \\ X_1 \\ X_3 \end{pmatrix} = \begin{pmatrix} X_{(1)} \\ X_{(2)} \end{pmatrix}. \end{aligned}$$

Finalmente, como las variables aleatorias, X_2 y X_3, tienen distribución normal y la covarianza entre éstas es cero, se concluye que son independientes. ☑

2.2.2 Correlación parcial

Con la misma partición para X, $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, contemplada en la propiedad 2.2.3, se mide la *dependencia* entre las p_1 variables del subvector $X_{(1)}$, manteniendo fijos o "controlados" los valores de las restantes p_2 variables contenidas en $X_{(2)}$. Es decir, la población se "estratifica" de acuerdo con los valores fijos en los cuales se mantienen tales variables; se busca la asociación entre las demás variables dentro del "estrato" definido. De esta manera, la covarianza entre las variables X_i y X_j (ambas en $X_{(1)}$) dado que $X_{(2)} = c_2$, corresponde al elemento (i, j) de la matriz (2.2b), se nota por $\sigma_{ij|p_1+1,\ldots,p}$. La escritura más explícita de (2.2b) es

$$\begin{aligned} \boldsymbol{\Sigma}_{11\cdot 2} &= \left(\sigma_{ij|p_1+1,\ldots,p}\right) \\ &= \begin{pmatrix} \sigma_{11|p_1+1,\ldots,p} & \sigma_{12|p_1+1,\ldots,p} & \cdots & \sigma_{1p_1|p_1+1,\ldots,p} \\ \sigma_{12|p_1+1,\ldots,p} & \sigma_{22|p_1+1,\ldots,p} & \cdots & \sigma_{2p_1|p_1+1,\ldots,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p_1|p_1+1,\ldots,p} & \sigma_{2p_1|p_1+1,\ldots,p} & \cdots & \sigma_{p_1p_1|p_1+1,\ldots,p} \end{pmatrix}. \end{aligned} \tag{2.3}$$

El coeficiente de *correlación parcial* entre las variables i y j de $X_{(1)} = (X_1, \ldots, X_{p_1})$, manteniendo las variables de $X_{(2)=(X_{p_1+1},\ldots,X_p)}$ constantes, está definido por

$$\rho_{ij|p_1+1,\ldots,p} = \frac{\sigma_{ij|p_1+1,\ldots,p}}{\sqrt{\sigma_{ii|p_1+1,\ldots,p} \cdot \sigma_{jj|p_1+1,\ldots,p}}}.$$

Similar al coeficiente de correlación de Pearson (producto-momento), el coeficiente de correlación parcial satisface

$$-1 \leq \rho_{ij|p_1+1,\ldots,p} \leq 1.$$

El coeficiente de correlación parcial para dos variables puede ser definido como la correlación de errores después de ajustar la regresión sobre el segundo conjunto de variables.

Ejemplo 2.2 Las variables tórax (X_1), abdomen (X_2), circunferencia craneana (X_3), longitud del brazo (X_4) y longitud de la pierna (X_5) (medidas en cm.), se pueden asumir, para un grupo humano específico, como una distribución normal conjunta 5-variante. Con propósitos ilustrativos supóngase que la distribución está caracterizada por el vector de medias, la matriz de covarianzas y la matriz de correlaciones siguientes:

$$\boldsymbol{\mu} = \begin{pmatrix} 93.92 \\ 75.18 \\ 57.14 \\ 69.07 \\ 94.18 \end{pmatrix}, \; \boldsymbol{\Sigma} = \begin{pmatrix} 68.51 & 64.16 & 19.23 & 6.19 & 18.53 \\ 64.16 & 86.44 & 21.19 & 9.61 & 12.30 \\ 19.23 & 21.19 & 13.68 & 7.32 & 13.45 \\ 6.19 & 9.61 & 7.32 & 45.47 & 31.28 \\ 18.53 & 12.30 & 13.45 & 31.28 & 47.63 \end{pmatrix},$$

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.83 & 0.63 & 0.11 & 0.32 \\ 0.83 & 1.00 & 0.62 & 0.15 & 0.19 \\ 0.63 & 0.62 & 1.00 & 0.29 & 0.53 \\ 0.11 & 0.15 & 0.29 & 1.00 & 0.67 \\ 0.32 & 0.19 & 0.53 & 0.67 & 1.00 \end{pmatrix}.$$

Las cinco variables antropomórficas anteriores se pueden dividir en dos grupos. Las tres primeras se pueden asociar con la contextura corporal y las dos últimas con las extremidades. El vector $X' = (X_1, X_2, X_3, X_4, X_5)$ se particiona en los subvectores $X'_{(1)} = (X_1, X_2, X_3)$ y $X'_{(2)} = (X_4, X_5)$ ($p_1 = 3$ y $p_2 = 2$). Análogamente, se particionan el vector de medias y la matriz de covarianzas. La matriz de covarianza particionada es

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \vdots & \boldsymbol{\Sigma}_{12} \\ \cdots & & \cdots \\ \boldsymbol{\Sigma}_{21} & \vdots & \boldsymbol{\Sigma}_{22} \end{pmatrix} = \left(\begin{array}{ccc:cc} 68.51 & 64.16 & 19.23 & 6.19 & 18.53 \\ 64.16 & 86.44 & 21.19 & 9.61 & 12.30 \\ 19.23 & 21.19 & 13.68 & 7.32 & 13.45 \\ \hdashline 6.19 & 9.61 & 7.32 & 45.47 & 31.28 \\ 18.53 & 12.30 & 13.45 & 31.28 & 47.63 \end{array}\right).$$

De acuerdo con las propiedades anteriores se puede concluir, entre otras cosas, las siguientes:

1. Los vectores $X_{(i)} \sim N_{p_i}(\boldsymbol{\mu}_{(i)}, \boldsymbol{\Sigma}_{ii})$, con $i = 1, 2$. En particular, la longitud de las extremidades se ajusta a distribución normal bivariada de media $\boldsymbol{\mu}'_{(2)} = (69.07, 94.18)$ y matriz de covarianzas $\boldsymbol{\Sigma}_{22} = \left(\begin{smallmatrix} 45.47 & 31.28 \\ 31.28 & 47.63 \end{smallmatrix}\right)$.
2. Los vectores $X_{(1)}$ y $X_{(2)}$ no son independientes, pues $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21} \neq \mathbf{0}$
3. Cada una de las variables tiene distribución normal con media, la respectiva componente de $\boldsymbol{\mu}$, y varianza el correspondiente elemento de la diagonal de la matriz $\boldsymbol{\Sigma}$. Específicamente, $X_2 \sim n(75.18; 86.44)$.

4. La matriz de covarianzas parcial, asociada a $X_{(1)}$, dado $X_{(2)}$, que está dada por la ecuación (2.2b) es

$$\boldsymbol{\Sigma}_{X_{(1)}|X_{(2)}} = \boldsymbol{\Sigma}_{11\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\Sigma}_{12})' = \begin{pmatrix} 64.65 & 64.52 & 14.72 \\ 64.52 & 89.83 & 19.23 \\ 14.72 & 19.23 & 10.57 \end{pmatrix},$$

la cual mide el grado de asociación lineal entre el tórax, el abdomen y la circunferencia craneana, manteniendo fijos la longitud de los brazos y de las piernas.

5. La correlación parcial entre el tórax y el abdomen, manteniendo fijas la longitud de los brazos y la longitud de las piernas, de acuerdo con la última matriz es

$$\begin{aligned} \rho_{12|4,5} &= \frac{\sigma_{12|4,5}}{\sqrt{\sigma_{11|4,5}\cdot\sigma_{22|4,5}}} \\ &= \frac{64.52}{\sqrt{64.65\cdot 89.83}} = \boxed{0.84} \end{aligned}$$

Es decir, existe una alta relación lineal entre estas dos variables, en personas de este grupo humano, quienes tienen una determinada longitud de sus extremidades. La matriz de correlación parcial completa es

$$(\rho_{ij|4,5}) = \begin{pmatrix} 1.00 & \boxed{0.84} & 0.56 \\ 0.84 & 1.00 & 0.62 \\ 0.56 & 0.62 & 1.00 \end{pmatrix}.$$

Se observa que la correlación parcial entre las variables tórax X_1 y abdomen X_2, notada por $\rho_{12|4,5}$, es aproximadamente igual a la correlación ρ_{12}; es decir, desde estos datos, se puede afirmar que la correlación entre las variables tórax y abdomen es casi la misma, independientemente de la longitud de las extremidades de tales personas. Una lectura e interpretación similar se puede hacer para la correlación parcial y no parcial entre las variable abdomen X_2 y circunferencia craneana X_3. No obstante, la correlación parcial entre la variable tórax X_1 y la variable circunferencia craneana X_3 es menor que la correlación ordinaria; a partir de estas correlaciones se puede afirmar que las personas cuyas extremidades tienen la longitud registrada muestran una asociación menor entre estas variables que cuando estas longitudes no se tienen en cuenta. ☑

2.3 Distribuciones asociadas a la normal multivariante

En esta sección se presentan, condensadamente, las distribuciones de uso más frecuente en el análisis estadístico multivariado.

2.3.1 Distribución ji-cuadrado no central

Sea X un vector de tamaño $(p \times 1)$ distribuido $N_p(\boldsymbol{\mu}, I)$, si se define $U = X'X$, la cual tiene distribución *ji-cuadrado no central*, si su función de densidad de probabilidad está

dada por:

$$\chi^2(u) = \begin{cases} \sum_{j=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^j}{j!} \right) \left(\frac{u^{(p+2j-2)/2} e^{-u/2}}{\Gamma\left(\frac{p+2j}{2}\right) 2^{j+(p/2)}} \right), & \text{para } u > 0 \\ 0, & \text{en otra parte,} \end{cases} \tag{2.4}$$

con $\boldsymbol{\lambda} = \boldsymbol{\mu}'\boldsymbol{\mu} \geq 0$. Se define $\lambda^j = 1$ para $\lambda = j = 0$.

Observaciones:

- En (2.4) aparece el término $(e^{-\lambda}\lambda^j/j!)$, que es la función de probabilidad de una variable aleatoria tipo Poisson con parámetro $\boldsymbol{\lambda}$. Cada término de la suma es el producto entre un término de una distribución tipo Poisson y el respectivo de ji-cuadrado central, con $(p + 2\boldsymbol{\lambda})$ grados de libertad; es decir, es una combinación lineal de ji-cuadrados centrales con coeficientes Poisson.
- La cantidad p corresponde a los grados de libertad de la distribución ji-cuadrado no central. y a $\boldsymbol{\lambda}$ se le denomina el *parámetro de no centralidad*.
- Si $\boldsymbol{\lambda} = 0$, entonces (2.4) se reduce a una función de densidad tipo *ji-cuadrado central*.
- Se nota $\chi^2(p, \boldsymbol{\lambda})$ para referirse a una distribución ji-cuadrado no central con p grados de libertad y parámetro de no centralidad $\boldsymbol{\lambda}$.

Algunas de las propiedades más importantes de la distribución ji-cuadrado no central se reseñan en seguida:

Propiedad 2.3.1 Si la variable aleatoria U tiene distribución $\chi^2(p, \boldsymbol{\lambda})$ entonces la media y la varianza de U son $(p + 2\boldsymbol{\lambda})$ y $2(p + 2\boldsymbol{\lambda})$, respectivamente. Nótese que para la distribución central ($\boldsymbol{\lambda} = 0$), la media es p y la varianza es $2p$.

Propiedad 2.3.2 Sean U_1 y U_2 dos variables aleatorias independientes con distribución ji-cuadrado no central de parámetros de no centralidad $\boldsymbol{\lambda}_1$ y $\boldsymbol{\lambda}_2$, con p_1 y p_2 grados de libertad respectivamente. La variable aleatoria suma $U = U_1 + U_2$, se distribuye como ji-cuadrado con parámetros de no centralidad $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)$, y $p = (p_1 + p_2)$ grados de libertad.

Propiedad 2.3.3 Sea X un vector aleatorio de tamaño $(p \times 1)$ distribuido $N_p(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, con $ran(\boldsymbol{\Sigma}) = p$, entonces la variable aleatoria $U = X'\boldsymbol{\Sigma}^{-1}X$ tiene distribución ji-cuadrado no central, con $\boldsymbol{\lambda} = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

2.3.2 Distribución t-Student no central

Se define la distribución *t-Student no central* a través de dos formas equivalentes:

∘ *Primera forma.* Sea Z una variable aleatoria distribuida $n(0, 1)$, U una variable aleatoria distribuida $\chi^2(p)$ y δ una constante, si Z y U son independientes, entonces la variable

$$T = \frac{(Z + \delta)}{(U/p)^{1/2}} \tag{2.5}$$

se distribuye como una *t-Student no central*, con p grados de libertad y parámetro de no centralidad δ.

∘ *Segunda forma.* Sea X una variable aleatoria distribuida normalmente con media μ y varianza σ^2, y sea Y/σ^2 una variable aleatoria distribuida ji-cuadrado con p grados de libertad e independiente de X, entonces:

$$t = \frac{\sqrt{p}X}{\sqrt{Y}} \tag{2.6}$$

tiene distribución t-Student no central con p grados de libertad y parámetro de no centralidad $\boldsymbol{\lambda} = \mu/\sigma$.

Los grados de libertad están asociados con el tamaño del vector que compone la forma cuadrática, éstos se notan por p para mantener una sola notación respecto al tamaño del vector aleatorio, pero se puede modificar sin pérdida de generalidad.

2.3.3 Distribución F no central

Sea U_1 una variable aleatoria distribuida $\chi^2(n_1, \boldsymbol{\lambda})$ y U_2 una variable aleatoria distribuida $\chi^2(n_2, 0)$, donde U_1 y U_2 son independientes. La variable aleatoria:

$$W = (U_1/n_1)/(U_2/n_2),$$

cuya función de densidad de probabilidad es:

$$F(w) = \begin{cases} \sum_{j=0}^{\infty} \left(\frac{e^{-\boldsymbol{\lambda}}\boldsymbol{\lambda}^j}{j!}\right) \frac{\Gamma\left(\frac{2j+n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{(n_1+2j)/2} w^{(n_1+2j-2)/2}}{\Gamma\left(\frac{n_2}{2}\right)\Gamma\left(\frac{2j+n_1}{2}\right)\left(1+\frac{n_1 w}{n_2}\right)^{(n_1+n_2+2j)/2}}, & w > 0 \\ 0, & w \leqslant 0, \end{cases} \tag{2.7}$$

tiene distribución *F no central* con n_1 y n_2 grados de libertad y parámetro de no centralidad $\boldsymbol{\lambda}$. De otra forma, una F no central es el cociente de una ji-cuadrado no central y una ji-cuadrado central.

Igual que en (2.4), se define $\boldsymbol{\lambda}^j = 1$, si $\boldsymbol{\lambda} = j = 0$. Para $\boldsymbol{\lambda} = 0$ se tiene la clásica distribución F central.

$F(n_1, n_2, \boldsymbol{\lambda})$ señala una distribución F con n_1 y n_2 grados de libertad en el numerador y denominador respectivamente y parámetro de no centralidad $\boldsymbol{\lambda}$.

Una de las aplicaciones más frecuentes de la distribución F no central es la determinación de la potencia en algunas pruebas de hipótesis, dentro del *análisis de varianza* y *diseño experimental* (capítulo 3); se requiere evaluar la integral:

$$\boldsymbol{\Pi}(\boldsymbol{\lambda}) = \int_{F(n_1,n_2,p)}^{\infty} F(n_1, n_2, \boldsymbol{\lambda})dw, \tag{2.8}$$

con n_1, n_2, y α valores fijos ($\alpha = P(ErrorTipoI)$). La cantidad $F(n_1, n_2, \alpha)$ es el percentil $(1-\alpha)\%$ de una distribución F central así:

$$\alpha = \int_{F(n_1,n_2,\alpha)}^{\infty} F(n_1, n_2, 0)dw, \tag{2.9}$$

para valores conocidos de n_1, n_2, α.

Los valores para $F(n_1, n_2, \alpha)$ se encuentran en tablas, lo mismo que el valor de $\left(1-\prod(\boldsymbol{\lambda})\right)$ para valores fijos de n_1, n_2, α y algunos valores de $\boldsymbol{\lambda}$. En lugar de $\boldsymbol{\lambda}$ se escribe ϕ, con:

$$\phi = \sqrt{\frac{2\boldsymbol{\lambda}}{n_1+1}}. \tag{2.10}$$

La última expresión es útil para determinar el tamaño de muestra o el número de *replicaciones* en diseños experimentales (Díaz y López, 1992).

2.3.4 Distribución de Wishart

La distribución de *Wishart* se asocia con la distribución muestral para estadísticas de la forma $\sum_i (X_i - \overline{\boldsymbol{X}})(X_i - \overline{\boldsymbol{X}})'$, con X_i vector aleatorio de tamaño $(p \times 1)$; ella equivale a la suma de cuadrados en el caso univariado $\sum_i (x_i - \bar{x})^2$.

Si $X_1, \ldots, X_n$, son vectores aleatorios independientes de tamaño $(p \times 1)$, con $n > p$ normalmente distribuidos; es decir, si $X_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces $\mathcal{W} = \mathbf{X}\mathbf{X}'$, con $\mathbf{X} = (X_1, \ldots, X_n)$, es una matriz de tamaño $(p \times p)$ de *Wishart*, con n grados de libertad, matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ y parámetro de no centralidad $\boldsymbol{\lambda}$, donde

$$\boldsymbol{\lambda} = 1/2\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.$$

Se nota $\mathcal{W} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n, \boldsymbol{\lambda})$

En el Capítulo 4 se presenta la forma funcional de esta distribución para caracterizar la distribución muestral de la matriz $\mathbf{S}$.

Un caso particular de la distribución de Wishart es la distribución ji-cuadrado central. Recuérdese que se define como: $\mathcal{W}_1 = \chi^2 = Z_1^2 + \cdots + Z_p^2 = Z'Z$, con $Z' = (Z_1, \ldots, Z_p)$ y las $Z_i \sim n(0, 1)$ e independientes, para $i = 1, \ldots, p$.

La distribución de Wishart central está ligada a la distribución de $\sum_{i=1}^{n} Z_i Z_i'$ donde los vectores aleatorios Z_i son independientes y distribuidos $N_p(0, \boldsymbol{\Sigma})$. Cuando $\boldsymbol{\Sigma} = \mathbf{I}_p$ la distribución está en forma estándar.

2.4 Distribución de formas cuadráticas

En la sección anterior se presentaron las distribuciones de algunas formas cuadráticas, en esta parte se tratan casos más generales, y se dan algunas condiciones para establecer la independencia tanto entre formas lineales y cuadráticas, como entre formas cuadráticas y ellas mismas. Las formas cuadráticas resultan en algunos métodos inferenciales tales como la estadística T^2 de Hotelling, el análisis de varianza, como también en el cálculo de distancias; casos en los cuales se debe determinar su distribución o garantizar el cumplimento de algunas propiedades.

Propiedad 2.4.1 *Distribución.* La siguiente proposición muestra de manera amplia la distribución de formas cuadráticas ligadas a distribuciones normales.

Sea X un vector de tamaño $(p \times 1)$ distribuido $N(0, I)$. La forma cuadrática $X'AX$ tiene distribución ji-cuadrado central, con k grados de libertad, si y sólo si, A es una matriz simétrica e idempotente, de rango k.

Propiedad 2.4.2 *Independencia entre forma cuadrática y lineal.* Sea X un vector aleatorio de tamaño $(p \times 1)$ con distribución $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ de rango p. La forma cuadrática $X'\boldsymbol{A}X$ es independiente de la forma líneal $\boldsymbol{B}X$, con $\boldsymbol{B}$ matriz de tamaño $(q \times p)$, si $\boldsymbol{B\Sigma A} = 0$.

Propiedad 2.4.3 *Independencia entre formas cuadráticas.* Sea X un vector aleatorio con distribución $\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con $\boldsymbol{\Sigma}$ matriz de rango p. Las formas cuadráticas $X'\boldsymbol{A}X$ y $X'\boldsymbol{B}X$ son independientes si $\boldsymbol{A\Sigma B} = 0$.

Propiedad 2.4.4 *Valor esperado de una forma cuadrática.* Sea X el vector aleatorio de tamaño $(p \times 1)$ con $\mathcal{E}(X) = \boldsymbol{\mu}$ y $\mathrm{Cov}(X) = \boldsymbol{\Sigma}$. Entonces

$$\mathcal{E}(X'AX) = tra(A\boldsymbol{\Sigma}) + \boldsymbol{\mu}'A\boldsymbol{\mu}.$$

Para cualquier matriz A de tamaño $p \times p$.

Esta propiedad se usa frecuentemente para simplificar expresiones que aparecen en el análisis multivariado o en los modelos lineales. Se deja para que el lector intente su demostración.

2.5 Ajuste a multinormalidad y transformaciones

Se ofrecen algunas herramientas útiles para diagnosticar el ajuste a la distribución normal multivariante, junto con algunas transformaciones que "normalizan" los datos. Los diagnósticos se hacen mediante gráficos y algunas pruebas estadísticas. Se resume en esta sección una parte de estos precedimientos, pues la literatura al respecto es bastante amplia. En primer lugar se hace una sinopsis de las técnicas univariadas, ya que como se ha mostrado si no se garantiza la normalidad univariada de un conjunto de datos tampoco se puede sostener la normalidad multivariada, aunque el recíproco no es siempre verdad; es decir, datos con distribución normal univariada no necesariamente tienen distribución normal multivariada (sección (2.2.1)).

2.5.1 Contrastes de multinormalidad

Para variables aleatorias *unidimensionales* no se deben descartar los gráficos del análisis exploratorio y descriptivo tales como: histogramas, diagramas de tallos y hojas, diagramas de cajas, entre otros, para advertir acerca del comportamiento normal de un conjunto de datos. Para el diagnóstico específico de normalidad se han desarrollado varias estrategias gráficas que, de manera visual, alertan sobre la normalidad o no de un conjunto de datos. La estrategia más usada consiste en graficar las cuantilas de los datos muestrales frente a las cuantilas de la distribución normal univariada; estos gráficos se conocen con el nombre de gráficos tipo $Q \times Q$. El papel probabilístico para la distribución normal está hecho de forma que si los datos se ajustan a esta distribución, los puntos se ubican en una línea recta; desviaciones de esta línea recta indican no normalidad (al menos en muestras de tamaño grande).

Las cuantilas son similares a los percentiles, los cuales se muestran en términos de porcentajes. Las cuantilas son expresadas en términos de fracciones o proporciones. Un gráfico $Q \times Q$ se obtiene como sigue:

1. Se ordenan las observaciones $x_1, \ldots, x_n$ en la forma $x_{(1)} \leq \cdots \leq x_{(n)}$. Así, el punto $x_{(i)}$ es la cuantila muestral $\frac{i}{n}$. Por ejemplo, si $n = 50$, el punto $x_{(8)}$ es la cuantila $\frac{8}{50} = 0.16$, porque el 0.16 (16%) de la muestra es menor o igual que $x_{(8)}$. La fracción $\frac{i}{n}$ se reemplaza frecuentemente por $(i - \frac{1}{2})/n$ para remover discontinuidad.

   Las cuantilas poblacionales se definen similarmente con relación a $(i - \frac{1}{2})/n$. Si se notan por $q_1, \ldots, q_n$, entonces q_i es el valor por debajo del cual una proporción $(i - \frac{1}{2})/n$ de observaciones poblacionales quedan ubicadas; es decir, $(i - \frac{1}{2})/n$ es la probabilidad de obtener una observación menor o igual a q_i. Formalmente, q_i se encuentra a partir de la distribución normal estándar al resolver

$$\Phi(q_i) = P(x < q_i) = \frac{i - \frac{1}{2}}{n} = p_i.$$

2. Se ubican entonces los pares $(q_i = \Phi^{-1}(p_i), x_{(i)})$ y se examina la linealidad del diagrama resultante $Q \times Q$.

El papel *probabilístico*, elimina la necesidad de encontrar los valores q_i. Únicamente se necesita ubicar los pares $[(i - \frac{1}{2})/n, x_{(i)})]$ y observar el ajuste de estos puntos a una línea recta.

Actualmente no se necesita papel probabilístico, pues los paquetes estadísticos tales como el SAS, el MINITAB o el SPSS, entre otros, suministran gráficos de los pares $(q_i = \Phi^{-1}(p_i), x_{(i)})$, con $p_i = \Phi(z_i) = P(X \leq x_{(i)})$ y los datos ordenados en la forma $x_{(1)} \leq \cdots \leq x_{(n)}$; esto equivale a una hoja probabilística.

Para el caso univariado, además de los recursos gráficos, desarrollados cada vez más en los programas de computación, se debe echar mano de las estadísticas, que de manera necesaria pero no suficiente, sugerirán el comportamiento frecuentista normal de unos datos. Tal es el caso de la media, la mediana, la desviación típica, los percentiles o cuantiles, el coeficiente de asimetría, el coeficiente de curtosis; con los cuales se puede hipotetizar el ajuste a una distribución normal de un conjunto de datos. Así por ejemplo: un distanciamiento apreciable entre la media y la mediana; un coeficiente de asimetría en valor absoluto grande o coeficiente de curtosis distante de 3.0, ponen en tela de juicio la normalidad de los datos.

Los contrastes usuales de normalidad univariada son los siguientes:

1. *Ji-cuadrado*, compara las frecuencias $O_1, \ldots, O_k$ observados en k-clases

$$[x_0, x_1), \ldots, [x_{k-1}, x_k),$$

con las frecuencias esperadas bajo el modelo probabilístico supuesto. Para la distribución normal se tiene que las frecuencias esperadas son: $E_i = np_i$, donde $p_i = P(x_{i-1} \leq X < x_i) = \Phi(z_i) - \Phi(z_{i-1})$. La discrepancia entre las frecuencias observadas y las esperadas por el modelo, se miden a través de la estadística

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

la cual se distribuye aproximadante como χ^2 si el modelo supuesto es el correcto. Los grados de libertad son $k-r-1$, con r el número de parámetros que se estiman (generalmente, para normalidad $r = 2$, pues se estiman la media y la varianza).

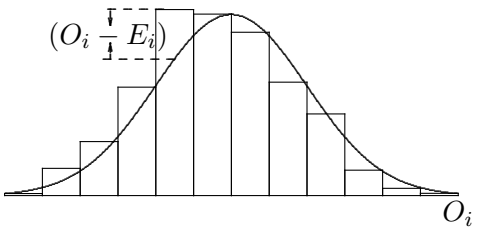


Figura 2.1 *Contraste Ji-cuadrado para normalidad.*

No sobra advertir, que la verificación del ajuste de un conjunto de datos a otro modelo probabilístico, se hace mediante el cálculo de los p_i con la respectiva dis-

tribución. La figura 2.1 muestra la distribución observada (histograma) y la distribución normal (curva suave); se observa que la estadística χ^2 mide la distancia entre estas dos distribuciones.

2. *Kolmogorov-Smirnov*, calcula la distancia entre la función de distribución empírica de la muestra $F_n(x)$ y la teórica; en este caso la normal; es decir, $F(x) = \Phi(x)$. La función de distribución empírica es

$$F_n(x) = \begin{cases} 0, & \text{si } x < x_{(1)} \\ \frac{r}{n}, & \text{si } x_{(r)} \leq x < x_{(r+1)} \\ 1, & \text{si } x \geq x_{(n)}. \end{cases}$$

donde $x_{\min} = x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} = x_{\max}$, son los valores muestrales ordenados. La figura 2.2 muestra esta estadística.

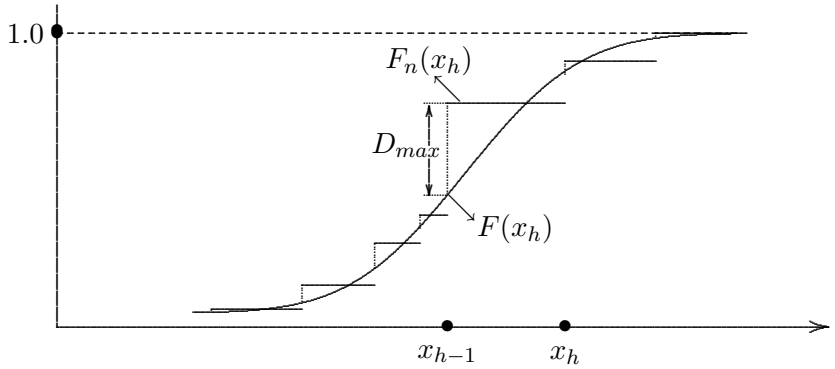


Figura 2.2 *Contraste de Kolmogorov-Smirnov.*

El estadístico de prueba es

$$D_n = \text{máx}\{|\, F_n(x) - F(x)\, |\},$$

cuya distribución exacta, bajo la hipótesis nula, se ha tabulado o se encuentra en los paquetes estadísticos. Si el máximo no existe, se usa el "supremun" o mínima cota superior.

3. Una prueba de ajuste de los datos ubicados en papel probabilístico a una recta, se conoce como el contraste de *Shapiro y Wilk*. Se rechaza la hipótesis de normalidad para valores pequeños del estadístico

$$\mathcal{W} = \frac{1}{ns^2}\left[\sum_{j=1}^{h} a_{j,n}(x_{(n-j+1)} - x_{(j)})\right]^2,$$

donde s^2 es la varianza muestral y h igual a $n/2$ para n par e igual a $(n-1)/2$ para n impar. Los coeficientes de $a_{j,n}$ han sido tabulados, y $x_{(j)}$ es el j-ésimo valor ordenado de la muestra. El estadístico $\mathcal{W}$ es similar a un coeficiente de determinación, mide el ajuste a una línea recta. Se rechaza la hipótesis de normalidad para valores pequeños de éste.

4. Contrastes basados en los coeficientes de *asimetría y curtosis*, los cuales se denotan por

$$\sqrt{b_1} = \sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 / \{\sum_{i=1}^{n}(x_i - \bar{x})^2\}^{\frac{3}{2}},$$

y

$$b_2 = n\sum_{i=1}^{n}(x_i - \bar{x})^4 / \left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^2.$$

Estos coeficientes son invariantes bajo transformaciones de localización y escala. Si la población es normal, los parámetros respectivos $\sqrt{\beta}_1$ y β_2 toman los valores de 0 y 3, respectivamente.

Para muestras de tamaño superior a 50 datos, la distribución de $\sqrt{b_1}$ es aproximadamente normal con media y varianza

$$\mathcal{E}(\sqrt{b_1}) = 0, \qquad \text{var}(\sqrt{b_1}) = \frac{6}{n},$$

respectivamente. De manera que con la estadística

$$Z = \frac{\sqrt{b_1}\sqrt{n}}{\sqrt{6}} \sim n(0,1),$$

se puede desarrollar el contraste para la hipótesis de simetría, $H_0 : \sqrt{\beta}_1 = 0$, de los datos.

Para muestras con un número de observaciones superior a 200, la distribución de b_2 es asintóticamente normal con

$$\text{media } \mathcal{E}(b_2) = 3 \text{ y varianza } \text{var}(b_2) = \frac{24}{n}.$$

D'Agostino y Pearson (1973) presentaron un estadístico que combina las dos medidas (asimetría y apuntamiento) para generar una prueba *omnibus* de normalidad. Por omnibus se entiende que la prueba es *capaz de detectar desviaciones de la normalidad*, sea por asimetría o por apuntamiento. El estadístico es

$$\boldsymbol{\chi}_2^2 = \frac{nb_1}{6} + \frac{n(b_2 - 3)^2}{24},$$

el cual se distribuye asintóticamente como una χ^2 con *dos* grados de libertad. Se rechaza la hipótesis de normalidad (sea por sesgo o por curtosis) para valores superiores a un valor crítico $\chi^2_{(2,\alpha)}$.

Para *contrastar multinormalidad* no es suficiente con probar la normalidad de las distribuciones marginales, puesto que se estaría dejando de lado la asociación lineal entre las variables; la cual se refleja a través de la matriz de covarianzas. Un ejemplo sobre la afirmación anterior se puede consultar en Hogg y Craig (1978, pág. 121). En resumen, la normalidad marginal no implica la distribución normal multivariada conjunta. La idea para contrastar multinormalidad es una extensión de alguna de las pruebas univariadas.

- Mardia (1970) define los coeficientes de simetría y curtosis multivariados, para un vector X de tamaño $(p \times 1)$ con media $\boldsymbol{\mu}$ y matriz de dispersión $\boldsymbol{\Sigma}$, mediante las siguientes expresiones

$$\begin{aligned}\boldsymbol{\beta}_{1,p} &= \mathcal{E}\left[\left\{(X - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(Y - \boldsymbol{\mu})\right\}^3\right] \\ \boldsymbol{\beta}_{2,p} &= \mathcal{E}\left[\left\{(X - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu})\right\}^2\right],\end{aligned} \tag{2.11}$$

donde X y Y son independientes e idénticamente distribuidos. Estas medidas son invariantes por transformaciones lineales. Si $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces, los coeficientes de simetría y curtosis son, respectivamente, $\boldsymbol{\beta}_{1,p} = 0$ y $\boldsymbol{\beta}_{2,p} = p(p+2)$.

La generalización de las medidas se observa porque $\sqrt{\boldsymbol{\beta}_{1,1}} = \sqrt{\beta_1}$ y $\boldsymbol{\beta}_{2,1} = \beta_2$. Se pueden contrastar las hipótesis sobre estos valores empleando los siguientes estimadores muestrales

$$b_{1,p} = \frac{1}{n^2}\sum_{h=1}^{n}\sum_{i=1}^{n} g_{hi}^3 \quad \text{y} \quad b_{2,p} = \frac{1}{n}\sum_{i=1}^{n} g_{ii}^2, \text{ con}$$
$$g_{hi} = (x_h - \bar{x})'\Big(\frac{A}{n}\Big)^{-1}(x_i - \bar{x}), \quad \text{y} \quad A = \sum_i (x_i - \bar{x})(x_i - \bar{x})'. \tag{2.12}$$

Mardia (1970) demuestra que bajo la hipótesis de distribución normal multivariante, se tiene la distribución asintótica de

$$\mathbf{B}_1 = \frac{n}{6} b_{1,p} \ \sim \ \chi_f^2 \quad \text{donde } f = \frac{1}{6}p(p+1)(p+2), \tag{2.13a}$$

Se rechaza la hipótesis de simetría entorno a la media ($H_0 : \sqrt{\boldsymbol{\beta}_{1,1}} = 0$) si $B_1 \geqslant \chi^2_{\alpha,f}$.

Para verificar que el coeficiente de curtosis no es significativamente diferente de $p(p+2)$ se emplea la estadística

$$\mathbf{B}_2 = \frac{b_{2,p} - p(p+2)}{[8p(p+2)/n]^{1/2}} \ \sim \ n(0,1). \tag{2.13b}$$

Estas estadísticas se pueden emplear a manera de prueba omnibus en muestras con tamaño de al menos 50 datos. Mardia (1974) presenta tablas para valores de $n \geqslant 10$ y $p = 2$. El procedimiento consiste en verificar si el conjunto de datos es simétrico respecto al vector de medias y si su coeficiente de curtosis es $p(p+2)$; de lo contrario se tendrá evidencia suficiente para rechazar la hipótesis de multinormalidad. Al final de este capítulo se ofrece un programa con el procedimiento IML del SAS para desarrollar esta prueba.

Rencher (1995, págs. 110-114) resume, entre otros, dos procedimientos de manejo sencillo para juzgar la multinormalidad de un conjunto de datos.

- *El primer procedimiento* se basa en la distancia de Mahalanobis de cada observación x_i al centroide de los datos $\bar{x}$,

  $$D_i^2 = (x_i - \bar{x})'\boldsymbol{S}^{-1}(x_i - \bar{x}).$$

  Si los x_i proceden de una distribución normal multivariada, se demuestra que

  $$u_i = \frac{nD_i^2}{(n-1)^2}$$

  tiene distribución beta. Para obtener un gráfico del tipo $Q \times Q$, los valores $u_1, \ldots, u_n$ son ordenados en la forma $u_{(1)} \leq, \cdots, \leq u_{(n)}$, y se grafican los pares $(v_i, u_{(i)})$, donde las cuantilas v_i de la distribución beta son dados por

  $$v_i = \frac{i - \alpha}{n - \alpha - \beta + 1}, \text{ con } \alpha = \frac{p-2}{2p} \text{ y } \beta = \frac{n-p-2}{2(n-p-1)}.$$

Si la nube de puntos se aleja de una línea recta, se advierte acerca de un posible distanciamiento de la normalidad en este conjunto de datos multivariados. Una prueba formal de significación es evaluada con la estadística $D^2_{(n)} = \max_i\{D^2_i\}$, para la cual se disponen tablas al 1% y 5%, $p = 2, 3, 4, 5$ y algunos valores de $n \geq 5$ (tabla C.3).

- *El segundo procedimiento*, se basa en las propiedades (2.2.3) y (2.2.6); consiste en graficar cada par de variables. La propiedad (2.2.3) garantiza distribución normal bivariada para cada par de variables, mientras que la propiedad (2.2.6) asegura que cada par de variables se ajusta a una línea recta, siempre que la distribución conjunta de donde procedan las variables sea normal multivariada. Si la forma de la nube de puntos, para alguno de los $\binom{p}{2}$ gráficos (como los que se muestran en la figura 1.4), no muestra ajuste a una línea recta, esto es una señal de no multinormalidad para el conjunto de datos particular. Este procedimiento puede extenderse para tres variables, con un programa de gráficas adecuado se pueden hacer diagramas en tres dimensiones; mediante rotaciones y proyecciones adecuadas, resultan de alta utilidad para el diagnóstico de normalidad multivariada.

  Las siguientes estrategias para detectar normalidad multivariada se basan en algunas de las definiciones de distribución normal multivariada. Es decir, el vector X, de tamaño $(p \times 1)$, tiene distribución normal p variante, si y sólo si, $\boldsymbol{a}'X$ es normal univariado, para todo vector $\boldsymbol{a}$ de tamaño $(p \times 1)$.

- Una generalización multivariada de la prueba de Shapiro y Wilks consiste en definir $z_i = \boldsymbol{c}'X_i$, para $i = 1, \ldots, n$, con $\boldsymbol{c}$ vector de constantes de tamaño $(p \times 1)$ y

$$\mathcal{W}(\boldsymbol{c}) = \frac{\sum_{i=1}^{n} a_i(z_{(i)} - \bar{z})^2}{\sum_{i=1}^{n} a_i(z_i - \bar{z})^2},$$

  donde $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n)}$, y los a_i son coeficientes tabulados (Shapiro y Wilks, 1965). La hipótesis de multinormalidad no se rechaza si

$$\max_{\boldsymbol{c}}[\mathcal{W}(c)] \geq \alpha,$$

  con α el nivel de significancia dispuesto.

  Nótese que si $\boldsymbol{c}$ es un vector canónico, por ejemplo, $\boldsymbol{e}_j = (0, \ldots, 1, \ldots, 0)'$, entonces $\boldsymbol{e}_i x_i$ es la i-ésima observación de la j-ésima variable. Entonces, el problema se reduce a verificar la normalidad de la variable X_j.

- *Contraste de normalidad direccional*: Sean $X_1, \ldots, X_n$ una muestra aleatoria de vectores de tamaño $p \times 1$ y sean $Z_i = (\boldsymbol{S}^{\frac{1}{2}})^{-1}(X_i - \overline{X})$, con $i = 1, \ldots, n$ los vectores "estandarizados", donde $\boldsymbol{S}^{\frac{1}{2}}$ es la raíz cuadrada de la matriz $\boldsymbol{S}$ (A2.34). Cada vector Z_i es multiplicado por un vector direccional $\boldsymbol{d}_k$ (aunque parezca redundante) para obtener $v_i = \boldsymbol{d}_k Z_i$, para $i = 1, \ldots, n$. Los v_i son aproximadamente normales univariados siempre que los vectores X tengan distribución multinormal. Se debe intentar con distintos vectores $\boldsymbol{d}_k$ para verificar normalidad en diferentes direcciones, con base en diferentes pruebas de normalidad univariada.

  Naturalmente, encontrar la dirección en la cual no hay normalidad necesita de un poco de paciencia, experiencia y buena suerte; pues que en algunas direcciones no se registre "anormalidad" no es suficiente garantía para asegurar la normalidad de manera isotrópica.

- Finalmente, Andrews y colaboradores (1973) sugieren usar la ransformación en versión multivariada de Box-Cox (2.15) para contrastar multinormalidad. Si los

datos proceden de una población normal multivariante, no es necesario transformar los datos y $\boldsymbol{\lambda} = \mathbf{1}_p$. La prueba se hace mediante el estadístico

$$2[L_{max}(\hat{\boldsymbol{\lambda}}) - L_{max}(\mathbf{1}_p)]$$

el cual se distribuye, aproximadamente, como χ^2_p, cuando $\boldsymbol{\lambda} = \mathbf{1}_p$.

2.5.2 Transformaciones para obtener normalidad

El modelo probabilístico normal es la base de muchos de los procedimientos de inferencia estadística y en algunos métodos multivariados. Cuando, a través de alguno de los procedimientos anteriores, se observa que los datos se apartan del modelo normal, una estrategia es la transformación de los datos, siempre que sea posible, para "acercarlos" a la normalidad.

Peña (1998, pág. 374) advierte que para distribuciones de los datos unimodales y simétricas, el camino es transformarlos en "normales"; en cambio, cuando la distribución sea bimodal o muestre la presencia de observaciones atípicas, las transformaciones a la normalidad pueden resultar infructuosas; casos en los que se debe optar por métodos robustos o no paramétricos.

En esta seccción se revisan primero algunas transformaciones en el caso univariado, para presentar luego algunas transformaciones para el campo multivariado.

▶ Transformaciones univariadas

La siguiente transformación es de uso frecuente

$$x^{(\lambda)} = \begin{cases} x^{\lambda}, & \lambda \neq 0 \\ \ln x, & \lambda = 0 \text{ y } x > 0. \end{cases} \tag{2.14}$$

A partir de (2.14) se pueden obtener las transformaciones: logaritmo, raíz cuadrada, inversa multiplicativa, entre otras, mediante valores adecuados de λ. Para valores $\mid \lambda \mid \leq 1$, se tiene la familia de transformaciones de Tukey (1957). Una modificación de la transformación anterior que remueve la discontinuidad en $\lambda = 0$, es la propuesta por Box y Cox (1964):

$$x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln x, & \lambda = 0 \text{ y } x > 0. \end{cases} \tag{2.15}$$

El problema que se debe enfrentar, para un conjunto de datos específico, es la determinación de un valor adecuado para λ. El procedimiento se puede resumir en los siguientes pasos:

1. Asumir que las "nuevas" observaciones $x_i^{(\lambda)}$ se distribuyen independientemente conforme a una $n(\mu, \sigma^2)$ y obtener los estimadores máximo verosímiles para μ y σ^2.
2. Reemplazar los valores obtenidos y buscar el valor de λ que maximice el logaritmo de la función de verosimilitud.
3. La maximización puede encontrarse resolviendo, de manera iterativa, la ecuación $dL_{max}(\lambda)/d\lambda = 0$, o buscando en una gráfica de $L(\lambda)$ frente a λ el valor $\hat{\lambda}$ que se aproxime al óptimo. En esta parte se puede hacer uso de los métodos numéricos

para optimizar funciones, por ejemplo el método del gradiente ("mayor o menor pendiente"). La figura 2.3 ilustra el proceso de búsqueda del $\hat{\lambda}$ óptimo.

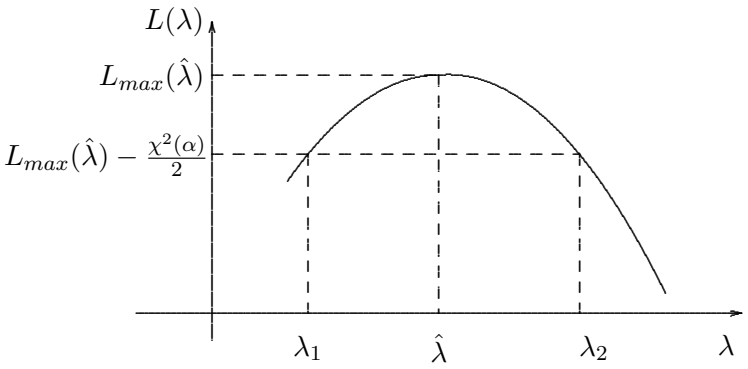


Figura 2.3 *Estimación gráfica de* λ.

El procedimiento anterior proporciona, además, intervalos de confianza para el valor de λ, y en consecuencia, una prueba de normalidad. La distribución del logaritmo de la razón de verosimilitud es asintóticamente χ^2, y por tanto, para el verdadero valor de λ, la distribución de

$$2(L_{max}(\lambda) - L(\lambda))$$

es $\chi^2_{(1)}$ con un sólo grado de libertad, pues se trata de un único parámetro.

A un nivel de confianza $(1 - \alpha)$, se puede construir un intervalo de confianza para el valor de la función de verosimilitud en el verdadero valor de λ. Sea $\chi^2_{\alpha,1}$ el valor de la distribución $\chi^2_{(1)}$ con un grado de libertad que deja una probabilidad a la izquierda de α, entonces

$$L_{max}(\lambda) - L(\lambda) \leq \frac{1}{2}\chi^2_1,$$

luego

$$L(\lambda) \geq L_{max}(\lambda) - \frac{1}{2}\chi^2_1,$$

la cual interseca la función $L(\lambda)$ a la altura $L_{max}(\lambda) - \frac{1}{2}\chi^2_1$, las proyecciones de estos valores sobre el eje horizontal determinan los extremos del intervalo de confianza $[\lambda_1, \lambda_2]$ para λ. Si el valor $\lambda = 1$ está incluido en este intervalo, no se rechaza la hipótesis de normalidad de los datos a un nivel de significación α, mientras que si está fuera del intervalo, se rechaza la hipótesis de normalidad. La figura 2.3 muestra la construcción del intervalo de confianza para el parámetro λ (Gnanadesikan 1998, pág. 167).

▶ Transformaciones multivariadas

Se pueden aplicar cada una de las transformaciones anteriores a los componentes del vector aleatorio. Andrews y colaboradores (1971) generalizaron la transformación de Box-Cox al caso vectorial. La transformación contempla un vector de parámetros $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$, de manera que el vector $X_i^{(\lambda)} = (X_{i1}^{(\lambda_1)}, \ldots, X_{ip}^{(\lambda_p)})'$ resulte distribuido

$N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, para $i = 1, \ldots, n$. La función de máxima verosimilitud para $\boldsymbol{\lambda}$ es

$$L_{Max}(\boldsymbol{\lambda}) = -\frac{1}{2} n \ln \mid \hat{\boldsymbol{\Sigma}} \mid + \sum_{j=1}^{p} (\lambda_j - 1) \sum_{i=1}^{n} \ln x_{ij} \tag{2.16}$$

donde x_{ij} es el elemento (i, j) de la matriz de datos y $\hat{\boldsymbol{\Sigma}}$ el estimador máximo verosímil de $\boldsymbol{\Sigma}$; es decir,

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (x_i^{(\boldsymbol{\lambda})} - \bar{x}^{(\boldsymbol{\lambda})})(x_i^{(\boldsymbol{\lambda})} - \bar{x}^{(\boldsymbol{\lambda})})'.$$

Se escoge el $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$ que maximice $L_{Max}(\boldsymbol{\lambda})$. Se puede contrastar la hipótesis $H_0 : \boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ y construir regiones de confianza para la expresión $2[L_{Max}(\hat{\boldsymbol{\lambda}}) - L_{Max}(\boldsymbol{\lambda}_0)]$, la cual se distribuye, bajo H_0, aproximadamente como χ_p^2. Esta inferencia (prueba de hipótesis o región de confianza) sobre $\boldsymbol{\lambda}$, orienta acerca del valor apropiado de $\boldsymbol{\lambda}$ para efectuar la transformación más adecuada.

Ejemplo 2.3 Se registraron medidas sobre hornos micro-ondas respecto a la radiación emitida fueron registradas en 42 de éstos, tanto con la puerta abierta como con la puerta cerrada. Los datos se consignan en la tabla 2.1.

Tabla 2.1 Radiación emitida por hornos micro-ondas

Puerta cerrada (X_1)				Puerta abierta (X_2)			
Horno	Radia.	Horno	Radia.	Horno	Radia.	Horno	Radia.
1	0.15	22	0.05	1	0.30	22	0.10
2	0.09	23	0.03	2	0.09	23	0.05
3	0.18	24	0.05	3	0.30	24	0.05
4	0.10	25	0.15	4	0.10	25	0.15
5	0.05	26	0.10	5	0.10	26	0.30
6	0.12	27	0.15	6	0.12	27	0.15
7	0.08	28	0.09	7	0.09	28	0.09
8	0.05	29	0.08	8	0.10	29	0.09
9	0.08	30	0.18	9	0.09	30	0.28
10	0.10	31	0.10	10	0.10	31	0.10
11	0.07	32	0.20	11	0.07	32	0.10
12	0.02	33	0.11	12	0.05	33	0.10
13	0.01	34	0.30	13	0.01	34	0.30
14	0.10	35	0.02	14	0.45	35	0.12
15	0.10	36	0.20	15	0.12	36	0.25
16	0.10	37	0.20	16	0.20	37	0.20
17	0.02	38	0.30	17	0.04	38	0.40
18	0.10	39	0.30	18	0.10	39	0.33
19	0.01	40	0.40	19	0.01	40	0.32
20	0.40	41	0.30	20	0.60	41	0.12
21	0.10	42	0.05	21	0.12	42	0.12

Fuente: Johnson y Wichern (1998, págs. 192 y 212)

Para cada una de las variables, de acuerdo con el procedimiento mostrado en la sección anterior, los valores de las potencias adecuadas para cada variable son $\hat{\lambda}_1 = 0.3$ y $\hat{\lambda}_2 = 0.3$,

respectivamente. Estas potencias se determinan para las distribuciones marginales de X_1 y X_2 en forma independiente, siguiendo el procedimiento que se ilustra en la figura 2.3.

Ahora, como se trata de determinar los valores (λ_1, λ_2) tales que la distribución conjunta de $(X_1^{(\lambda_1)}, X_2^{(\lambda_2)})$ sea normal bivariada, se debe maximizar $L(\lambda_1, \lambda_2)$ de acuerdo con la expresión (2.16) respecto a λ_1 y λ_2 conjuntamente.

Se hacen los cálculos de la función de verosimilitud $L(\lambda_1, \lambda_2)$ en una serie de valores de (λ_1, λ_2) que cubre la región $\{0 \leq \lambda_1 \leq 0.50$ y $0 \leq \lambda_2 \leq 0.50\}$ y se construyen las curvas de nivel como se presenta en la figura 2.4. Se observa que el máximo es aproximadamente 225.9 y ocurre en $(\hat{\lambda}_1, \hat{\lambda}_2) = (0.16, 0.16)$. Así, se deben transformar los datos elevando a estas potencias los valores de X_1 y X_2, respectivamente.

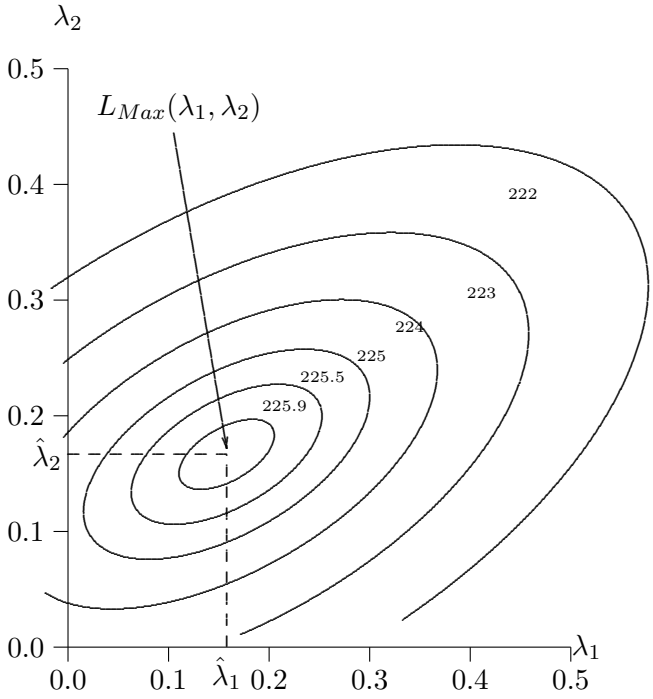


Figura 2.4 *Curvas de nivel para $L(\lambda_1, \lambda_2)$ con los datos de radiación.*

2.6 Visión geométrica de la distribución normal multivariante

El exponente $(x - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x - \boldsymbol{\mu})$ de la función densidad normal multivariada dada por (2.1), corresponde a la ecuación de un elipsoide en el espacio p dimensional cuando éste es igual a una constante C positiva. La figura 2.5 muestra la función de densidad para un vector $X = (X_1, X_2)'$ con distribución normal bivariada.

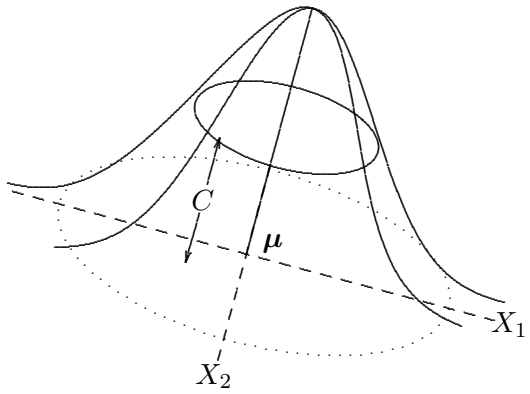


Figura 2.5 *Densidad constante en una normal bivariada.*

La familia de elipsoides concéntricos, generados al variar C, tiene su centro común $\boldsymbol{\mu}$. El eje principal de cada elipsoide está en la línea que pasa a través de los puntos más distantes de la elipse; es decir, es el segmento principal de la elipse (o diámetro) el cual pasa por $\boldsymbol{\mu}$ y tiene sus extremos en la superficie de un elipsoide, éste tiene las coordenadas que maximizan el cuadrado de la mitad de su longitud. Así,

$$\|(x-\boldsymbol{\mu})\|^2 = (x-\boldsymbol{\mu})'(x-\boldsymbol{\mu}) \tag{2.17}$$

es la distancia entre x y $\boldsymbol{\mu}$, que debe maximizarse bajo la restricción:

$$C = (x-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}) \tag{2.18}$$

la cual dice que el punto x pertenece al elipsoide. Para que la longitud sea el máximo valor, es necesario que su derivada con respecto a los elementos de x sea igual a cero. La restricción (2.18) se introduce mediante la adición del respectivo multiplicador de Lagrange, entonces la función a maximizar es:

$$g(x) = (x-\boldsymbol{\mu})'(x-\boldsymbol{\mu}) - \lambda\big((x-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}) - C\big); \tag{2.19}$$

su vector de primeras derivadas parciales es:

$$\begin{aligned}\frac{\partial g(x)}{\partial x} &= 2(x-\boldsymbol{\mu}) - 2\lambda\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}) = 0\\ &= (x-\boldsymbol{\mu}) - \lambda\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}) = 0,\end{aligned}$$

entonces:

$$(I - \lambda\boldsymbol{\Sigma}^{-1})(x-\boldsymbol{\mu}) = 0 \tag{2.20}$$

puesto que $\boldsymbol{\Sigma}$ es no singular, una expresión alterna a (2.20) es:

$$\big(\boldsymbol{\Sigma} - \lambda I\big)(x-\boldsymbol{\mu}) = 0. \tag{2.21}$$

De esta forma, las coordenadas asociadas al primer eje principal son proporcionales a los elementos de un vector característico de $\boldsymbol{\Sigma}$; así,

$$(x-\boldsymbol{\mu}) = \lambda\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})$$

Premultiplicando la ecuación (2.20) por $4(x-\boldsymbol{\mu})'$ (pues (2.17) es el cuadrado de la longitud del semieje mayor) y de (2.18) resulta:

$$\begin{aligned} 4(x-\boldsymbol{\mu})'(x-\boldsymbol{\mu}) &= 4\lambda(x-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}) \\ 4(x-\boldsymbol{\mu})'(x-\boldsymbol{\mu}) &= 4\lambda C. \end{aligned} \tag{2.22}$$

De tal manera que la longitud dada por (2.18), implicada en (2.22), es máxima en el valor de λ más grande; es decir, para la más grande raíz característica de $\boldsymbol{\Sigma}$. En resumen: los ejes principales conservan el orden decreciente de sus longitudes de acuerdo con el mismo orden de los correspondientes valores propios; es decir, si los valores propios se ordenan como:

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0,$$

la magnitud de los ejes está unívocamente determinada por los vectores propios, y además, puesto que $\lambda_i \neq \lambda_j$ para $i \neq j$, entonces los vectores propios l_i y l_j asociados a λ_i y a λ_j, son ortogonales; esto es $\langle l_i, l_j \rangle = 0$, en consecuencia, los ejes donde están contenidos son mutuamente perpendiculares.

Las coordenadas respecto a los "nuevos" ejes (principales) conforman el vector $Y = (Y_1, \ldots, Y_p)$, y se relacionan con las variables originales por medio de:

$$Y = L'(X - \boldsymbol{\mu}), \tag{2.23}$$

donde L está constituida por los respectivos vectores normalizados l_i. La ortogonalidad de la matriz L ($L'L = \mathbf{I}$), implica que la transformación consiste en una rotación "rígida" de los ejes originales sobre los ejes principales del elipsoide seguida por la traslación del origen a $\boldsymbol{\mu}$, el centro del elipsoide.

La matriz de covarianzas de Y es:

$$L'\boldsymbol{\Sigma}L \tag{2.24}$$

y la varianza de la variable del i-ésimo eje principal es:

$$\text{var}(Y_i) = l_i'\boldsymbol{\Sigma}l_i = \lambda_i,$$

y la covarianza entre Y_i y Y_j es

$$\text{cov}(Y_i, Y_j) = 0 \text{ para } i \neq j. \tag{2.25}$$

Un resultado importante, es que la transformación dada por (2.23) genera variables no correlacionadas cuyas varianzas son proporcionales a la longitud de los ejes de algún elipsoide de concentración.

La transformación de Box-Cox es generalizada por Yeo y Johnson (2000), quienes proponen una familia de transformaciones por potencias definida sobre toda la recta numérica la cual resulta apropiada para reducir el sesgo y aproximar los datos a la normalidad.

Para visualizar y reforzar lo tratado en esta sección, se presenta a continuación la función de densidad normal bivariada.

2.7 Distribución normal bivariada

Sea $X' = (X_1, X_2)$ un vector aleatorio de tamaño (1×2), la función de densidad conjunta de X_1 y X_2 es:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ \frac{-1}{2(1-\rho^2)} \left(\frac{(x_1-\boldsymbol{\mu}_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\boldsymbol{\mu}_1)(x_2-\boldsymbol{\mu}_2)}{\sigma_1\sigma_2} + \frac{(x_2-\boldsymbol{\mu}_2)^2}{\sigma_2^2} \right) \right\} \tag{2.26}$$

donde ρ es el coeficiente de correlación entre X_1 y X_2; σ_1 y σ_2 son las desviaciones estándar de X_1 y X_2, respectivamente. De (2.18) se obtiene para este caso:

$$\frac{(x_1-\boldsymbol{\mu}_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\boldsymbol{\mu}_1)(x_2-\boldsymbol{\mu}_2)}{\sigma_1\sigma_2} + \frac{(x_2-\boldsymbol{\mu}_2)^2}{\sigma_2^2} = (1-\rho^2)C, \tag{2.27}$$

las matrices de covarianzas y de correlación son las siguientes, respectivamente:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \text{ y } \boldsymbol{\rho} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Los valores propios de $\boldsymbol{\rho}$ son: $\lambda_1 = 1+\rho$ y $\lambda_2 = 1-\rho$. Los vectores propios normalizados

$$l_1' = (1/2\sqrt{2}, 1/2\sqrt{2}) \text{ y } l_2' = (-1/2\sqrt{2}, 1/2\sqrt{2}).$$

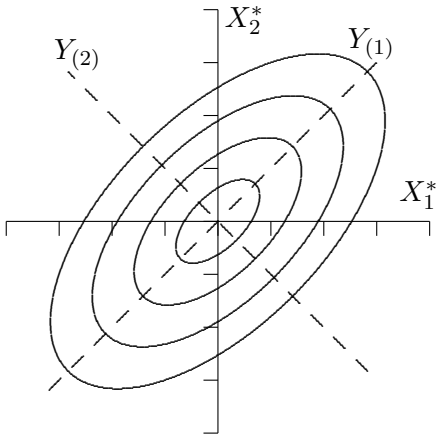


Figura 2.6 *Ejes principales.*

El eje principal tiene pendiente positiva o negativa de acuerdo con el signo positivo o negativo de la correlación ρ. Si el coeficiente de correlación ρ es cero, la elipse es un círculo[3].

[3]Distribución normal esférica, $N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$.

La figura 2.6 muestra varias elipses para $\rho = 0.6$ respecto a las variables estandarizadas X_1^* y X_2^* de acuerdo con diferentes valores de C.

La matriz

$$P = (l_1, l_2) = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix},$$

permite rotar los ejes X_1^* y X_2^* un ángulo $\theta = 45^0$ $(\pi/4)$, para producir el sistema de coordenadas $Y_{(1)}$ y $Y_{(2)}$.

2.8 Detección de datos atípicos ("outliers")

La traducción más cercana del término *"outlier"* es observación *atípica, discordante, anómala o contaminante*. En el texto se mantienen estos términos con el mismo significado. Intuitivamente un valor atípico es una observación extrema que se aparta bastante de los demás datos. Ésta es una caracterización apropiada para el caso univariado, pues allí existe un orden natural de los datos, con el cual se puede establecer cuando una distancia es extrema con respecto a un punto como la media o la mediana, entre otras. Se debe distinguir entre observación *atípica* e *influyente*. La segunda es una observación que tiene un alto impacto sobre los valores de predicción a través de los parámetros estimados, o en general sobre los componentes de un modelo estadístico; tal es el caso de un modelo de regresión o un modelo de series temporales. Un outlier, en cambio, es una observación que discrepa de lo esperado y que tal vez se genera desde una población no considerada.

La presencia de datos multivariados atípicos en un conjunto de datos son más problemáticos que en el caso univariado. Uno de tales problemas es que estos datos pueden distorsionar no sólo las medidas de localización y escala sino las de asociación u orientación. Un segundo problema es que es más difícil caracterizar y descubrir que un dato univariado atípico. Un tercer problema es que un dato multivariado por el hecho mismo de ser un vector conformado por varias datos univariados, la atipicidad puede deberse a un error extremo en alguna de sus componentes o a la ocurrencia de errores sistemáticos en varias (sino en todas) sus componentes.

En el tratamiento de estos datos hay dos aspectos. El primero consiste en su *detección* o *identificación*. Para esto se dispone de una serie de herramientas gráficas y de cálculo, con las cuales se puede evidenciar la presencia de estas observaciones en un conjunto de datos. El segundo aspecto corresponde al *tratamiento* dado a las observaciones declaradas como outliers. Esto implica la posible modificación de los datos o de los métodos de análisis o de modelamiento. Se procede a una modificación de los datos, sea por su exclusión o modificación, cuando se descubre que los datos atípicos se deben a errores de medición, de registro o de concepto. Los métodos *robustos*, en los cuales se reduce la influencia de datos atípicos, es el caso más común de modificación del análisis (Wilcox, 1997). En la segunda parte del texto se comentan varias alternativas de análisis robusto en algunas técnicas multivariadas.

La detección de datos atípicos en el caso multivariado no es igualmente sencilla. Algunas diferencias con respecto al caso univariado son las siguientes:

- Para más de dos variables ($p > 2$), las gráficas se hacen más complejas o imposibles.
- Las observaciones multivariadas no se pueden ordenar como en el caso univariado.

- Un vector de observaciones puede ser un outlier debido a que alguna de sus componentes lo es.
- Un dato atípico multivariado puede reflejar *"desfase"* o corrimiento (slippage) en la media, la varianza o la correlación. Esto se entiende como un pequeño corrimiento en la media o la varianza, lo cual puede provocar un desajuste lineal.

Una forma útil de detectar datos atípicos es a través de la distancia entre cada observación y el centro de los datos, ésta se calcula con la distancia de Mahalanobis. Cada observación x_i puede ordenarse de acuerdo con la distancia

$$D_i^2 = (x_i - \bar{x})' \boldsymbol{S}^{-1} (x_i - \bar{x}). \tag{2.28}$$

Valores grandes de D_i^2 advierten sobre la posibilidad de que la observación sea un dato anómalo.

Un procedimiento equivalente es el cómputo de la razón de varianzas generalizadas

$$r_{(i)}^2 = \frac{\mid \boldsymbol{S}_{(i)} \mid}{\mid \boldsymbol{S} \mid}, \tag{2.29}$$

donde $\boldsymbol{S}_{(i)}$ significa la matriz de covarianzas de los datos sin la observación x_i. Un valor relativamente pequeño de $r_{(i)}^2$ indica que la observación x_i es un potencial outlier.

Otro procedimiento útil en la identificación de un dato atípico, el cual sirve también para juzgar multinormalidad de los datos, se basa en la estadística de Wilks

$$\omega = \max_i \frac{\mid (n-2)\boldsymbol{S}_{(i)} \mid}{\mid (n-1)\boldsymbol{S} \mid} \tag{2.27}$$

$$= 1 - \frac{n D_{(n)}^2}{(n-1)^2}, \tag{2.30}$$

donde $D_{(n)}^2 = \max_i (x_i - \bar{x})' \boldsymbol{S}^{-1} (x_i - \bar{x})$. Así, la prueba para detectar un outlier (uno sólo) se basa en la estadística D_i^2, que se muestra en la sección (2.5.1) y con la cual se diagnostica gráficamente multinormalidad. La tabla C.3 contiene los valores críticos al 5% y 1% para la estadística $D_{(n)}^2$ junto con algunos valores de n y p.

Yan y Lee (1987) suministran una estadística F asociada a ω, definida por

$$F_i = \frac{n-p-1}{p} \left[\frac{1}{1 - n D_i^2/(n-1)^2} - 1 \right], \text{ para } i = 1, \ldots, n. \tag{2.31}$$

Dado que los F_i son independientes e idénticamente distribuidos como $F_{p,n-p-1}$, la prueba puede construirse en términos del máximo de los F_i, así:

$$\begin{aligned} P\left(\max_i F_i > f \right) &= 1 - P(\text{Todo} F_i \leq f) \\ &= 1 - [P(F \leq f)]^n. \end{aligned}$$

Por tanto, la prueba puede desarrollarse empleando la tabla de la estadística F (tabla C.8). De la ecuación (2.30) se obtiene

$$\max_i F_i = F_{(n)} = \frac{n-p-1}{p} \left(\frac{1}{\omega} - 1 \right).$$

Gnanadesikan y Kattenring (1972) proponen las siguientes estadísticas dentro de una clase general de éstas. Se supone que $x_1, \ldots, x_n$ es una muestra aleatoria multivariada

$$
\begin{aligned}
q_i^2 &= (x_i - \bar{x})'(x_i - \bar{x}), \quad i = 1, \ldots, n \\
t_i^2 &= (x_i - \bar{x})'\boldsymbol{S}(x_i - \bar{x}), \quad i = 1, \ldots, n \\
u_i^2 &= \frac{(x_i - \bar{x})'\boldsymbol{S}(x_i - \bar{x})}{(x_i - \bar{x})'(x_i - \bar{x})}, \quad i = 1, \ldots, n \\
v_i^2 &= \frac{(x_i - \bar{x})'\boldsymbol{S}^{-1}(x_i - \bar{x})}{(x_i - \bar{x})'(x_i - \bar{x})}, \quad i = 1, \ldots, n \\
d_{i0}^2 &= (x_i - \bar{x})'\boldsymbol{S}^{-1}(x_i - \bar{x}), \quad i = 1, \ldots, n \\
d_{ij}^2 &= (x_i - x_j)'\boldsymbol{S}^{-1}(x_i - x_j), \quad i < j = 1, \ldots, n.
\end{aligned}
$$

Cada una de estas estadísticas identifica la contribución de cada observación sobre algunos aspectos característicos de los datos tales como localización, escala u orientación (correlación), entre otras. Así:

- q_i^2 permite identificar las observaciones que estan "infladas" excesivamente sobre la escala global.
- t_i^2 muestra cuales observaciones tienen la mayor influencia sobre la orientación y escala; la cual resulta de mucha utilidad para identificar datos atípicos en la matriz de covarianzas y por ende en componentes principales (capítulo 5).
- u_i^2 pone más énfasis en la orientación que en la escala. Nótese que ésta es igual a: t_i^2/q_i^2.
- v_i^2 mide la contribución relativa de las observaciones sobre la orientación de las últimas componentes principales.
- $d_{i0}^2 = D_i^2$ muestra las observaciones que, con esta distancia, "caen" lejos del grupo de datos.
- d_{ij}^2 además del objetivo anterior provee con algún detalle la separación entre observaciones.

Los mismos autores sugieren graficar las estadísticas q_i^2, t_i^2, d_{i0}^2 y d_{ij}^2 en escalas probabilísticas tipo beta y para u_i^2 y v_i^2 en escala probabilística tipo F. Para una revisión más amplia sobre estas estadísticas y otros métodos para detectar outliers multivariados el libro de Gnanadesikan (1997, págs. 305-317) hace una buena presentación de este tema.

Ejemplo 2.4 La tabla 2.2 muestra los datos sobre longitud de huesos registrados en 20 jóvenes a los 8, 8.5, 9 y 9.5 años, respectivamente (Rencher 1995, pág. 90).

Además, la tabla 2.2 contiene los valores de las estadísticas D_i^2 y F_i para detectar posibles outliers y los pares $(v_i, u_{(i)})$ definidos en la sección (2.5.1) con los cuales se puede verificar el ajuste a multinormalidad.

De acuerdo con los valores D_9^2, D_{12}^2 y D_{20}^2, se observa que las respectivas observaciones (9, 12 y 20) son datos potencialmente atípicos. Esto es confirmado con la estadística F_i, ya que para un $\alpha = 0.05$, F_9, F_{12} y $F_{20} > F_{(0.05,4,15)} = 3.06$ (tabla C.8).

La figura 2.7, que representa los pares $(v_i, u_{(i)})$, muestra la influencia de estos tres datos en el desajuste a la multinormalidad, pues se apartan de la línea ajustada para los demás datos.

Tabla 2.2 Longitud de huesos en 20 jóvenes

Ind.	8 años (X_1)	8.5 años (X_2)	9 años (X_3)	9.5 años (X_4)	$D_i^2(i)^*$	$(v_i, u_{(i)})$	F_i
1	47.8	48.8	49.0	49.7	0.7588 (3)	(0.136,0.042)	0.165
2	46.4	47.3	47.7	48.4	1.2980 (7)	(0.333,0.072)	0.291
3	46.3	46.8	47.8	48.5	1.7591 (8)	(0.382,0.097)	0.405
4	45.1	45.3	46.1	47.2	3.8539 (13)	(0.629,0.214)	1.018
5	47.6	48.5	48.9	49.3	0.8706 (5)	(0.234,0.048)	0.190
6	52.5	53.2	53.3	53.7	2.8106 (11)	(0.530,0.156)	0.692
7	51.2	53.0	54.3	54.5	4.2915 (14)	(0.678,0.238)	1.170
8	49.8	50.0	50.3	52.7	7.9897 (17)	(0.826,0.443)	2.978
9	48.1	50.8	52.3	54.4	11.0301 (20)	(0.974,0.611)	5.892
10	45.0	47.0	47.3	48.3	5.3519 (16)	(0.776,0.297)	1.581
11	51.2	51.4	51.6	51.9	2.8301 (12)	(0.579,0.157)	0.697
12	48.5	49.2	53.0	55.5	10.5718 (19)	(0.924,0.586)	5.301
13	52.1	52.8	53.7	55.0	2.5941 (10)	(0.481,0.144)	0.629
14	48.2	48.9	49.3	49.8	0.6594 (2)	(0.086,0.037)	0.142
15	49.6	50.4	51.2	51.8	0.3246 (1)	(0.037,0.018)	0.069
16	50.7	51.7	52.7	53.3	0.8321 (4)	(0.185,0.046)	0.181
17	47.2	47.7	48.4	49.5	1.1083 (6)	(0.283,0.061)	0.245
18	53.3	54.6	55.1	55.3	4.3633 (15)	(0.727,0.242)	1.195
19	46.2	47.5	48.1	48.4	2.1088 (9)	(0.432,0.117)	0.496
20	46.3	47.6	51.3	51.8	10.0931 (18)	(0.875,0.559)	4.757

*puesto u orden respecto a los otros D_i^2.

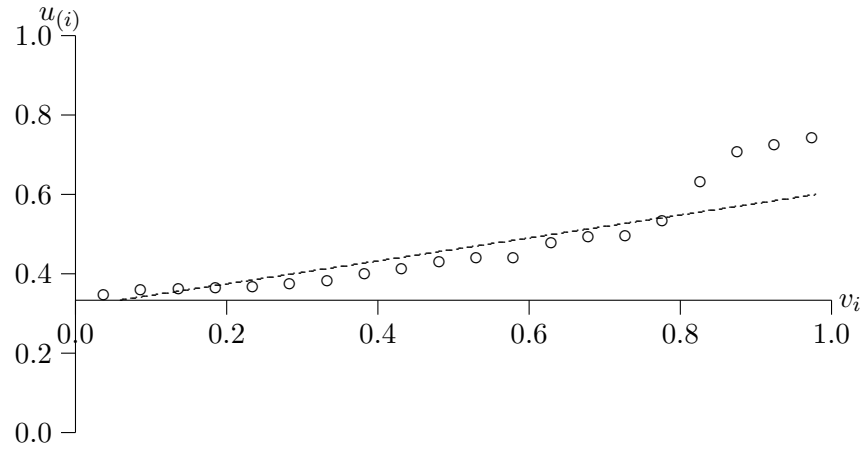


Figura 2.7 *Gráfico $Q \times Q$ de v_i y $u_{(i)}$.*

Además, para $\alpha = 0.05$, $n = 20$ y $p = 4$, la tabla C.3 suministra el valor crítico 11.63, el cual no es excedido por $D^2_{(20)} = 11.0301$. Esto no debe sorprender, pues la prueba está hecha para detectar outliers individualmente, mientras que en este caso hay tres.

2.9 Rutina SAS para Generar muestras multinormales

El siguiente programa ilustra la generación de muestras a partir de una población normal multivariante. El programa se hace mediante el procedimiento IML (Interactive Matrix Language). La sintaxis se escribe en mayúsculas, esto no es necesario, simplemente se hace para resaltar los comandos SAS. Al frente (o debajo) de cada instrucción se explica su propósito dentro de los símbolos /∗ y ∗/.

```
PROC IML;
SEED=552154123; /* semilla  */
N=20; /* muestra de tamaño N=20  */
SIGMA=\{4 2 1, 2 3 1, 1 1 5\}; /*  matriz de covarianzas */
MU={1,3, 0}; /* vector de medias */
P=NROW(SIGMA); /* número de variables */
M=REPEAT(MU',N,1); /*vector MU'n veces por fila y 1 por columna */
G=ROOT(SIGMA); /* descomposición de Cholesky */
Z=NORMAL(REPEAT(SEED,N,P)); /* genera n vectores Np(0,Ip) */
Y=Z*G+M; /* genera n vectores Np(MU,SIGMA) */
PRINT Y; /* imprime la matriz Y, de tamaño (20 x 3) */
```

2.10 Rutina SAS para la prueba de multinormalidad de Mardia

Para ilustrar la prueba de multinormalidad se consideran los datos de la tabla 3.4.

```
PROC IML; /* invocación del procedimiento IML */
Y={72 66 76 77, 60 53 66 63, 56 57 64 58, 41 29 36 38, 32 32 35  36,
   30 35 34 26, 39 39 31 27, 42 43 31 25, 37 40 31 25, 33 29 27  36,
   32 30 34 28, 63 45 74 63, 54 46 60 52, 47 51 52 43, 91 79 100 75,
   56 68 47 50, 79 65 70 61, 81 80 68 58, 78 55 67 60, 46 38 37  38,
   39 35 34 37, 32 30 30 32, 60 50 67 54, 35 37 48 39,
   39 36 39 31, 50 34 37 40, 43 37 39 50, 48 54 57 43};
                        /* matriz de datos Y de tamaño (28 x 4) } */
N=NROW(Y); /* No. de filas de Y  */
P=NCOL(Y); /* No. de columnas de Y */
GL\_CHI=(P)*(P+1)*(P+2)/6; /* grados de libertad */
Q=I(N)-(1/N)*j(N,N,1); /*  calcula Ip-1/n·1n·1n' */
S=(1/(N))*Y'*Q*Y; /* matriz de covarianzas muestral */
S\_INV=INV(S); /*  inversa de la martiz S */
G\_MATRIZ=Q*Y*S\_INV*Y'*Q; /* cálculo de la matriz (g_hi) de la */
                          /* ecuación (2.12) */
b\_1=( SUM(G\_MATRIZ\#G\_MATRIZ\#G\_MATRIZ) )/(N*N); /* cálculo de
                                        /* la simetría b_(1,p) */
b\_2=TRACE(G\_MATRIZ\#G\_MATRIZ)/N; /* cálculo de curtosis b_(2,p) */
EST\_b\_1=N*b\_1/6; /* cálculo de la estadística B1, ec. (2.13a) */
EST\_b\_2=(b\_2-P*(P+2))/SQRT(8*P*(P+2)/N); /* estadística B2, ec.(2.13b)*/
```

```
PVAL\_ses=1-PROBCHI(EST\_b\_1,GL\_CHI); /* valor p=Pr(B1>= EST_b_1) */
PVAL\_cur=2*(1-PROBNORM(ABS(EST\_b\_2))); /* p=Pr(|B2|>= EST_b_2) */
PRINT  b\_1 b\_2 EST\_b\_1 EST\_b\_2 PVAL\_ses PVAL\_cur; /* imprime
                                          /*  estas estadísticas */
RUN; /* ejecuta el programa */
```

Los coeficientes de simetría y curtosis, las estadísticas para probar las hipótesis respecto a la simetría y curtosis junto con los valores p, se muestran a continuación

The SAS System

b_1	b_2	EST_b_1	EST_b_2	$PVAL\_ses$	$PVAL\_cur$
4.4763816	22.95687	20.889781	−0.398352	0.4036454	0.6903709

De acuerdo con los p valores no se rechaza la hipótesis respecto a la procedencia de una distribución normal 4 variante de los datos.

2.11 Procesamiento de datos con R

Para la generación de muestras a partir de una distribución normal multivariante se traduce el código de `SAS` que usted tiene en el libro a código de `R` pero se hace de una manera mucho más interesante; como una función. Es usuario invoca a la función entregándole n , μ , Σ y opcionalmente una semilla y la función regresa la matriz de datos generados con esa media y covarianza.

```
gmultinorm<-function(n,sigma=diag(1,n),mu=matrix(0,nrow=n),
 semilla=NULL){
 if(!is.null(semilla)){set.seed(semilla)}
 # vector de medias
 p<-nrow(sigma)        # numero de variables
 unos<-matrix(1,nrow=n)  #
 # repite el vector mu n veces por fila y 1 vez por columnas
 M<-t(mu)%x%unos
 G<-chol(sigma)  # descomposición de cholesky
 Z<-matrix(rnorm(p*n),ncol=p)  # genera n vectores $N_p(0,Ip)
 # genera n vectores $N_p(mu ,sigma)
 Y<-Z%*%G+M
 Y
 }
 # el llamado a la función
 n<-20  # muestra de tamaño n=20
 # vector de medias
 media<-matrix(c(1,3,0))
 # mat de covarianzas
 S<-matrix(c(4,2,1,2,3,1,1,1,5),nrow=3,byrow=TRUE)
 gmultinorm(20,mu=media,sigma=S,semilla=552154123)
 # datos de una normal con media cero y covarianza identidad
 gmultinorm(20,mu=matrix(0,nrow=3),sigma=diag(1,3))
```

Alternativamente, para la generación de datos multinomiales se puede usar la función `mvrnorm()` de la librería `MASS`, de la siguiente forma `library(MASS)`

```
mvrnorm(n, media, S)
 # Usando una semilla
 set.seed(552154123)
mvrnorm(20, media, S)
```

Tenga en cuenta que esta función hace la descomposición de la matriz S, vía `eigen` mientras que la función creada arriba lo hace mediante Cholesky.

Para la prueba de multinormalidad de Mardia se traduce el código `SAS` a código de `R` pero como una función, de tal forma que el usuario entrega una matriz de datos y la función regresa las estadísticas

```
 mardia.test<-function(Y){
 n<-nrow(Y) # numero de filas de Y
 p<-ncol(Y) # numero de columnas de Y
 gl_chi<-p*(p+1)*(p+2)/6 # grados de libertad
 Q<-diag(n)-(1/n)*matrix(1,n,n)  # I_p-(1/n)1_n1'_n
 S<-(1/n)*t(Y)%*%Q%*%Y  # matriz de covarianzas muestral
 # Matriz g_hi de la ecuación 2.12
 G_MATRIZ<- Q%*%Y%*%solve(S)%*%t(Y)%*%Q
 b_1<-sum(G_MATRIZ^3)/(n^2) # cálculo de la simetía
 b_2<-sum(diag(G_MATRIZ^2))/n  # calculo de la curtosis b_(2,p)
 EST_b_1<-n*b_1/6   # calculo de la estadística B1 ec. (2.13a)
 # calculo de la  estadística B1 ec. (2.13a)
 EST_b_2<-(b_2-p*(p+2))/sqrt(8*p*(p+2)/n)
 PVAL_ses<-1-pchisq(EST_b_1,gl_chi)
 PVAL_cur<-2*(1-pnorm(abs(EST_b_2)))
 cat("b_1=",b_1,"b_2=",b_2,"EST_b_1=",EST_b_1,
 "EST_b_2=",EST_b_2,"\n")
 cat("PVAL_ses=",PVAL_ses,"PVAL_cur=",PVAL_cur,"\n") }
```

El llamado a la función considerando los datos de la tabla 3.4

```
Y1<-scan()
72 66 76 77 60 53 66 63 56 57 64 58 41 29 36 38 32 32 35  36
30 35 34 26 39 39 31 27 42 43 31 25 37 40 31 25 33 29 27  36
32 30 34 28 63 45 74 63 54 46 60 52 47 51 52 43 91 79 100 75
56 68 47 50 79 65 70 61 81 80 68 58 78 55 67 60 46 38 37  38
39 35 34 37 32 30 30 32 60 50 67 54 35 37 48 39 39 36 39  31
50 34 37 40 43 37 39 50 48 54 57 43
Y<-matrix(Y1,ncol=4,byrow=TRUE)
 mardia.test(Y)
```

Test de normalidad multivariada usando la prueba de Shapiro-Wilk

```
#prueba de Shapiro-Wilk multivariada, requiere la
 #librería mvnormtest
(mvnormtest)
mshapiro.test(t(Y))
```

Capítulo 3

Inferencia sobre el vector de medias

3.1 Introducción

En el capítulo anterior se tratan las características más relevantes de la distribución normal multivariada, se indica que ésta queda completamente definida por el vector de medias y la matriz de varianzas y covarianzas. En este capítulo se presentan algunos tópicos acerca de: la estimación de sus parámetros, distribución muestral, propiedades de los estimadores, y verificación de hipótesis sobre el vector de medias; para los casos donde la matriz de varianzas y covarianzas sea conocida o desconocida, respectivamente. En la segunda parte se trata el análisis de varianza multivariado junto con algunas aplicaciones al campo del diseño experimental tales como análisis de perfiles, medidas repetidas y curvas de crecimiento.

3.2 Estimación

A partir de una muestra aleatoria de una población normal p variante se obtienen los estimadores de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, por el método de *máxima verosimilitud* (MV)[1]. Es decir, se buscan los valores de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ que maximizan la probabilidad de que la muestra aleatoria $X_1, X_2, \ldots, X_n$ proceda de esta población.

Supóngase una muestra aleatoria de n observaciones obtenida de una población que se distribuye $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; esto es $X_1, \ldots, X_n$, con $n > p$ (cada X_i es un vector aleatorio de tamaño $(p \times 1)$). La *función de verosimilitud* es:

$$\boldsymbol{L} = \prod_{\alpha=1}^{n} N(X_\alpha|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{pn/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left(-\frac{1}{2}\sum_{\alpha=1}^{n}(X_\alpha - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(X_\alpha - \boldsymbol{\mu})\right) \quad (3.1)$$

[1] Verosimilitud aquí es sinónimo de probabilidad.

En la ecuación (3.1), los vectores $X_1, \ldots, X_n$ son valores muestrales fijos y $\boldsymbol{L}$ es una función de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$. Como la función logarítmica es continua y creciente, los valores de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ que maximizan a $\boldsymbol{L}$ son los mismos que maximizan a $\ln \boldsymbol{L}$

$$\boldsymbol{l}(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \ln \boldsymbol{L} = -\frac{1}{2} pn \ln(2\pi) - \frac{1}{2} n \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{\alpha=1}^{n} (X_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (X_\alpha - \boldsymbol{\mu}). \tag{3.2}$$

Los estimadores de máxima verosimilitud se obtienen al resolver el sistema de ecuaciones siguiente

$$\begin{aligned} \frac{\partial \boldsymbol{l}}{\partial \boldsymbol{\mu}} &= 0 \\ \frac{\partial \boldsymbol{l}}{\partial \boldsymbol{\Sigma}^{-1}} &= 0. \end{aligned} \tag{3.3}$$

Para resolver este sistema se presentan algunas identidades con las cuales se simplifica la solución.

La forma cuadrática contenida en el exponente de (3.1) es equivalente a

$$\begin{aligned} (X_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (X_\alpha - \boldsymbol{\mu}) &= [(X_\alpha - \bar{\boldsymbol{X}}) + (\bar{\boldsymbol{X}} - \boldsymbol{\mu})]' \boldsymbol{\Sigma}^{-1} [(X_\alpha - \bar{\boldsymbol{X}}) + (\bar{\boldsymbol{X}} - \boldsymbol{\mu})] \\ &= (X_\alpha - \bar{\boldsymbol{X}})' \boldsymbol{\Sigma}^{-1} (X_\alpha - \bar{\boldsymbol{X}}) + (\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \\ &\quad + 2(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (X_\alpha - \bar{\boldsymbol{X}}); \end{aligned}$$

al sumar sobre el subíndice α, el último término de la identidad anterior se anula, de donde resulta la expresión

$$\sum_{\alpha=1}^{n} (X_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (X_\alpha - \boldsymbol{\mu}) = \sum_{\alpha=1}^{n} (X_\alpha - \bar{\boldsymbol{X}})' \boldsymbol{\Sigma}^{-1} (X_\alpha - \bar{\boldsymbol{X}}) + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}). \tag{3.4}$$

La forma cuadrática $(X_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (X_\alpha - \boldsymbol{\mu})$ es un escalar, luego es igual a su traza y como $\operatorname{tra}(\boldsymbol{AB}) = \operatorname{tra}(\boldsymbol{BA})$, entonces

$$(X_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (X_\alpha - \boldsymbol{\mu}) = \operatorname{tra}\Big\{ \boldsymbol{\Sigma}^{-1} (X_\alpha - \boldsymbol{\mu})(X_\alpha - \boldsymbol{\mu})' \Big\},$$

en consecuencia la igualdad (3.4) es equivalente a

$$\sum_{\alpha=1}^{n} (X_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (X_\alpha - \boldsymbol{\mu}) = (n) \operatorname{tra}\Big\{ \boldsymbol{\Sigma}^{-1} \boldsymbol{A} \Big\} + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}), \tag{3.5}$$

con $\sum_{\alpha=1}^{n}(X_\alpha - \bar{\boldsymbol{X}})(X_\alpha - \bar{\boldsymbol{X}})' = \boldsymbol{A} = n\boldsymbol{S}$. Al reemplazar en la ecuación (3.2) se obtiene

$$\boldsymbol{l}(\boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{np}{2}\ln(2\pi) + \frac{n}{2}\ln|\boldsymbol{\Sigma^{-1}}| - \frac{n}{2}\operatorname{tra}\Big\{\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\Big\} - \frac{n}{2}\operatorname{tra}\Big\{\boldsymbol{\Sigma^{-1}}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\Big\}. \tag{3.6}$$

La diferenciación que se muestra en las ecuaciones (3.3), se obtiene mediante la aplicación de las propiedades dadas en las igualdades (A2.44 a A2.53a) junto con la identidad (3.5). Así:

$$\begin{aligned}
\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{\mu}} &= n\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \\
\frac{\partial \ln|\boldsymbol{\Sigma}^{-1}|}{\partial \boldsymbol{\Sigma}^{-1}} &= 2\boldsymbol{\Sigma} - \operatorname{Diag}\Big\{\boldsymbol{\Sigma}\Big\} \\
\frac{\partial \operatorname{tra}\Big\{\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\Big\}}{\partial \boldsymbol{\Sigma}^{-1}} &= 2\boldsymbol{A} - \operatorname{Diag}\Big\{\boldsymbol{A}\Big\} \\
\frac{\partial \operatorname{tra}\Big\{\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\Big\}}{\partial \boldsymbol{\Sigma}^{-1}} &= 2(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' - \operatorname{Diag}\Big\{(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\Big\}.
\end{aligned} \tag{3.7}$$

Al sustituir la última igualdad de (3.7) en el sistema de ecuaciones (3.3), se observa que

$$\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{\mu}} = \boldsymbol{\Sigma}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) = 0. \tag{3.8}$$

Luego el estimador de máxima verosimilitud de $\boldsymbol{\mu}$ es $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$.

Similarmente,

$$\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{n}{2}2\Big(\boldsymbol{\Sigma} - \boldsymbol{S} - (\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' - \operatorname{Diag}\Big\{\boldsymbol{\Sigma} - \boldsymbol{S} - (\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\Big\}\Big) = 0; \tag{3.9}$$

la última expresión implica que

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} + (\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})', \tag{3.10}$$

como $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$, el estimador máximo verosímil de $\boldsymbol{\Sigma}$ es $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S}$.

Realmente, hasta ahora tan sólo se ha encontrado que $\hat{\boldsymbol{\mu}}$ y $\widehat{\boldsymbol{\Sigma}}$ corresponden a un punto crítico de la función de verosimilitud. Resta por demostrar que $\bar{\boldsymbol{X}}$ y $\boldsymbol{S}$ maximizan la función de verosimilitud sobre todos los valores. La demostración se encuentra en Anderson (1984, págs. 62-64).

En resumen los estimadores de máxima verosimilitud para $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son:

$$\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{pmatrix}, \quad y \quad \widehat{\boldsymbol{\Sigma}} = \left(\frac{1}{n}\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)\right), \; i, j = 1, \ldots, p \tag{3.11}$$

Además, los estimadores para las varianzas de cada una de las variables y de los coeficientes de correlación (de Pearson) entre cada par de variables, son respectivamente:

$$\widehat{\sigma}_i^2 = \frac{1}{n}\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)^2, \quad i = 1, \ldots, p$$

y

$$\widehat{\rho}_{ij} = \frac{\sum\limits_{k=1}^{n}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\left(\sum\limits_{k=1}^{n}(X_{ik} - \bar{X}_i)^2\right)^{1/2}\left(\sum\limits_{k=1}^{n}(X_{jk} - \bar{X}_j)^2\right)^{1/2}} = \frac{\widehat{\sigma}_{ij}}{\widehat{\sigma}_i\widehat{\sigma}_j}, \; i, j = 1, \ldots, p. \tag{3.12}$$

Si se asume que las 10 observaciones sobre los manzanos del ejemplo 1.3 (sección (1.4.2)) son una muestra aleatoria de una población $N_4(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, entonces, el vector de medias $\bar{X}$ y la matriz de covarianzas $\boldsymbol{S}$ serían los respectivos estimadores máximo verosímiles para $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$.

3.3 Propiedades de los estimadores MV de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$

De la teoría estadística para el caso univariado, la media $\bar{X}$ de una muestra aleatoria tiene distribución normal y es independiente de la varianza muestral, siempre que la muestra sea obtenida de una población normal. De manera análoga, el vector de medias obtenido en (3.11) tiene distribución normal multivariada y es independiente de $\widehat{\boldsymbol{\Sigma}}$.

Las siguientes propiedades, conllevan a la distribución de $\widehat{\boldsymbol{\mu}}$ y $\widehat{\boldsymbol{\Sigma}}$.

Suponga que $X_1, \ldots, X_n$ son vectores independientes, donde cada uno de los X_i se distribuye $N_p(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, $i = 1, \ldots, n$; es decir, se dispone de una *muestra aleatoria* de una población normal p variante.

Como

$$\mathcal{E}(\bar{\boldsymbol{X}}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{E}(X_i) = \boldsymbol{\mu}, \tag{3.13}$$

se concluye que $\bar{\boldsymbol{X}}$ es un estimador insesgado del vector de medias poblacional $\boldsymbol{\mu}$.

De otra parte, como $\text{Cov}(X_i) = \boldsymbol{\Sigma}$ y $\text{Cov}(X_i, X_{i'}) = 0$ para $i \neq i'$, entonces

$$\begin{aligned}\text{Cov}(\bar{\boldsymbol{X}}) &= \text{Cov}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} \\ &= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\text{Cov}(X_i) + \sum_{i\neq i'}\text{Cov}(X_i, X_{i'})\right\} \\ &= \frac{1}{n}\boldsymbol{\Sigma}. \end{aligned} \tag{3.14}$$

El resultado anterior es equivalente a $\text{Cov}(\bar{\boldsymbol{X}}) = \mathcal{E}\left\{(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\right\} = \frac{1}{n}\boldsymbol{\Sigma}$.

Ahora,

$$\begin{aligned}\mathcal{E}\left\{\widehat{\boldsymbol{\Sigma}}\right\} &= \mathcal{E}\Big\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})(X_i - \bar{\boldsymbol{X}})'\Big\} \\ &= \frac{1}{n}\mathcal{E}\left\{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})(X_i - \boldsymbol{\mu})'\right\} - \mathcal{E}\left\{(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\right\} \\ &= \boldsymbol{\Sigma} - \frac{1}{n}\boldsymbol{\Sigma} \\ &= \frac{n-1}{n}\boldsymbol{\Sigma}, \end{aligned} \tag{3.15}$$

con esto se demuestra que $\widehat{\boldsymbol{\Sigma}}$ es un estimador *sesgado* de $\boldsymbol{\Sigma}$. Si se define a

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})(X_i - \bar{\boldsymbol{X}})' = \boldsymbol{S}, \tag{3.16}$$

se obtiene que $\boldsymbol{S}$ es un estimador *insesgado* de $\boldsymbol{\Sigma}$.

Observación:

> Hasta ahora no se ha hecho distinción respecto a la notación del estimador de $\boldsymbol{\Sigma}$, sea éste sesgado o insesgado. En adelante se asume, a menos que se diga lo contrario, que es un estimador insesgado y se notará por $\boldsymbol{S}$.

La media muestral $\bar{\boldsymbol{X}}$ se puede escribir como una función lineal sobre el vector $\boldsymbol{X}$, de la forma $\bar{\boldsymbol{X}}' = \frac{1}{n}\mathbf{1}'\boldsymbol{X}$, donde $\boldsymbol{X} = (X_1, \ldots, X_n)'$ es la matriz de datos de una población $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y $\mathbf{1}$ es el vector de unos de tamaño $(n \times 1)$. De acuerdo con las propiedades enunciadas en la sección (2.2) y

con las ecuaciones (3.14-15), $\bar{\boldsymbol{X}}$ tiene distribución normal p variante con media $\boldsymbol{\mu}$ y matriz de covarianzas $\frac{1}{n}\boldsymbol{\Sigma}$.

La matriz de covarianzas $\widehat{\boldsymbol{\Sigma}}$ se puede escribir de la siguiente manera:

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\mathbb{X}'(\boldsymbol{I}-\frac{1}{n}\mathbf{1}'\mathbf{1})\mathbb{X} \\ &= \frac{1}{n}\begin{pmatrix} X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1j} & X_{2j} & \cdots & X_{nj} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n} \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}.\end{aligned}$$

El vector de medias $\bar{\boldsymbol{X}}' = \frac{1}{n}\mathbf{1}'\boldsymbol{X}$ es una forma lineal con $\boldsymbol{B} = \frac{1}{n}\mathbf{1}'$, la matriz de covarianzas muestral es una forma cuadrática $\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}'(\boldsymbol{I}-\frac{1}{n}\mathbf{1}\mathbf{1}')\boldsymbol{X}$, con $\boldsymbol{A} = (\boldsymbol{I}-\frac{1}{n}\mathbf{1}\mathbf{1}')$. Por la propiedad (2.4.2), se obtiene que $\boldsymbol{BSA} = 0$; de donde se puede afirmar que $\bar{\boldsymbol{X}}$ y $\boldsymbol{S}$ son *estadísticas independientes*.

Las proposiciones anteriores se resumen a continuación.

1. La media $\bar{\boldsymbol{X}}$ de una muestra aleatoria de tamaño n tomada de una población $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, se distribuye $N_p(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma})$ y es independiente de $\widehat{\boldsymbol{\Sigma}}$, el estimador máximo verosímil de $\boldsymbol{\Sigma}$.

2. La distribución de la matriz de covarianzas muestral está ligada a una *Wishart*. En el capítulo 4 se muestra que $n\widehat{\boldsymbol{\Sigma}}$ se distribuye como $\mathcal{W}_p(\boldsymbol{\Sigma}, n-1)$.

3. Una propiedad útil para desarrollar pruebas de hipótesis sobre el vector de medias $\boldsymbol{\mu}$, derivada a partir de la propiedad 2.2.5, es que

$$n(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}), \tag{3.17}$$

   tiene distribución j i–cuadrado central con p grados de libertad.

   Aunque una muestra aleatoria no proceda de una población normal, el vector de medias $\bar{\boldsymbol{X}}$, para tamaños de muestra grandes, tiene distribución normal. Esto se contempla en la siguiente proposición.

4. *Teorema del Límite Central*. Sea $X_1, X_2, \ldots$ una sucesión infinita de vectores aleatorios idénticamente distribuidos de una población con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$. Entonces

$$n^{-1/2}\sum_{\alpha=1}^{n}(X_\alpha-\boldsymbol{\mu}) = n^{1/2}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}) \xrightarrow{D} N_p(0,\boldsymbol{\Sigma}), \quad \text{en tanto } n\to\infty.$$

El símbolo $\xrightarrow{D}$ significa *convergencia en distribución*. Un resultado equivalente es la distribución asintótica de $\bar{\boldsymbol{X}}$, la cual es normal tal como se escribe a continuación

$$\bar{\boldsymbol{X}} \overset{D}{\sim} N_p(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma}).$$

5. En el caso multidimensional la definición de *consistencia* es semejante a la presentada en las ecuaciones (B.28 y B.29). De cualquier modo las definiciones son equivalentes.

   Giri (1977) demuestra que $\bar{\boldsymbol{X}}$ converge estocásticamente a $\boldsymbol{\mu}$ y que:

   $$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})(X_i - \bar{\boldsymbol{X}})'$$

   converge estocásticamente a $\boldsymbol{\Sigma}$. El procedimiento se basa en demostrar la consistencia de cada uno de los elementos del vector $\bar{\boldsymbol{X}}$ y de la matriz $\widehat{\boldsymbol{\Sigma}}$.

6. La función de densidad de probabilidad ligada a una muestra aleatoria de n vectores $X_1, \ldots, X_n$, de tamaño $(p \times 1)$, es el producto de las *fdp* de cada X_i, $i = 1, \ldots, n$; es decir,

   $$\begin{aligned} f(x_1, \ldots, x_n) &= (2\pi)^{-np/2}|\boldsymbol{\Sigma}|^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(X_i - \boldsymbol{\mu})\right) \\ &= (2\pi)^{-np/2}|\boldsymbol{\Sigma}|^{-n/2}\exp\Big\{(-1/2)\Big(\mathrm{tra}(A\boldsymbol{\Sigma}^{-1}) + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})\Big)\Big\} \\ &= g(\bar{\boldsymbol{X}}, \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot h(X_1, \ldots, X_n), \end{aligned}$$

   donde

   $$A = (n)\widehat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})(X_i - \bar{\boldsymbol{X}})'.$$

   Por el Teorema de factorización (B.30) se concluye que $\widehat{\boldsymbol{\mu}}$ y $\widehat{\boldsymbol{\Sigma}}$ son estadísticas suficientes, con $h(X_1, \ldots, X_n) = 1$.

En resumen [2], dada una muestra aleatoria $X_1, \ldots, X_n$, de vectores de tamaño $(p \times 1)$ sobre una población $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con las estadísticas $\bar{\boldsymbol{X}}$ y S estimadores de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ respectivamente, entonces:

[2] Que no agota la existencia de otras propiedades.

1. $\bar{\boldsymbol{X}}$ y $n\boldsymbol{S}$ tienen distribución $N(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$ y $\mathcal{W}(\boldsymbol{\Sigma}, n-1)$, respectivamente.
2. $\bar{\boldsymbol{X}}$ y $\boldsymbol{S}$ son independientes.
3. $\bar{\boldsymbol{X}}$ y $\boldsymbol{S}$ son estimadores insesgados de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, respectivamente.
4. $\bar{\boldsymbol{X}}$ y $\boldsymbol{S}$ son consistentes.
5. $\bar{\boldsymbol{X}}$ y $\boldsymbol{S}$ son estadísticas suficientes.

Ejemplo 3.1 Los datos de la tabla 3.1, tomados de Anderson (1984), corresponden al incremento en horas sueño debido al uso de dos medicamentos $\boldsymbol{A}$ y $\boldsymbol{B}$. El experimento se realizó sobre diez pacientes.

Tabla 3.1 Incremento en horas de sueño

Paciente	Medicina $\boldsymbol{A}$ X_1	Medicina $\boldsymbol{B}$ X_2
1	1.9	0.7
2	0.8	-1.6
3	1.1	-0.2
4	0.1	-1.2
5	-0.1	-0.1
6	4.4	3.4
7	5.5	3.7
8	1.6	0.8
9	4.6	0.0
10	3.4	2.0

Si se asume que el par (X_{i1}, X_{i2}) es una observación de una población $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con $i = 1, \ldots, 10$ se obtienen los estimadores de máxima verosimilitud para $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ respectivamente.

$$\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} = (1/n)\mathbf{1}'\boldsymbol{X}$$

donde $\mathbf{1}'$ denota el vector de unos de tamaño (1×10) y $\boldsymbol{X}$ es la matriz de datos de tamaño (10×2) contenida en la tabla 3.1

$$\bar{\boldsymbol{X}}' = (1,\ldots,1)/10 \begin{pmatrix} 1.9 & 0.7 \\ 0.8 & -1.6 \\ 1.1 & -0.2 \\ 0.1 & -1.2 \\ -0.1 & -0.1 \\ 4.4 & 3.4 \\ 5.5 & 3.7 \\ 1.6 & 0.8 \\ 4.6 & 0.0 \\ 3.4 & 2.0 \end{pmatrix} = (2.33, 0.75)$$

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}} &= (1/n)(\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{X}})'(\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{X}}) \\ &= (1/n)\boldsymbol{X}_c'\boldsymbol{X}_c \\ &= \begin{pmatrix} 3.61 & 2.56 \\ 2.56 & 2.88 \end{pmatrix}. \end{aligned}$$

Un estimador insesgado para la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ es:

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}_c'\boldsymbol{X}_c = \begin{pmatrix} 4.01 & 2.85 \\ 2.85 & 3.20 \end{pmatrix}$$

Un estimador de la matriz de correlación es:

$$R = \begin{pmatrix} 1 & 0.7956 \\ 0.7956 & 1 \end{pmatrix}.$$

Es decir, los medicamentos tienen efectos semejantes sobre los pacientes, en términos de aumentar o disminuir las horas sueño de éstos. ☑

3.4 Contraste de hipótesis y regiones de confianza sobre $\boldsymbol{\mu}$

En esta sección se desarrolla la inferencia estadística (regiones de confianza y verificación de hipótesis) sobre el vector de medias en una y dos poblaciones; para los casos donde la matriz de covarianzas se conoce o se desconoce, respectivamente. Se consideran algunas aplicaciones de la estadística T^2 en problemas de medidas repetidas, análisis de perfiles y en control estadístico de calidad.

En el contexto multivariado los contrastes de hipótesis son más complejos que los univariados. La distribución normal p–variante tiene p–medias, p–varianzas y $\binom{p}{2}$ covarianzas, así, el número total de parámetros es $\frac{1}{2}p(p+3)$. Es decir, se puede formular este número de hipótesis, por ejemplo, si $p = 5$ se deben desarrollar pruebas sobre 20 parámetros univariados; 5 para las medias, 5 para la varianzas y 10 para las covarianzas.

Además del inconveniente anterior, hay, entre otros, cuatro argumentos a favor de las pruebas multivariadas frente a las univariadas, éstos son:

1. El desarrollo de p–pruebas univariadas incrementa la tasa de error Tipo I, mientras que con las pruebas multivariadas ésta se mantiene. Por ejemplo, si se hacen separadamente $p = 10$ pruebas univariadas a un nivel de significación $\alpha = 0.05$, la probabilidad de tener al menos un rechazo es mayor que 0.05. Si las variables son independientes (situación poco común), bajo H_0, se tiene que

$$\begin{aligned} P(\text{al menos un rechazo}) &= 1 - P(\text{no rechazar las 10 pruebas}) \\ &= 1 - (0.95)^{10} = 0.40. \end{aligned}$$

   En general, supóngase que se está interesado en desarrollar k-pruebas simultáneas y sea E_i $(i = 1, \cdots, k)$ el evento "la i-ésima hipótesis no es rechazada, dado que es verdadera". Se debe encontrar un nivel crítico apropiado para cada prueba, de manera que la probabilidad de que ellas sea aceptadas simultáneamente sea igual a $1 - \alpha$, bajo el supuesto de que todas son verdaderas, es decir:

$$P\left(\bigcap_{i=1}^{k} E_i\right) = 1 - \alpha.$$

   Así, para el caso de pruebas sobre $\boldsymbol{\mu}' = (\mu_1, \cdots, \mu_p)$ en lugar de una hipótesis global $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, el interés puede dirigirse a verificar de manera individual hipótesis de la forma $H_{0i} : \mu_i = \mu_{0i}$ $(i = 1, \cdots, p = k)$. En este caso $H_0 = \bigcap_{i=1}^{p} H_{0i}$. La diferencia entre pruebas simultáneas y la prueba global es que con la primera se puede evidenciar cuál de las hipótesis no se sostiene.

   Para obviar un poco el problema de la "inflación" del error Tipo I, ilustrado con el ejemplo anterior de las $k = \binom{5}{2} = 10$-pruebas simultáneas, se emplea un nivel de significancia α^* tal que

$$P\left(\bigcap_{i=1}^{k} E_i\right) \geq 1 - \alpha^* = 1 - \alpha,$$

de aquí se encuentra que $\alpha^* = \alpha/k$, la cual es una cota inferior que garantiza el alcance de al menos un nivel de probabilidad igual a $1-\alpha$ con las pruebas individuales desarrolladas simultáneamente. El sustento de esta igualdad se encuentra en la desigualdad de Bonferroni, la cual se expresa a continuación:

$$P\left(\bigcap_{i=1}^{k} E_i\right) \geq 1 - \sum_{i=1}^{k} P(E_i^c),$$

donde E_i^c es el complemento de E_i.

2. Las pruebas univariadas no consideran la posible correlación existente entre las variables, en contraposición, las pruebas multivariadas emplean esta información contenida en la matriz de covarianzas.

3. En la mayoría de los casos las pruebas multivariadas han mostrado ser más potentes que las univaridas. Esto se debe a que los pequeños efectos de algunas variables se combinan conjuntamente. Para un tamaño de muestra dado, hay un límite en el número de variables para que una prueba multivariada mantenga la potencia en cierto nivel (sección (3.4.3)).

4. Muchas pruebas multivariadas involucran medias de combinaciones lineales de variables, las cuales pueden resultar más reveladoras de la forma como las variables "se unen" para rechazar la hipótesis.

3.4.1 Matriz de varianzas y covarianzas conocida

▶ Una población

En el caso multivariado, tanto la verificación de hipótesis como la construcción de regiones de confianza, se basan en que la diferencia entre el vector de medias muestral y el poblacional está normalmente distribuido con vector de medias cero y matriz de varianzas y covarianzas conocida.

La expresión (3.17) indica que si $\sqrt{n}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})$ se distribuye $N(0,\boldsymbol{\Sigma})$ entonces:

$$n(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}), \tag{3.18}$$

tiene distribución ji–cuadrado central, con p–grados de libertad.

La última expresión es la distancia de Mahalanobis o medida de discrepancia entre el vector de medias muestral y el vector de medias poblacional; con (3.18) se construyen las regiones de confianza y se busca detectar la

existencia de posibles diferencias entre el vector de medias muestral y el vector de medias supuesto. Para verificar la hipótesis

$$H_0 : \ \boldsymbol{\mu} = \boldsymbol{\mu}_0, \tag{3.19}$$

donde $\boldsymbol{\mu}_0$ es un vector específico, se usa como región crítica el conjunto de puntos tales que:

$$\chi_0^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \geq \chi^2_{(\alpha,p)} \tag{3.20}$$

donde $\chi^2_{(\alpha,p)}$, es el número tal que

$$P(\chi^2_{(p)} > \chi^2_{(\alpha,p)}) = \alpha$$

Así, una muestra que cumpla la desigualdad (3.20), provoca el rechazo de la hipótesis $H_0 : \ \boldsymbol{\mu} = \boldsymbol{\mu}_0$.

La *función de potencia* de la prueba dada por (3.20) se deriva del hecho de que

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$$

se distribuye ji–cuadrado no central con parámetro de no centralidad

$$\boldsymbol{\lambda} = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0), \tag{3.21}$$

y p grados de libertad. La función de potencia para la prueba dada en (3.20) tiene el valor mínimo α (nivel de significación) cuando $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, y su potencia es más grande que α cuando $\boldsymbol{\mu}$ es diferente de $\boldsymbol{\mu}_0$.

Para una media muestral $\bar{\boldsymbol{X}}$, la desigualdad

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}^*)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}^*) \leq \chi^2_{(\alpha,p)}, \tag{3.22}$$

se satisface con una probabilidad $(1 - \alpha)$, para una muestra con tamaño n, extraida de una población $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. El conjunto de valores de $\boldsymbol{\mu}^*$ que satisfacen (3.22) es una *región de confianza* para $\boldsymbol{\mu}$, con un coeficiente de confiabilidad $(1 - \alpha)$. Esta expresión representa el interior y la superficie de un elipsoide con centro en $\boldsymbol{\mu} = \bar{\boldsymbol{X}}$, cuya forma y tamaño dependen de $\boldsymbol{\Sigma}$ y $\chi^2_{(\alpha,p)}$; así por ejemplo, si $\boldsymbol{\Sigma} = I_p$; la región de confianza es una esfera.

Ejemplo 3.2 En la tabla 3.2 se registra la estatura X_1 (en pulgadas) y el peso X_2 (en libras) para una muestra de 20 estudiantes de educación media. Se asume que esta muestra es generada en una población normal bivariada $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde

$$\boldsymbol{\Sigma} = \begin{pmatrix} 20 & 100 \\ 100 & 1000 \end{pmatrix}$$

Supóngase que se quiere verificar la hipótesis que la estatura media es 70 y el peso medio es 170; es decir, $H_0: \boldsymbol{\mu} = (70, 170)'$, en este tipo de personas. De la matriz de datos contenida en la tabla 3.2, se tiene que $\bar{x_1} = 71.45$ $\bar{x_2} = 164.7$.

Tabla 3.2 Estatura y peso en una muestra de 20 estudiantes

Estudiante	Estatura (X_1)	Peso (X_2)	Estudiante	Estatura (X_1)	Peso (X_2)
1	69	153	11	72	140
2	74	175	12	79	265
3	68	155	13	74	185
4	70	135	14	67	112
5	72	172	15	66	140
6	67	150	16	71	150
7	66	115	17	74	165
8	70	137	18	75	185
9	76	200	19	75	210
10	68	130	20	76	220

Fuente: Rencher (1995, pág. 51)

De acuerdo con la estadística dada por (3.20),

$$\begin{aligned}
\chi_0^2 &= n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \\
&= (20)\begin{pmatrix} 71.45 - 70 \\ 164.7 - 170 \end{pmatrix}' \begin{pmatrix} 20 & 100 \\ 100 & 1000 \end{pmatrix}^{-1} \begin{pmatrix} 71.45 - 70 \\ 164.7 - 170 \end{pmatrix} \\
&= (20)\begin{pmatrix} 1.45 & -5.3 \end{pmatrix} \begin{pmatrix} 0.1 & -0.01 \\ -0.01 & 0.002 \end{pmatrix} \begin{pmatrix} 1.45 \\ -5.3 \end{pmatrix} \\
&= 8.4026.
\end{aligned}$$

Para $\alpha = 0.05$, $\chi^2_{(0.05,2)} = 5.99$, se rechaza la hipótesis $H_0: \boldsymbol{\mu} = (70, 170)'$, pues
$\chi_0^2 = 8.4026 > 5.99$ (tabla C.7).

En la figura 3.1 se muestra que la región de rechazo para $\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2)'$ está sobre o fuera de la elipse; es decir, χ_0^2 es mayor que 5.99, si y sólo si, $\bar{\boldsymbol{X}}$ está fuera de la elipse. Si $\bar{\boldsymbol{X}}$ se ubica dentro de la elipse, H_0 no se rechaza. Es decir, tanto la distancia a $\boldsymbol{\mu}_0$ como la dirección deben ser considerados. Nótese que la distancia es "estandarizada" por $\boldsymbol{\Sigma}$, de manera que todos los puntos que están sobre la elipse (puntos para los cuales $\chi^2 = 5.99$) son equidistantes (estadísticamente) del centro de la elipse.

Esta prueba es sensible a la estructura de la matriz de covarianzas. Si la covarianza entre X_1 y X_2 fuese negativa, el eje principal de la elipse tendría una pendiente negativa; es decir, la elipse tendría una orientación

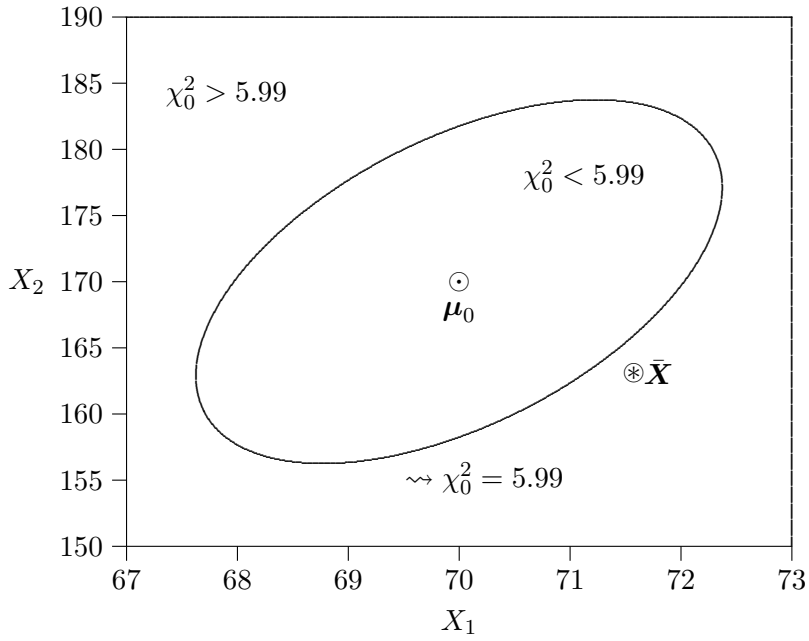


Figura 3.1 *Región de no rechazo bivariada.*

diferente. En ese caso, $\bar{\boldsymbol{X}}$ se ubicaría dentro de la región de aceptación. Esto advierte sobre la influencia de la correlación entre las variables en la decisión estadística sugerida por los datos.

Se presentan a continuación los resultados de las pruebas para cada parámetro por separado; es deicr, $H_{0_1}: \mu_1 = 70$ y $H_{0_2}: \mu_2 = 170$. Se emplea $Z_{\alpha/2} = 1.96$ para $\alpha = 0.05$ (tabla C.5). Cada una de las variables aleatorias es normal, puesto que conjuntamente tienen distribución normal bivariada, las pruebas estadísticas están dadas por

$$z_1 = \frac{\bar{x}_1 - \mu_{01}}{\sigma_1/\sqrt{n}} = \frac{71.45 - 70}{\sqrt{20}/\sqrt{20}} = 1.450 < 1.96$$

y

$$z_2 = \frac{\bar{x}_2 - \mu_{02}}{\sigma_2/\sqrt{n}} = \frac{164.5 - 170}{\sqrt{1000}/\sqrt{20}} = -0.7495 > -1.96.$$

De esta manera, en los dos casos no se rechazan las respectivas hipótesis nulas. Así, ninguna de las medias muestrales, $\bar{x}_1$ y $\bar{x}_2$, está suficientemente alejada del valor supuesto como para provocar su rechazo. Debido a la correlación positiva entre X_1 y X_2 las posibles discrepancias que existan respecto a cada una de las componentes del vector $\boldsymbol{\mu}_0$ se "combinan" para

causar el rechazo de H_0. Esto se anotó como la tercera ventaja de las pruebas multivariadas y se evidencia para el conjunto de datos de la muestra contenida en la tabla 3.2.

La figura 3.2 muestra la región de no rechazo ("aceptación") para las pruebas univariadas (rectángulo) y la correspondiente a la prueba multivariada (interior de la elipse), además se señalan las dos regiones que sugieren decisiones estadísticas en "contra vía". De una parte, está la zona donde se rechaza la hipótesis multivariada pero se aceptan las hipótesis univariadas (zona (1)). En la otra región, se acepta la hipótesis multivariada y se rechazan las univariadas (zona (2)). El rectángulo se obtiene como el producto cartesiano de las dos zonas de no rechazo; esto es,

$$\mu_{01} - 1.96\frac{\sigma_1}{\sqrt{n}} < \bar{x}_1 < \mu_{01} + 1.96\frac{\sigma_1}{\sqrt{n}} \text{ y}$$
$$\mu_{02} - 1.96\frac{\sigma_2}{\sqrt{n}} < \bar{x}_2 < \mu_{02} + 1.96\frac{\sigma_2}{\sqrt{n}}. \text{ ☑}$$

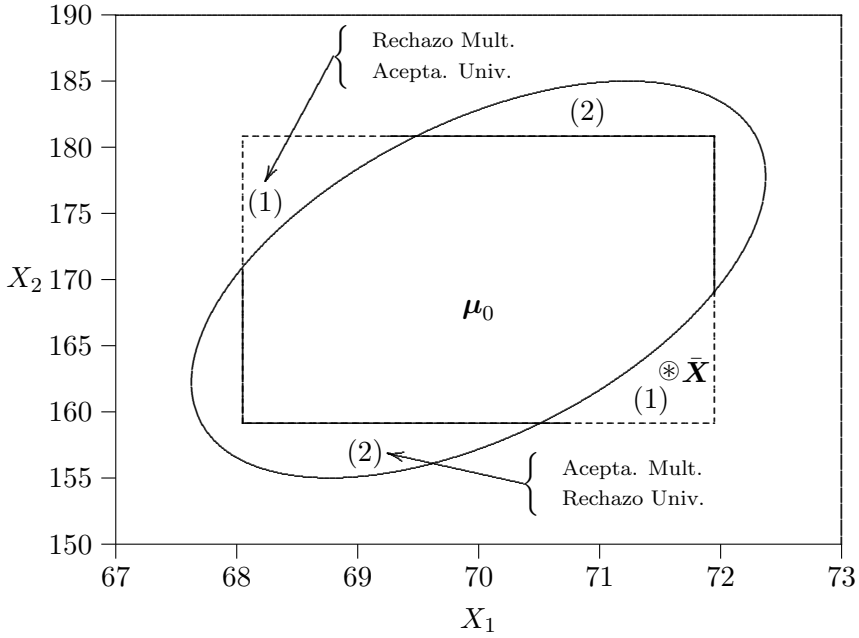


Figura 3.2 *Regiones de rechazo y no rechazo para pruebas univariadas y multivariadas.*

Ejemplo 3.3 Se ha observado, después de varios estudios en niños de alrededor dos años de edad, que la estatura (X_1), la longitud toráxica (X_2) y la circunferencia media del antebrazo (X_3), tienen aproximadamente una

distribución normal. Las siguientes mediciones fueron realizadas en seis de estos niños, la tabla 3.3 contiene los datos (tomados de Chatfield y Collins 1986, pág. 116)

Tabla 3.3 Estatura, tórax y antebrazo en niños

Niño	Estatura X_1 (cm)	Tórax X_2 (cm)	Antebrazo X_3 (cm)
1	78	60.6	16.5
2	76	58.1	12.5
3	92	63.2	14.5
4	81	59.0	14.0
5	81	60.8	15.5
6	84	59.5	14.0

Se sabe que la matriz de covarianzas del vector $\boldsymbol{X}' = (X_1, X_2, X_3)$ es

$$\boldsymbol{\Sigma} = \begin{pmatrix} 29.64 & 8.59 & 0.38 \\ 8.59 & 3.47 & 1.22 \\ 0.38 & 1.22 & 2.04 \end{pmatrix}.$$

Con esta información se desea probar la hipótesis $H_0: \ \boldsymbol{\mu}' = (90, 58, 16)$.

El vector de medias es $\bar{\boldsymbol{X}}' = (82.0, 60.0, 14.5)$, $(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' = (-8.0, 2.2, -1.5)$. El valor de la estadística (3.20) es

$$\chi_0^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) =$$

$$6 \cdot (-8.0, 2.2, -1.5) \begin{pmatrix} 29.64 & 8.59 & 0.38 \\ 8.59 & 3.47 & 1.22 \\ 0.38 & 1.22 & 2.04 \end{pmatrix}^{-1} \begin{pmatrix} -8.0 \\ 2.2 \\ -1.5 \end{pmatrix} =$$

$$6 \cdot (-8.0, 2.2, -1.5) \begin{pmatrix} 0.24697 & -0.75369 & 0.40473 \\ -0.75369 & 2.66491 & -1.45333 \\ 0.40473 & -1.45333 & 1.28395 \end{pmatrix} \begin{pmatrix} -8.0 \\ 2.2 \\ -1.5 \end{pmatrix} = 464.57402.$$

Este valor es mayor que $\chi^2_{(0.01,3)} = 11.34$ (tabla C.7), de donde se concluye que hay una evidencia fuerte contra la hipótesis de que las medias de estatura, longitud toráxica y la circunferencia del antebrazo son, respectivamente iguales a 90, 58 y 16. En estas situaciones convendría hacer caso omiso de la correlación entre las variables, para proceder a efectuar contrastes univariados sobre cada una de las medias, con el fin de verificar cuáles variables provocan el rechazo de H_0: otra alternativa es la verificación de hipótesis sobre subgrupos de variables de las que se tenga interés.☑

▶ Dos poblaciones

Para el caso de dos poblaciones p–dimensionales normales e independientes, con vectores de medias $\boldsymbol{\mu}_1$ y $\boldsymbol{\mu}_2$ respectivamente y la misma matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ conocida; se considera el problema de contrastar la hipótesis

$$H_0 : \ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2, \quad \text{equivalente a } H_0 : \ \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = 0. \tag{3.23}$$

Supóngase que se tienen dos muestras (X_{α_i}) para $\alpha_i = 1, \ldots, n_i$ de $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, para $i = 1, 2$. Con esta notación α indica la observación e i la población; así X_{α_i} es la observación α–ésima dentro de la i–ésima población.

Las medias muestrales son:

$$\bar{\boldsymbol{X}}_1 = \frac{1}{n_1} \sum_{\alpha_1=1}^{n_1} X_{\alpha_1} \ \text{ y } \ \bar{\boldsymbol{X}}_2 = \frac{1}{n_2} \sum_{\alpha_2=1}^{n_2} X_{\alpha_2}. \tag{3.24}$$

Estas medias son independientes y se distribuyen $N_p(\boldsymbol{\mu}_i, (1/n_i)\boldsymbol{\Sigma})$ para $i = 1, 2$. La diferencia entre las dos medias $\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$ se distribuye $N_p(\boldsymbol{\mu}, (n_1^{-1} + n_2^{-1})\boldsymbol{\Sigma})$ con $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

La región crítica para contrastar la hipótesis (3.23) es determinada por los puntos que satisfacen:

$$\chi_0^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) > \chi^2_{(\alpha,p)} \tag{3.25}$$

De otra forma, se rechaza H_0 con un nivel de significancia $(1 - \alpha)$ si se cumple (3.25).

Una región de confianza para estimar la diferencia entre los dos vectores de medias poblacionales es:

$$\frac{n_1 n_2}{n_1 + n_2} (\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \leq \chi^2_{(\alpha,p)}, \tag{3.26}$$

la cual es un elipsoide con centro $\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$, cuya forma y tamaño dependen de $\boldsymbol{\Sigma}$.

La cantidad:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{3.27}$$

es la *Distancia de Mahalanobis* (sección (1.4.6)), y mide la distancia al cuadrado entre los centros $\boldsymbol{\mu}_1$ y $\boldsymbol{\mu}_2$ de las dos poblaciones que tienen la misma matriz de varianzas y covarianzas.

▶ q–poblaciones

Ahora, para q–poblaciones normales p–variantes con la misma matriz de varianzas y covarianzas conocida, considérense q–muestras aleatorias independientes de tamaño n_i, $i = 1, \ldots, q$. Sea $\bar{\boldsymbol{X}}_i$ el vector de medias de la i-ésima muestra. La hipótesis para contrastar es:

$$H_0 : \sum_{i=1}^{q} l_i \boldsymbol{\mu}_i = \boldsymbol{\mu}_0, \tag{3.28}$$

donde los l_i son constantes conocidas y $\boldsymbol{\mu}_0$ es un vector p-dimensional también conocido.

Nótese que un estimador de $\sum_{i=1}^{q} l_i \boldsymbol{\mu}_i$ es $\sum_{i=1}^{q} l_i \bar{\boldsymbol{X}}_i$. La matriz de covarianzas del estimador $\sum_{i=1}^{q} l_i \bar{\boldsymbol{X}}_i$ es $\mathrm{Cov}\Big(\sum_{i=1}^{q} l_i \bar{\boldsymbol{X}}_i\Big) = \sum_{i=1}^{q} \Big(\frac{l_i^2}{n_i} \boldsymbol{\Sigma}\Big)$.

Para contrastar la combinación lineal de los vectores $\boldsymbol{\mu}_i$ dada en (3.28) se utiliza como región de rechazo a:[3]

$$C\Big(\sum_{i=1}^{q} l_i \bar{\boldsymbol{X}}_i - \boldsymbol{\mu}_0\Big)' \boldsymbol{\Sigma}^{-1} \Big(\sum_{i=1}^{q} l_i \bar{\boldsymbol{X}}_i - \boldsymbol{\mu}_0\Big) \geq \chi^2_{(\alpha,p)}, \tag{3.29}$$

donde C es una constante dada por:

$$C^{-1} = \sum_{i=1}^{q} l_i^2 / n_i.$$

Dada $\bar{\boldsymbol{X}}_i$, $i = 1, \ldots, q$, una región de confianza del $(1-\alpha)\%$ para el vector de medias poblacional $\boldsymbol{\mu} = \sum_{i=1}^{q} l_i \boldsymbol{\mu}_i$ está determinada por el elipsoide:

$$C\Big(\sum_{i=1}^{q} l_i \bar{\boldsymbol{X}}_i - \boldsymbol{\mu}\Big)' \boldsymbol{\Sigma}^{-1} \Big(\sum_{i=1}^{q} l_i \bar{\boldsymbol{X}}_i - \boldsymbol{\mu}\Big) \leq \chi^2_{(\alpha,p)}, \tag{3.30}$$

con centro en

$$\sum_{i=1}^{q} l_i \bar{\boldsymbol{X}}_i.$$

Para el caso en el que $\sum_{i=1}^{q} l_i = 0$ se tienen los llamados *contrastes lineales*.

[3] La determinación se hace a través de la razón de máxima verosimilitud generalizada.

3.4.2 Matriz de covarianzas desconocida: Estadística $\mathbf{T}^2$ de Hotelling

En la mayoría de las situaciones prácticas, rara vez se conoce la matriz de covarianzas. Se desarrolla ahora, un contraste de hipótesis y una estimación de regiones de confianza para el vector de medias $\boldsymbol{\mu}$ de una población normal p variante con matriz de varianzas y covarianzas desconocida.

En una población normal univariada, el problema de verificar si la media es igual a cierto valor específico, cuando se desconoce la varianza, se realiza mediante la variable aleatoria:

$$t = \sqrt{n}(\bar{X} - \mu_0)/s, \tag{3.31}$$

la cual tiene distribución t–Student con $(n-1)$ grados de libertad (n el tamaño de muestra).

Una expresión análoga a (3.31) se obtiene para el campo multivariado, ésta se conoce como la estadística T^2 *de Hotelling* (Hotelling (1931)).

▶ Obtención de la estadística $\mathbf{T}^2$ mediante la razón de máxima verosimilitud

Sea $X_1, \ldots, X_n$ $(n > p)$ una muestra aleatoria de una distribución normal p variante con media $\boldsymbol{\mu}$ y matriz de varianzas y covarianzas desconocida $\boldsymbol{\Sigma}$. Con base en esta muestra se quiere contrastar la hipótesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$.

Se deriva ahora la estadística de prueba pertinente. Este problema se conoce como el problema de Hotelling; puesto que fue quien primero la propuso para abordar el problema de dos muestras multivariadas, junto con su distribución bajo la hipótesis nula.

Para la muestra $X_1, \ldots, X_n$, la función de verosimilitud es:

$$\boldsymbol{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-pn/2}|\boldsymbol{\Sigma}|^{-n/2} \exp\left(-1/2 \sum_{\alpha=1}^{n} (x_\alpha - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (x_\alpha - \boldsymbol{\mu}) \right). \tag{3.32}$$

Del criterio de la razón de máxima verosimilitud:

$$\boldsymbol{\lambda} = \frac{\max\limits_{\boldsymbol{\Sigma}} \boldsymbol{L}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{\max\limits_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \boldsymbol{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}. \tag{3.33}$$

El numerador de (3.33) es el máximo de la función de verosimilitud para ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) en el espacio de parámetros restringido por la hipótesis nula (Ω_0), y el denominador es el máximo sobre todo el espacio de parámetros (Ω). Nótese

que como el conjunto de parámetros restringido por la hipótesis nula (Ω_0) está contenido en el espacio de parámetros completo (Ω), el numerador es menor que el denominador (pues si $A \subseteq B$ entonces $P(A) \leq P(B)$), de manera que $\boldsymbol{\lambda}$ es un número entre 0 y 1. Valores de $\boldsymbol{\lambda}$ cercanos a 1 provocan decisiones en favor de H_0, en tanto que valores cercanos a 0 sugieren el rechazo de H_0.

El máximo del denominador se obtiene en

$$\widehat{\boldsymbol{\mu}}_\Omega = \bar{\boldsymbol{X}} \quad \text{y} \quad \widehat{\boldsymbol{\Sigma}}_\Omega = \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{\alpha=1}^{n}(X_\alpha - \bar{\boldsymbol{X}})(X_\alpha - \bar{\boldsymbol{X}})'$$

respectivamente; mientras que el máximo del numerador, en el espacio restringido por la hipótesis nula, se obtiene en

$$\widehat{\boldsymbol{\Sigma}}_{\Omega_0} = \frac{1}{n}\sum_{\alpha=1}^{n}(X_\alpha - \boldsymbol{\mu}_0)(X_\alpha - \boldsymbol{\mu}_0)'.$$

Después de algunas consideraciones sobre la maximización y de adecuadas transformaciones algebraicas se obtiene:

$$\begin{aligned}\boldsymbol{\lambda} &= \frac{|\widehat{\boldsymbol{\Sigma}}_\Omega|^{\frac{n}{2}}}{|\widehat{\boldsymbol{\Sigma}}_{\Omega_0}|^{\frac{n}{2}}} = \frac{|\sum_\alpha (X_\alpha - \bar{\boldsymbol{X}})(X_\alpha - \bar{\boldsymbol{X}})'|^{\frac{n}{2}}}{|\sum_\alpha (X_\alpha - \boldsymbol{\mu}_0)(X_\alpha - \boldsymbol{\mu}_0)'|^{\frac{n}{2}}} \\ &= \frac{|\boldsymbol{A}|^{\frac{n}{2}}}{|\boldsymbol{A} + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'|^{\frac{n}{2}}}\end{aligned}$$

donde $\boldsymbol{A} = \sum_\alpha (X_\alpha - \bar{\boldsymbol{X}})(X_\alpha - \bar{\boldsymbol{X}})' = (n-1)\boldsymbol{S}$. Finalmente, de las propiedades para el cálculo de determinantes expresados en (A2.41a) y (A2.41b), se obtiene que:

$$\begin{aligned}\boldsymbol{\lambda}^{\frac{2}{n}} &= \frac{|\boldsymbol{A}|}{|\boldsymbol{A} + [\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)][\sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)]'|} \\ &= \frac{1}{1 + \left(\frac{1}{n-1}\right) n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)' S^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)} = \frac{1}{1 + T^2/(n-1)}\end{aligned}$$

con

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)' S^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \tag{3.34}$$

$\boldsymbol{S}$ la matriz de varianzas y covarianzas muestral.

La distribución de la estadística T^2 fue obtenida por Hotelling (1931), bajo H_0 y asumiendo que la muestra proviene de una distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. La distribución de T^2 es determinada por el valor p y los grados de libertad $\nu = n - 1$. La tabla C.1 contiene los valores críticos superiores para la

distribución exacta de la estadística T^2, con $\alpha = 0.05$ y $\alpha = 0.01$, para valores de p entre 1 y 10 con incrementos de 1 (se nota $p = 1\ (1)\ 10$).

Así como se muestra que la estadística univariada t-Student es un caso especial de la distribución F a través de la relación $t^2_{(n)} = F_{(1,n)}$, la distribución de la estadística T^2 de Hotelling se relaciona con la F; es decir, $T^2 \sim k \cdot F$, con k una constante, la cual junto con los grados de libertad se determina más adelante.

La región crítica para la prueba es el conjunto de valores muestrales que satisfacen la desigualdad $T^2 \geq T_0^2$, donde:

$$T_0^2 = (n-1)(\boldsymbol{\lambda}_0^{-2/n} - 1). \tag{3.35}$$

El valor de $\boldsymbol{\lambda}_0$ escogido es tal que

$$P(\boldsymbol{\lambda} \leq \boldsymbol{\lambda}_{(0)} | H_0) = \alpha. \tag{3.36}$$

Para la distribución de T^2 (necesaria en (3.34)), considérese $T^2 = Y' S^{-1} Y$, donde Y se distribuye $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y $(n-1)S$, se distribuye como $\mathcal{W}(\boldsymbol{\Sigma}, n-1)$. Para la estadística T^2 definida en (3.34) así:

$$Y = \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \qquad y \qquad \boldsymbol{\mu} = \sqrt{n}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$$

y bajo las anteriores consideraciones se determina que la distribución de $\frac{T^2}{n-1}\left(\frac{n-p}{p}\right)$ es:

$$\frac{T^2}{n-1}\left(\frac{n-p}{p}\right) \sim F_{(p,n-p)} \tag{3.37}$$

que corresponde a una distribución F no central con p y $(n-p)$ grados de libertad y parámetro de no centralidad $\boldsymbol{\lambda} = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$; ahora, si $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ entonces F es central (Muirhead, 1982, pág. 98).

▶ Obtención de la estadística $\mathbf{T}^2$ mediante el principio de unión–intersección

El *principio de unión–intersección* (UI) es un procedimiento para la construcción de pruebas desarrollado por Roy (1957). El propósito de esta sección es mostrar que para $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ frente a $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, con el principio de UI se logra una estadística aproximadamente tipo T^2.

La hipótesis H_0 es cierta, si y sólo si $H_{0a} : \ \boldsymbol{a}'\boldsymbol{\mu} = \boldsymbol{a}'\boldsymbol{\mu}_0$ es cierta *para todo* $\boldsymbol{a} \in \boldsymbol{R}^p$; nótese que si $\boldsymbol{a} = \boldsymbol{e}_i$ se trata de la hipótesis sobre una de las componentes del vector $\boldsymbol{\mu}$, es decir que $\mu_i = \mu_{i0}$, con $i = 1, 2, \ldots, p$. La hipótesis es falsa, si y sólo si, $H_{1a} : \ \boldsymbol{a}'\boldsymbol{\mu} \neq \boldsymbol{a}'\boldsymbol{\mu}_0$ *para al menos un* $\boldsymbol{a} \in \boldsymbol{R}^p$.

Se rechaza $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ si se encuentra al menos un vector $\boldsymbol{a} \in \boldsymbol{R}^p$ tal que la hipótesis univariada, $\boldsymbol{a}'\boldsymbol{\mu} = \boldsymbol{a}'\boldsymbol{\mu}_0$, sea rechazada. Por tanto, la región de rechazo para H_0 es la *unión* de las regiones de rechazo para las hipótesis univariadas asociadas con los $\boldsymbol{a} \in \boldsymbol{R}^p$. Similarmente, no se rechaza $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, únicamente si cada hipótesis univariada $\boldsymbol{a}'\boldsymbol{\mu} = \boldsymbol{a}'\boldsymbol{\mu}_0$ no es rechazada. La región de no rechazo es entonces la *intersección* sobre todos los $\boldsymbol{a} \in \boldsymbol{R}^p$ de las regiones de no rechazo ligadas con las hipótesis univariadas.

En símbolos, dadas (H_0, H_1) hipótesis nula y alterna, respectivamente, entonces

$$H_0 = \bigcap_a H_{0a}, \quad H_1 = \bigcup_a H_{1a}$$

donde (H_{0a}, H_{1a}) forman un par natural, pues la una es el complemento de la otra; éstas se subindizan con a para resaltar la dependencia del vector $\boldsymbol{a}$ que define la respectiva combinación lineal.

Como $\bar{\boldsymbol{X}}$ se distribuye $N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$, entonces $\boldsymbol{a}'\bar{\boldsymbol{X}}$ se distribuye $n(\boldsymbol{a}'\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$, así, se puede verificar la hipótesis $H_{oa} : \boldsymbol{a}'\boldsymbol{\mu} = \boldsymbol{a}'\boldsymbol{\mu}_0$ mediante la estadística t

$$t(\boldsymbol{a}) = \frac{\boldsymbol{a}'\bar{\boldsymbol{X}} - \boldsymbol{a}'\bar{\boldsymbol{\mu}}_0}{\sqrt{\frac{1}{n}\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}}$$

la cual, para un $\boldsymbol{a} \in \boldsymbol{R}^p$ dado, tiene región de no rechazo

$$|t(\boldsymbol{a})| < c$$

donde c se toma de acuerdo con un valor adecuado α para la prueba. La región de no rechazo para $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ es entonces

$$\bigcap_a \{|t(\boldsymbol{a})| < c\}$$

donde la intersección sobre todos los $\boldsymbol{a} \in \boldsymbol{R}^p$ define el intervalo más pequeño que contiene todos los $t(\boldsymbol{a})$ y es acotado por $\mp\, c$. La región de rechazo para $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ es

$$\bigcup_a \{|t(\boldsymbol{a})| \geq c\}$$

y se rechaza H_0 si algún $|t(\boldsymbol{a})|$ es mayor o igual que c, es decir, si $\max_a\{|t(\boldsymbol{a})| \geq c\}$.

Para encontrar $\max_a\{|t(\boldsymbol{a})| \geq c\}$, es más conveniente trabajar con $t^2(\boldsymbol{a})$, el cual puede escribirse como

$$t^2(\boldsymbol{a}) = \frac{n[\boldsymbol{a}'(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)]^2}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}}.$$

A través de cálculo diferencial se encuentra que el máximo corresponde al punto en donde se anula la primera derivada. Esto se hace resolviendo el sistema de ecuaciones que representa los puntos donde se anula la primera derivada respecto al vector $\boldsymbol{a}$. Así, el máximo de la expresión anterior es

$$\max_a t^2(\boldsymbol{a}) = \max_a \frac{n[\boldsymbol{a}'(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)]^2}{\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}} = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'S^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) = T^2.$$

En resumen, la razón de máxima verosimilitud y el principio de unión-intersección suministran la misma estadística de prueba para hipótesis sobre el vector de medias cuando la muestra es extraida de una población con distribución normal multivariada.

3.4.3 Aplicaciones de la Estadística $\mathbf{T}^2$

▶ Contraste de hipótesis sobre la media en una población

La región crítica para verificar la hipótesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, con base en una muestra aleatoria de una población $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, viene dada por el conjunto de puntos muestrales $\{X_\alpha : \ T^2 \geq T_0^2\}$. Si el nivel de significación es α, entonces, el percentil $(1-\alpha)\%$ de la distribución F se considera así:

$$T_0^2 = \frac{(n-1)p}{(n-p)}F(p, n-p)(\alpha) = T^2_{(\alpha,p,n-1)}. \tag{3.38}$$

▶ Región de confianza para el vector de medias

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una población normal p variante con media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$, ambas desconocidas.

La expresión $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$ tiene distribución T^2 de Hotelling donde $\nu = (n-1)$ grados de libertad. Sea $T^2_{(\alpha)}$, tal que $P\big(T^2 \geq T^2_{(\alpha)}\big) = \alpha$; entonces la probabilidad de extraer una de estas muestras tal que:

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' S^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \leq T^2_{(p;\nu)}(\alpha), \tag{3.39}$$

es $(1 - \alpha)$. En consecuencia, para una muestra $X_1, \ldots, X_n$ una región del $(1-\alpha)\%$ de confianza para estimar el vector $\boldsymbol{\mu}$, consta de todos los vectores $\boldsymbol{m}$ que satisfacen

$$n(\bar{\boldsymbol{X}} - \boldsymbol{m})' S^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{m}) \leq T^2_{(p;\nu)}(\alpha). \tag{3.40}$$

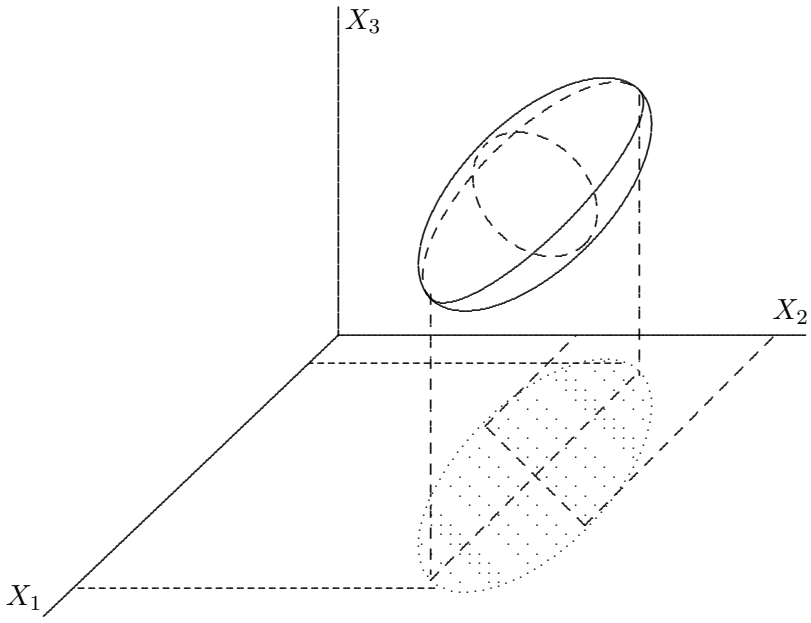


Figura 3.3 *Región de confianza.*

La desigualdad (3.40) representa el interior y la superficie de un elipsoide en el espacio p-dimensional de $\boldsymbol{m}$, con centro en $\bar{\boldsymbol{X}}$, cuyo tamaño y forma dependen de $\boldsymbol{S}^{-1}$ y α. La estructura (forma, tamaño y orientación) del elipsoide están determinados por la magnitud de los elementos dispuestos en la diagonal principal (varianza de cada variable) de la matriz de covarianzas y por la magnitud y signo de los elementos ubicados fuera de la diagonal principal (covarianzas) de la misma. La figura 3.3 muestra la región de confianza, que para el caso tridimensional es un elipsoide, la cual al proyectarse sobre el plano $X_1 \times X_2$ determina una elipse.

La figura 3.4 muestra más explícitamente la región de confianza para el caso bivariado; se advierte la existencia de puntos tales como B, C y D,

para los cuales la pertenencia a alguno de los intervalos de confianza univariados no implica la pertenencia a la región de confianza multivariada. Así: el punto A se encuentra dentro de la región de confianza multivariada pero fuera de los intervalos univariados ; los puntos B y C se ubican en sólo uno de los intervalos univariados y fuera de la región de confianza multivariada; el punto D se ubica dentro de los intervalos de confianza univariados pero fuera de la región de confianza multivariada; y, el punto E se ubica dentro de todas las tres regiones de confianza. Esta consideración es importante tenerla en cuenta cuando se hacen contrastes de hipótesis de manera independiente para cada variable, pues los resultados univariados no siempre coinciden con los resultados multivariados, y, recíprocamente, los resultados multivariados no implican los univariados. Esto se puede apreciar en las figuras 3.2 y 3.4. Una observación semejante a la anterior se tiene en el trabajo con *cartas para control de calidad* multivariadas, las cuales se presentan y comentan más adelante.

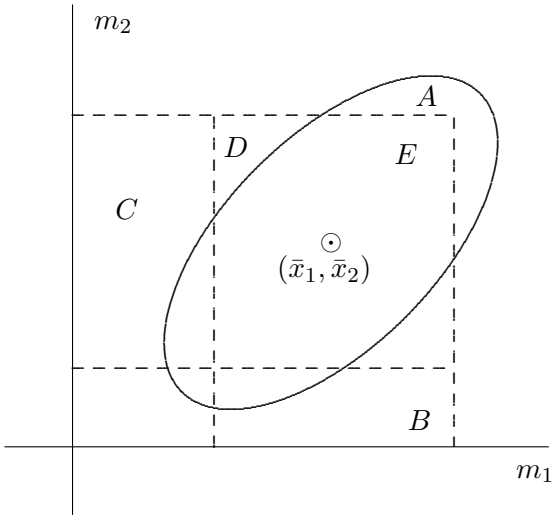


Figura 3.4 *Región de confianza bivariada.*

► Contrastes sobre combinación lineal de medias ($C\boldsymbol{\mu} = \mathbf{0}$)

La hipótesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ determina a cada μ_i completamente; pues esto implica que $\mu_i = \mu_{0i}$ para todo $i = 1, \ldots, p$. Algunas combinaciones entre los μ_i son de interés frecuente, siempre que las variables $X_1, \ldots, X_p$, del vector aleatorio X, sean *conmensurables*, es decir, con las mismas unidades y con varianzas comparables. Varias combinaciones lineales pueden exami-

narse desde la expresión $\boldsymbol{C\mu} = 0$, por ejemplo, la hipótesis

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_p$$

equivale a $H_{0^*}: \ \mu_1 - \mu_i = 0$ para todo $i = 2, \ldots, p$, o también a $H_{0^*}: \mu_i - \mu_{i+1} = 0$ para todo $i = 1, \ldots, p-1$. Estas expresiones equivalen a combinaciones lineales de los μ_i, las cuales pueden escribirse como $H_0: \boldsymbol{C}_1\boldsymbol{\mu} = 0$ o $H_0: \ \boldsymbol{C}_2\boldsymbol{\mu} = 0$, donde

$$\boldsymbol{C}_1 = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & -1 \end{pmatrix}, \quad \boldsymbol{C}_2 = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix},$$

entre otras. Esto indica que la matriz $\boldsymbol{C}$ no es única. La matriz $\boldsymbol{C}$ es una matriz de tamaño $((p-1) \times p)$, de rango completo, tal que $\boldsymbol{C1} = 0$ (filas suman cero). La hipótesis puede extenderse a la forma más general $H_0: \ \boldsymbol{\mu} = \boldsymbol{\gamma}$ para un valor de γ específico (más adelante se consideran las medidas repetidas como un caso especial de ésta).

Para probar $H_0: \ \boldsymbol{\mu} = 0$, se deben transformar los datos mediante $Y = \boldsymbol{C}X$. De esta manera, el vector de medias muestral para Y es $\bar{Y} = \boldsymbol{C}\bar{X}$ y su matriz de covarianzas $\boldsymbol{S}_Y = \boldsymbol{C}\boldsymbol{S}_X\boldsymbol{C}'$. Si el vector X tiene distribución $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, entonces, por la propiedad (2.2.2), $Y = \boldsymbol{C}X$ tiene distribución $N_{p-1}(\boldsymbol{C\mu}, \boldsymbol{C\Sigma C}')$.

La estadística para verificar esta hipótesis está dada por

$$T^2 = n(\bar{Y})'\boldsymbol{S}_Y^{-1}(\bar{Y}) = n(\boldsymbol{C}\bar{X})'(\boldsymbol{C}\boldsymbol{S}_X\boldsymbol{C}')^{-1}(\boldsymbol{C}\bar{X}), \tag{3.41}$$

la cual se distribuye, bajo H_0, como $T^2_{(p-1,n-1)}$. Se rechaza $H_0: \ \boldsymbol{\mu} = 0$ si $T^2 \geq T^2_{(\alpha,p-1,n-1)}$. Nótese que la dimensión $(p-1)$ es el número de filas de $\boldsymbol{C}$, y son las variables resultantes de la transformación $Y = \boldsymbol{C}X$.

Para una hipótesis más general $H_0: \ \boldsymbol{C\mu} = \boldsymbol{\gamma}$, donde $\boldsymbol{C}$ es una matriz de tamaño $(k \times p)$ y de rango k, se usa

$$T^2 = n(\boldsymbol{C}\bar{X} - \boldsymbol{\gamma})'\left(\boldsymbol{C}\boldsymbol{S}_X\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\bar{X} - \boldsymbol{\gamma}), \tag{3.42}$$

la cual se distribuye como $T^2_{(k,n-1)}$ (bajo $H_0: \ \boldsymbol{\mu} = \boldsymbol{\gamma}$). La relación con la estadística F es:

$$F = [(n-k)/(k(n-1))]T^2,$$

la cual, bajo H_0, tiene distribución $F_{(k,n-k)}$.

Ejemplo 3.4 Los datos contenidos en la tabla 3.4 corresponden a los pesos (en centigramos) del corcho encontrado en muestras tomadas en la

dirección norte (N), este (E), sur (S) y oeste (O) del tronco de 28 árboles cultivados en una parcela experimental. En este caso las variables corresponden al peso de las cuatro muestras tomadas sobre cada árbol.

Tabla 3.4 Pesos de corcho

(N)	(E)	(S)	(O)	(N)	(E)	(S)	(O)
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	30	32
37	40	31	25	60	50	67	54
33	29	27	36	35	37	48	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	43	48	54	57	43

Fuente: Krzanowski–Marriot (1994, pág. 165)

El vector de medias y la matriz de covarianzas muestral son, respectivamente,

$$\bar{x}' = (50.535, 46.179, 49.679, 45.179) \text{ y}$$

$$\boldsymbol{S} = \begin{pmatrix} 290.41 & 223.75 & 288.49 & 226.27 \\ 223.75 & 219.93 & 229.06 & 171.37 \\ 288.49 & 229.06 & 350.00 & 259.54 \\ 226.27 & 171.37 & 259.54 & 226.00 \end{pmatrix}.$$

Se quiere verificar si las medias de los pesos de corcho son iguales en la dirección norte–sur (N–S) y en la dirección este–oeste (E–O). Esto equivale a contrastar la hipótesis $H_0 : \ \mu_1 = \mu_3, \ \text{y} \ \mu_2 = \mu_4$. La hipótesis H_0 se puede expresar como $\boldsymbol{C\mu} = 0$, donde

$$\boldsymbol{C} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}.$$

Las expresiones $(\boldsymbol{C}\bar{X})$ y $(\boldsymbol{CS}_X\boldsymbol{C}')$, calculadas de acuerdo con los datos disponibles, son respectivamente

$$(\boldsymbol{C}\bar{X})' = (0.857, 1.000) \ \text{ y } \ \boldsymbol{CS}_X\boldsymbol{C}' = \begin{pmatrix} 61.27 & 26.96 \\ 26.96 & 99.50 \end{pmatrix}.$$

Mediante la ecuación (3.41), la estadística T^2 toma el valor

$$\begin{aligned} T^2 &= n(\boldsymbol{C}\bar{X})'(\boldsymbol{C}\boldsymbol{S}_X\boldsymbol{C}')^{-1}(\boldsymbol{C}\bar{X}) \\ &= (28)(0.857, 1.000)\begin{pmatrix} 61.27 & 26.96 \\ 26.96 & 99.50 \end{pmatrix}^{-1}\begin{pmatrix} 0.857 \\ 1.000 \end{pmatrix} \\ &= (28)(0.01641) \\ &= 0.4594. \end{aligned}$$

El valor para $F_{(5\%,2,26)} \cong 3.38$ (tabla C.8), de manera que estos datos no provocan el rechazo de la hipótesis nula. Es decir, en estas direcciones el contenido medio de corcho, en estos troncos, no es significativamente diferente. ☑

▶ Comparación de dos poblaciones asumiendo $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$

Como en la sección (3.4.1), considérense dos muestras de poblaciones normales p-variantes e independientes. Supóngase que (X_{α_1}), es una muestra de tamaño n_1 de una población $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ y (X_{α_2}) es una segunda muestra de tamaño n_2 de una población $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, con $\alpha_i = 1, \ldots, n_i$ e i=1,2. En estas condiciones la estadística T^2 puede emplearse para contrastar la hipótesis que la media de una población es igual a la media de la otra; donde la matriz de covarianzas, aunque desconocida, se supone igual.

El vector de medias muestral $\bar{\boldsymbol{X}}_i$ tiene distribución $N_p(\boldsymbol{\mu}_i, \frac{1}{n_i}\boldsymbol{\Sigma})$, para $i = 1, 2$. Así, el vector aleatorio

$$[n_1 n_2/(n_1 + n_2)]^{1/2}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$$

se distribuye como $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$; la deducción es similar a la anterior, aplicando adecuadamente las propiedades de la Sección (3.3). La matriz de covarianzas $\boldsymbol{\Sigma}$, se estima en forma mancomunada con las matrices de covarianzas muestrales; así,

$$\boldsymbol{S}_p = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{n_1 + n_2 - 2}.$$

La estadística

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)'\boldsymbol{S}_p{}^{-1}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2), \tag{3.43}$$

se distribuye como T^2 con dimensión p y $\nu = (n_1 + n_2 - 2)$ grados de libertad. La región crítica para contrastar la hipótesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ es

$$T^2 > \frac{\nu p}{(\nu - p + 1)}F_{(p,\nu-p+1)}(\alpha), \tag{3.44}$$

con un nivel de significación igual a α. Una región de confianza para $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, con un nivel de confiabilidad de $(1-\alpha)\%$, es el conjunto de vectores $\boldsymbol{m}$ que satisfacen:

$$((\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - \boldsymbol{m})' \boldsymbol{S}_p{}^{-1} ((\bar{\boldsymbol{X}}_1 - \bar{X}_2) - \boldsymbol{m}) \leq \frac{n_1 + n_2}{n_1 n_2} T^2_{(\nu)}(\alpha)$$
$$= \frac{n_1 + n_2}{n_1 n_2} \frac{(\nu)p}{(\nu - p + 1)} F_{(p,\nu-p+1)}(\alpha).$$

Ejemplo 3.5 Cuatro pruebas psicológicas fueron aplicadas sobre 32 hombres y 32 mujeres. Las variables a considerar son

X_1 : inconsistencias pictóricas $\quad$ X_2 : reconocimiento de herramientas

X_3 : forma de emplear el papel $\quad$ X_4 : Vocabulario

Se asume que cada grupo de personas es una muestra aleatoria de una población tetra-variante, con distribución normal de media $\boldsymbol{\mu}_i \quad (i = 1, 2)$ y matriz de covarianza $\boldsymbol{\Sigma}$, igual y desconocida para las dos poblaciones. El experimento se llevó a cabo de tal forma que las poblaciones (hombres y mujeres) resultaran independientes. El interés se dirige a contrastar la hipótesis: "mujeres y hombres tienen respuestas, en promedio, igual con respecto a cada uno de los cuatro atributos considerados"; en un lenguaje más técnico se escribe,

$$H_0 : \ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2.$$

Aquí $n_1 = n_2 = 32$, luego $\nu = n_1 + n_2 - 2 = 62$.

Los respectivos vectores de medias y matrices de covarianzas son

$$\bar{\boldsymbol{X}}_1 = \begin{pmatrix} 15.97 \\ 15.91 \\ 27.19 \\ 22.75 \end{pmatrix} \qquad \boldsymbol{S}_1 = \begin{pmatrix} 5.192 & 4.545 & 6.522 & 5.250 \\ 4.545 & 13.18 & 6.760 & 6.266 \\ 6.522 & 6.760 & 28.67 & 14.47 \\ 5.250 & 6.266 & 14.47 & 16.65 \end{pmatrix}$$

$$\bar{\boldsymbol{X}}_2 = \begin{pmatrix} 12.34 \\ 13.91 \\ 16.59 \\ 21.94 \end{pmatrix} \qquad \boldsymbol{S}_2 = \begin{pmatrix} 9.136 & 7.549 & 5.531 & 4.151 \\ 7.549 & 18.60 & 5.446 & 5.446 \\ 5.531 & 5.446 & 13.55 & 13.55 \\ 4.151 & 5.446 & 13.55 & 28.00 \end{pmatrix}.$$

Se asume que las matrices de covarianzas muestrales no reflejan una diferencia notoria con relación a las respectivas matrices de covarianzas poblacionales. (Una prueba que permite ratificar este supuesto se desarrolla en la sección (4.3.3) respecto a la hipótesis $H_0 : \ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$). La matriz de covarianzas muestral y mancomunada es

$$\boldsymbol{S}_p = \frac{1}{32 + 32 - 2} \left[(32-1)\boldsymbol{S}_1 + (32-1)\boldsymbol{S}_2 \right] = \begin{pmatrix} 7.64 & 6.047 & 6.027 & 4.701 \\ 6.047 & 15.89 & 8.747 & 5.586 \\ 6.027 & 8.747 & 29.46 & 14.01 \\ 4.701 & 5.586 & 14.01 & 22.32 \end{pmatrix}.$$

La estadística de prueba, por (3.43), es

$$T^2 = \Big(\frac{n_1 n_2}{n_1 + n_2}\Big)(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)' \boldsymbol{S}_p{}^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) = 97.61497,$$

entonces, por la transformación a la estadística F dada por la ecuación en (3.44) y como $F_{(4,59,5\%)} \approx 2.53$ (tabla C.8), se tiene que

$$T^2 > \frac{\nu p}{(\nu - p + 1)} F_{(5\%, p, \nu - p + 1)} \approx \frac{(62)(4)}{59}(2.53) = 10.6346,$$

y por tanto se rechaza $H_0 :\ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

Al final del capítulo se muestra el programa SAS del procedimiento IML con el cual se calcula la estadística T^2 para estos datos, junto con el valor p correspondiente a la estadística F asociada a éste (expresión (3.44)). Una decisión similar se obtiene empleando la tabla C.1, puesto que el valor de $T^2_{(62,\%)}$ es aproximadamente 10.6 y $T^2 = 97.61497 > 10.6$ (Rencher 1995, págs. 140-142). ☑

▶ Contrastes sobre observaciones pareadas

Supóngase que se tienen dos muestras, las cuales no son independientes porque existe un apareamiento natural entre la observación X_i de la primera muestra con la observación Y_i de la segunda muestra para todo i. Por ejemplo, cuando se aplica un tratamiento a un individuo y se observa su respuesta "pre" (X_i) y su respuesta "post" (Y_i) al tratamiento; otra situación es cuando los objetos son mezclados de acuerdo con algún criterio de homogeneidad, por ejemplo, individuos con un mismo cociente intelectual (CI) o con los mismos rasgos familiares. Con tales pares, el procedimiento es frecuentemente referido como *observaciones pareadas* o *pares mezclados*.

Se denotan las muestras por $X_1, \ldots, X_n$ y $Y_1, \ldots, Y_n$. Las dos muestras son correlacionadas; es decir, $\mathrm{Cov}(X_i, Y_i) \neq 0$, se puede trabajar directamente con las diferencias dentro de cada par de observaciones, $d_i = Y_i - X_i$. De esta forma, los n pares de observaciones se reducen a una sola muestra de n diferencias d_i, $i = 1, \ldots, n$. La hipótesis de igualdad de vectores de medias, $H_0 :\ \boldsymbol{\mu}_X = \boldsymbol{\mu}_Y$, es equivalente a $H_0 :\ \boldsymbol{\mu}_d = 0$. Para verificar H_0, se calcula

$$\bar{\boldsymbol{d}} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad \boldsymbol{S}_d = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{\boldsymbol{d}})(d_i - \bar{\boldsymbol{d}})',$$

de donde se obtiene

$$T^2 = n\bar{\boldsymbol{d}}' \boldsymbol{S}_d^{-1} \bar{\boldsymbol{d}}. \tag{3.45}$$

Si la hipótesis H_0 es cierta, la estadística T^2 se distribuye como $T^2_{(p,n-1)}$. Se rechaza la hipótesis H_0 si $T^2 \geq T^2_{(\alpha,p,n-1)}$. Se puede también transformar la estadística T^2, conforme a como se muestra en la ecuación (3.38), de manera la resultante esté asociada con la estadística F.

Aquí el supuesto de igualdad de matrices de covarianzas, $\boldsymbol{\Sigma}_{XX} = \boldsymbol{\Sigma}_{YY}$, no se requiere porque $\boldsymbol{S}_d$ estima a $\text{Cov}(X_i, Y_i) = \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XY} - \boldsymbol{\Sigma}_{YX} + \boldsymbol{\Sigma}_{YY}$; las cuales, como se observa, están contenidas en ésta.

Ejemplo 3.6 Se desea comparar dos tipos de esmalte para la resistencia a la corrosión, 15 piezas de tubería fueron cubiertas con cada tipo de esmalte. Dos tuberías, cada una con esmalte diferente, se enterraron y se dejaron durante el mismo período de tiempo en 15 lugares distintos; esto corresponde a un par de observaciones en condiciones semejantes, excepto por el tipo de cubrimiento. El efecto por la corrosión en el primer tipo de esmalte fue medido a través de las siguientes variables:

X_1**:** profundidad máxima de la picadura por corrosión (en milésimas de pulgada),

X_2**:** número de picaduras por corrosión.

Para el segundo tipo de esmalte se midieron las mismas variables notadas por Y_1 y Y_2. La tabla 3.5 contiene los respectivos datos. Para estas diferencias se obtiene

$$\bar{\boldsymbol{d}} = \begin{pmatrix} 8.000 \\ 3.067 \end{pmatrix} \text{ y } \boldsymbol{S}_d = \begin{pmatrix} 121.571 & 17.071 \\ 17.071 & 21.781 \end{pmatrix}.$$

De acuerdo con (3.46)

$$T^2 = (15)(8.000, 3.067) \begin{pmatrix} 121.571 & 17.071 \\ 17.071 & 21.781 \end{pmatrix}^{-1} \begin{pmatrix} 8.000 \\ 3.067 \end{pmatrix} = 10.819.$$

De la relación entre la estadística T^2 y la estadística F, mostrada en la ecuación (3.38), resulta

$$F_{(p,n-p)} = \frac{T^2}{n-1}\Big(\frac{n-p}{p}\Big) = \frac{10.819}{14}\Big(\frac{13}{2}\Big) = 5.02311.$$

Como $5.02311 > F_{(5\%,2,14)} = 3.74$ se rechaza H_0; es decir, los tipos de esmaltes tienen efectos significativamente diferentes, bajo las condiciones experimentales señaladas, respecto al control de la corrosión en tales tuberías. ☑

Tabla 3.5 Profundidad y número de picaduras por corrosión en tubos

	Esmalte 1		Esmalte 2		Diferencia	
Localidad	X_1	X_2	Y_1	Y_2	d_{E_1}	d_{E_2}
1	73	31	51	35	22	-4
2	43	19	41	14	2	5
3	47	22	43	19	4	3
4	53	26	41	29	12	-3
5	58	36	47	34	11	2
6	47	30	32	26	15	4
7	52	29	24	19	28	10
8	38	36	43	37	-5	-1
9	61	34	53	24	8	10
10	56	33	52	27	4	6
11	56	19	57	14	-1	5
12	34	19	44	19	-10	0
13	55	26	57	30	-2	-4
14	65	15	40	7	25	8
15	75	18	68	13	7	5

Fuente: Rencher (1995, pág. 152)

▶ Comparación de dos poblaciones asumiendo $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Para el caso univariado ($p = 1$), el problema de contrastar $H_0 : \ \mu_1 = \mu_2$ cuando $\sigma_1^2 \neq \sigma_2^2$, para muestras independientes, se conoce con el nombre de *problema de Behrens-Fisher*. En estas situaciones la variable aleatoria t (ecuación 3.31) no tiene distribución t–Student. Entre las aproximaciones propuestas, se tiene la solución debida a Welch (1937,1947). Si $\sigma_1^2 \neq \sigma_2^2$, entonces $\text{var}(\bar{X}_1 - \bar{X}_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$ (para muestras independientes), su estimador es $s_1^2/n_1 + s_2^2/n_2$. Cuando se emplea esto,

$$t_\nu = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

tiene distribución t con ν grados de libertad, donde

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1+1)] + (s_2^2/n_2)^2/(n_2+1)} - 2.$$

El correspondiente caso multivariado consiste en verificar $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ para $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. En esta prueba se asume que las dos muestras $X_{11}, \ldots, X_{1n_1}$ y $X_{21}, \ldots, X_{2n_1}$ de $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ y de $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, son independientes, respectivamente, con $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. Para estos casos la estadística T^2 asociada a

(3.43) no tiene distribución T^2 de Hotelling. A continuación se desarrollan la pruebas para los casos de tamaño de muestra igual y para muestras de tamaños desiguales.

- Tamaño de muestra igual ($n_1 = n_2$).

  Si $n_1 = n_2 = n$, se puede emplear la prueba para observaciones pareadas presentada en la sección anterior, puesto, como se advirtió allí, el supuesto que $\mathbf{\Sigma}_{XX} = \mathbf{\Sigma}_{YY}$ no es requerido en (3.45). La conformación de parejas (pareamiento) se hace mediante la asignación aleatoria de una pareja a cada observación de la primera muestra. Una vez que se han conformado las parejas se procede a desarrollar la prueba para observaciones pareadas conforme la estadística (3.45). El procedimiento produce una estadística con distribución T^2 exacta; aunque, tiene la desventaja de tener $\nu = n_1$ grados de libertad en lugar de $2(n-1)$. La pérdida de grados de libertad afecta la perdida de potencia en la prueba, se puede tomar como alternativa la prueba que se muestra a continuación.

- Tamaño de muestra desigual ($n_1 \neq n_2$).

  Una primera solución al problema de Behrens-Fisher es la conocida aproximación de Bennet. Ésta suministra una estadística con distribución T^2 exacta, pero excluye $(n_2 - n_1)$ observaciones de X_{2i} (si $n_2 > n_1$) al desarrollar los cálculos de la estadística. De aquí, se advierten dos desventajas de este procedimiento: (i) hay una pérdida en la potencia de la prueba si n_1 es bastante menor que n_2 y (ii) los resultados varían de acuerdo con las observaciones excluidas X_{2i}, este procedimiento se torna muy subjetivo. Por estas razones no se presentan los cálculos para el procedimiento de Bennet. En cambio, se muestra una solución multivariada aproximada al problema de Behrens-Fisher, dada por Johansen (1980), Nel y van der Merwe (1986) y Kim (1992), citados por Rencher (1998, pág. 101).

Si $\mathbf{\Sigma}_1$ y $\mathbf{\Sigma}_2$ fueran conocidas, la estadística

$$Z^2 = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)' \left(\frac{\mathbf{\Sigma}_1}{n_1} + \frac{\mathbf{\Sigma}_2}{n_2} \right)^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$$

tiene distribución $\chi^2_{(p)}$ bajo H_0. La versión muestral es

$$T^{*2} = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)' \left(\frac{\boldsymbol{S}_1}{n_1} + \frac{\boldsymbol{S}_2}{n_2} \right)^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2). \tag{3.46}$$

La aproximación dada por Nel y van der Merwe (1986) usa la estadística T^{*2}, la cual se distribuye aproximadamente como $T^2_{(p,\nu)}$ donde

$$\nu = \frac{\text{tra}(\boldsymbol{S}_e)^2 + [\text{tra}(\boldsymbol{S}_e)]^2}{(n_1-1)^{-1}\{\text{tra}(\boldsymbol{V}_1)^2 + [\text{tra}(\boldsymbol{V}_1)]^2\} + (n_2-1)^{-1}\{\text{tra}(\boldsymbol{V}_2)^2 + [\text{tra}(\boldsymbol{V}_2)]^2\}}, \quad (3.47)$$

y

$$\boldsymbol{V}_i = \frac{\boldsymbol{S}_i}{n_i},\ i = 1, 2 \text{ y}$$
$$\boldsymbol{S}_e = \boldsymbol{V}_1 + \boldsymbol{V}_2.$$

Para el desarrollo de la prueba de Kim (1992), se emplea la notación

$$\boldsymbol{A}^{-2} = (\boldsymbol{A}^{-1})^2, \qquad \boldsymbol{D} = \text{Diag}(d_1, \ldots, d_p)$$
$$\boldsymbol{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_p), \qquad \boldsymbol{w} = \boldsymbol{Q}'(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) \text{ y}$$
$$r = \Big(\prod_{j=1}^{p} d_j\Big)^{1/2p},$$

donde los d_j y los $\boldsymbol{q}_j$ son los valores y vectores propios de $\boldsymbol{V}_2^{-1}\boldsymbol{V}_1$. Entonces

$$\frac{\nu - p + 1}{bc\nu}\boldsymbol{w}'(\boldsymbol{D}^{1/2} + r\boldsymbol{I})^{-2}\boldsymbol{w}, \quad (3.48)$$

se distribuye aproximadamente como $F_{(c,\nu-p+1)}$, donde

$$b = \Big(\sum_{j=1}^{p} a_j^2\Big)/\Big(\sum_{j=1}^{p} a_j\Big),$$
$$c = \Big(\sum_{j=1}^{p} a_j\Big)^2/\Big(\sum_{j=1}^{p} a_j^2\Big),$$
$$a_j = (d_j + 1)/\big(d_j^{1/2} + r\big)^2,$$

y

$$\frac{1}{\nu} = \frac{1}{n_1 - 1}\left[\frac{\boldsymbol{w}'\boldsymbol{D}(\boldsymbol{D}+\boldsymbol{I})^{-2}\boldsymbol{w}}{\boldsymbol{w}'(\boldsymbol{D}+\boldsymbol{I})^{-1}\boldsymbol{w}}\right]^2 + \frac{1}{n_2 - 1}\left[\frac{\boldsymbol{w}'\boldsymbol{D}(\boldsymbol{D}+\boldsymbol{I})^{-2}\boldsymbol{w}}{\boldsymbol{w}'(\boldsymbol{D}+\boldsymbol{I})^{-1}\boldsymbol{w}}\right]^2. \quad (3.49)$$

Ejemplo 3.7 Se compararon dos tipos de suelos, uno de los cuales contiene un tipo de bacterias y el otro no. Las variables medidas fueron X_1 el pH , X_2 la cantidad de fosfato y X_3 el contenido de nitrógeno. La tabla 3.6 contiene estos datos. Se quiere verificar la hipótesis acerca de la similitud entre estos suelos, en términos de las medias asociadas con las variables medidas.

Tabla 3.6 Comparación de suelos

Con la bateria			Sin la bacteria		
X_1	X_2	X_3	X_1	X_2	X_3
8.0	60	58	6.2	49	30
8.0	156	68	5.6	31	23
8.0	90	37	5.8	42	22
6.1	44	27	5.7	42	14
7.4	207	31	6.2	40	23
7.4	120	32	6.4	49	18
8.4	65	43	5.8	31	17
8.1	237	45	6.4	31	19
8.3	57	60	5.4	62	26
7.0	94	43	5.4	42	16
8.5	86	40			
8.4	52	48			
7.9	146	52			

Fuente: Rencher (1998, pág. 103)

Los vectores de medias y las matrices de covarianzas son

$$\bar{x}_1 = \begin{pmatrix} 7.81 \\ 108.70 \\ 44.92 \end{pmatrix}, \qquad \bar{x}_2 = \begin{pmatrix} 5.89 \\ 41.90 \\ 20.80 \end{pmatrix}$$

$$\boldsymbol{S}_1 = \begin{pmatrix} 0.461 & 1.18 & 4.49 \\ 1.18 & 3776.4 & -17.35 \\ 4.49 & -17.35 & 147.24 \end{pmatrix}, \quad \boldsymbol{S}_2 = \begin{pmatrix} 0.148 & -0.679 & 0.209 \\ -0.679 & 96.10 & 20.20 \\ 0.209 & 20.20 & 24.18 \end{pmatrix}$$

$$\boldsymbol{V}_1 = \begin{pmatrix} 0.035 & 0.090 & 0.345 \\ 0.090 & 290.4 & -1.335 \\ 0.345 & -1.335 & 11.326 \end{pmatrix}, \quad \boldsymbol{V}_2 = \begin{pmatrix} 0.0148 & -0.0679 & 0.0209 \\ -0.0679 & 9.610 & 2.020 \\ 0.0209 & 2.020 & 2.418 \end{pmatrix}.$$

Asumir igualdad de matrices de covarianzas para este caso no es muy plausible, en el capítulo 4 se muestra la técnica para verificar este supuesto. El valor de la estadística T^{*2}, de acuerdo con (3.47), es

$$T^{*2} = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)' \left(\frac{\boldsymbol{S}_1}{n_1} + \frac{\boldsymbol{S}_2}{n_2} \right)^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) = 96.818.$$

Para aplicar la aproximación de Nel y Merwe se calculan los grados de libertad ν mediante (3.48), así, $\nu = 12.874 \approx 13$. Se rechaza la hipótesis de igualdad de medias, puesto que $T^{*2} = 96.818 > T^2_{(0.05,3,13)} = 12.719$ (tabla C.1). Así, los suelos difieren en la media de alguna de estas variables.

Para la matriz

$$\boldsymbol{V}_2{}^{-1}\boldsymbol{V}_1 = \begin{pmatrix} 2.18125 & 236.68005 & 9.5458894 \\ -0.001517 & 39.345995 & -1.259947 \\ 0.1250939 & -35.46755 & 5.6540877 \end{pmatrix},$$

se calculan los vectores propios, con los cuales se determina el valor de c asociado con la estadística F para la aproximación de Kim. De la expresión (3.50) se tiene $\nu = 16.97$, por la expresión (3.49), $F = 26.958$, con la cual también se rechaza la hipótesis nula, pues el p–valor es 3.08×10^{-6}; es decir, estos suelos difieren significativamente en términos de las medias para las variables pH, cantidad de fosfato y contenido de nitrógeno. ☑

▶ Potencia y tamaño de muestra

Se define como *potencia* de una estadística la probabilidad de rechazar H_0 cuando H_0 es falsa. En las pruebas consideradas hasta ahora la potencia se incrementa al aumentar cualquiera de las siguientes cantidades: (1) El valor de α, (2) el tamaño de muestra(s) y (3) la separación entre el verdadero valor del parámetro y el valor del parámetro supuesto en H_0.

La diferencia en el numeral (3) es medida por un parámetro de *no centralidad*, es un indicador de como la distribución supuesta difiere de la actual. Para la prueba T^2 el parámetro de no centralidad se obtiene desde la prueba estadística al reemplazar los estimadores muestrales por los correspondientes parámetros poblacionales. En el caso de una muestra el parámetro de no centralidad es

$$\boldsymbol{\lambda} = n(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0), \tag{3.50}$$

para dos muestras

$$\boldsymbol{\lambda} = \frac{n_1 n_2}{n_1 + n_2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{3.51}$$

En (3.52), $\boldsymbol{\Sigma}$ es la matriz de covarianzas común para las dos poblaciones.

De acuerdo con la relación mostrada entre la estadística T^2 y la estadística F, se puede encontrar la potencia para la prueba T^2. El parámetro de no centralidad para la estadística F es el mismo que el de la estadística T^2, puesto que ambas se relacionan con el parámetro de no centralidad de la ji–cuadrado. Tiku (1967) sumunistra tablas del tipo $\beta = 1 - potencia$ de la prueba F. Para usar estas tablas se calcula el parámetro de no centralidad conforme a (3.51) o a (3.52) y los respectivos grados de libertad,

$$\nu_1 = p$$
$$\nu_2 = \begin{cases} n - p, & \text{para una muestra} \\ n_1 + n_2 - p - 1, & \text{para dos muestras.} \end{cases}$$

Lo anterior conlleva a una tabla de cuatro entradas (α, $\boldsymbol{\lambda}$, ν_1 y ν_2),; esto se obvia combinando $\boldsymbol{\lambda}$ y ν_1 en la forma

$$\boldsymbol{\phi} = \sqrt{\frac{\boldsymbol{\lambda}}{\nu_1 + 1}}. \tag{3.52}$$

Con estas tablas, que contienen los valores de $\beta = 1 - potencia$, se debe restar este valor de 1 para hallar la potencia de la prueba.

La tabla asociada con el error Tipo II (β) o a la potencia de la prueba ($1 - \beta$), puede emplearse en dos sentidos: (1) para encontrar la potencia en una situación experimental particular y (2) para encontrar el *tamaño de la muestra* necesario para lograr cierta potencia en una prueba (Díaz y López, 1992). Para estimar la potencia de una prueba con un conjunto particular de datos, se puede usar valores muestrales en lugar de parámetros poblacionales en el parámetro de no centralidad $\boldsymbol{\lambda}$. Esta estimación de la potencia resulta interesante en pruebas que no rechazan la hipótesis, pues si los resultados indican baja potencia para la prueba, esto advierte que no se debe estar muy confiado sobre la cercanía entre $\boldsymbol{\mu}$ y $\boldsymbol{\mu}_0$ o entre $\boldsymbol{\mu}_1$ y $\boldsymbol{\mu}_2$.

Otro uso de estas tablas es la determinación del tamaño de la muestra requerido para lograr cierta potencia, de acuerdo con una diferencia $(\boldsymbol{\mu}_0 - \boldsymbol{\mu})$ o $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ sobre la cual el investigador esté interesado. La matriz de covarianzas $\boldsymbol{\Sigma}$ puede estimarse desde un estudio piloto o preliminar. Se emplea el mismo valor n para el caso de dos muestras. Para una selección particular de n, se calcula $\boldsymbol{\phi}$ mediante (3.53) y se lee la *potencia* $(1 - \beta)$ desde la tablas mencionadas. Este procedimiento se hace ensayando con valores de n que suministren la potencia deseada.

Algunos paquetes estadísticos proveen distribuciones tales como la F no central, los cuales reemplazan el uso de tablas. Por ejemplo, el paquete SAS contiene la función *PROBF*, de manera que

$$Potencia = 1 - PROBF(F_{(\alpha,\nu_1,\nu_2),\boldsymbol{\lambda})} = P(F > F_{(\alpha)})$$

donde $F_{(\alpha)}$ es un valor crítico de la distribución F no central.

▶ Contrastes sobre información adicional

Cuando el número de variables es grande, una inquietud para el investigador es si con un número más pequeño de variables se puede mantener la separación que se muestra entre los grupos cuando se consideran todas las variables. Se empieza con un vector X de tamaño $(p \times 1)$ que contiene las medidas sobre cada unidad muestral, el problema es: si un vector adicional Y, de tamaño $(q \times 1)$, de otras medidas sobre las mismas unidades

muestrales incrementa significativamente la separación entre los grupos. En otras palabras, la pregunta es si las q variables adicionales contribuyen en la separación de los grupos. El procedimiento puede desarrollarse observando q variables adicionales, o q variables seleccionadas entre las p variables iniciales.

Se asume que las dos muestras provienen de poblaciones multinormales con matriz de covarianzas común $\boldsymbol{\Sigma}$; es decir,

$$\begin{pmatrix} X_{11} \\ Y_{11} \end{pmatrix}, \ldots, \begin{pmatrix} X_{1n_1} \\ Y_{1n_1} \end{pmatrix} \text{ son de } N_{p+q}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \text{ y}$$
$$\begin{pmatrix} X_{21} \\ Y_{21} \end{pmatrix}, \ldots, \begin{pmatrix} X_{2n_2} \\ Y_{2n_2} \end{pmatrix} \text{ son de } N_{p+q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}).$$

El vector de medias y la matriz de covarianzas muestral son particionados de una manera conveniente en la forma:

$$\begin{pmatrix} \bar{\boldsymbol{X}}_1 \\ \bar{\boldsymbol{Y}}_1 \end{pmatrix}, \quad \begin{pmatrix} \bar{\boldsymbol{X}}_2 \\ \bar{\boldsymbol{Y}}_2 \end{pmatrix} \text{ y } \boldsymbol{S}_p = \begin{pmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{pmatrix},$$

donde $\boldsymbol{S}_p$ es la matriz de covarianzas para las dos poblaciones.

Se quiere verificar la hipótesis de que las q variables en Y_1 y en Y_2 no brindan una información adicional (extra) y significativa, respecto a la que ofrecen X_1 y X_2, en la separación de los grupos.

Si los Y son independientes de los X, se puede emplear la estadística $T^2_{(p+q)} = T^2_{(p)} + T^2_{(q)}$, en general esto no siempre se tiene, pues los dos conjuntos de variables son correlacionados. La idea es comparar la estadística $T^2_{(p+q)}$ para el conjunto completo de variables $(X_1, \ldots, X_p, Y_1, \ldots, Y_q)$ con la estadística $T^2_{(p)}$ basada en el conjunto de variables $(X_1, \ldots, X_p)$.

Por definición, la estadística T^2, sobre un conjunto de $(p+q)$ variables está dada por

$$T^2_{(p+q)} = \frac{n_1 n_2}{n_1 + n_2} \left[\begin{pmatrix} \bar{\boldsymbol{X}}_1 \\ \bar{\boldsymbol{Y}}_1 \end{pmatrix} - \begin{pmatrix} \bar{\boldsymbol{X}}_2 \\ \bar{\boldsymbol{Y}}_2 \end{pmatrix} \right]' \boldsymbol{S}_p{}^{-1} \left[\begin{pmatrix} \bar{\boldsymbol{X}}_1 \\ \bar{\boldsymbol{Y}}_1 \end{pmatrix} - \begin{pmatrix} \bar{\boldsymbol{X}}_2 \\ \bar{\boldsymbol{Y}}_2 \end{pmatrix} \right],$$

y la estadística T^2 para el conjunto reducido a las p–variables (las X) es

$$T^2_{(p)} = \frac{n_1 n_2}{n_1 + n_2} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)' \boldsymbol{S}^{-1}_{XX} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2).$$

Se rechaza la hipótesis de redundancia (no información "extra") de las Y si

$$F = \frac{(\nu - p - q + 1)}{q} \frac{T^2_{(p+q)} - T^2_{(p)}}{\nu + T^2_{(p)}} \geq F_{(\alpha, q, \nu - p - q + 1)}, \tag{3.53}$$

o alternamente, si

$$T^2 = (\nu - p)\frac{T^2_{(p+q)} - T^2_{(p)}}{\nu + T^2_{(p)}} \geq T^2_{(\alpha,q,\nu-p)}, \tag{3.54}$$

donde $\nu = (n_1 + n_2 - 2)$. Nótese que en ambos casos los primeros grados de libertad son q.

► Comparación de varias poblaciones

Se trata ahora de verificar la hipótesis (como en 3.28)

$$H_0 : \sum_{i=1}^{q} l_i \boldsymbol{\mu}_i = \boldsymbol{\mu}_0, \tag{3.55}$$

donde los l_i son constantes conocidas y $\boldsymbol{\mu}_0$ es un vector p–dimensional conocido también.

Para las q–poblaciones normales p–variantes e independientes, con igual matriz de varianzas y covarianzas pero desconocida, sea $X_{\alpha i}$ la i–ésima muestra $i = 1, \ldots, q$, con $\alpha = 1, \ldots, n_i$. El criterio para verificar la última hipótesis es:

$$T^2 = C\left(\sum_{i=1}^{q} l_i \bar{\boldsymbol{X}}_i - \boldsymbol{\mu}_0\right)' \boldsymbol{S}^{-1} \left(\sum_{i=1}^{q} l_i \bar{\boldsymbol{X}}_i - \boldsymbol{\mu}_0\right)$$

donde

$$\bar{\boldsymbol{X}}_i = \frac{1}{n_i}\sum_{\alpha=1}^{n_i} x_\alpha; \quad C = \left(\sum_{i=1}^{q} \frac{l_i^2}{n_i}\right)^{-1} \quad y$$

$$\left(\sum_{i=1}^{q} n_i - q\right) S = \sum_{i=1}^{q}\sum_{\alpha=1}^{n_i} (X_{\alpha i} - \bar{\boldsymbol{X}}_i)(X_{\alpha i} - \bar{\boldsymbol{X}}_i)', \tag{3.56}$$

la variable aleatoria T^2, se distribuye conforme a una T^2 con ν grados de libertad, donde $\nu = \sum_{i=1}^{q}(n_i - 1) = \left(\sum_{i=1}^{q} n_i - q\right)$. Como en los casos anteriores la distribución de la estadística T^2 puede aproximarse a la distribución F, hecho que facilita los cálculos para los respectivos p valores.

► Cartas de control de calidad multivariadas

Una de las herramientas más potentes en el control estadístico de calidad son las cartas de control. Las cartas de control son diseñadas para detectar

desviaciones significativas del nivel de un proceso respecto de su estándar o patrón.

Estas cartas han sido diseñadas para monitorear un proceso en el que intervienen una o varias características medidas sobre un objeto o producto. En el caso de una sola variable, que es el más desarrollado, se construye una carta de control univariada. Ésta consiste en un gráfico elaborado sobre un plano cartesiano, donde, sobre el eje vertical se ubica el valor estándar del parámetro y a su lado los valores extremos, superior e inferior, admisibles, y sobre el eje horizontal el tiempo o el espacio correspondiente a la observación o muestra seleccionada. De esta manera, resulta un gráfico con tres líneas horizontales paralelas; en los extremos las líneas de control superior e inferior (LCS y LCI) y en el centro la línea base o estándar (LC). Un proceso se dice estar *bajo control* si el valor de la estadística se ubica dentro de la franja determinada por las dos líneas de control (dentro de LCI y LCS).

Hay muchas situaciones en las cuales es necesario monitorear de manera simultánea varias características de calidad de un producto. Tales problemas son referidos como *control de calidad multivariado*.

Una técnica para monitorear un proceso con base en la media de varias variables, involucra el uso de la estadística χ^2 o de la estadística T^2. Sea X un vector aleatorio de p-medidas sobre las cuales se quiere hacer un control estadístico. Si se asume que X tiene una media objetivo $\boldsymbol{m}$ y una matriz de varianzas y covarianzas conocida $\boldsymbol{\Sigma}$, entonces

$$\chi^2_{obs} = (X - \boldsymbol{m})'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{m}),$$

bajo multinormalidad, se distribuye como ji-cuadrado con p grados de libertad. En este caso se establece como límite de control superior (o una señal de alarma), con una probabilidad de falsa alarma igual a $100\alpha\%$, al valor $\chi^2_{(alpha,p)}$, el límite inferior es el eje horizontal.

Para controlar una observación X en un momento dado, se puede emplear la estadística

$$T^2 = (X - \boldsymbol{m})'\boldsymbol{S}^{-1}(X - \boldsymbol{m}),$$

donde $\boldsymbol{m}$ es el valor objetivo o estándar. En este caso $n = 1$, que corresponde a X, la cual coincide con el valor de la media. Sin embargo, $\boldsymbol{S}$ puede calcularse mediante algunas observaciones anteriores sobre el proceso, por ejemplo k de ellas; así, la estadística T^2 anterior tiene distribución $T^2_{(p,k-1)}$. La carta de control para la media tiene como límite de control superior el valor $T^2_{(\alpha,p,k-1)}$ (no es necesario un límite inferior, pues $T^2 \geq 0$).

En la figura 3.4 se muestran varias situaciones notadas como A, B, C, y D, en las cuales se advierte sobre los problemas en que se puede incurrir cuando se hace una carta de control para cada atributo en forma separada. El caso A indica que el proceso está bajo control en forma conjunta pero fuera de control por cada variable, los casos B y D están bajo control en una de las variables pero fuera de control en la otra, y el caso C está bajo control en ambas variables separadamente pero no conjuntamente. Los casos A y C muestran la importancia de considerar la asociación entre las variables para efectos de ejercer un control estadístico sobre ellas.

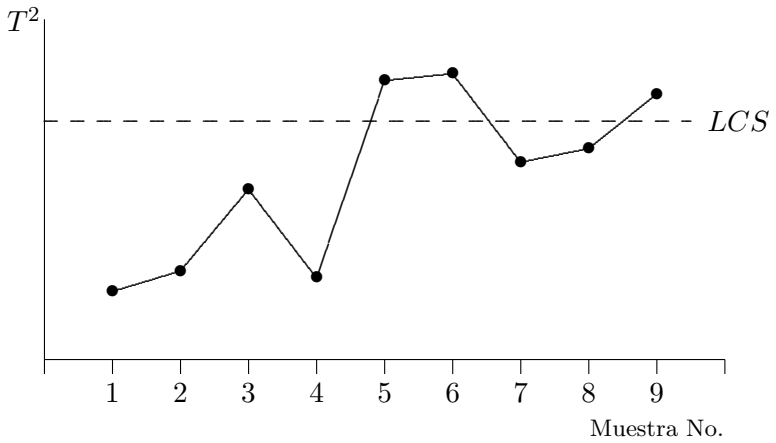


Figura 3.5 *Carta de control* T^2.

Si se usa el vector de medias $\bar{\boldsymbol{X}}$ de una muestra de tamaño n, en lugar de un vector de observaciones individuales, entonces la estadística T^2 es:

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{m})' {\boldsymbol{S}_p}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{m}),$$

la cual se distribuye como $T^2_{(p,k(n-1))}$, donde $\boldsymbol{S}_p = \sum_{i=1}^{k} \boldsymbol{S}_i/k$.

La figura 3.5 muestra una carta de control tipo T^2, donde se advierte una "señal de fuera de control" con relación a las muestras No. 5, 6, y 9.

Una vez que se ha determinado que el proceso se salió de control, el problema es identificar que característica o grupo de características provocan esta situación; Mason, Tracy y Young (1995) ofrecen una estrategia para la identificación de las variables o atributos que ponen fuera de control un proceso determinado. Para esto emplean la estadística que mide la contribución de cada variable en la estadística T^2 (ecuaciones 3.54 o 3.55).

▶ Medidas Repetidas

Muchas situaciones experimentales son conducidas de manera que a una misma unidad experimental se le aplican sucesivamente varios tratamientos; de donde resultan valores repetidos de una respuesta sobre la misma unidad u objeto. Los tratamientos pueden ser dietas, dosis de un fármaco, diferentes estímulos, entre otros. Por ejemplo:

- A un animal se le aplican varios medicamentos en diferentes ocasiones o tiempos, luego se le registra su tiempo de pastoreo.
- En pacientes, la tensión arterial sistólica es medida en intervalos de tiempo fijos, como respuesta a un fármaco desde la administración del mismo hasta que aquélla se estabilice.
- Pruebas sobre lectura son administradas a estudiantes en diferentes estadios de su educación, se registran los respectivos puntajes.
- Medidas tales como la alzada y el peso es registrado sobre un tipo de bovino en diferentes edades.
- Medidas sobre la composición del suelo se toman a diferentes profundidades, sobre un terreno experimental.

La información anterior se puede disponer en una matriz $\boldsymbol{X} = (x_{ij})$, donde x_{ij} representa la respuesta a la j–ésima medición (tratamiento) sobre la i-ésima unidad. Las observaciones por fila de esta matriz pueden estar correlacionadas por corresponder a mediciones hechas sobre un mismo sujeto. Si los tratamientos son tales que el orden (temporal o espacial) de aplicación sobre los sujetos puede variarse, entonces la asignación debe aleatorizarse para evitar problemas de sesgo.

Usualmente los individuos pertenecen a grupos distintos o reciben tratamientos diferentes, de manera que uno de los propósitos es estimar o determinar el efecto de los tratamientos sobre las respuestas.

Si los sujetos son medidos en puntos sucesivos en el tiempo, resulta necesario buscar el grado del polinomio que mejor se ajuste a los datos, esta técnica se conoce con el nombre de *curvas de crecimiento* y es abordada en la sección (3.5).

Asumiendo que cada fila, de la matriz anterior, es independientemente distribuida respecto a los otras de acuerdo con una normal p–variante con vector de medias

$$\boldsymbol{\mu}' = (\mu_1, \ldots, \mu_p),$$

y matriz de covarianzas $\mathbf{\Sigma}$, se verifica la hipótesis de igualdad de efectos debido a los p–tratamientos; es decir,

$H_0:\ \mu_1 = \cdots = \mu_p$ frente a: $H_1:\ \mu_i \neq \mu_j$ para algún par $i \neq j = 1, \ldots, p$.

Una expresión equivalente a la hipótesis anterior es:

$$H_0: \begin{pmatrix} \mu_1 - \mu_2 \\ \vdots \\ \mu_{p-1} - \mu_p \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{frente a } H_1: \begin{pmatrix} \mu_1 - \mu_2 \\ \vdots \\ \mu_{p-1} - \mu_p \end{pmatrix} \neq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

En escritura matricial, $H_0 = C\boldsymbol{\mu}$, donde C es la matriz de tamaño $((p-1) \times p)$,

$$C = \begin{pmatrix} 1 & -1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & -1 \end{pmatrix}.$$

Esto sugiere también que se debe hacer una transformación a los datos del tipo $Y = CX$.

La estadística T^2 computada sobre la transformación Y viene dada por

$$T^2 = n(C\bar{\boldsymbol{X}})'(CSC')^{-1}(C\bar{\boldsymbol{X}})$$

la cual se distribuye como $T^2_{(p-1,n-1)}$. Nótese que la dimesión es $(p-1)$ porque $C\bar{\boldsymbol{X}}$ es de tamaño $((p-1) \times 1)$.

Observaciones:

- La matriz C no es única, y se llama *matriz de contrastes*, porque sus $(p-1)$ filas son linealmente independientes y cada una es un contraste $(\sum_{j=1}^{p} c_{ij} = 0, \text{ para } i = 1, \ldots, p-1)$.
- Se rechaza H_0 si

$$T^2 = n(C\bar{\boldsymbol{X}})'(CSC')^{-1}C\bar{\boldsymbol{X}} > \frac{(n-1)(p-1)}{n-p+1} F_{(\alpha,p-1,n-p+1)},$$

  donde $F_{(\alpha,p-1,n-p+1)}$ es el percentil $(1-\alpha)\%$ de la distribución F con $(p-1)$ y $(n-p+1)$ grados de libertad (tabla C.8).

Tabla 3.7 Ritmo cardíaco en perros

Perro	Tratamiento			
	T_1	T_2	T_3	T_4
1	426	609	556	600
2	253	236	392	395
3	359	433	349	357
4	432	431	522	600
5	405	426	513	513
6	324	438	507	539
7	310	312	410	456
8	326	326	350	504
9	375	447	547	548
10	286	286	403	422
11	349	382	473	497
12	429	410	488	547
13	348	377	447	514
14	412	473	472	446
15	347	326	455	468
16	434	458	637	524
17	364	367	432	469
18	420	395	508	531
19	397	556	645	625

Fuente: Johnson y Wichern (1998, pág. 300)

Ejemplo 3.8 Se probó un anestésico en perros con el fin de observar el tiempo entre cada latido cardíaco (medido en milisegundos). A cada uno de estos 19 animales se le suministró cuatro tipos de anestésicos diferentes (tratamientos). Se quiere analizar el efecto de los anestésicos sobre el ritmo cardíaco. Como cada animal recibió sucesiva y adecuadamente cada una de las sustancias, éste se puede considerar como un caso de medidas repetidas; el experimento fue conducido de tal forma que entre cada tratamiento se deja un espacio de tiempo adecuado para eliminar los posibles efectos residuales, los cuales afectarían los resultados de los tratamientos.

Los tratamientos se notarán por T_i y cada uno corresponde a la siguiente preparación:

T_1**:** CO_2 a presión alta sin halotano.

T_2**:** CO_2 a presión baja sin halotano.

T_3**:** CO_2 a presión alta con halotano.

T_4**:** CO_2 a presión baja con halotano.

Las hipótesis que se desean contrastar, simultáneamente, son las siguientes:

1: "Efecto de la presencia de halotano".

2: "Efecto de la presión".

3: "Influencia del halotano sobre las diferencias de presión".

Las hipótesis anteriores se pueden escribir en la forma:

$$H_0: \begin{pmatrix} \mu_3 + \mu_4 \\ \mu_1 + \mu_3 \\ \mu_1 + \mu_4 \end{pmatrix} = \begin{pmatrix} \mu_1 + \mu_2 \\ \mu_2 + \mu_4 \\ \mu_2 + \mu_3 \end{pmatrix},$$

en términos de una matriz de contrastes C, la hipótesis anterior se escribe como:

$$\begin{pmatrix} -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

con los datos de la tabla 3.7 y de la matriz C se calculan las siguientes estadísticas

$$\bar{\boldsymbol{X}} = \begin{pmatrix} 368.21 \\ 404.63 \\ 479.26 \\ 502.89 \end{pmatrix} \text{ y } \boldsymbol{S} = \begin{pmatrix} 2819.29 & 3568.42 & 2943.49 & 2295.35 \\ 3568.42 & 7963.14 & 5303.98 & 4065.44 \\ 2943.49 & 5303.98 & 6851.32 & 4499.63 \\ 2295.35 & 4065.44 & 4499.63 & 4878.99 \end{pmatrix}$$

también

$$C\bar{\boldsymbol{X}} = \begin{pmatrix} 209.31 \\ -60.05 \\ -12.79 \end{pmatrix}; \quad C\boldsymbol{S}C' = \begin{pmatrix} 9432.32 & 1098.92 & 927.62 \\ 1098.92 & 5195.84 & 914.54 \\ 927.62 & 914.54 & 7557.44 \end{pmatrix},$$

de donde

$$T^2 = n(C\bar{\boldsymbol{X}})'(C\boldsymbol{S}C')^{-1}(C\bar{\boldsymbol{X}}) = 116.$$

Para un nivel de significación $\alpha = 0.05$, $F_{(0.05,3,16)} = 3.24$ (tabla C.8);

$$\frac{(n-1)(p-1)}{n-p+1} F_{(p-1,n-p+1)}(\alpha) = \frac{18(3)}{16}(3.24) = 10.94;$$

en conclusión, como $T^2 = 116 > 10.94$ se rechaza la hipótesis $H_0: C\boldsymbol{\mu} = 0$. Así, se puede afirmar, desde los datos disponibles, que existe un efecto sobre

el ritmo cardíaco de acuerdo con los niveles de presión, alto o bajo, con CO_2 y la presencia o no del halotano como anestésico. ☑

▶ Análisis de perfiles

Si $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ y las variables de X están en las mismas unidades de medición (conmensurables) con varianza aproximadamente igual, se pueden comparar la medias $\mu_1, \ldots, \mu_p$ que conforman a $\boldsymbol{\mu}$. Este caso puede ser de interés, como el citado anteriormente, para diseños de *medidas repetidas* o para *curvas de crecimiento*.

A manera de ilustración, considérese el caso en el que se quiere observar el efecto de dos fármacos A y B sobre el la tensión arterial sistólica (TAS) en un grupo de pacientes. Al cabo de dos minutos de aplicado el fármaco o el placebo se observó en intervalos de cinco minutos la TAS para los pacientes de cada grupo. La atención se dirige a dar cuenta sobre el tipo de perfil (en términos del tiempo y del fármaco) que se genera con los datos disponibles. Éste es uno de las problemas de los cuales se ocupa el análisis de los perfiles; aclarando que se hace referencia tan sólo a una variable respuesta, lo cual no sinifica la imposibilidad de abordar el problema para más de una variable respuesta.

Se presenta el análisis de perfiles para una y dos poblaciones. El caso de varias poblaciones se trata en la sección (3.5).

El patrón geométrico que se obtiene al ubicar $\mu_1, \mu_2, \ldots, \mu_p$ en las ordenadas y conectarlas en este orden mediante líneas, se llama *perfil*; éste se conforma por la línea poligonal que une los puntos $(1, \mu_1), (2, \mu_2), \ldots, (p, \mu_p)$.

El análisis de perfiles se desarrolla para una, dos o varias muestras. Este análisis contempla tanto la construcción, la indagación acerca de la forma o topología de un perfil, como la comparación entre los perfiles ligados a cada una de varias poblaciones multivariadas.

• *Análisis de perfiles en una muestra*

Se considera un vector de medias $\boldsymbol{\mu}$ de una población. Un diagrama de perfiles sobre $\boldsymbol{\mu}$ se muestra en la figura 3.6, allí se ubican y conectan los puntos $(1, \mu_1), (2, \mu_2), \ldots, (p, \mu_p)$.

Para comparar las medias $\mu_1, \mu_2, \ldots, \mu_p$ de $\boldsymbol{\mu}$, la hipótesis básica es que el perfil está en posición *horizontal*:

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_p \text{ frente a } H_1: \ \mu_i \neq \mu_j, \text{ para } i \neq j.$$

La igualdad de las p-medias equivale a expresar la nulidad de las $(p-1)$-diferencias siguientes:

$$H_0 : \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \\ \vdots \\ \mu_{p-1} - \mu_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{3.58a}$$

o también, equivalente a:

$$H_0 : \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{3.58b}$$

Las dos expresiones anteriores pueden escribirse en la forma: $\boldsymbol{C}_1\boldsymbol{\mu} = 0$ o $\boldsymbol{C}_2\boldsymbol{\mu} = 0$, donde las matrices $\boldsymbol{C}_1$ y $\boldsymbol{C}_2$ son de tamaño $(p-1) \times p$:

$$\boldsymbol{C}_1 = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}, \quad \boldsymbol{C}_2 = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

Cualquier matriz $\boldsymbol{C}$ de tamaño $(p-1)\times p$ y de rango $(p-1)$ tal que $\boldsymbol{C}\mathbf{1} = 0$, puede emplearse para verificar la hipótesis anterior, donde $\mathbf{1}$ es un vector de $1's$. Si $\boldsymbol{C}\mathbf{1} = 0$, los elementos de cada fila de $\boldsymbol{C}$ suman *cero*, entonces $\boldsymbol{C}\boldsymbol{\mu}$ es un conjunto de $(p-1)$ *contrastes* en los $\mu's$.

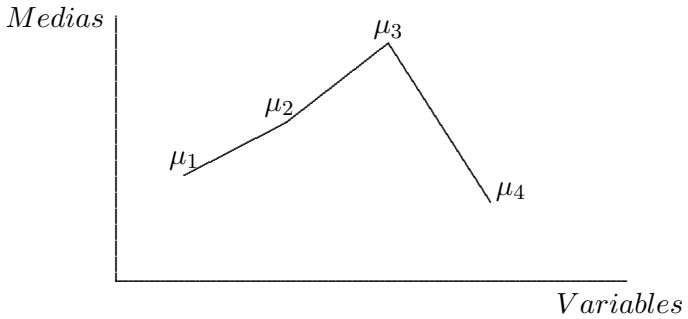


Figura 3.6 *Perfil de medias,* $p = 4$.

A partir de una muestra aleatoria $X_1, X_2, \ldots, X_n$ se obtienen los estimadores $\bar{\boldsymbol{X}}$ y $\boldsymbol{S}$ de los parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$. Tal como se muestra en la ecuación (3.41), la hipótesis de que las p-medias son iguales se verifica a través de

$$T^2 = n(\boldsymbol{C}\bar{\boldsymbol{X}})'(\boldsymbol{CSC}')^{-1}(\boldsymbol{C}\bar{\boldsymbol{X}}).$$

Se rechaza $H_0: \ \boldsymbol{C\mu} = 0$, si $T^2 > T^2_{(\alpha,p-1,n-1)}$.

Si las variables tienen un orden natural se puede probar una tendencia lineal o polinómica en las medias con base en una selección adecuada de las filas de $\boldsymbol{C}$.

• *Análisis de perfiles en dos muestras*

Supóngase que dos grupos (muestras) indepedientes reciben los mismos p tratamientos. En lugar de probar la hipótesis $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, se quieren comparar los perfiles obtenidos al conectar los puntos (i, μ_{1i}), $i = 1, \ldots, p$, y (i, μ_{2i}), $i = 1, \ldots, p$, respectivamente. Hay tres hipótesis de interés en la comparación de los perfiles ligados a dos muestras; éstas son: perfiles paralelos, perfiles en el mismo nivel (coincidentes) y los perfiles planos.

◦ La primera es "*¿Son los dos perfiles similares, o más precisamente, son paralelos?*". Si son paralelos, pero no coincidentes, entonces un grupo es uniformemente mejor que el otro en términos de medias. Las figuras 3.7a y 3.7b ilustran el caso para el cual H_{01} es verdadera y el caso para el cual es falsa, respectivamente.

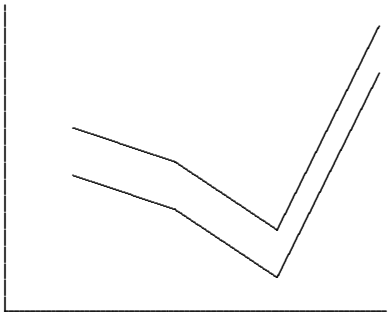


Figura 3.6a *Hipótesis H_{01} verdadera.*

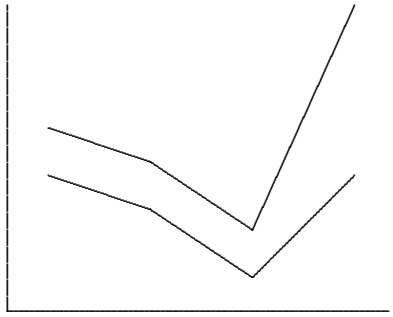


Figura 3.6b *Hipótesis H_{01} falsa.*

El paralelismo puede ser definido en términos de las pendientes. Dos perfiles son paralelos si las pendientes de los segmentos correspondientes a cada par de abcisas son iguales; es decir, los incrementos son los mismos para los respectivos pares de medias. Esto se puede expresar a través de la hipótesis

$$H_{01}: \ \mu_{1i} - \mu_{1,i-1} = \mu_{2i} - \mu_{2,i-1}, \ \text{para } i = 2, 3, \ldots, p,$$

o equivalentemente

$$H_{01}: \begin{pmatrix} \mu_{12} - \mu_{11} \\ \mu_{13} - \mu_{12} \\ \vdots \\ \mu_{1p} - \mu_{1,p-1} \end{pmatrix} = \begin{pmatrix} \mu_{22} - \mu_{21} \\ \mu_{23} - \mu_{22} \\ \vdots \\ \mu_{2p} - \mu_{2,p-1} \end{pmatrix}.$$

La cual puede escribirse como $H_{01}: \ \boldsymbol{C}\boldsymbol{\mu}_1 = \boldsymbol{C}\boldsymbol{\mu}_2$, donde la matriz de contrastes es

$$\boldsymbol{C} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

Mediante las dos muestras $X_{11}, X_{12}, \ldots, X_{1n_1}$ y $X_{21}, X_{22}, \ldots, X_{2n_2}$, se obtienen los vectores de medias $\bar{\boldsymbol{X}}_1$, $\bar{\boldsymbol{X}}_2$ y la matriz de covarianzas pareada $\boldsymbol{S}_p$; los cuales son estimadores de $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ y $\boldsymbol{\Sigma}$, respectivamente. Como en el caso de dos poblaciones, se emplea la estadística T^2 asumiendo que cada X_{1i} en la primera muestra es $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, y que cada X_{2i} en la segunda muestra es $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. La estadística T^2 toma la forma

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\boldsymbol{C}\bar{\boldsymbol{X}}_1 - \boldsymbol{C}\bar{\boldsymbol{X}}_2)'\Big[\boldsymbol{C}\boldsymbol{S}_p\boldsymbol{C}'\Big]^{-1}(\boldsymbol{C}\bar{\boldsymbol{X}}_1 - \boldsymbol{C}\bar{\boldsymbol{X}}_2)$$

la cual se distribuye como $T^2_{(p-1, n_1+n_2-2)}$.

Si se rechaza la hipótesis H_{01}, las pruebas univariadas sobre las componentes de $\boldsymbol{C}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ indican cuales variables son las posibles responsables de tal rechazo.

- La segunda hipótesis de interés es: "*¿ Están las dos poblaciones o grupos en el mismo nivel?*". Se puede expresar esta hipótesis en términos del nivel promedio del grupo 1 comparada con el nivel promedio del grupo 2:

  $$H_{02}: \ \frac{\mu_{11} + \mu_{12} + \cdots + \mu_{1p}}{p} = \frac{\mu_{21} + \mu_{22} + \cdots + \mu_{2p}}{p},$$

  o equivalentemente: $H_{02}: \ \mathbf{1}'\boldsymbol{\mu}_1 = \mathbf{1}'\boldsymbol{\mu}_2$. Si H_{02} es cierta se puede asociar con el gráfico 3.7a, de lo contrario con el 3.7b.

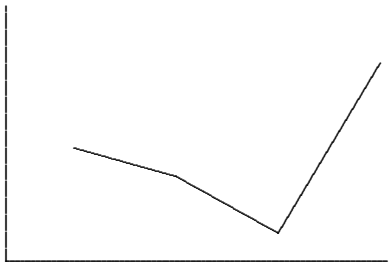


Figura 3.7a *Hipótesis H_{02} verdadera.*

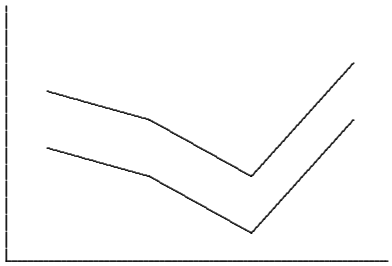


Figura 3.7b *Hipótesis H_{02} falsa.*

La hipótesis H_{02} puede ser verdadera sin que H_{01} lo sea; es decir, los niveles promedio pueden ser iguales y los perfiles ser no paralelos, como se muestra en la figura 3.8. En este caso el "grupo de efectos principales" es algo más complejo de interpretar, como ocurre en el *análisis de varianza para diseños de doble vía de clasificación*, donde los efectos principales son más difíciles de describir cuando la interacción está presente significativamente.

Para verificar la hipótesis $H_{02} : \mathbf{1}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$, se emplea la estadística $\mathbf{1}'(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$ como estimador de $\mathbf{1}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, el cual tiene distribución univariada

$n(0, \mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}[1/n_1 + 1/n_2])$, bajo H_{02}.

Se utiliza la estadística

$$t = \frac{\mathbf{1}'(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)}{\sqrt{\mathbf{1}'\boldsymbol{S}_p\mathbf{1}\Big(1/n_1 + 1/n_2\Big)}},$$

se rechaza H_{02} si $|t| > t_{(\alpha/2, n_1+n_2-2)}$.

- La tercera hipótesis de interés, se relaciona con la pregunta "*¿Son los perfiles planos*?". Asumiendo paralelismo horizontal (H_{01} es cierta), se puede dibujar esta hipótesis para los dos casos, verdadera y falsa. La figura 3.9a y 3.9b muestra esta situación.

  La tercera hipótesis se puede escribir en la forma:

$$H_{03} : \ \frac{1}{2}(\mu_{11} + \mu_{21}) = \frac{1}{2}(\mu_{12} + \mu_{22}) = \cdots = \frac{1}{2}(\mu_{1p} + \mu_{2p}),$$

  o también

$$H_{03} : \ \boldsymbol{C}\left(\frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right) = 0,$$

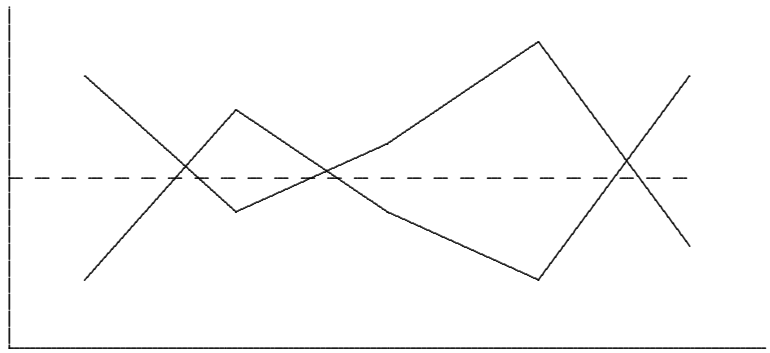


Figura 3.8 *Hipótesis H_{02}: "igual efecto sin paralelismo".*

donde $\boldsymbol{C}$ es una matriz de tamaño $((p-1)\times p)$ tal que $\boldsymbol{C}\mathbf{1}=0$. La figura 3.9a sugiere que H_{03} puede expresarse como $\mu_{11}=\cdots=\mu_{1p}$ y $\mu_{21}=\cdots=\mu_{2p}$, o también en la forma

$$H_{03}:\ \boldsymbol{C}\boldsymbol{\mu}_1=0 \text{ y } \boldsymbol{C}\boldsymbol{\mu}_2=0.$$

Para estimar $\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$, se emplea la media muestral general ponderada; es decir,

$$\bar{\boldsymbol{X}}=\frac{n_1\bar{\boldsymbol{X}}_1+n_2\bar{\boldsymbol{X}}_2}{n_1+n_2}.$$

Se demuestra que $\boldsymbol{C}\bar{\boldsymbol{X}}$, bajo las hipótesis H_{03} y H_{01}, tiene distribución $N_{p-1}(\mathbf{0},\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}'/(n_1+n_2))$. La estadística pertinente para contrastar la hipótesis nula H_{03} es

$$T^2=(n_1+n_2)(\boldsymbol{C}\bar{\boldsymbol{X}})'(\boldsymbol{C}\boldsymbol{S}_p\boldsymbol{C}')^{-1}(\boldsymbol{C}\bar{\boldsymbol{X}})$$

y se distribuye como $T^2_{(p-1,n_1+n_2-2)}$.

3.5 Análisis de varianza multivariado

Hasta aquí se ha considerado la verificación de hipótesis respecto al vector de medias de una o dos poblaciones. En esta sección se extiende la comparación de poblaciones, a través de los respectivos vectores de medias al caso de dos o más poblaciones. Por ejemplo:

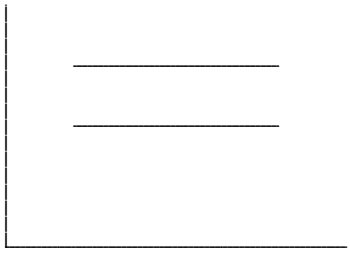


Figura 3.9a *Hipótesis H_{03} verdadera.*

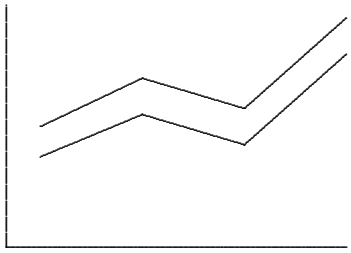


Figura 3.9b *Hipótesis H_{03} falsa.*

1. Comparar el efecto de cuatro tratamientos sobre la respuesta media de algunas variables fisiológicas en animales específicos
2. Indagar acerca de la efectividad de tres metodologías de enseñanza, en términos de logros cognoscitivos obtenidos por un grupo de estudiantes.
3. Determinar el efecto de tres fertilizantes (efectos fijos) y de la variedad (efecto aleatorio) sobre la calidad de un fruto, en términos de algunas variables observadas sobre éstos.

Se considera ahora, para este tipo de problemas, el análisis de varianza multivariado (llámese *ANAVAMU*), con lo cual se busca verificar la igualdad de vectores de medias ligados a varias poblaciones. La técnica es un caso especial de la *hipótesis lineal general multivariada*. Dada la similitud con el modelo de regresión múltiple, se desarrollan algunos aspectos teóricos en modelos de regresión para luego ser tomados en el modelo lineal general multivariada. La teoría de los mínimos cuadrados, empleada en la generalización, esencialmente es la misma del caso univariado.

3.5.1 Modelo lineal general multivariado

La distinción entre los modelos lineales multivariados y los modelos univariados es, como su nombre lo señala, que el modelo multivariado involucra más de una variable dependiente o respuesta.

Considérese que las observaciones multivariadas $Y_1, \ldots, Y_n$, conforman un conjunto de observaciones independientes de una población normal p-variante; es decir, $Y_\alpha \sim N_p(X_\alpha \boldsymbol{\beta}, \boldsymbol{\Sigma})$, para $\alpha = 1, \ldots, n$. Los vectores X_α de tamaño $(1 \times q)$ son conocidos. Tanto la matriz $\boldsymbol{\Sigma}_{p\times p}$, como la matriz $\boldsymbol{\beta}_{q\times p}$ son desconocidas.

- Los Y_α corresponden a las variables respuesta en un modelo de regresión (dependientes), mientras que las X_α son las variables regresoras o explicativas. En tales condiciones los vectores se pueden relacionar a través de un *modelo lineal general multivariado*, tal como el siguiente:

$$\begin{pmatrix} y_{11} & \cdots & y_{1p} \\ y_{21} & \cdots & y_{2p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1q} \\ x_{21} & \cdots & x_{2q} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nq} \end{pmatrix} \begin{pmatrix} \beta_{11} & \cdots & \beta_{1p} \\ \beta_{21} & \cdots & \beta_{2p} \\ \vdots & \ddots & \vdots \\ \beta_{q1} & \cdots & \beta_{qp} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1p} \\ \varepsilon_{21} & \cdots & \varepsilon_{2p} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{np} \end{pmatrix}. \tag{3.57}$$

En forma condensada, el modelo lineal multivariado anterior se escribe de la manera siguiente:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \tag{3.59a}$$

La matriz $\boldsymbol{X}$ conforma, en la mayoría de los casos, la matriz de diseño o la matriz de variables regresoras. $\boldsymbol{\beta}$ es la matriz de parámetros desconocidos y la matriz aleatoria $\boldsymbol{\mathcal{E}}$ contiene los errores.

Para los propósitos de este texto, se propone, estima e infiere sobre los modelos ligados una estructura de una y dos vías de clasificación, mediante una conformación adecuada de la matriz de diseño $\boldsymbol{X}$ y de la matriz de parámetros $\boldsymbol{\beta}$. Además, se extiende el análisis de perfiles, de medidas repetidas y de curvas de crecimiento, para el caso de varias poblaciones multivariadas.

Tal como en el modelo lineal clásico ($q = 1$), los estimadores de máxima verosimilitud para $\boldsymbol{\beta}$ y $\boldsymbol{\Sigma}$ son:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \Big(\sum_{\alpha=1}^{n} X_\alpha' X_\alpha\Big)^{-1}\Big(\sum_{\alpha=1}^{n} X_\alpha' Y_\alpha\Big) \\ \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{\alpha=1}^{n}(Y_\alpha - X_\alpha\widehat{\boldsymbol{\beta}})(Y_\alpha - X_\alpha\widehat{\boldsymbol{\beta}})'. \end{aligned} \tag{3.58}$$

Observaciones:

- Se puede deducir con estos estimadores los correspondientes a la regresión lineal múltiple, donde $q = 1$. El estimador máximo verosímil $\widehat{\boldsymbol{\beta}}$, dado en (3.59) tiene distribución normal con vector de medias $\boldsymbol{\beta}$ y matriz de varianzas y covarianzas la resultante del producto directo

o Kronecker (ecuación A2.43) entre $\mathbf{\Sigma}$ y A^{-1}; es decir,

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \mathbf{\Sigma} \otimes A^{-1} = \begin{pmatrix} \sigma_{11}A^{-1} & \sigma_{12}A^{-1} & \dots & \sigma_{1p}A^{-1} \\ \sigma_{21}A^{-1} & \sigma_{22}A^{-1} & \dots & \sigma_{2p}A^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1}A^{-1} & \sigma_{p2}A^{-1} & \dots & \sigma_{pp}A^{-1} \end{pmatrix}, \quad (3.59)$$

donde

$$A = \sum_{\alpha=1}^{n} X'_\alpha X_\alpha.$$

- Se nota la similitud con el modelo de regresión lineal, donde se asume que los errores tienen matriz de covarianzas $\mathbf{\Sigma} = \sigma^2 \boldsymbol{I}$, así que $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1}$, es un caso especial de la última expresión.
- De manera similar, el estimador máximo verosímil $n\widehat{\mathbf{\Sigma}}$ es distribuido como $\mathcal{W}(\mathbf{\Sigma}, n-q)$, e independiente de $\widehat{\boldsymbol{\beta}}$, con q el número de componentes de X_α.

Para obtener un estimador insesgado de $\mathbf{\Sigma}$ se debe hacer $\boldsymbol{S} = \Big(n/(n-q)\Big)\widehat{\mathbf{\Sigma}}$.

3.5.2 Contraste de hipótesis

Supóngase que se particiona la matriz de parámetros $\boldsymbol{\beta}$ como:

$$\boldsymbol{\beta} = (\boldsymbol{\beta_1} : \boldsymbol{\beta_2}), \quad (3.60)$$

con $\boldsymbol{\beta_1}$ de q_1 columnas y $\boldsymbol{\beta_2}$ de q_2 columnas ($q_1 + q_2 = q$). La razón de máxima verosimilitud para probar la hipótesis

$$H_0: \quad \boldsymbol{\beta_1} = \boldsymbol{\beta_1}^*, \quad (3.61)$$

se obtiene en forma semejante a como se procedió con la estadística T^2; ésta es,

$$\boldsymbol{\lambda} = \frac{|\widehat{\mathbf{\Sigma}}_\Omega|^{n/2}}{|\widehat{\mathbf{\Sigma}}_{\Omega_0}|^{n/2}}. \quad (3.62)$$

La matriz $\widehat{\mathbf{\Sigma}}_\Omega$ corresponde al estimador máximo verosímil en el espacio global de parámetros. La matriz $\widehat{\mathbf{\Sigma}}_\omega$ es el estimador de máxima verosimilitud en el espacio de parámetros restringido por la hipótesis nula (3.62); con:

$$\mathbf{\Sigma}_\Omega = \frac{1}{n}\left(\sum_{\alpha=1}^{n}(Y_\alpha - X_\alpha\boldsymbol{\beta}_1^*)(Y_\alpha - X_\alpha\boldsymbol{\beta}_1^*)'\right) - \boldsymbol{\beta}_{2\omega}A_{22}\boldsymbol{\beta}'_{2\omega}, \quad (3.63b)$$

$\boldsymbol{\beta}_2$ y $\boldsymbol{A}_{22}$ se obtienen mediante una partición apropiada de $\boldsymbol{\beta}$ y $\boldsymbol{A}$, respectivamente.

Se rechaza la hipótesis H_0, si $\boldsymbol{\lambda} < \boldsymbol{\lambda}_0$, para $\boldsymbol{\lambda}_0$ un número escogido adecuadamente de acuerdo con la distribución de $\boldsymbol{\lambda}$ y el nivel de significancia α.

Un caso especial de (3.63) es la estadística T^2 de Hotelling; la cual se obtiene al hacer $q = q_1 = 1$, $q_2 = 0$, $X_\alpha = 1$ para $\alpha = 1, \ldots, n$ y $\boldsymbol{\beta} = \boldsymbol{\beta}_1 = \boldsymbol{\mu}$.

Bajo la hipótesis nula, la razón de máxima verosimilitud (3.63) puede transformarse en

$$\Lambda = \boldsymbol{\lambda}^{2/n} = \frac{|\widehat{\boldsymbol{\Sigma}}_\Omega|}{|\widehat{\boldsymbol{\Sigma}}_{\Omega_0}|} = \frac{|n\widehat{\boldsymbol{\Sigma}}_\Omega|}{\left|n\widehat{\boldsymbol{\Sigma}}_\Omega + \left(\widehat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right) A_{11.2} \left(\widehat{\boldsymbol{\beta}}_{1\Omega} - \boldsymbol{\beta}_1^*\right)\right|}, \tag{3.63}$$

donde $A_{11.2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$.

La variable Λ es el cociente de dos varianzas generalizadas, las cuales están ligadas a la distribución $\mathcal{W}$ de Wishart; esto es

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}|} \tag{3.64}$$

donde $\boldsymbol{E} = n\widehat{\boldsymbol{\Sigma}}$, se distribuye de acuerdo con una $\mathcal{W}(\boldsymbol{\Sigma}, n - q)$ y $\boldsymbol{E} + \boldsymbol{H} = n\boldsymbol{\Sigma}_{\Omega_0}$, con $\boldsymbol{H}$ distribuida $\mathcal{W}(\boldsymbol{\Sigma}, q_1)$. La estadística Λ se conoce con el nombre de *lambda de Wilks*, es el equivalente a la estadística F para contrastar la igualdad de las medias asociadas a varias poblaciones independientes con distribución normal univariada. La tabla 3.8 muestra la distribución exacta de Λ para algunos casos especiales respecto al número de variables p y al número de poblaciones q. Más adelante se presenta la distribución asintótica (para tamaños de muestra grandes) de esta estadística .

Las matrices $\boldsymbol{E}$ y $\boldsymbol{H}$ contienen las sumas de cuadrados, en términos vectoriales, *dentro* y *entre* grupos respectivamente, las cuales se escriben para los modelos de una y de dos vías de clasificación.

3.5.3 Análisis de varianza multivariado

Desde un punto de vista práctico, el análisis de varianza multivariado es una técnica con la cual se puede verificar la igualdad de los vectores de medias ligados a varias poblaciones multivariadas.

Muchas hipótesis en el campo multivariado pueden expresarse como las hipótesis concernientes al análisis de regresión esquematizado anteriormente.

Dentro de este estilo, se presenta la técnica del análisis de varianza para arreglos de una y dos vías de clasificación.

▶ Modelos de una vía de clasificación

Cosidérese que Y_{ij} es una observación de una población $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ con $i = 1, \ldots, n_i$, y $j = 1, \ldots, q$. Los datos se pueden visualizar de la siguiente forma

Población	*Muestra*	*Media muestral*
Pob. 1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$	$\bar{\boldsymbol{Y}}_{1.}$
Pob. 2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$	$\bar{\boldsymbol{Y}}_{2.}$
$\vdots$	$\vdots$	$\vdots$
Pob. q	$Y_{q1}, Y_{q2}, \ldots, Y_{qn_q}$	$\bar{\boldsymbol{Y}}_{q.}$

Observación:

> Nótese que se han considerado n_i observaciones en cada población, éste es el caso más general. Si los n_i son diferentes se dice que se trata de un diseño experimental *desbalanceado*; cuando $n_1 = \cdots = n_q = n$ se dice que el diseño es *balanceado*.

La media $\bar{\boldsymbol{Y}}_{i.}$ en cada muestra se obtiene mediante

$$\bar{\boldsymbol{Y}}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_i} \boldsymbol{Y}_{i.}, \quad \text{para } i = 1, \ldots, q.$$

La media general $\bar{\boldsymbol{Y}}_{..}$ se obtiene de

$$\bar{\boldsymbol{Y}}_{..} = \frac{1}{N} \sum_{i=1}^{q} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{N} \sum_{i=1}^{q} \bar{\boldsymbol{Y}}_{i.}$$

con $N = \sum_{i=1}^{q} n_i$, el número total de observaciones.

El modelo que relaciona las observaciones con los parámetros $\boldsymbol{\mu}_i$ es de la forma

$$Y_{ij} = \boldsymbol{\mu}_i + \mathcal{E}_{ij}, \text{ con } \mathcal{E}_{ij} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), \text{ para } i = 1, \ldots, q \text{ y } j = 1, \ldots, n_i.$$

El modelo anterior, escrito en forma matricial, es:

$$\begin{pmatrix} Y'_{11} \\ Y'_{12} \\ \vdots \\ Y'_{1n_1} \\ -- \\ \vdots \\ -- \\ Y'_{q1} \\ Y'_{q2} \\ \vdots \\ Y'_{qn_q} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{n_q} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}'_1 \\ \boldsymbol{\mu}'_2 \\ \vdots \\ \boldsymbol{\mu}'_q \end{pmatrix} + \begin{pmatrix} \varepsilon'_{11} \\ \varepsilon'_{12} \\ \vdots \\ \varepsilon'_{1n_1} \\ -- \\ \vdots \\ -- \\ \varepsilon'_{q1} \\ \varepsilon'_{q2} \\ \vdots \\ \varepsilon'_{qn_q} \end{pmatrix}$$

$$\boldsymbol{Y} = \bigoplus_{i=1}^{q} \mathbf{1}_{n_i} \boldsymbol{\mu}_i + \boldsymbol{\mathcal{E}}.$$

La hipótesis a verificar es la igualdad de los vectores de medias de las q-poblaciones; es decir,

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_q. \tag{3.65}$$

Una expresión equivalente con (3.62) es

$$\begin{aligned} \boldsymbol{\beta}_1 &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_q, \ldots, \boldsymbol{\mu}_{q-1} - \boldsymbol{\mu}_q) \\ \boldsymbol{\beta}_2 &= \boldsymbol{\mu}_q. \end{aligned} \tag{3.66}$$

La hipótesis planteada en (3.66) se puede escribir en la forma

$$\begin{aligned} H_0 : \; & \boldsymbol{\mu}_1 - \boldsymbol{\mu}_q = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_q = \cdots = \boldsymbol{\mu}_{q-1} - \boldsymbol{\mu}_q = \mathbf{0} \\ : \; & \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}'_1 \\ \boldsymbol{\mu}'_2 \\ \vdots \\ \boldsymbol{\mu}'_q \end{pmatrix} = \mathbf{0}. \end{aligned} \tag{3.67}$$

La ecuación (3.64) se utiliza para contrastar esta hipótesis. La región de rechazo a un nivel de significación α es

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}|} = \frac{|N\widehat{\boldsymbol{\Sigma}}|}{|N\widehat{\boldsymbol{\Sigma}}_\omega|} < \boldsymbol{\Lambda}_{(\alpha, p, \nu_H, \nu_E)} \tag{3.68}$$

donde $\nu_H = q - 1$ son los grados de libertad para la hipótesis, $\nu_E = N - q$ son los grados de libertad del error ($N = \sum_{i=1}^{q} n_i$).

Las matrices $\widehat{\mathbf{\Sigma}}$ y $\widehat{\mathbf{\Sigma}}_\omega$ se calculan de

$$N\widehat{\mathbf{\Sigma}} = \sum_{i,j}(Y_{ij} - \bar{\boldsymbol{Y}}_{i.})(Y_{ij} - \bar{\boldsymbol{Y}}_{i.})', \tag{3.69}$$

con

$$\bar{\boldsymbol{Y}}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

y

$$N\widehat{\mathbf{\Sigma}}_\omega = \sum_{i=1}^{q} n_i(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})' + N\widehat{\mathbf{\Sigma}}. \tag{3.70}$$

La tabla C.2 contiene los valores de la estadística $\mathbf{\Lambda}_{(\alpha,p,\nu_H,\nu_E)}$ (valores críticos inferiores), para diferentes valores de p, ν_H, ν_E y α. Se rechaza la hipótetsis nula para valores observados de $\mathbf{\Lambda}$ menores que el valor $\mathbf{\Lambda}_{(\alpha,p,\nu_H,\nu_E)}$ de la tabla C.2.

El modelo anterior permite hacer la siguiente descomposición del vector Y_{ij}

$$Y_{ij} = \bar{\boldsymbol{Y}}_{..} + (\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..}) + (Y_{ij} - \bar{\boldsymbol{Y}}_{i.}), \tag{3.72a}$$

o también

$$(Y_{ij} - \bar{\boldsymbol{Y}}_{..}) = (\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..}) + (Y_{ij} - \bar{\boldsymbol{Y}}_{i.}). \tag{3.72b}$$

La desagregación presentada en (3.72a) o en (3.72b), semejante al caso univariado, permite mostrar como la *variabilidad total* es igual a la *variabilidad entre las poblaciones* más la *variabilidad dentro* de las poblaciones. Naturalmente que estando en el caso multivariado las identidades anteriores (3.72$a-b$) no miden la variabilidad en forma apropiada, pero al multiplicar por los respectivos vectores transpuestos y sumar sobre los subíndices i y j se obtiene la siguiente identidad, semejante a la del caso univariado,

$$\begin{aligned}\sum_{i=1}^{q}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\boldsymbol{Y}}_{..})(Y_{ij} - \bar{\boldsymbol{Y}}_{..})' &= \sum_{i=1}^{q}\sum_{j=1}^{n_i}[(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..}) + (Y_{ij} - \bar{\boldsymbol{Y}}_{i.})][(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..}) + (Y_{ij} - \bar{\boldsymbol{Y}}_{i.})]' \\ &= \sum_{i=1}^{q} n_i(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})' + \sum_{i=1}^{q}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\boldsymbol{Y}}_{i.})(Y_{ij} - \bar{\boldsymbol{Y}}_{i.})'.\end{aligned}$$

En la simplificación interviene el hecho que $\sum_{j=1}^{n_i}(Y_{ij} - \bar{\boldsymbol{Y}}_{i.}) = \mathbf{0}$.

En resumen,

$$\underbrace{\sum_{i=1}^{q}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\boldsymbol{Y}}_{..})(Y_{ij} - \bar{\boldsymbol{Y}}_{..})'}_{\text{Covariabilidad total}} = \underbrace{\sum_{i=1}^{q} n_i(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})'}_{\text{Covariabilidad entre}} + \underbrace{\sum_{i=1}^{q}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\boldsymbol{Y}}_{i.})(Y_{ij} - \bar{\boldsymbol{Y}}_{i.})'}_{\text{Covariabilidad dentro}} \tag{3.71}$$

El término variabilidad se emplea por tener como referencia al caso univariado, porque en realidad la descomposición es sobre la información contenida en la matriz de covarianzas; que corresponde a variabilidad y asociación lineal (covarianza o covariabilidad).

En el caso univariado, la identidad para el análisis de varianza es

$$SC.Total = SC.Modelo \quad + \quad SC.Error$$

$$SC.Total = SC.Entre \quad + \quad SC.Dentro.$$

La estadística de prueba $F = \frac{N-q}{q1}\frac{SC.Entre}{SC.Dentro}$, se puede transformar a:

$$\frac{1}{[q/(N-q)]F+1} = \frac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2};$$

de manera que Λ corresponde, en forma semejante, al cociente de la suma de cuadrados dentro y la suma de cuadrados total; $\boldsymbol{E}$ y $\boldsymbol{E}+\boldsymbol{H}$ hacen tal papel. Más explícitamente

$$\begin{aligned}
\boldsymbol{H} &= \sum_{i=1}^{q} n_i(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})' \\
\boldsymbol{E} &= \sum_{i=1}^{q}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\boldsymbol{Y}}_{i.})(Y_{ij} - \bar{\boldsymbol{Y}}_{i.})' \\
\boldsymbol{E}+\boldsymbol{H} &= \sum_{i=1}^{q}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\boldsymbol{Y}}_{..})(Y_{ij} - \bar{\boldsymbol{Y}}_{..})'. \qquad (3.72)
\end{aligned}$$

Esta escritura de $\boldsymbol{E}$ permite encontrar un estimador insesgado de $\boldsymbol{\Sigma}$. De esta manera:

$$\boldsymbol{E} = (n_1-1)\boldsymbol{S}_1 + \cdots + (n_q-1)\boldsymbol{S}_q = \sum_{i=1}^{q}(n_i-1)\boldsymbol{S}_i, \qquad (3.73)$$

donde $\boldsymbol{S}_i$ es la matriz de covarianzas de la i-ésima muestra. Así, la matriz de varianzas y covarianzas estimada, puesto que las poblaciones se han considerado con igual matriz de covarianzas, es:

$$\boldsymbol{S}_p = \frac{1}{\sum_{i=1}^{q}(n_i-1)}\boldsymbol{E} = \frac{\sum_{i=1}^{q}(n_i-1)\boldsymbol{S}_i}{\sum_{i=1}^{q}(n_i-1)}.$$

Es inmediato que para $p=1$ (caso univarido), la razón de máxima verosimilitud se reduce a la conocida estadística F; así se rechaza H_0 si:

$$\frac{\sum_i n_i(\bar{Y}_i - \bar{Y})^2}{\sum_\alpha (Y_{\alpha_i} - Y_i)^2}\left(\frac{N-q}{q-1}\right) > F_{(\alpha,q-1,N-q)}. \qquad (3.74)$$

La distribución exacta de Λ ha sido obtenida para algunos casos especiales, la tabla 3.8 los resume.

Tabla 3.8 Relación entre las estadísticas Λ y F

No. Variables	No. Grupos	Transformación	Distribución F
$p = 1$	$q \geq 2$	$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{N-q}{q-1}\right)$	$F_{(q-1,N-q)}$
$p = 2$	$q \geq 2$	$\left(\frac{1-\Lambda^{1/2}}{\Lambda^{1/2}}\right)\left(\frac{N-q-1}{q-1}\right)$	$F_{(2(q-1),2(N-q-1))}$
$p \geq 1$	$q = 2$	$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{N-p-1}{p}\right)$	$F_{(p,N-p-1)}$
$p \geq 1$	$q = 3$	$\left(\frac{1-\Lambda^{1/2}}{\Lambda^{1/2}}\right)\left(\frac{N-p-2}{p}\right)$	$F_{(2p,2(N-p-2))}$

Para muestras de tamaño grande se tiene la estadística de Bartlett

$$\boldsymbol{V} = -\left(N - 1 - \frac{(p+q)}{2}\right)\ln\Lambda = -\left(N - 1 - \frac{(p+q)}{2}\right)\ln\left(\frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|}\right), \tag{3.75}$$

la cual tiene aproximadamente una distribución ji-cuadrado con $p(q-1)$ grados de libertad. Se rechaza H_0 para valores de $\boldsymbol{V}$ mayores que $\chi^2_{(\alpha,p(q-1))}$.

► Otras estadísticas aproximadas para el ANAVAMU

En esta parte se abordan otras estadísticas equivalentes al lambda de Wilks.

Se demuestra que

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}+\boldsymbol{H}|} = \prod_{i=1}^{p}(1+l_i)^{-1},$$

donde los l_i son las raices de

$$|\boldsymbol{H} - l\boldsymbol{E}| = 0$$

que corresponden a los valores propios de $\boldsymbol{H}\boldsymbol{E}^{-1}$. No es difícil intuir que se rechaza H_0 para valores de l_i grandes; puesto que estos hacen pequeño a Λ. En esta misma dirección se han desarrollado algunos criterios para el *ANAVAMU*.

- *La traza de Lawley–Hotelling*

  Lawley (1938) y Hotelling (1947, 1951) propusieron la suma de las raices características de $\boldsymbol{H}\boldsymbol{E}^{-1}$ como estadístico de prueba. Dado

que la suma de las raices características es igual a la traza de la matriz; es decir,

$$\boldsymbol{U} = \sum l_i = \text{tra}(\boldsymbol{H}\boldsymbol{E}^{-1}),$$

se rechaza la hipótesis nula si este valor es más grande que una cantidad que depende de p, N y q. La distribución exacta de la estadística $\boldsymbol{U} = \text{tra}(\boldsymbol{H}\boldsymbol{E}^{-1})$ no es sencilla, bajo la hipótesis nula la distribución límite de $N\,\text{tra}(\boldsymbol{H}\boldsymbol{E}^{-1})$ es *ji-cuadrado* con pq-grados de libertad. Con la distribución límite se toma la decisión de no rechazar o rechazar H_0.

- *La traza de Bartlett–Nanda–Pillai*

  El criterio propuesto por Bartlett (1939), Nanda (1950) y finalmente Pillai (1955), es

  $$\boldsymbol{V} = \sum_{i=1}^{p} \frac{l_i}{1+l_i} = \text{tra}(\boldsymbol{H}(\boldsymbol{E}+\boldsymbol{H})^{-1}).$$

  Asintóticamente, Anderson (1984) demuestra que $\boldsymbol{V}$ tiene distribución *ji–cuadrado* con pq–grados de libertad.

  Mijares (1990) obtiene una aproximación a la distribución normal con sus dos primeros momentos exactos. Los valores obtenidos a 5% y 1% son bastante aproximados a los obtenidos en otras tablas.

- *Criterio del máximo valor propio de Roy*

  Roy (1953) propuso al máximo valor propio de $\boldsymbol{H}\boldsymbol{E}^{-1}$ como estadístico de prueba, denótese por l_1. Se rechaza la hipótesis nula si l_1 es más grande que cierto valor o, equivalentemente si

  $$\boldsymbol{R} = \frac{l_1}{(1+l_1)}$$

  es más grande que un número $r_{\alpha,p,q,n}$ tal que

  $$P\{R \geq r_{p,q,N}(\alpha)\} = \alpha.$$

  Anderson (1984) obtuvo tablas para la estadística Nl_1/q, las cuales permiten emplear esta estadística para algunos valores particulares de N, p y q.

En resumen las cuatro estadísticas son las siguientes

- Lambda de Wilks: $\Lambda = \prod_{i=1}^{p} \frac{1}{(1+l_i)}$
- Traza de Lawley–Hotelling: $\boldsymbol{U} = \sum l_i = \text{tra}(\boldsymbol{H}\boldsymbol{E}^{-1})$
- Traza de Bartlett–Nanda–Pillai: $\boldsymbol{V} = \sum_{i=1}^{p} \frac{l_i}{1+l_i} = \text{tra}(\boldsymbol{H}(\boldsymbol{E}+\boldsymbol{H})^{-1})$
- Máximo valor propio de Roy: $\boldsymbol{R} = \frac{l_1}{(1+l_1)}$

Cabe anotar que paquetes como SAS, SPSS, MINITAB, entre otros, desarrollan los cálculos para el análisis de varianza multivariado y suministran el *p valor* para cada una de las estadísticas anteriores. Por ejemplo, para la estadística lambda de Wilks, p es un valor (conocido como el *"p-valor"*) tal que $P(\Lambda_{p,\nu_H,\nu_E} < \Lambda) = p$; de manera que si $p < \alpha$ se rechaza H_0. Esto nos hace menos dependientes de las tradicionales tablas estadísticas.

Ejemplo 3.9 Con los siguientes datos se quiere establecer si tres métodos de enseñanza producen el mismo rendimiento promedio en matemáticas y escritura en niños de características similares.

Es éste un problema de análisis de varianza multivariado, con $p = 2$ que corresponde a los puntajes en matemáticas y escritura por estudiante. El número de poblaciones es $q = 3$; es decir, las tres metodologías. Los resultados del experimento se muestran a continuación.

Método 1	$\binom{69}{75}$	$\binom{69}{70}$	$\binom{71}{73}$	$\binom{78}{82}$	$\binom{79}{81}$	$\binom{73}{75}$	
Método 2	$\binom{69}{70}$	$\binom{68}{74}$	$\binom{75}{80}$	$\binom{78}{85}$	$\binom{68}{68}$	$\binom{63}{68}$	$\binom{72}{74}$
	$\binom{63}{66}$	$\binom{71}{76}$	$\binom{72}{78}$	$\binom{71}{73}$	$\binom{70}{73}$	$\binom{56}{59}$	$\binom{77}{83}$
Método 3	$\binom{72}{79}$	$\binom{64}{65}$	$\binom{74}{74}$	$\binom{72}{75}$	$\binom{82}{84}$	$\binom{69}{68}$	$\binom{76}{76}$
	$\binom{68}{65}$	$\binom{78}{79}$	$\binom{70}{71}$	$\binom{60}{61}$			

Fuente: Freund, Litell and Spector (1986)

Se hará el análisis de varianza univariado (ANDEVA); es decir, para cada una de las dos variables, y el análisis de varianza multivariado (ANAVAMU) que se sugiere en este capítulo.

Las tablas 3.9 y 3.10 corresponden al análisis de varianza para cada una de las variables en forma separada.

Tabla 3.9 ANDEVA para matemáticas

Fuentes de Variación	G. L.	S. C.	C. M.	Valor F	$Pr > F$
Métodos	2	60.6051	30.3025	0.91	0.4143
Error	28	932.8788	33.3171		
Total	30	993.4839			

G.L.: Grados de libertad, S.C.: Suma de cuadrados y C.M.: Cuadrados medios

De los resultados mostrados en la tabla 3.9 se puede afirmar que las metodologías no producen rendimientos promedios diferentes en matemáticas, en esta clase de niños.

Una conclusión similar se puede extraer de la tabla 3.10 para la variable escritura

Tabla 3.10 ANDEVA para escritura

Fuentes de Variación	G. L.	S. C.	C. M.	Valor F	$Pr > F$
Métodos	2	49.7359	24.8679	0.56	0.5776
Error	28	1243.9416	44.4265		
Total	30	1293.6775			

G.L.: Grados de libertad, S.C.: Suma de cuadrados y C.M.: Cuadrados medios

Ahora se desarrolla, sobre los mismos datos, el análisis de varianza multivariado. El modelo es

$$Y_{ij} = \boldsymbol{\mu} + \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij} \quad \text{con } i = 1, \ldots, 3 \quad j = 1, \ldots, n_i.$$

En este caso $n_1 = 6$, $n_2 = 14$, $n_3 = 11$ y $N = 31$.

Mediante la hipótesis nula se afirma que los métodos producen un rendimiento en promedio igual en matemáticas y en escritura; es decir,

$$H_0 : \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3.$$

Las matrices de sumas de cuadrados (covariabilidad) dentro y entre tratamientos se obtienen aplicando (3.72)

$$\boldsymbol{E} = \begin{pmatrix} 932.87879 & 1018.6818 \\ 1018.6818 & 1243.9416 \end{pmatrix} \qquad \boldsymbol{H} = \begin{pmatrix} 60.6050 & 31.5117 \\ 31.5117 & 49.7358 \end{pmatrix}.$$

El valor del lambda de Wilks es

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}|} = \frac{\begin{vmatrix} 932.8788 & 1018.6818 \\ 1018.6818 & 1243.9416 \end{vmatrix}}{\begin{vmatrix} 993.4838 & 1050.1935 \\ 1050.1935 & 1293.6774 \end{vmatrix}} = 0.6731.$$

De la tabla 3.8 y como $p = 2$ y $q = 3$, se puede utilizar la estadística

$$\left(\frac{1 - \Lambda^{1/2}}{\Lambda^{1/2}}\right)\left(\frac{n - q - 1}{q - 1}\right) = \left(\frac{1 - \sqrt{0.6731}}{\sqrt{0.6731}}\right)\left(\frac{31 - 3 - 1}{3 - 1}\right) = 2.954851.$$

El valor anterior comparado con $F_{(5\%,2(3-1),2(31-3-1))} = F_{(5\%,4,54)} \approx 2.5$ (tabla C.8), permite afirmar que el puntaje promedio no es el mismo para las tres metodologías.

¡*El resultado no es el mismo que se obtuvo con los análisis de varianza univariados*!

¿Qué ocurre? Pues bien, nótese que en el primer análisis no se considera la relación que pueda haber entre las dos variables, algunos pedagogos podrán afirmar que la correlación entre la habilidad matemática y la escritura es alta; de manera que hay información en los datos que se está desaprovechando. La matriz de covarianzas estimada es

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S}_p = \begin{pmatrix} 33.3171 & 36.3815 \\ 36.3815 & 44.4265 \end{pmatrix}.$$

La matriz $\boldsymbol{S}_p$ muestra la relación entre las variables rendimiento en matemáticas y escritura. Con el primer tipo de análisis de varianza se está descartando esta asociación lineal de las variables; hecho que explica la diferencia de los procedimientos.

Las estadísticas ligadas a los valores propios de $\boldsymbol{E}$ y de $\boldsymbol{H}\boldsymbol{E}^{-1}$ se calculan a continuación, como una herramienta adicional para desarrollar el ANAVAMU de estos datos. De las matrices $\boldsymbol{E}$ y $\boldsymbol{H}$, calculadas anteriormente, se obtiene

$$\boldsymbol{H}\boldsymbol{E}^{-1} = \begin{pmatrix} 0.3527037 & -0.2635020 \\ -0.093424 & 0.11648850 \end{pmatrix}.$$

Los valores propios de $\boldsymbol{H}\boldsymbol{E}^{-1}$ son la solución de la ecuación

$$\begin{aligned} |\boldsymbol{H}\boldsymbol{E}^{-1} - \lambda I| &= 0 \\ (0.3527037 - \lambda)(0.1164885 - \lambda) - (0.093424)(0.2635020) &= 0 \\ \lambda^2 - 0.4692\lambda + 0.0165 &= 0 \end{aligned}$$

de donde, después de redondear las cifras, las soluciones de esta ecuación son: $\lambda_1 = 0.4309$ y $\lambda_2 = 0.0382$.

La traza de *Lawley–Hotelling* es igual a $\lambda_1 + \lambda_2 = 0.4691$; es decir, se rechaza la hipótesis de igualdad de vectores de medias en las tres poblaciones, puesto que

$$n \cdot \text{tra}(\boldsymbol{H}\boldsymbol{E}^{-1}) = 31(0.4309) = 13.5579$$

es mayor que el percentil 95 de la estadística ji-cuadrado con 6 grados de libertad; éste es, según la tabla C.7, $\chi^2_{(5\%,6)} = 12.59$.

Las estadísticas de *Roy* y de *Bartlett-Nanda-Pillai*, sobre estos datos, toman los valores de 0.4309 y 0.3379, respectivamente. De acuerdo con la distribución ya expuesta para estas estadísticas, se sugiere tomar decisiones similares respecto a la hipótesis H_0, en consecuencia se rechaza la hipótesis de igualdad de vectores de medias en las tres poblaciones. ☑

▶ Modelos de doble vía de clasificación

Otro caso a desarrollar es el plan experimental asociado a un modelo de *doble vía de clasificación*. Se puede pensar en un conjunto de datos dispuesto en una tabla de doble entrada, donde las filas (o columnas) representan los niveles de un primer factor (notado por A) y las columnas (o filas) los niveles de un segundo factor (notado por B); las celdas corresponden a los tratamientos. En cada celda estarán las observaciones por cada tratamiento.

De esta manera, sea Y_{ijk}, con $i = 1, \ldots, f$; $j = 1, \ldots, c$ y $k = 1, \ldots, n_{ij}$ un conjunto de vectores aleatorios p-dimensionales e independientes. El modelo que relaciona la respuesta Y_{ijk} con el factor A, el factor B y la interacción entre A y B es

$$Y_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij} + \boldsymbol{\varepsilon}_{ijk}, \tag{3.76}$$

donde

- $\boldsymbol{\alpha}_i$ es el efecto debido al i–ésimo nivel del factor A,
- $\boldsymbol{\beta}_j$ es el efecto debido al j–ésimo nivel del factor B, y
- $\boldsymbol{\gamma}_{ij}$ es el efecto debido a la interacción entre el i–ésimo nivel del factor A y el j–ésimo nivel del factor B.

Las hipótesis sobre la significación del factor A, del factor B y de la interacción AB, son respectivamente, las siguientes:

$$\begin{aligned} (A)\ & H_0 : \boldsymbol{\alpha}_i = 0 \text{ para } \ i = 1, \ldots, f \\ (B)\ & H_0 : \boldsymbol{\beta}_j = 0 \text{ para } \ j = 1, \ldots, c \\ (AB)\ & H_0 : \boldsymbol{\gamma}_{ij} = 0 \text{ para } \ i = 1, \ldots, f \quad j = 1, \ldots, c. \end{aligned} \tag{3.77}$$

Observaciones:

- Es común encontrar en la literatura del análisis de varianza los nombres de factor A para el primer factor, B para el segundo y AB para la interacción.

- Si tan sólo se dispone de una observación por cada tratamiento ($n_{ij} = 1$), no será posible estimar el efecto de la respectiva interacción.

Con la misma álgebra empleada para modelos de una vía de clasificación se hace el análisis de varianza para modelos de doble vía de clasificación. Para esto, sean $\boldsymbol{Y}_{...}$, $\boldsymbol{Y}_{i..}$, $\boldsymbol{Y}_{.j.}$ y $\boldsymbol{Y}_{ij.}$, el total general, el total por fila, el total por columna y el total por celda, respectivamente.

La razón de máxima verosimilitud para contrastar alguna de las tres hipótesis expresadas en (3.78) es similar a (3.65). Las matrices $\boldsymbol{H}_A$, $\boldsymbol{H}_B$, $\boldsymbol{H}_{AB}$ y $\boldsymbol{E}$ representan las sumas de cuadrados para los factores principales, la interacción y el error. Éstas son:

$$\begin{aligned}
\boldsymbol{H}_A &= c\sum_i n_{i\cdot}(\bar{\boldsymbol{Y}}_{i..} - \bar{\boldsymbol{Y}}_{...})(\bar{\boldsymbol{Y}}_{i..} - \bar{\boldsymbol{Y}}_{...})' \\
\boldsymbol{H}_B &= f\sum_j n_{\cdot j}(\bar{\boldsymbol{Y}}_{.j.} - \bar{\boldsymbol{Y}}_{...})(\bar{\boldsymbol{Y}}_{.j.} - \bar{\boldsymbol{Y}}_{...})' \\
\boldsymbol{H}_{AB} &= \sum_{i,j} n_{ij\cdot}(\bar{\boldsymbol{Y}}_{ij.} - \bar{\boldsymbol{Y}}_{i..} - \bar{\boldsymbol{Y}}_{.j.} + \bar{\boldsymbol{Y}}_{...})(\bar{\boldsymbol{Y}}_{ij.} - \bar{\boldsymbol{Y}}_{i..} - \bar{\boldsymbol{Y}}_{.j.} + \bar{\boldsymbol{Y}}_{...})' \\
\boldsymbol{E} &= \sum_{i,j,k}(Y_{ijk} - \bar{\boldsymbol{Y}}_{ij.})(Y_{ijk} - \bar{\boldsymbol{Y}}_{ij.})'. \qquad (3.78)
\end{aligned}$$

Los *lambda de Wilks* para contrastar cada una de las hipótesis son, respectivamente,

$$\begin{aligned}
\Lambda_A &= \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}_A|}, \\
\Lambda_B &= \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}_B|} \\
\Lambda_{AB} &= \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}_{AB}|}. \qquad (3.79)
\end{aligned}$$

De acuerdo con los valores p y q se pueden emplear las transformaciones resumidas en la tabla 3.8. Para tamaños de muestra grandes, se sugiere utilizar la aproximación de Bartlett, la cual para cada una de las hipótesis planteadas en (3.78) es la siguiente:

1. Se rechaza la hipótesis de que no existe efecto debido al factor A; es decir,
   $\boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_f = 0$, si

$$-\left(cf(N-1) - \frac{p+1-(f-1)}{2}\right)\ln\Lambda_A > \chi^2_{(\alpha,(f-1)p)},$$

donde $\chi^2_{(\alpha,(f-1)p)}$ corresponde al percentil $(1-\alpha)$ de la distribución *ji–cuadrado* con $(f-1)p$ grados de libertad.

El tratamiento para las hipótesis sobre el efecto del factor B y de la interacción AB es similar; así:

2. Se rechaza la hipótesis de que no existe efecto debido al factor B; es decir, $H_0: \boldsymbol{\beta}_j = 0$ para $j = 1, \ldots, c$, si

$$-\left(cf(N-1) - \frac{p+1-(c-1)}{2}\right) \ln \Lambda_B > \chi^2_{(\alpha,(c-1)p)}.$$

3. Se rechaza la hipótesis de que no existe efecto debido a la interacción entre A y B; es decir, $H_0: \boldsymbol{\gamma}_{ij} = 0$ para $i = 1, \ldots, f$ y $j = 1, \ldots, c$, si

$$-\left(cf(N-1) - \frac{p+1-(f-1)(c-1)}{2}\right) \ln \Lambda_{AB} > \chi^2_{(\alpha,(f-1)(c-1)p)}.$$

Ejemplo 3.10 Los datos de la tabla 3.11 indican la producción de cinco variedades de cebada (factor A) para dos años consecutivos en seis localidades diferentes (factor B). Las columnas indican las variedades y las filas las localidades; en cada localidad hay dos vectores que corresponden a la producción de cada año para las cinco variedades.

De acuerdo con el desarrollo hecho en la sección (3.5.3) y con las expresiones contenidas en (3.79) y (3.80) se obtienen los siguientes resultados:

$$\begin{aligned}
\sum_{i.j} Y_{ij} Y'_{ij} &= \begin{pmatrix} 380.944 & 315.381 \\ 315.381 & 277.625 \end{pmatrix} \\
\sum_{j} (6Y_{.j})(6Y_{.j})' &= \begin{pmatrix} 2157.924 & 1844.346 \\ 1844.346 & 1579.583 \end{pmatrix} \\
\sum_{i} (5Y_{i.})(5Y_{i.})' &= \begin{pmatrix} 1874.386 & 1560.145 \\ 1560.145 & 1353.727 \end{pmatrix} (30Y..)(30Y..)' \\
&= \begin{pmatrix} 10705.986 & 9145.240 \\ 9145.240 & 7812.025 \end{pmatrix}.
\end{aligned}$$

La matriz correspondiente al error es:

$$\boldsymbol{E} = \begin{pmatrix} 3279 & 802 \\ 802 & 4017 \end{pmatrix}.$$

Tabla 3.11 Producción de cebada por variedad, año y localidad

Localización	Variedad (A)					
(B)	V_1	V_2	V_3	V_4	V_5	Y_i
L_1	81	105	120	110	98	514
	81	82	80	87	84	414
L_2	147	142	151	192	146	778
	100	116	112	148	108	584
L_3	82	77	78	131	90	458
	103	105	117	140	130	595
L_4	120	121	124	141	125	631
	99	62	96	126	76	459
L_5	99	89	69	89	104	450
	66	50	97	62	80	355
L_6	87	77	79	102	96	441
	68	67	67	92	94	388
$Y_{.j}$	616	611	621	765	659	3272
	517	482	569	655	572	2795

Las sumas de cuadrados por fila son:

$$5\boldsymbol{\Sigma}(Y_{i.} - Y_{..})(Y_{i.} - Y_{..})' = \begin{pmatrix} 18.011 & 7.188 \\ 7.188 & 10.345 \end{pmatrix}.$$

Las sumas de cuadrados por columna (entre tratamientos) son:

$$\boldsymbol{H}_A = \begin{pmatrix} 2788 & 2550 \\ 2550 & 2863 \end{pmatrix}.$$

La estadística de prueba, de acuerdo con (3.80), es:

$$\Lambda_A = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}_A|} = \frac{\begin{vmatrix} 3279 & 802 \\ 802 & 4017 \end{vmatrix}}{\begin{vmatrix} 6067 & 3352 \\ 3352 & 6880 \end{vmatrix}} = 0.4107.$$

Por el resultado contenido en la tabla 3.8 (segunda línea) se tiene

$$\left(\frac{1 - \Lambda_A^{1/2}}{\Lambda_A^{1/2}}\right)\left(\frac{N - q - 1}{q - 1}\right) \sim F_{(2(q-1),2(N-q-1))} \frac{1 - \sqrt{0.4107}}{\sqrt{0.4107}} \cdot \frac{19}{4} = 2.66,$$

para el caso $p = 2$, $N - q = (f-1)(c-1) = 20$ y $q = c - 1 = 4$

$$\frac{1 - \sqrt{0.4107}}{\sqrt{0.4107}} \cdot \frac{19}{4} = 2.66,$$

el cual comparado con el percentil 95 de una distribución $F_{(8,38)}$, es decir, con $F_{(5\%,8,38)} \approx 2.18$ (tabla C.8), es significativo. Resultado que muestra la diferencia en rendimiento entre las variedades de cebada para los dos años considerados. ☑

► Contrastes

Una vez que se ha rechazado la hipótesis nula, viene la pregunta ¿Cuáles son las variables que provocan el rechazo de la hipótesis? Varias han sido las estrategias consideradas para resolver esta inquietud, los *contrastes* es una de ellas , los cuales, en la mayoría de las veces, son comparaciones entre las medias, planeadas por el investigador o sugeridas por los datos.

• Caso univariado.

En el caso univariado, un contraste de las medias poblacionales es una combinación lineal de la forma

$$\delta = c_1\mu_1 + \cdots + c_q\mu_q,$$

donde los coeficientes satisfacen: $\sum_{i=1}^{q} c_i = 0$. Un estimador insesgado de δ es

$$\hat{\delta} = c_1\bar{Y}_{1.} + \cdots + c_q\bar{Y}_{q.}.$$

Como los $\bar{Y}_{i.}$ son independientes con varianza σ^2/n_i, la varianza de los $\hat{\delta}$ es

$$\text{var}(\hat{\delta}) = \sigma^2 \sum_{i=1}^{q} \frac{c_i^2}{n_i},$$

la cual puede estimarse por

$$s_{\hat{\delta}}^2 = CME \sum_{i=1}^{q} \frac{c_i^2}{n_i},$$

donde CME es el cuadrado medio del error. Una estadística para verificar la hipótesis asociada con el contraste, $H_0 : \boldsymbol{\delta} = c_1\mu_1 + \cdots + c_q\mu_q = 0$, es

$$F = \frac{\hat{\delta}^2}{s^2_{\hat{\delta}}} = \frac{\left(\sum_{i=1}^q c_i \bar{Y}_{i.}\right)^2}{CME \sum_{i=1}^q c_i^2/n_i}$$
$$= \frac{\left(\sum_{i=1}^q c_i \bar{Y}_{i.}\right)^2 / \sum_{i=1}^q c_i^2/n_i}{CME},$$

la cual tiene distribución $F_{(1,N-q)}$, donde $N = \sum_{i=1}^q n_i$.

Si dos contrastes sobre las medias $\delta = \sum_{i=1}^q a_i\mu_i$ y $\gamma = \sum_{i=1}^q b_i\mu_i$ son tales que $\sum_{i=1}^q a_i b_i/n_i = 0$, los contrastes se denominan *ortogonales*. Si el diseño es balanceado es suficiente con que $\sum_{i=1}^q a_i b_i = 0$.

• Caso multivariado.

En la sección (3.4) se han considerado hipótesis de la forma $H_0 : \boldsymbol{C\mu} = 0$. Cada fila de la matriz $\boldsymbol{C}$ suma cero, así, $\boldsymbol{C\mu}$ es un conjunto de contrastes entre las medias $\mu_1, \ldots \mu_p$ de $\boldsymbol{\mu}$. En esta sección se hacen contrastes donde se comparan vectores de medias y no sus elementos dentro de ellos.

Un contraste entre los vectores de medias asociados a q–poblaciones está definido por

$$\boldsymbol{\delta} = c_1\boldsymbol{\mu}_1 + \cdots + c_q\boldsymbol{\mu}_q, \tag{3.80}$$

donde $\sum_{i=1}^q c_i = 0$. Un estimador insesgado de $\boldsymbol{\delta}$ es la correspondiente combinación lineal de las medias muestrales:

$$\hat{\boldsymbol{\delta}} = c_1\bar{\boldsymbol{Y}}_{1.} + \cdots + c_q\bar{\boldsymbol{Y}}_{q.}. \tag{3.81}$$

Los vectores de medias muestrales $\bar{\boldsymbol{Y}}_{1.}, \ldots, \bar{\boldsymbol{Y}}_{q.}$ se definen como se mostró al comienzo de esta sección; es decir, $\bar{\boldsymbol{Y}}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$, los cuales se asumen independientes y con matriz de covarianzas $\mathrm{Cov}(\bar{\boldsymbol{Y}}_{i.}) = \boldsymbol{\Sigma}/n_i$. De esta manera, la matriz de covarianzas para $\hat{\boldsymbol{\delta}}$ es

$$\mathrm{Cov}(\hat{\boldsymbol{\delta}}) = c_1^2\frac{\boldsymbol{\Sigma}}{n_1} + \cdots + c_q^2\frac{\boldsymbol{\Sigma}}{n_i} = \boldsymbol{\Sigma}\sum_{i=1}^q \frac{c_i^2}{n_i},$$

la cual se estima mediante

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\delta}}) = \boldsymbol{S}_p\sum_{i=1}^q \frac{c_i^2}{n_i}$$

con

$$\boldsymbol{S}_p = \frac{1}{\sum_{i=1}^{q}(n_i-1)}\boldsymbol{E} = \frac{\sum_{i=1}^{q}(n_i-1)\boldsymbol{S}_i}{\sum_{i=1}^{q}(n_i-1)}.$$

La hipótesis a verificar mediante el contraste que involucra los vectores de medias poblacionales, es: $\boldsymbol{\delta} = c_1\boldsymbol{\mu}_1 + \cdots + c_q\boldsymbol{\mu}_q = 0$. Por ejemplo, para $q = 3$, $2\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 - \boldsymbol{\mu}_3$ es equivalente a

$$\boldsymbol{\mu}_1 = \frac{1}{2}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3).$$

Naturalmente, esto implica que los elementos de $\boldsymbol{\mu}_1$ son iguales a los respectivos elementos de $\frac{1}{2}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3)$; es decir,

$$\begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(\mu_{21} + \mu_{31}) \\ \frac{1}{2}(\mu_{22} + \mu_{32}) \\ \vdots \\ \frac{1}{2}(\mu_{2p} + \mu_{3p}) \end{pmatrix}.$$

Bajo el supuesto de que los datos se distribuyen conforme a un modelo multinormal, la hipótesis $H_0: \boldsymbol{\delta} = c_1\boldsymbol{\mu}_1 + \cdots + c_q\boldsymbol{\mu}_q = 0$ se verifica con la estadística

$$\begin{aligned} T^2 &= \hat{\boldsymbol{\delta}}'\left(\boldsymbol{S}_p \sum_{i=1}^{q} \frac{c_i^2}{n_i}\right)^{-1}\hat{\boldsymbol{\delta}} \\ &= \frac{1}{\sum_{i=1}^{q} c_i^2/ni}\left(\sum_{i=1}^{q} c_i\bar{\boldsymbol{Y}}_{i.}\right)'\left(\frac{\boldsymbol{E}}{N-q}\right)^{-1}\left(\sum_{i=1}^{q} c_i\bar{\boldsymbol{Y}}_{i.}\right), \end{aligned} \tag{3.82}$$

la cual se distribuye como $T^2_{(p,N-q)}$, con $N = \sum_{i=1}^{q} n_i$.

Una prueba equivalente para la hipótesis H_0 sobre el contraste $\boldsymbol{\delta}$ se construye mediante el lambda de Wilks.

3.5.4 Análisis de perfiles en q–muestras

En la sección (3.4) se trató el análisis de perfiles en una y dos muestras, se considera en esta sección el caso de q–grupos o muestras independientes. Como en los casos anteriores se asume que las variables para cada una de las p–respuestas son conmensurables.

El modelo asociado corresponde a un ANAVAMU, de una vía de clasificación balanceado; es decir,

$$Y_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij}, \text{ para } i = 1, \ldots, q \text{ y } j = 1, \ldots, n.$$

Se quiere verificar la hipótesis $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_q$. Con variables conmensurables, la hipótesis anterior puede orientarse más específicamente a los q perfiles generados al graficar los vectores $\boldsymbol{\mu}_i$. El interés se dirige sobre las mismas hipótesis anteriores. Éstas son:

- H_{01}: Los q perfiles son *paralelos*.
- H_{02}: Los q perfiles están en el *mismo nivel (coinciden)*.
- H_{03}: Los q perfiles son *planos*.

▶ Perfiles paralelos

Se denominan perfiles paralelos a los que corresponden a líneas poligonales que no se cruzan o intersecan (isoclinos); significa que la tasa (pendiente) de variación, en el tiempo, entre los dos medias particulares es la misma, cualquiera que sea la población.

Se debe aclarar que en el ambiente estadíistico la idea de paralelismo no es estrictamente la misma que la geométrica, pues el paralelismo es declarado por la estadística con la cual se verifique esta hipótesis en términos del rechazo o no rechazo de la hipótesis con cierta grado de incertidumbre (probabilidad).

La hipótesis es una extensión del caso de dos muestras, así,

$$H_{01} : \boldsymbol{C}\boldsymbol{\mu}_1 = \cdots = \boldsymbol{C}\boldsymbol{\mu}_q,$$

donde $\boldsymbol{C}$ es una matriz de tamaño $(p-1) \times p$ y de rango $(p-1)$, tal que $\boldsymbol{C}\mathbf{1} = 0$. Como se ha advertido, esta matriz no es única, por ejemplo,

$$\boldsymbol{C} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

La hipótesis anterior es equivalente a $H_{01} : \boldsymbol{\mu}_{Z_1} = \cdots = \boldsymbol{\mu}_{Z_q}$, ésta se verifica mediante un ANAVAMU en un diseño a una vía de clasificación sobre las variables transformadas mediante $Z_{ij} = \boldsymbol{C}Y_{ij}$. De acuerdo con la propiedad (2.2.2) el vector $Z_{ij} \sim N_{p-1}(\boldsymbol{C}\boldsymbol{\mu}_i, \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}')$. Como la matriz $\boldsymbol{C}$ tiene $p-1$ filas, $\boldsymbol{C}Y_{ij}$ es de tamaño $((p-1) \times 1)$, $\boldsymbol{C}\boldsymbol{\mu}_i$ es de tamaño $((p-1) \times 1)$, y el tamaño de $\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}'$ es $(p-1) \times (p-1)$.

Las matrices asociadas con la covariación "entre" y "dentro" son, respectivamente,

$$\boldsymbol{H}_Z = \boldsymbol{C}\boldsymbol{H}\boldsymbol{C}' \quad \text{y} \quad \boldsymbol{E}_Z = \boldsymbol{C}\boldsymbol{E}\boldsymbol{C}'.$$

La estadística de prueba es

$$\boldsymbol{\Lambda}_1 = \frac{|\boldsymbol{CEC}'|}{|\boldsymbol{CEC}' + \boldsymbol{CHC}'|} = \frac{|\boldsymbol{CEC}'|}{|\boldsymbol{C}(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{C}'|},$$

la cual se distribuye como $\boldsymbol{\Lambda}_{(p-1,q-1,q(n-1))}$. Las otras tres pruebas estadísticas se obtienen mediante los valores propios de la matriz

$(\boldsymbol{CEC}')^{-1}(\boldsymbol{CHC}')$. En el caso de diseños desbalanceados los cálculos de las matrices $\boldsymbol{H}$ y $\boldsymbol{E}$ se hacen conforme a las fórmulas mostradas en las ecuaciones (3.74).

▶ Perfiles en el mismo nivel

La hipótesis de que los q perfiles están en el mismo nivel se escribe como:

$$H_{02}: \ \mathbf{1}'\boldsymbol{\mu}_1 = \cdots = \mathbf{1}'\boldsymbol{\mu}_q.$$

La expresión $\mathbf{1}'Y_{ij} = z_{ij}$ transforma los vectores Y_{ij} en escalares z_{ij}. Se puede emplear la prueba F de un ANDEVA a una vía de clasificación sobre los z_{ij} para comparar las q–muestras. También se puede emplear la estadística

$$\boldsymbol{\Lambda}_2 = \frac{|\mathbf{1}'\boldsymbol{E}\mathbf{1}|}{|\mathbf{1}'\boldsymbol{E}\mathbf{1} + \mathbf{1}'\boldsymbol{H}\mathbf{1}|},$$

que se distribuye como $\boldsymbol{\Lambda}_{(1,q-1,q(n-1))}$.

Se rechaza la hipótesis de que "los perfiles están en el mismo nivel" si el valor de $\boldsymbol{\Lambda}_2 < \boldsymbol{\Lambda}_{(1,q-1,q(n-1),\alpha)}$.

Esta estadística se relaciona con la estadística F sobre los $\mathbf{1}'Y_{ij} = z_{ij}$, de acuerdo con las tabla 3.8 (primera línea), mediante

$$F = \frac{1-\boldsymbol{\Lambda}}{\boldsymbol{\Lambda}} \frac{q(n-1)}{q-1} \sim F_{(q-1,q(n-1))}.$$

▶ Perfiles planos

Se quiere establecer si la media de las p variables es la misma. Esto equivale a establecer la hipótesis de que el promedio de las medias en los q grupos es el mismo para cada variable; es decir,

$$H_{03} = \frac{\mu_{11} + \cdots + \mu_{q1}}{q} = \cdots = \frac{\mu_{1p} + \cdots + \mu_{qp}}{q},$$

o también que

$$\frac{\boldsymbol{C}(\boldsymbol{\mu}_1 + \cdots + \boldsymbol{\mu}_q)}{q} = \mathbf{0},$$

donde la matriz $\boldsymbol{C}$ es una matriz cuyas entradas en cada fila definen un contraste de las $\boldsymbol{\mu}_j's$, ésta se construye como se muestra al comienzo de esta sección.

La hipótesis de "horizontalidad" o "planitud" de los perfiles establece que las medias de las p variables en cada grupo son iguales; es decir, $\mu_{i1} = \cdots = \mu_{ip}$, para $i = 1, \ldots, q$.

La verificación de la hipótesis H_{03} se hace mediante la estadística T^2. Un estimador puntual de $(\boldsymbol{\mu}_1 + \cdots + \boldsymbol{\mu}_q)/q$ es $\bar{Y}_{\cdot\cdot} = \sum_{ij} Y_{ij}/qn$. Bajo la hipótesis H_{03} (y H_{01}), la estadística $\boldsymbol{C}\bar{Y}_{\cdot\cdot}$ se distribuye como $N_{p-1}(\boldsymbol{0}, \boldsymbol{C\Sigma C'}/qn)$; en consecuencia la hipótesis de que los perfiles son planos, es decir H_{03}, se puede verificar mediante la estadística

$$T^2 = qn(\boldsymbol{C}\bar{Y}_{\cdot\cdot})'\left(\frac{\boldsymbol{CEC'}}{q(n-1)}\right)^{-1}(\boldsymbol{C}\bar{Y}_{\cdot\cdot}).$$

Cuando las hipótesis H_{01} y H_{03} son ciertas, la estadística T^2 se distribuye como $T^2_{(p-1,q(n-1))}$.

Ejemplo 3.11 Se quiere evidenciar el efecto de la dosis de la vitamina E sobre el peso (ganacia o pérdida) de animales. Para este propósito a un grupo de animales experimentales se les suministró tres suplementos de vitamina E en los niveles cero o placebo (1), bajo (2) y alto (3); los cuales corresponden a los tratamientos. Cada tratamiento fue asignado y suministrado, de manera aleatoria, a cinco animales, a los cuales se les registró el peso (en gramos) al final de las semanas 1, 3, 4, 5, 6 y 7, respectivamente.

La tabla 3.12 contiene los pesos de cada uno de los 15 animales, sometidos a uno de los tres tratamientos, en cada punto de tiempo decidido; así, los valores en cada fila corresponden a las medidas repetidas de cada animal.

Éste es un caso típico de datos longitudinales, pues se trata de un diseño balanceado donde todos los animales son medidos en las mismas ocasiones y no hay datos faltantes.

El objetivo es comparar los perfiles asociados con cada uno de los tres tratamientos durante estas siete semanas.

Los vectores de medias muestrales para cada uno de los tres tratamientos, y el vector de medias general, son respectivamente,

$$\begin{aligned}
\bar{Y}'_{1\cdot} &= (466.4, 519.4, 568.4, 561.6, 546.6, 572.0),\\
\bar{Y}'_{2\cdot} &= (494.4, 551.0, 574.2, 567.0, 603.0, 644.0),\\
\bar{Y}'_{3\cdot} &= (497.8, 534.6, 579.8, 571.8, 588.2, 623.2),\\
\bar{Y}'_{\cdot\cdot} &= (486.2, 535.0, 574.3, 566.8, 579.3, 613.1).
\end{aligned}$$

Tabla 3.12 Peso de animales bajo 3 niveles de vitamina E

Nivel	Animal	Sem. 1	Sem. 3	Sem. 4	Sem. 5	Sem. 6	Sem. 7
1	1	455	460	510	504	436	466
1	2	467	565	610	596	542	587
1	3	445	530	580	597	582	619
1	4	485	542	594	583	611	612
1	5	480	500	550	528	562	576
2	6	514	560	565	524	552	597
2	7	440	480	536	584	567	569
2	8	495	570	569	585	576	677
2	9	520	590	610	637	671	702
2	10	503	555	591	605	649	675
3	11	496	560	622	622	632	670
3	12	498	540	589	557	568	609
3	13	478	510	568	555	576	605
3	14	545	565	580	601	633	649
3	15	472	498	540	524	532	583

Fuente: Crowder y Hand (1990, págs. 21-29)

La figura 3.10 muestra los tres perfiles de las medias para estas semanas. Se observa un alto grado de "paralelismo" entre los tres perfiles, con excepción de la semana 6 para el grupo de animales que recibió cero vitamina E.

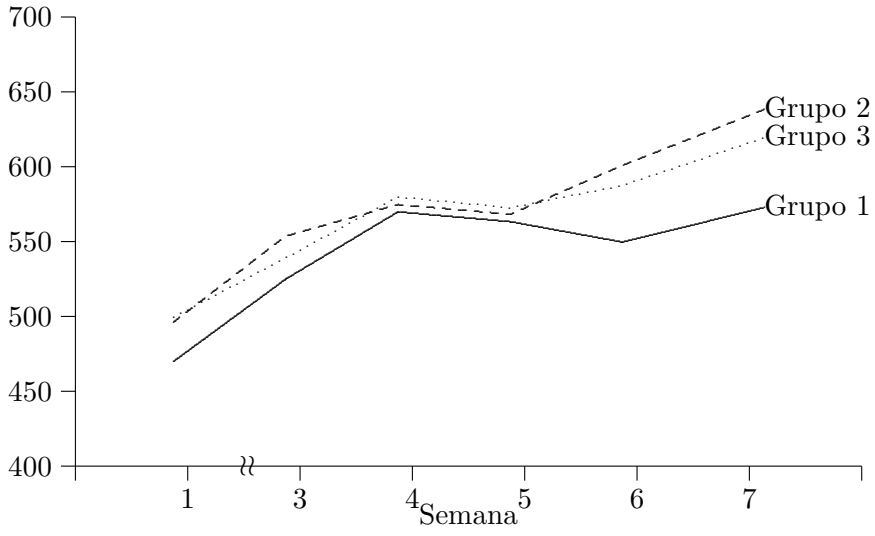


Figura 3.10 *Perfiles de los tres grupos de animales experimentales.*

Las matrices de covariación "dentro" y "entre", $\boldsymbol{E}$ y $\boldsymbol{H}$, son las siguientes:

$$\boldsymbol{E} = \begin{pmatrix} 8481.2 & 8538.8 & 4819.8 & 8513.6 & 8710.0 & 8468.2 \\ 8538.8 & 17170.0 & 13293.0 & 19476.4 & 17034.2 & 20035.4 \\ 4819.8 & 13293.0 & 12992.4 & 17077.4 & 17287.8 & 17697.2 \\ 8513.6 & 19476.4 & 17077.4 & 28906.0 & 26226.4 & 28625.2 \\ 8710.0 & 17034.2 & 17287.8 & 26226.4 & 36898.0 & 31505.8 \\ 8468.2 & 20035.4 & 17697.2 & 28625.2 & 31505.8 & 33538.8 \end{pmatrix}$$

$$\boldsymbol{H} = \begin{pmatrix} 2969.2 & 2177.2 & 859.4 & 813.0 & 4725.2 & 5921.6 \\ 2177.2 & 2497.6 & 410.0 & 411.6 & 4428.8 & 5657.6 \\ 859.4 & 410.0 & 302.5 & 280.4 & 1132.1 & 1392.5 \\ 813.0 & 411.6 & 280.4 & 260.4 & 1096.4 & 1352.0 \\ 4725.2 & 4428.8 & 1132.1 & 1096.4 & 8550.9 & 13830.9 \\ 5921.6 & 5657.6 & 1392.5 & 1352.0 & 10830.9 & 13730.1 \end{pmatrix}.$$

La prueba de paralelismo se hace con la matriz $\boldsymbol{C}$ anterior

$$\boldsymbol{C} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix},$$

así,

$$\boldsymbol{\Lambda}_1 = \frac{|\boldsymbol{CEC}'|}{|\boldsymbol{C}(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{C}'|} = \frac{3.8238 \times 10^{18}}{2.1355 \times 10^{19}} = 0.1791.$$

Como $0.1791 > \boldsymbol{\Lambda}_{(5\%,5,2,12)} = 0.152793$ (tabla C.2), no se rechaza la hipótesis de paralelismo; es decir, el peso promedio de los animales cambia en igual proporción, de una semana a la otra, para los tres tratamientos (vitaminas).

Para verificar la hipótesis de que los perfiles están en el mismo nivel, se emplea la estadística

$$\boldsymbol{\Lambda}_2 = \frac{|\mathbf{1}'\boldsymbol{E}\mathbf{1}|}{|\mathbf{1}'\boldsymbol{E}\mathbf{1} + \mathbf{1}'\boldsymbol{H}\mathbf{1}|} = \frac{632605.2}{632605.2 + 111288.1} = 0.8504.$$

Dado que $0.8504 > \boldsymbol{\Lambda}_{(5\%,1,2,12)} = 0.6070$ (tabla C.2), no se rechaza la hipótesis; es decir, se puede afirmar que los tres tratamientos están al mismo nivel para cada una de las medias. Como se advierte en la figura 3.10 los perfiles hasta la semana 6 se confunden un poco; la prueba estadística no detecta estas diferencias.

Para la prueba de "planitud" se tiene

$$
\begin{aligned}
T^2 &= qn(\boldsymbol{C}\bar{\boldsymbol{Y}}_{\cdot\cdot})'(\boldsymbol{CEC}'/q(n-1))^{-1}(\boldsymbol{C}\bar{\boldsymbol{Y}}_{\cdot\cdot}) \\
&= 15 \begin{pmatrix} -48.80 \\ -39.27 \\ 7.47 \\ -12.47 \\ -33.80 \end{pmatrix}' \begin{pmatrix} 714.5 & -13.2 & 207.5 & -219.9 & 270.2 \\ -13.2 & 298.1 & -174.9 & 221.0 & -216.0 \\ 07.5 & -174.9 & 645.3 & -240.8 & 165.8 \\ -219.9 & 221.0 & -240.8 & 1112.6 & -649.2 \\ 270.2 & -216.0 & 165.8 & -649.2 & -618.8 \end{pmatrix}^{-1} \begin{pmatrix} -48.80 \\ -39.27 \\ 7.47 \\ -12.47 \\ -33.80 \end{pmatrix} \\
&= 297.13.
\end{aligned}
$$

Como $297.13 > T^2_{(1\%,5,12)} = 49.739$ (tabla C.1), se rechaza la hipótesis de planitud.

3.5.5 Medidas repetidas en q–muestras

El diseño de medidas repetidas implica un modelo de una vía de clasificación de la forma $Y_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij}$. Desde los q–grupos, de n-observaciones cada uno, se calcula $\bar{\boldsymbol{Y}}_1, \ldots, \bar{\boldsymbol{Y}}_q$ y la matriz de errores $\boldsymbol{E}$. Los datos se disponen conforme a una tabla que contiene los factores A y B, en columnas y filas respectivamente y se consideran los siguientes tres casos: *El primero* considera cada uno de los niveles del factor B como grupo o población y se hace el análisis para las medidas repetidas ante los niveles del factor A (columnas); un *segundo* análisis es hecho entre los niveles del factor B (filas), y finalmente; un *tercer* análisis es desarrollado para verificar las interacciones entre columnas y filas. De esta forma se consigue un análisis semejante al que se desarrolla para un modelo de doble vía de clasificación. En la tabla (3.13) se tienen muestras sobre q poblaciones (factor B), las cuales consisten en p-medidas efectuadas en n–individuos diferentes para cada muestra, cada medida es la respuesta de un individuo ante un nivel del factor A (tratamiento). Así, el arreglo $(Y_{ij1}, Y_{ij2}, \ldots, Y_{ijp})'$ corresponde a las p medidas repetidas sobre el indivuo $j = 1, \ldots, n$ en la muestra (nivel del factor B) $i = 1, \ldots, q$.

Para verificar el efecto del factor A, dentro de cada uno de los sujetos, se comparan las medias de las variables $Y_1, \ldots, Y_p$ dentro del vector Y a través de las q-muestras. Se puede emplear la estadística T^2 como en el caso de una muestra (sección (3.4)). En el modelo $Y_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij}$, los vectores de medias $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_q$ corresponden a las medias en las q poblaciones, las cuales se estiman mediante $\bar{\boldsymbol{Y}}_{1\cdot}, \ldots, \bar{\boldsymbol{Y}}_{q\cdot}$. Para comparar las medias de $Y_1, \ldots, Y_p$ promediadas a través de las q muestras, se usa $\bar{\boldsymbol{\mu}}_{\cdot} = \sum_{i=1}^{q} \boldsymbol{\mu}_i/q$. La hipótesis $H_0 : \mu_{\cdot 1} = \cdots = \mu_{\cdot p}$, que contrasta la media de las respuestas ante los niveles del factor A (tratamientos), puede expresarse mediante contrastes así:

$$H_0 : \boldsymbol{C}\bar{\boldsymbol{\mu}} = \boldsymbol{0}, \tag{3.83}$$

donde $\boldsymbol{C}$ es una matriz de contrastes con tamaño $((p-1)\times p)$ y de rango completo; es decir, $\boldsymbol{C}\mathbf{1}=\mathbf{0}$. Esto equivale a probar la hipótesis de *"perfiles planos"*. Un estimador de $\boldsymbol{C}\bar{\boldsymbol{\mu}}$ es $\boldsymbol{C}\bar{\boldsymbol{Y}}_{\cdot\cdot}$, donde $\bar{\boldsymbol{Y}}_{\cdot\cdot}=\sum_{i=1}^{q}\bar{Y}_{i\cdot}/q$ es el vector de medias global. Bajo la hipótesis nula H_0, el vector $\boldsymbol{C}\bar{\boldsymbol{Y}}_{\cdot\cdot}$ se distribuye $N_{p-1}(\mathbf{0},\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}'/N)$ donde $N=\sum_{i=1}^{q}n_i$ para un diseño con estructura de datos desbalanceada y $N=qn$ para el caso balanceado. Se verifica la hipótesis nula mediante

$$T^2=N(\boldsymbol{C}\bar{\boldsymbol{Y}}_{\cdot\cdot})'(\boldsymbol{C}'\boldsymbol{S}_p\boldsymbol{C})^{-1}(\boldsymbol{C}\bar{\boldsymbol{Y}}_{\cdot\cdot}),$$

donde $\boldsymbol{S}_p=\boldsymbol{E}/(N-q)$. La anterior estadística T^2 se distribuye, bajo H_0, como $T^2_{(p-1,N-q)}$. Nótese que la dimensión de T^2 es $(p-1)$, pues $\boldsymbol{C}\bar{\boldsymbol{Y}}_{\cdot\cdot}$ es de tamaño $((p-1)\times 1)$.

Tabla 3.13 Medidas repetidas en q–grupos

		Factor A(Medidas repetidas)				
Factor B Grupos	Sujeto	A_1	A_2	$\cdots$	A_p	
B_1	S_{11}	$(Y_{111}$	Y_{112}	$\cdots$	$Y_{11p})$	$=Y'_{11}$
	S_{12}	$(Y_{121}$	Y_{122}	$\cdots$	$Y_{12p})$	$=Y'_{12}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	S_{1n}	$(Y_{1n1}$	Y_{1n2}	$\cdots$	$Y_{1np})$	$=Y'_{1n}$
B_2	S_{21}	$(Y_{211}$	Y_{212}	$\cdots$	$Y_{21p})$	$=Y'_{21}$
	S_{22}	$(Y_{221}$	Y_{222}	$\cdots$	$Y_{22p})$	$=Y'_{22}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	S_{2n}	$(Y_{2n1}$	Y_{2n2}	$\cdots$	$Y_{2np})$	$=Y'_{2n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_q	S_{q1}	$(Y_{q11}$	Y_{q12}	$\cdots$	$Y_{q1p})$	$=Y'_{q1}$
	S_{q2}	$(Y_{q21}$	Y_{q22}	$\cdots$	$Y_{q2p})$	$=Y'_{q2}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	S_{qn}	$(Y_{qn1}$	Y_{qn2}	$\cdots$	$Y_{qnp})$	$=Y'_{qn}$

Para comparar las medias de los q-niveles del factor B, se toman las medias en cada grupo. Éstas son el promedio sobre cada uno de los niveles del factor A; es decir, $\sum_{j=1}^{p}\mu_{ij}/p=\mathbf{1}'\boldsymbol{\mu}_i/p$. La hipótesis se escribe como

$$H_0:\ \mathbf{1}'\boldsymbol{\mu}_1=\cdots=\mathbf{1}'\boldsymbol{\mu}_q, \tag{3.84}$$

la cual es equivalente a probar que los perfiles fila están en el mismo nivel. Las expresiones $\mathbf{1}'\boldsymbol{\mu}_i$, para $i = 1, \ldots, q$ son escalares, luego esta hipótesis puede verificarse mediante la estadística F, como en un análisis de varianza univariado a una vía de clasificación sobre $Z_{ij} = \mathbf{1}Y_{ij}$, para $i = 1, \ldots, q$ y $j = 1, \ldots, n_i$. De esta manera, a cada sujeto S_{ij} se le hace corresponder el escalar Z_{ij}. Es decir, cada observación vectorial para cada sujeto o individuo se reduce a una observación de tipo escalar, luego, mediante un análisis de varianza univariado (ANDEVA) se comparan las medias $\mathbf{1}'\bar{Y}_{1.}, \ldots, \mathbf{1}'\bar{Y}_{q.}$.

La hipótesis sobre la interacción AB es equivalente a la hipótesis de "paralelismo" mostrada en el análisis de perfiles

$$H_0 : \boldsymbol{C}\boldsymbol{\mu}_1 = \cdots = \boldsymbol{C}\boldsymbol{\mu}_q. \tag{3.85}$$

Así, las diferencias o contrastes entre los niveles del factor A son los mismos a través de los niveles del factor B. Este resultado se prueba fácilmente mediante un análisis de varianza multivariado (ANAVAMU) a una vía de clasificación sobre $Z_{ij} = \boldsymbol{C}Y_{ij}$, con

$$\boldsymbol{\Lambda} = \frac{|\boldsymbol{C}\boldsymbol{E}\boldsymbol{C}'|}{|\boldsymbol{C}(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{C}'|},$$

la cual se distribuye como $\boldsymbol{\Lambda}_{(p-1,q-1,N-q)}$.

Observación:

El cálculo de las estadísticas de prueba para medidas repetidas puede hacerse mediante las matrices $\boldsymbol{H}$ y $\boldsymbol{E}$ del ANAVAMU. Otra forma consiste en transformar los datos de acuerdo con $Z_{ij} = \boldsymbol{C}Y_{ij}$. Para la hipótesis (3.84) asociada al factor A, por ejemplo para $p = 4$,

$$\boldsymbol{C} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

así, cada observación de $Y' = (Y_1, Y_2, Y_3, Y_4)$ se transforma por medio de $Z' = (Y_1 - Y_2, Y_2 - Y_3, Y_3 - Y_4)$. De esta forma se verifica la hipótesis $H_0 : \boldsymbol{\mu}_Z = 0$ mediante la estadística para una muestra

$$T^2 = N\bar{Z}'\boldsymbol{S}_Z^{-1}\bar{Z}$$

con $N = \sum_{i=1} qn_i$, $\bar{Z} = \sum_{ij} Z_{ij}/N$ y $\boldsymbol{S}_Z = \boldsymbol{E}_Z/(N-q)$. Se rechaza la hipótesis H_0 si $T^2 \geq T^2_{(\alpha,p-1,N-q)}$.

Para verificar la hipótesis (3.85) en el factor B, se suman las componentes de cada vector de observaciones, se obtiene

$$Z_{ij} = \mathbf{1}'Y_{ij} = Y_{ij1} + \cdots + Y_{ijp},$$

luego se comparan las medias $\bar{Z}_1, \ldots, \bar{Z}_q$ mediante una estadística F en un ANDEVA a una vía de clasificación.

Para la hipótesis (3.86), de interacción entre los factores A y B, se transforma cada Y_{ij} en $Z_{ij} = \boldsymbol{C}Y_{ij}$, empleando las filas de la matriz $\boldsymbol{C}$ anterior. El vector Z_{ij} resultante es un vector de tamaño $(p-1) \times 1$. Así, se debe hacer un ANAVAMU sobre Z_{ij} para obtener

$$\boldsymbol{\Lambda} = \frac{|\boldsymbol{E}_Z|}{|\boldsymbol{E}_Z + \boldsymbol{H}_Z|}.$$

▶ Medidas repetidas con dos factores dentro de sujetos y un factor entre sujetos

Este modelo corresponde a un diseño de una vía de clasificación multivariada, en la cual cada vector de observaciones incluye medidas de un arreglo de tratamientos tipo factorial a dos vías. Cada sujeto recibe todos los tratamientos, los cuales corresponden a las combinaciones de los niveles de los dos factores A y B. Los niveles del factor entre sujetos (C) determinan los grupos de sujetos, a los cuales se les aplican los tratamientos resultantes de los dos factores A y B.

En la tabla 3.14 cada vector Y_{ij}, que identifica al sujeto S_{ij}, tiene nueve elementos, los cuales corresponden a los tratamientos:

$$A_1B_1,\ A_1B_2,\ A_1B_3,\ A_2B_1,\ A_2B_2,\ A_2B_3,\ A_3B_1,\ A_3B_2 \text{ y } A_3B_3.$$

El interés se dirige a probar una hipótesis semejante a la que se prueba en diseños de *"parcelas divididas"*, pero ahora en versión multivariada. El modelo para estas observaciones es de la forma

$$Y_{ij} = \boldsymbol{\mu} + \boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij}, \tag{3.86}$$

donde $\boldsymbol{\gamma}_i$ es el efecto debido al i-ésimo nivel del factor C.

Tabla 3.14 Medidas repetidas con dos factores "dentro" y un factor "entre" sujetos

		Factores dentro de sujetos (A y B)								
Entre Suj.		A_1			A_2			A_3		
(C)	Obs.	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
C_1	$Y_{11} =$	$(Y_{111}$	Y_{112}	Y_{113}	Y_{114}	Y_{115}	Y_{116}	Y_{117}	Y_{118}	$Y_{119})$
	$Y_{12} =$	$(Y_{121}$	Y_{122}	Y_{123}	Y_{124}	Y_{125}	Y_{126}	Y_{127}	Y_{128}	$Y_{129})$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$Y_{1n_1} =$	$(Y_{1n_11}$	Y_{1n_12}	Y_{1n_13}	Y_{1n_14}	Y_{1n_15}	Y_{1n_16}	Y_{1n_17}	Y_{1n_18}	$Y_{1n_19})$
C_2	$Y_{21} =$	$(Y_{211}$	Y_{212}	Y_{213}	Y_{214}	Y_{215}	Y_{216}	Y_{217}	Y_{218}	$Y_{219})$
	$Y_{22} =$	$(Y_{221}$	Y_{222}	Y_{223}	Y_{224}	Y_{225}	Y_{226}	Y_{127}	Y_{228}	$Y_{229})$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$Y_{2n_1} =$	$(Y_{2n_11}$	Y_{2n_12}	Y_{2n_13}	Y_{2n_14}	Y_{2n_15}	Y_{2n_16}	Y_{2n_17}	Y_{2n_18}	$Y_{2n_19})$
C_3	$Y_{31} =$	$(Y_{311}$	Y_{312}	Y_{313}	Y_{314}	Y_{315}	Y_{316}	Y_{317}	Y_{318}	$Y_{319})$
	$Y_{32} =$	$(Y_{321}$	Y_{322}	Y_{323}	Y_{324}	Y_{325}	Y_{326}	Y_{327}	Y_{328}	$Y_{329})$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$Y_{3n_1} =$	$(Y_{3n_11}$	Y_{3n_12}	Y_{3n_13}	Y_{3n_14}	Y_{3n_15}	Y_{3n_16}	Y_{3n_17}	Y_{3n_18}	$Y_{3n_19})$

Para verificar hipótesis sobre el factor A, el factor B y la interacción AB, se emplean contrastes entre los Y_{ij}. Algunos de estos contrastes, por ejemplo, se presentan a través de las siguientes matrices

$$\boldsymbol{A} = \begin{pmatrix} 2 & 2 & 2 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \end{pmatrix},$$

$$\boldsymbol{B} = \begin{pmatrix} 2 & -1 & -1 & 2 & -1 & -1 & 2 & -1 & -1 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \end{pmatrix},$$

$$\boldsymbol{P} = \begin{pmatrix} 4 & -2 & -2 & -2 & 1 & 1 & -2 & 1 & 1 \\ 0 & 2 & -2 & 0 & -1 & 1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 2 & -1 & -1 & -2 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix}.$$

Las filas de la matriz $\boldsymbol{A}$ corresponden a contrastes ortogonales entre los niveles del factor A, los cuales comparan, los siguientes niveles:

i) El nivel A_1 frente a los niveles A_2 y A_3 conjuntamente, y

ii) El nivel A_2 frente al nivel A_3.

En forma semejante, las filas de la matriz $\boldsymbol{B}$ contienen los contrastes

i) El nivel B_1 frente a los niveles B_2 y B_3 conjuntamente, y

ii) El nivel B_2 frente al nivel B_3.

Se advierte que es posible construir otros contrastes ortogonales para el factor A y el factor B. La matriz $\boldsymbol{P}$ está asociada con las interacciones entre los dos factores, y se obtiene como el producto entre los respectivos elementos de las filas de la matriz $\boldsymbol{A}$ y los de las filas de $\boldsymbol{B}$.

Como en el caso anterior, se calcula $\bar{Y}_{\cdot\cdot} = \sum_{ij} Y_{ij}/N$, $\boldsymbol{S}_p = \boldsymbol{E}/(N-q)$, $N = \sum_i n_i$. Si el factor C tiene q niveles con medias $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_q$, entonces $\bar{\boldsymbol{\mu}} = \sum_i \boldsymbol{\mu}_i/k$, los efectos principales de A asociados con $H_0: \ \boldsymbol{A}\bar{\boldsymbol{\mu}} = 0$, se verifican con la siguiente estadística

$$T_A^2 = N(\boldsymbol{A}\bar{Y}_{\cdot\cdot})'(\boldsymbol{A}'\boldsymbol{S}_p\boldsymbol{A})^{-1}(\boldsymbol{A}\bar{Y}_{\cdot\cdot}), \tag{3.87}$$

la cual se distribuye como $T^2_{(2,N-q)}$, donde 2 corresponde al número de filas de la matriz $\boldsymbol{A}$.

Las hipótesis $H_0: \ \boldsymbol{B}\bar{\boldsymbol{\mu}} = 0$ y $H_0: \ \boldsymbol{P}\bar{\boldsymbol{\mu}} = 0$, para los efectos principales de B y las interacciones entre A y B, se verifican de manera similar con las estadísticas

$$T_B^2 = N(\boldsymbol{B}\bar{Y}_{\cdot\cdot})'(\boldsymbol{B}'\boldsymbol{S}_p\boldsymbol{B})^{-1}(\boldsymbol{B}\bar{Y}_{\cdot\cdot}), \text{ y} \tag{3.88}$$

$$T_{AB}^2 = N(\boldsymbol{P}\bar{Y}_{\cdot\cdot})'(\boldsymbol{P}'\boldsymbol{S}_p\boldsymbol{P})^{-1}(\boldsymbol{P}\bar{Y}_{\cdot\cdot}), \tag{3.89}$$

las cuales se distribuyen como $T^2_{(2,N-q)}$ y $T^2_{(4,N-q)}$, respectivamente. En general, si el factor A tiene a niveles y el factor B tiene b niveles, entonces las matrices de contrastes $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{P}$ tienen $(a-1)$, $(b-1)$ y $(a-1)(b-1)$ filas, respectivamente. Las estadísticas de prueba se distribuyen, en general, como $T^2_{(a-1,N-q)}$, $T^2_{(b-1,N-q)}$ y $T^2_{((a-1)(b-1),N-q)}$, respectivamente.

Una prueba alternativa, para los efectos principales A y B y la interacción entre éstos es el lambda de Wilks ($\boldsymbol{\Lambda}$). Se particiona la "suma de cuadrados total" como: $\sum_{ij} Y_{ij}Y_{ij}' = \boldsymbol{E}+(\boldsymbol{H}+\boldsymbol{H}^*)$, donde $\boldsymbol{H}^* = N\bar{Y}_{\cdot\cdot}\bar{Y}_{\cdot\cdot}'$. La hipótesis de interés es $H_{0A}: \ \boldsymbol{A}\bar{\boldsymbol{\mu}} = 0$, la cual se contrasta mediante la estadística

$$\boldsymbol{\Lambda}_A = \frac{|\boldsymbol{A}\boldsymbol{E}\boldsymbol{A}'|}{|\boldsymbol{A}(\boldsymbol{E}+\boldsymbol{H}^*)\boldsymbol{A}'|}, \tag{3.90}$$

la cual, bajo H_0, se distribuye como $\boldsymbol{\Lambda}_{(a-1,1,N_q)}$, con a el número de niveles del factor A. Nótese que la dimensión es $(a-1)$ porque la matriz $\boldsymbol{A}\boldsymbol{E}\boldsymbol{A}'$

es de tamaño $((a-1)\times(a-1))$. Estadísticas similares se obtienen para verificar los efectos del factor B y las interacciones entre A y B.

Los efectos principales del factor C, como en el caso de *medidas repetidas en q muestras*, son equivalentes a verificar la hipótesis

$$H_{C_0}:\ \mathbf{1}'\boldsymbol{\mu}_1 = \cdots = \mathbf{1}'\boldsymbol{\mu}_q,$$

al igual que la hipótesis planteada en la igualdad (3.85), ésta se verifica con una estadística F univariada sobre los $Z_{ij} = \mathbf{1}'Y_{ij}$, en la forma de un ANDEVA a una vía de clasificación.

Las interacciones tipo AC, BC y ABC se prueban en la forma siguiente:

- *Interacción AC.* La interacción AC equivale a la hipótesis

  $$H_{AC_0}:\ \boldsymbol{A\mu}_1 = \cdots = \boldsymbol{A\mu}_q,$$

  la cual establece que los contrastes en el factor A son los mismos a través de todos los q niveles del factor C. Una estadística para verificar esta hipótesis es

  $$\boldsymbol{\Lambda}_{AC} = \frac{|\boldsymbol{AEA}'|}{|\boldsymbol{A}(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{A}'|} \tag{3.91}$$

  la cual se distribuye como $\boldsymbol{\Lambda}_{(a-1,q-1,N-q)}$. La hipótesis anterior se puede contrastar a través de un ANAVAMU para un modelo a una vía de clasificación, sobre los vectores de observaciones transformados a $Z_{ij} = \boldsymbol{A}Y_{ij}$.

- *Interacción BC.* La interacción BC se expresa a través de la hipótesis

  $$H_{BC_0}:\ \boldsymbol{B\mu}_1 = \cdots = \boldsymbol{B\mu}_q,$$

  la cual se verifica a través de la estadística

  $$\boldsymbol{\Lambda}_{BC} = \frac{|\boldsymbol{BEB}'|}{|\boldsymbol{B}(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{B}'|} \tag{3.92}$$

  que se distribuye como $\boldsymbol{\Lambda}_{(b-1,q-1,N-q)}$. También se puede verificar con un ANAVAMU sobre los $Z_{ij} = \boldsymbol{B}Y_{ij}$.

- *Interacción ABC.* La interacción ABC se expresa mediante la hipótesis

  $$H_{ABC_0}:\ \boldsymbol{P\mu}_1 = \cdots = \boldsymbol{P\mu}_q,$$

la cual se contrasta mediante la estadística

$$\boldsymbol{\Lambda}_{ABC} = \frac{|\boldsymbol{PEP}'|}{|\boldsymbol{P}(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{P}'|} \tag{3.93}$$

que se distribuye como $\boldsymbol{\Lambda}_{((a-1)(b-1),\ q-1,\ N-q)}$. También se puede verificar con un ANAVAMU sobre los $Z_{ij} = \boldsymbol{P}Y_{ij}$.

Las pruebas sobre los contrastes AC, BC o ABC se pueden desarrollar a través de los valores propios de las matrices asociadas a "covariación entre" y la "covariación dentro". Así por ejemplo, para la interacción tipo AC se obtienen los valores propios de la matriz $(\boldsymbol{AEA}')^{-1}(\boldsymbol{AHA}')$, y con ellos se calculan estadísticas como la traza de Lawley–Hotelling, la traza de Bartlett–Nanda-Pillai o el máximo valor propio de Roy.

Ejemplo 3.12 Los datos de la tabla 3.15 representan medidas repetidas correspondientes a un diseño con dos factores dentro de los sujetos y un factor entre los mismos. Como los factores se ajustan a la tabla 3.14 anterior, se pueden emplear las matrices $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{P}$ mostradas anteriormente.
El vector de medias general es

$$\bar{Y}'_{\cdot\cdot} = (46.45, 39.25, 31.70, 38.85, 45.40, 40.15, 34.55, 36.90, 39.15).$$

La prueba para el factor A está dada por la estadística (3.88), así:

$$\begin{aligned} T_A^2 &= N(\boldsymbol{A}\bar{Y}_{\cdot\cdot})'(\boldsymbol{A}'\boldsymbol{S}_p\boldsymbol{A})^{-1}(\boldsymbol{A}\bar{Y}_{\cdot\cdot}) \\ &= 20(-0.20, 13.80)\begin{pmatrix} 2138.4 & 138.6 \\ 138.6 & 450.4 \end{pmatrix}^{-1}\begin{pmatrix} -0.20 \\ 13.80 \end{pmatrix} = 8.645. \end{aligned}$$

Como el valor de $T_A^2 = 8.645 > T^2_{(0.05,2,18)} = 7.606$ (de la tabla C.1), se concluye que hay diferencia entre los niveles del factor A.

Para verificar la significancia del factor B, se emplea la estadística (3.89), resulta

$$\begin{aligned} T_B^2 &= N(\boldsymbol{B}\bar{Y}_{\cdot\cdot})'(\boldsymbol{B}'\boldsymbol{S}_p\boldsymbol{B})^{-1}(\boldsymbol{B}\bar{Y}_{\cdot\cdot}) \\ &= 20(7.15, 10.55)\begin{pmatrix} 305.7 & 94.0 \\ 94.0 & 69.8 \end{pmatrix}^{-1}\begin{pmatrix} 7.15 \\ 10.55 \end{pmatrix} = 37.438. \end{aligned}$$

De la tabla C.1, se obtiene que $T^2_{(1\%,2,18)} = 12.943$, se concluye entonces que el factor B influye significativamente en las respuestas, pues el valor de la estadística $T_B^2 = 37.438 > 12.943$.

Tabla 3.15 Datos con dos factores dentro y un factor entre sujetos

		Factores dentro de sujetos (A y B)								
Entre suj.		A_1			A_2			A_3		
(C)	Obs.	B_1	B_2	B_3	B_1	B_2	B_3	B_1	B_2	B_3
C_1	Y_{11}	20	21	21	32	42	37	32	32	32
	Y_{12}	67	48	29	43	56	48	39	40	41
	Y_{13}	37	31	25	27	28	30	31	33	34
	Y_{14}	42	40	38	37	36	28	19	27	35
	Y_{15}	57	45	32	27	21	25	30	29	29
	Y_{16}	39	39	38	46	54	43	31	29	28
	Y_{17}	43	32	20	33	46	44	42	37	31
	Y_{18}	35	34	34	39	43	39	35	39	42
	Y_{19}	41	32	23	37	51	39	27	28	30
	$Y_{1,10}$	39	32	24	30	35	31	26	29	32
C_2	Y_{21}	47	36	25	31	36	29	21	24	27
	Y_{22}	53	43	32	40	48	47	46	50	54
	Y_{23}	38	35	33	38	42	45	48	48	49
	Y_{24}	60	51	41	54	67	60	53	52	50
	Y_{25}	37	36	35	40	45	40	34	40	46
	Y_{26}	59	48	37	45	52	44	36	44	52
	Y_{27}	67	50	33	47	61	46	31	41	50
	Y_{28}	43	35	27	32	36	35	33	33	32
	Y_{29}	64	59	53	58	62	51	40	42	43
	$Y_{2,10}$	41	38	34	41	47	42	37	41	46

Fuente: Rencher (1995, pág. 240)

Para verificar la interacción AB, la estadística dada en (3.90) y calculada con estos datos toma el valor

$$T^2_{AB} = N(\boldsymbol{P}\bar{Y}_{\cdot\cdot})'(\boldsymbol{P}'\boldsymbol{S}_p\boldsymbol{P})^{-1}(\boldsymbol{P}\bar{Y}_{\cdot\cdot}) = 61.825,$$

la cual es mayor que $T^2_{(1\%,4,18)} = 23.487$ (tabla C.1).

Para verificar la significancia del factor C, se desarrolla un ANDEVA sobre los datos transformados a $Z_{ij} = \mathbf{1}'Y_{ij}/9$. La tabla que resulta es la siguiente

El valor de $F_{(1\%,1,18)} \approx 8.29$ (tabla C.8), luego como $F = 8.54 > 8.29$, se concluye que el factor C es significativo.

Fuente de var.	Suma de Cuad.	GL	Cuadrado medio	F
Entre grupos (C)	3042.22	1	3042.22	8.54
Error	6408.98	18	356.05	

Para calcular las estadísticas con las que se verifican las interacciones AC, BC y ABC es necesario calcular las matrices $\boldsymbol{E}$ y $\boldsymbol{H}$, las cuales son de tamaño (9×9). No se presentan estas matrices de manera explícita sino los resultados intermedios y finales asociados a éstas.

Para contrastar la hipótesis de interacción AC se calcula la estadística

$$\boldsymbol{\Lambda}_{AC} = \frac{|\boldsymbol{AEA'}|}{|\boldsymbol{A}(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{A'}|} = \frac{3.058 \times 10^8}{3.092 \times 10^8} = 0.9889.$$

De la tabla C.2 la estadística $\boldsymbol{\Lambda}_{(5\%,2,1,18)} = 0.703$, como el valor observado de la estadística es $\boldsymbol{\Lambda}_{AC} = 0.9889 > 0.703$, no se rechaza la hipótesis de no interacción entre los factores A y C sobre estas respuestas.

Para la interacción BC, la estadística evaluada en los datos es

$$\boldsymbol{\Lambda}_{BC} = \frac{|\boldsymbol{BEB'}|}{|\boldsymbol{B}(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{B'}|} = \frac{4.053 \times 10^6}{4.170 \times 10^6} = 0.9718.$$

Como $\boldsymbol{\Lambda}_{BC} = 0.9718 > 0.703$ (tabla C.2), se concluye que la interacción entre los factores B y C no es significativa.

Para la interacción ABC, se evalúa la estadística

$$\begin{aligned}\boldsymbol{\Lambda}_{ABC} &= \frac{|\boldsymbol{PEP'}|}{|\boldsymbol{P}(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{P'}|} \\ &= \frac{2.643 \times 10^{12}}{2.927 \times 10^{12}} = 0.9029.\end{aligned}$$

De acuerdo con la tabla C.2, $\boldsymbol{\Lambda}_{(5\%,4,1,18)} = 0.551$, y como el valor observado de la estadística es $\boldsymbol{\Lambda}_{ABC} = 0.9029 > 0.551$, se concluye que la interacción entre los factores A, B y C no es significativa. ☑

3.5.6 Curvas de crecimiento

Los modelos de *curvas de crecimiento* se consideran para datos registrados en varias ocasiones, sobre individuos que reciben diferentes tratamientos o que están divididos en varios grupos o clases, en las cuales cada registro se

conforma por medidas sobre un número de variables generalmente correlacionadas. Este caso es muy común cuando a un individuo se le hace un seguimiento durante un período de tiempo. Se considera el problema de estimación y prueba de hipótesis sobre la forma de la curva para el caso de una o varias muestras.

▶ Curvas de crecimiento en una muestra

Los datos para curvas de crecimiento de una muestra tienen una estructura semejante a la presentada en la tabla 3.13 para medidas repetidas, donde los niveles del factor A corresponden a los períodos de tiempo. La aproximación o ajuste de la curva se hace a través de un polinomio en función del tiempo. Si los períodos de tiempo están igualmente espaciados, la aproximación se puede hacer mediante *polinomios ortogonales*; cuando los períodos no son de igual longitud se emplea el método que se explica más adelante.

∘ *Un polinomio ortogonal* es un caso especial de contraste, empleado para verificar tendencias de orden lineal, cuadrático o superior en factores cuantitativos. Se presenta esta metodología mediante el estudio de un caso particular[4]. Supóngase que se suministra una droga a un grupo de pacientes y se observa su reacción cada 3 minutos en los tiempos 0, 3, 6, 9, 12 minutos, respectivamente ($p = 5$). Sean μ_1, μ_2, μ_3 μ_4 y μ_5 las medias de las respectivas respuestas . Para verificar la hipótesis de que no hay tendencia en las μ_i (perfiles, planos u horizontales); es decir, $H_0 : \ \mu_1 = \cdots = \mu_5$ se emplea la matriz de contrastes

$$\boldsymbol{C} = \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ 2 & -1 & -2 & -1 & 2 \\ -1 & 2 & 0 & -2 & 1 \\ 1 & -4 & 6 & -4 & 1 \end{pmatrix}.$$

Las filas de esta matriz corresponden a los coeficientes de los polinomios en la variable t, las cuales son ortogonales. Cada uno de estos polinomios prueba la tendencia lineal, cuadrática, cúbica o de cuarto grado en las medias. Se trata de encontrar algunas filas de la matriz $\boldsymbol{C}$ que se ajusten a la forma de la curva de respuesta.

Se han elaborado tablas que contienen los coeficientes asociados a los términos de cada polinomio. La tabla C.4 contiene los coeficientes hasta para $p = 10$ períodos o tratamientos asociados al tiempo de polinomios hasta de grado $(p - 1) = 9$.

[4]Rencher (1995), págs. 243-253.

Igual que en los contrastes ortogonales, cada fila de la matriz $\boldsymbol{C}$ suman cero y son mutuamente ortogonales. En cada fila los elementos están de acuerdo con el patrón mostrado por la media de las respuestas en cada punto del tiempo; es decir, crecen o decrecen. La primera fila de la matriz $\boldsymbol{C}$ los coeficientes $(-2,-1,0,1,2)$ crecen regularmente conforme en una tendencia en línea recta. Los de la segunda fila bajan y suben sobre una parábola. En la tercera fila se da un ascenso, luego un descenso profundo y luego un ascenso en una trayectoria cúbica de dos ramas. Finalmente, en la última fila los coeficientes se "curvan" tres veces siguiendo una curva de cuarto grado.

Para entender de qué manera los polinomios ortogonales reflejan la tendencia de las medias, considérense los siguientes tres patrones de medias: $\boldsymbol{\mu}_a' = (8,8,8,8,8)$, $\boldsymbol{\mu}_b' = (20,16,12,8,4)$ y $\boldsymbol{\mu}_c' = (5,12,15,12,5)$. Las filas de $\boldsymbol{C}$ se denotan por c_1', c_2', c_3' y c_4'. Se observa que $c_i\boldsymbol{\mu}_a = 0$ para $i = 1,2,3,4$. Si $\boldsymbol{\mu}$ es del tipo $\boldsymbol{\mu}_b$ anterior, solamente $c_1'\boldsymbol{\mu}_b$ es diferente de cero. Las otras filas no son sensibles a esta tendencia lineal, así,

$$\begin{aligned}
c_1'\boldsymbol{\mu}_b &= (-2)(20) + (-1)(16) + (0)(12) + (1)(8) + (2)(4) = -44\\
c_2'\boldsymbol{\mu}_b &= (2)(20) + (-1)(16) + (-2)(12) + (-1)(8) + (2)(4) = 0\\
c_3'\boldsymbol{\mu}_b &= (-1)(20) + (2)(16) + (0)(12) + (-2)(8) + (1)(4) = 0\\
c_4'\boldsymbol{\mu}_b &= (1)(20) + (-4)(16) + (6)(12) + (-4)(8) + (1)(4) = 0.
\end{aligned}$$

De esta manera, el polinomio dado por la primera fila de la matriz $\boldsymbol{C}$ se ajusta a la tendencia observada por las medias; es decir, la lineal.

La tendencia mostrada por $\boldsymbol{\mu}_c'$; es cuadrática, pues únicamente $c_2\boldsymbol{\mu}_c'$ es diferente de cero. Por ejemplo,

$$\begin{aligned}
c_1'\boldsymbol{\mu}_c &= (-2)(5) + (-1)(12) + (0)(15) + (1)(12) + (2)(5) = 0\\
c_2'\boldsymbol{\mu}_c &= (2)(5) + (-1)(12) + (-2)(15) + (-1)(12) + (2)(5) = -34.
\end{aligned}$$

Así, estos polinomios ortogonales siguen la trayectoria requerida. Cada uno de manera independiente detecta un tipo de curvatura y es diseñado para ignorar los otros tipos de tendencia. Naturalmente los datos experimentales no se comportan tan "juiciosamente" como los de este ejemplo, estos suelen mostrar curvaturas mezcladas. En la práctica el contraste dado por más de un polinomio ortogonal puede resultar significativo.

Para verificar hipótesis sobre la forma de la curva, se emplean algunas filas de la matriz $\boldsymbol{C}$. Para el caso de que se trata, supóngase que se tienen elementos suficientes para suponer que la curva tiene tendencia lineal y cuadrática combinadas. Así, la matriz $\boldsymbol{C}$ queda particionada como

$$\boldsymbol{C}_1 = \begin{pmatrix} c_1' \\ c_2' \end{pmatrix} = \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ 2 & -1 & -2 & -1 & 2 \end{pmatrix}$$

y

$$\boldsymbol{C}_2 = \begin{pmatrix} c_3' \\ c_4' \end{pmatrix} = \begin{pmatrix} -1 & 2 & 0 & -2 & 1 \\ 1 & -4 & 6 & -4 & 1 \end{pmatrix}.$$

La hipótesis $H_0: \boldsymbol{C}_1\boldsymbol{\mu} = 0$ se verifica mediante la estadística

$$T^2 = n(\boldsymbol{C}_1\bar{Y})'(\boldsymbol{C}_1\boldsymbol{S}\boldsymbol{C}_1')^{-1}(\boldsymbol{C}_1\bar{Y}),$$

la cual se distribuye como $T^2_{(2,n-1)}$, donde 2 corresponde al número de filas de $\boldsymbol{C}_1$ y n el número de sujetos de la muestra, $\bar{Y}$ el vector de medias y $\boldsymbol{S}$ la matriz de covarianzas muestral. Análogamente, la hipótesis $H_0: \boldsymbol{C}_2\boldsymbol{\mu} = 0$ se contrasta a través de

$$T^2 = n(\boldsymbol{C}_2\bar{Y})'(\boldsymbol{C}_2\boldsymbol{S}\boldsymbol{C}_2')^{-1}(\boldsymbol{C}_2\bar{Y}),$$

la cual se distribuye como $T^2_{(2,n-1)}$. Se espera rechazar la primera hipótesis y no rechazar la segunda.

Cuando no se tienen indicios o supuestos con relación a la forma de la curva, se debe proceder a realizar una prueba general del tipo $H_0: \boldsymbol{C}\boldsymbol{\mu} = 0$, si se rechaza esta hipótesis, se deben hacer pruebas sobre las filas o un grupo de filas de la matriz $\boldsymbol{C}$ separadamente. La estadística para contrastar esta hipótesis es

$$T^2 = n(\boldsymbol{C}\bar{Y})'(\boldsymbol{C}\boldsymbol{S}\boldsymbol{C}')^{-1}(\boldsymbol{C}\bar{Y}),$$

que se distribuye como $T^2_{(4,n-1)}$. Las pruebas sobre cada fila de $\boldsymbol{C}$ (polinomio), del tipo $c_i'\boldsymbol{\mu} = 0$, se hacen mediante

$$t_i = \frac{c_i'\bar{Y}}{\sqrt{c_i'\boldsymbol{S}c_i/n}}, \quad \text{para } i = 1, 2, 3, 4,$$

esta estadística se distribuye como una t-Student con $(n-1)$ grados de libertad.

◦ Ahora se considera el caso de puntos en el tiempo con *separación distinta*; es decir, períodos de longitud diferente. Supóngase que se observa una respuesta de un sujeto en los tiempos $t_1, \ldots, t_p$, y que la media de la respuesta $\boldsymbol{\mu}$, en cualquier punto del tiempo t, es un polinomio sobre t de grado $k < p$; es decir,

$$\mu = \beta_0 + \beta_1 t + \beta_2 t^2 \cdots + \beta_k t^k.$$

Esto se tiene para cada punto t_i con respuesta media μ_i. La hipótesis es entonces

$$H_0: \begin{pmatrix} \mu_1 = \beta_0 + \beta_1 t_1 + \beta_2 t_1^2 + \cdots + \beta_k t_1^k \\ \mu_2 = \beta_0 + \beta_1 t_2 + \beta_2 t_2^2 + \cdots + \beta_k t_2^k \\ \vdots \\ \mu_p = \beta_0 + \beta_1 t_p + \beta_2 t_p^2 + \cdots + \beta_k t_p^k \end{pmatrix},$$

que equivale a

$$H_0 : \ \boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\beta},$$

con

$$\boldsymbol{A} = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^k \\ 1 & t_2 & t_2^2 & \cdots & t_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & t_p^2 & \cdots & t_p^k \end{pmatrix}, \quad \text{y} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}. \tag{3.94}$$

El modelo $\boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\beta}$ es similar a un modelo de regresión lineal $\mathcal{E}(Y) = \boldsymbol{X}\boldsymbol{\beta}$. De manera análoga con la regresión lineal, se debe encontrar el valor de $\hat{\boldsymbol{\beta}}$ que haga mínima la distancia (tipo Mahalanobis) entre las observaciones y el modelo supuesto; esto es: $(\bar{Y} - \boldsymbol{A}\boldsymbol{\beta})'\boldsymbol{S}^{-1}(\bar{Y} - \boldsymbol{A}\boldsymbol{\beta})$. Después de aplicar cálculo diferencial se encuentra que el "óptimo" viene dado por:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}'\boldsymbol{S}^{-1}\boldsymbol{A})^{-1}(\boldsymbol{A}'\boldsymbol{S}^{-1}\bar{Y}).$$

Así, $H_0 : \ \boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\beta}$ se verifica a través de la estadística

$$\begin{aligned} T^2 &= n(\bar{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}})'\boldsymbol{S}^{-1}(\bar{Y} - \boldsymbol{A}\hat{\boldsymbol{\beta}}) \\ &= n(\bar{Y}'\boldsymbol{S}^{-1}\bar{Y} - \bar{Y}'\boldsymbol{S}^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}), \end{aligned} \tag{3.95}$$

la cual tiene distribución $T_{(p-k-1,n-1)}$.

▶ Curvas de crecimiento en q–muestras

Para varias muestras o grupos, los datos tienen la estructura que se muestra en la tabla 3.13, donde los p-niveles del factor A representan puntos en el tiempo. Es decir, se tienen $Y_{i1}, \ldots, Y_{in_i}$ vectores de p-medidas sobre n_i sujetos en el grupo i, para $i = 1, \ldots, q$.

Si los puntos en el tiempo están igualmente espaciados, se pueden emplear polinomios ortogonales en la matriz de contrastes $\boldsymbol{C}$ de tamaño $(p-1) \times p$ para expresar la hipótesis de la forma $\boldsymbol{C}\boldsymbol{\mu}_{\cdot} = \boldsymbol{0}$, donde $\boldsymbol{\mu}_{\cdot} = \sum_{i=1}^{q} \boldsymbol{\mu}_i/q$. Se denotan la medias muestrales de cada grupo por $\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{q\cdot}$, la media global por $\bar{Y}_{\cdot\cdot}$ y la matriz de covarianzas conjunta por $\boldsymbol{S}_p = \boldsymbol{E}/(N-q)$. La hipótesis $\boldsymbol{C}\boldsymbol{\mu}_{\cdot} = \boldsymbol{0}$, de no diferencia entre las medias $\mu_1, \ldots, \mu_p$, promediadas a través de los q grupos, se verifica con

$$T^2 = N(\boldsymbol{C}\bar{Y}_{\cdot\cdot})'(\boldsymbol{C}\boldsymbol{S}_p\boldsymbol{C}')^{-1}(\boldsymbol{C}\bar{Y}_{\cdot\cdot}), \tag{3.96}$$

la cual tiene distribución $T^2_{(p-1,N-q)}$, con $N = \sum_{i=1}^{q} n_i$. Una prueba que el promedio, sobre los grupos, de curvas de crecimiento tiene una forma

particular se puede desarrollar con una matriz $\boldsymbol{C}_1$ que contenga algunas filas de la matriz $\boldsymbol{C}$, mediante

$$T^2 = N(\boldsymbol{C}_1\bar{Y}_{\cdot\cdot})'(\boldsymbol{C}_1\boldsymbol{S}_p\boldsymbol{C}_1')^{-1}(\boldsymbol{C}_1\bar{Y}_{\cdot\cdot}), \tag{3.97}$$

cuya distribución es $T^2_{(r,N-q)}$, con r el número de filas de la matriz $\boldsymbol{C}_1$.

Las curvas de crecimiento para varios grupos pueden compararse a través de la prueba para interacción o paralelismo usando $\boldsymbol{C}$ o $\boldsymbol{C}_1$. Se desarrolla un ANAVAMU sobre los $\boldsymbol{C}Y_{ij}$ o sobre los $\boldsymbol{C}_1Y_{ij}$ a través de las estadísticas

$$\boldsymbol{\Lambda} = \frac{|\boldsymbol{C}\boldsymbol{E}\boldsymbol{C}'|}{|\boldsymbol{C}(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{C}'|} \quad \text{o} \quad \boldsymbol{\Lambda}_1 = \frac{|\boldsymbol{C}_1\boldsymbol{E}\boldsymbol{C}_1'|}{|\boldsymbol{C}_1(\boldsymbol{E}+\boldsymbol{H})\boldsymbol{C}_1'|},$$

las cuales se distribuyen $\boldsymbol{\Lambda}_{(p-1,q-1,N-q)}$ y $\boldsymbol{\Lambda}_{(r,q-1,N-q)}$, respectivamente.

Cuando los puntos en el tiempo no están igualmente espaciados, se procede conforme al caso de una muestra con el ajuste de polinomios de grado k (con $k < p$). Supóngase que todos los vectores Y_{ij}, con $i = 1, \ldots, q$, $j = 1, \ldots, n_i$, tienen la misma matriz de covarianzas $\boldsymbol{\Sigma}$. Si un polinomio de grado k se ajusta a la curva de crecimiento, se tiene una representación matricial semejante a la expresada en (3.95); es decir,

$$\boldsymbol{A} = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^k \\ 1 & t_2 & t_2^2 & \cdots & t_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & t_p^2 & \cdots & t_p^k \end{pmatrix} \quad \text{y} \quad \boldsymbol{\beta}_i = \begin{pmatrix} \boldsymbol{\beta}_{i0} \\ \boldsymbol{\beta}_{i1} \\ \vdots \\ \boldsymbol{\beta}_{ip} \end{pmatrix}. \tag{3.98}$$

Un estimador de $\boldsymbol{\beta}_i$ es

$$\hat{\boldsymbol{\beta}}_i = (\boldsymbol{A}'\boldsymbol{S}_p^{-1}\boldsymbol{A})^{-1}(\boldsymbol{A}'\boldsymbol{S}_p^{-1}\bar{Y}), \tag{3.99}$$

donde

$$\boldsymbol{S}_p = \frac{1}{N-q}\big((n_1-1)\boldsymbol{S}_1 + \cdots + (n_q-1)\boldsymbol{S}_q\big) = \frac{1}{N-q}\boldsymbol{E},$$

es el estimador de la matriz de covarianzas común $\boldsymbol{\Sigma}$, con $N = \sum_{i=1}^{q} n_i$, .

Una estadística tipo lambda de Wilks, para verificar que un polinomio de grado k se ajusta adecuadamente a las curvas de crecimiento de las p variables, se obtiene mediante la razón de máxima verosimilitud. Ésta es

$$\boldsymbol{\Lambda}_{cc} = \frac{|\boldsymbol{E}|}{|\boldsymbol{E}_k|}, \tag{3.100}$$

donde

$$\boldsymbol{E}_k = \sum_{i=1}^{q}\sum_{j=1}^{n_i}(Y_{ij} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_i)(Y_{ij} - \boldsymbol{A}\hat{\boldsymbol{\beta}}_i)'$$

para muestras de tamaño grande, la hipótesis nula, que establece la adecuación del polinomio de grado k, se rechaza si

$$-\Big(N - \frac{1}{2}(p-k+q)\Big) \ln \mathbf{\Lambda}_{cc} > \chi^2_{(\alpha,(p-k-1)q)}. \tag{3.101}$$

Ejemplo 3.13 La tabla 3.16 consigna las medidas sobre el contenido de calcio del hueso cúbito de mujeres de edad avanzada. Las mujeres se dividieron en dos grupos, uno de los grupos recibió una ayuda especial a través de una dieta y un programa de ejercicios físicos (tratamiento) y el otro no (control). Además de una medida inicial se hicieron mediciones durante tres años consecutivos. Para los datos de la tabla 3.16 se explora y verifica el ajuste de curvas de crecimiento conforme a un modelo cuadrático. Las estimaciones de los $\boldsymbol{\beta}$, de acuerdo con (3.100) son

Tabla 3.16 Contenido de calcio en cúbito

Grupo control					Grupo tratado				
Suj.	Año 0	Año 1	Año 2	Año 3	Suj.	Año 0	Año 1	Año 2	Año 3
1	87.3	86.9	86.7	75.5	1	83.8	85.5	86.2	81.2
2	59.0	60.2	60.0	53.6	2	65.3	66.9	67.0	60.6
3	76.7	76.5	75.7	69.5	3	81.2	79.5	84.5	75.2
4	70.6	76.1	72.1	65.3	4	75.4	76.7	74.3	66.7
5	54.9	55.1	57.2	49.0	5	55.3	58.3	59.1	54.2
6	78.2	75.3	69.1	67.6	6	70.3	72.3	70.6	68.6
7	73.7	70.8	71.8	74.6	7	76.5	79.9	80.4	71.6
8	61.8	68.7	68.2	57.4	8	66.0	70.9	70.3	64.1
9	85.3	84.4	79.2	67.0	9	76.7	79.0	76.9	70.3
10	82.3	86.9	79.4	77.4	10	77.2	74.0	77.8	67.9
11	68.6	65.4	72.3	60.8	11	67.3	70.7	68.9	65.9
12	67.8	69.2	66.3	57.9	12	50.3	51.4	53.6	48.0
13	66.2	67.0	67.0	56.2	13	57.7	57.0	57.5	51.5
14	81.0	82.3	86.8	73.9	14	74.3	77.7	72.6	68.0
15	72.3	74.6	75.3	66.1	15	74.0	74.7	74.5	65.7
					16	57.3	56.0	64.7	53.0
Media	72.38	73.29	72.47	64.79	Media	69.29	70.66	71.18	64.53

Fuente: Johnson y Wichern (1998, págs. 350-351)

$$(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) = \begin{pmatrix} 73.0701 & 70.1387 \\ 3.6444 & 4.0900 \\ -2.0274 & -1.8534 \end{pmatrix}.$$

Así, las curvas de crecimiento estimadas son

Grupo control: $73.0701 + 3.6444t - 2.0274t^2$

Grupo tratado: $70.1387 + 4.0900t - 1.8534t^2$.

donde

$$(\boldsymbol{A}'\boldsymbol{S}_p^{-1}\boldsymbol{A})^{-1} = \begin{pmatrix} 93.1744 & -5.8368 & 0.2184 \\ -5.8368 & 9.5699 & -3.0240 \\ 0.2184 & -3.0240 & 1.1051 \end{pmatrix}.$$

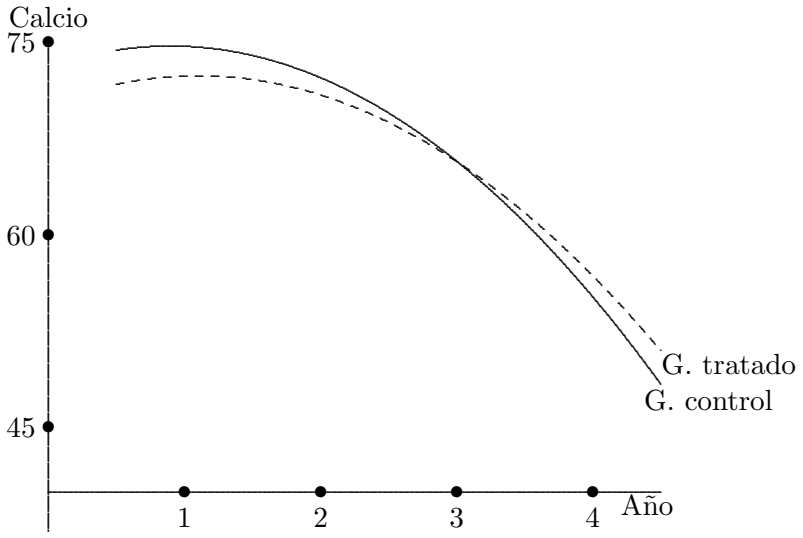


Figura 3.11 *Curvas de crecimiento, grupo control y tratamiento.*

El valor de la estadística lambda de Wilks para verificar la hipótesis que las curvas de crecimiento cuadráticas se ajustan a los datos es

$$\Lambda_{cc} = \frac{\boldsymbol{E}}{\boldsymbol{E}_2} = \frac{\begin{vmatrix} 2726.282 & 2660.749 & 2369.308 & 2335.912 \\ 2660.749 & 2756.009 & 2343.514 & 2327.961 \\ 2369.308 & 2343.514 & 2301.714 & 2098.544 \\ 2335.912 & 2327.961 & 2098.544 & 2277.452 \end{vmatrix}}{\begin{vmatrix} 2781.017 & 2698.589 & 2363.228 & 2362.253 \\ 2698.589 & 2832.430 & 2331.235 & 2381.160 \\ 2363.228 & 2331.235 & 2303.687 & 2089.996 \\ 2362.253 & 2381.160 & 2089.996 & 2314.485 \end{vmatrix}}$$
$$= 0.7627.$$

Para un $\alpha = 0.05$, el valor de la estadística dada en (3.102) es

$$-\Big(N - \frac{1}{2}(p-k+q)\Big)\ln \Lambda_{cc} > \chi^2_{(\alpha,p-k-1)} = -\Big(31 - \frac{1}{2}(4-2+2)\Big)\ln 0.7627$$
$$= 7.86 > \chi^2_{(0.05,(4-2-1)2)} = 5.991.$$

Luego los datos se ajustan a una curva de crecimiento cuadrática. Aunque, para $\alpha = 0.01$ ($\chi^2_{(0.01,(4-2-1)2)} = 9.21$) hay evidencia de que estos polinomios cuadráticos no se ajustan adecuadamente a los datos. De acuerdo con estas curvas (figura 3.11), ambas son decrecientes después del primer año de estudio, lo cual significa que existe una pérdida de calcio en ambos grupos. Sin considerar el ajuste cuadrático se puede hacer un análisis de perfiles para verificar el paralelismo o coincidencia en la pérdida de calcio a través del tiempo en estos grupos de mujeres. ☑

3.6 Rutina SAS para calcular la estadística T^2 de Hotelling

```
DATA EJEMP3\_5; /* archivo del ejemplo  3.5 */
INPUT sexo $ X1 X2 X3 X4 \@\@; /* variable sexo categórica y X1 a X4
                                                    numérica */
CARDS;  /*  para ingresar datos */%
1 15 17 24 14  1 17 15 32 26  1 15 14 29 23  1 13 12 10 16  1 20 17 26 28
1 15 21 26 21  1 15 13 26 22  1 13  5 22 22  1 14  7 30 17  1 17 15 30 27
1 17 17 26 20  1 17 20 28 24  1 15 15 29 24  1 18 19 32 28  1 18 18 31 27
1 15 14 26 21  1 18 17 33 26  1 10 14 19 17  1 18 21 30 29  1 18 21 34 26
1 13 17 30 24  1 16 16 16 16  1 11 15 25 23  1 16 13 26 16  1 16 13 23 21
1 18 18 34 24  1 16 15 28 27  1 15 16 29 24  1 18 19 32 23  1 18 16 33 23
1 17 20 21 21  1 19 19 30 28  2 13 14 12 21  2 14 12 14 26  2 12 19 21 21
2 12 13 10 16  2 11 20 16 16  2 12  9 14 18  2 10 13 18 24  2 10  8 13 23
2 12 20 19 23  2 11 10 11 27  2 12 18 25 25  2 14 18 13 26  2 14 10 25 28
```

```
2 13 16  8 14  2 14  8 13 25  2 13 16 23 28  2 16 21 26 26  2 14 17 14 14
2 16 16 15 23  2 13 16 23 24  2  2  6 16 21  2 14 16 22 26  2 17 17 22 28
2 16 13 16 14  2 15 14 20 26  2 12 10 12  9  2 14 17 24 23  2 13 15 18 20
2 11 16 18 28  2  7  7 19 18  2 12 15  7 28  2  6  5  6 13  /* datos,
                                              sexo=1 hombre y 2  mujer */
 ;
PROC IML; /* invoca el procedimiento IML */
USE EJEMP3\_5; /* toma los datos del archivo EJEMP3\_5 */
  READ ALL VAR {X1 X2 X3 X4} INTO X;  /* forma la matriz X con las variables
                                                                  X1 a X4 */
  X1 = X[1:32,]; /* toma los datos para hombres */
  X2 = X[33:64,]; /* toma los datos para mujeres */
  p=NCOL(X); /* número de variables (columnas) en la matriz de datos X */
  N1 = NROW(X1); /* número de observaciones en la submatriz hombres */
  N2 = NROW(X2); /* número de observaciones en la submatriz mujeres */
XMH = 1/N1*X1‘*J(N1,1); /* vector de medias en archivo hombres */
XMM = 1/N2*X2‘*J(N2,1); /* vector de medias en archivo mujeres */
SH = 1/(N1-1)*XMH‘*(I(N1)-1/N1*J(N1))*XMH; /* matriz de covarianzas archivo
                                                                 hombres */
SM = 1/(N2-1)*XMM‘*(I(N2)-1/N2*J(N2))*XMM; /* matriz de covarianzas archivo
                                                                  mujeres */
Sp = 1/(N1+N2-2)*((N1-1)*SH+(N2-1)*SM); /* matriz de covarianzas pareada */
T2 = N1*N2/(N1+N2)*(X1BAR-X2BAR)‘*INV(Spl)*(X1BAR-X2BAR); /* Est. T2 */
F0=((N1+N2-p-1)/((N1+N2-2)*p))*T2; /* transformación a la estadística F */
p\_val=1-PROBF(F,p,N1+N2-p-1); /* p valor asociado a F0 */
PRINT T2 p\_val; /* imprime el valor de T2 y el valor p */
RUN; /*  ejecución del programa */
```

3.7 PROCcedimiento GLM para el ANAVAMU

Con el procedimiento GLM (general linear models) se desarrollan los cálculos del análisis de varianza multivariado.

```
/* EJEMPLO 3.9 */ DATA EJEM3\_9; /* archivo del ejemplo  3.9 */
INPUT  METODO \$ MATEMAT ESCRIT  @@;
                /* método 1 y 2 (grupos) matem. y escrit. (resp.) */
CARDS; /* ingreso de datos */
1 69 75     1 69  70    1 71  73     1 78 82    1 79  81   1 73 75
2 69 70     2 68  74    2 75  80     2 78 85    2 68  68   2 63 68
2 72 74     2 63  66    2 71  76
2 72 78     2 71  73    2 70  73     2 56 59    2 77  83
3 72 79     3 64 65     3 74  74     3 72 75    3 82 84    3 69 68
3 76 76     3 68 65
3 78 79     3 70 71     3 60 61
;
PROC GLM; /* invocación del procedimiento GLM */
  CLASS METODO;  /* define la variable de clasificación (poblaciones) */
```

```
  MODEL MATEMAT ESCRIT = METODO;  /* modelo multivariado a una  */
                                 /*  vía de clasific. */
  MANOVA H=METODO/PRINTE PRINTH; /* H= hipótesis de acuerdo  */
                        /* co el modeloe imprime las matrices E y H */
RUN\,; /* ejecuta la rutina */
```

3.8 PROCcedimiento GLM para contrastes y medidas repetidas en el ANAVAMU

Se muestra, en forma resumida, la sintaxis del procedimiento GLM del paquete estadístico SAS, para desarrollar los cálculos necesarios en problemas de contrastes de tratamientos, medidas repetidas en análisis de varianza multivariado, de una o varias vías de clasificación. Una presentación más amplia de esta sintaxis se puede consultar en (SAS User's Guide, 1998). Al frente de cada instrucción se explica su propósito dentro de los símbolos /$*$ y $*$/.

```
 PROC GLM options; /* invocación del procedimiento GLM */
    CLASS lista de variables; /*  variables de clasificación  g */
    MODEL var. depends.= var. independs. / opciones; /* variables */
                             /* depend.  e indepen. en el modelo */
    CONTRAST  'rótulo' de valores para los efectos.../ opciones;
            /* especifica un vector o matriz de coeficientes */
            /* asociados a los contrastes */
    MANOVA H= efectos E= efectos M= ecuaciones...
         /* H= efecto de hipótesis, E= efecto del error M= ecuaciones */
         /* del o los modelos */
    MEANS efectos / opciones;
     /* efectos a la derecha de la ecuación del modelo */
    REPEATED /* nombre de los niveles de los factores (valor de niveles) */
      /* para variables dependientes que representan medidas repetidas */
      /* sobre la misma unidad */
RUN;
```

3.9 Procesamiento de datos con R

Se traduce el código SAS de la sección 3.6

```
 # lectura de datos ejemplo 3.5
datos<-scan()
 1   15  17  24  14 1 17 15 32 26 1 15 14 29 23
 1   13  12  10  16 1 20 17 26 28 1 15 21 26 21
 1   15  13  26  22 1 13 5  22 22 1 14 7  30 17
```

```
1   17  15  30  27 1 17 17 26 20 1 17 20 28 24
1   15  15  29  24 1 18 19 32 28 1 18 18 31 27
1   15  14  26  21 1 18 17 33 26 1 10 14 19 17
1   18  21  30  29 1 18 21 34 26 1 13 17 30 24
1   16  16  16  16 1 11 15 25 23 1 16 13 26 16
1   16  13  23  21 1 18 18 34 24 1 16 15 28 27
1   15  16  29  24 1 18 19 32 23 1 18 16 33 23
1   17  20  21  21 1 19 19 30 28 2 13 14 12 21
2   14  12  14  26 2 12 19 21 21 2 12 13 10 16
2   11  20  16  16 2 12 9  14 18 2 10 13 18 24
2   10  8   13  23 2 12 20 19 23 2 11 10 11 27
2   12  18  25  25 2 14 18 13 26 2 14 10 25 28
2   13  16  8   14 2 14 8  13 25 2 13 16 23 28
2   16  21  26  26 2 14 17 14 14 2 16 16 15 23
2   13  16  23  24 2 2  6  16 21 2 14 16 22 26
2   17  17  22  28 2 16 13 16 14 2 15 14 20 26
2   12  10  12  9  2 14 17 24 23 2 13 15 18 20
2   11  16  18  28 2 7  7  19 18 2 12 15 7  28
2   6   5   6   13

datos2<-matrix(datos,ncol=5,byrow=TRUE)
ejemp3_5<-data.frame(datos2)
colnames(ejemp3_5)<-c("sexo","X1","X2","X3","X4")

# toma los datos para hombres
hombres<-subset(ejemp3_5,ejemp3_5$sexo==1,select=c(2:5))
# toma los datos para mujeres
mujeres<-subset(ejemp3_5,ejemp3_5$sexo==2,select=c(2:5))
# numero de variables
p<-ncol(hombres)
# numero de hombres
n1<-nrow(hombres)
# numero de mujeres
n2<-nrow(mujeres)
# grados de libertad de Sp
v<-n1+n2-2
# vector de medias de hombres
XMH<-mean(hombres)
# vector de medias de mujeres
XMM<-mean(mujeres)
# Matriz de covarianzas de hombres
SH<-cov(hombres)
# Matriz de covarianzas de mujeres
SM<-cov(mujeres)
 # Matriz de covarianzas ponderadas
Sp<-1/v*( (n1-1)*SH+(n2-1)*SM )
# estadística T^2
T2<-(n1*n2/(n1+n2))*mahalanobis(XMH, XMM,  Sp)
# transformación a la estadística F
F0<-(v-p+1)/(v*p)*T2
# p valor asociado a F0
```

```
p_val<-pf(F0,p,n1+n2-p-1,lower.tail=FALSE)
# imprime resultados
cat("\n","T2= ",T2, " P_valor= ",p_val, "\n")
```

Análisis de varianza multivariado, corresponde a la sección 3.7 con los mismos datos y genera las salidas del ejemplo 3.9

```
# lectura de datos ejemplo 3.5
datos<-scan()
1  69 75 1  69 70 1  71 73 1  78 82 1  79 81 1  73
75 2  69 70 2  68 74 2  75 80 2  78 85 2  68 68 2
63 68 2  72 74 2  63 66 2  71 76 2  72 78 2  71 73
2  70 73 2  56 59 2  77 83 3  72 79 3  64 65 3  74
74 3  72 75 3  82 84 3  69 68 3  76 76 3  68 65 3
78 79 3  70 71 3  60 61

datos2<-matrix(datos,ncol=3,byrow=TRUE)
ejemp3_9<-data.frame(datos2)
colnames(ejemp3_9)<-c("metodo","matemat","escrit")
# se ubican las columnas y1, y2, en una matriz llamada Mdatos
Mdatos<-as.matrix(ejemp3_9[,-1])
# se define el factor  y se llama metodo
ejemp3_9$metodo<- as.factor(ejemp3_9$metodo)
# Análisis de varianza univariado
# para matemat
ajusteM<-lm(matemat~metodo,data=ejemp3_9)
anova(ajusteM)
# Análisis de varianza univariado
# para escritura
ajusteE<-lm(escrit~metodo,data=ejemp3_9)
anova(ajusteE)
# Ajustamos el modelo multivariado de una via
ajuste<-manova(Mdatos~metodo )
# Las diferentes estadisticas
summary(ajuste ,test="Wilks")
summary(ajuste,test="Pillai")
summary(ajuste ,test= "Hotelling-Lawley")
summary(ajuste ,test= "Roy")
# Las matrices E  y H
M<-summary(ajuste)$SS
H<-M$metodo
E<-M$Residuals
```

Las funciones `dmvnorm`, `mvnorm` contenidas en el paquete `mvtnorm` (el cual debe traerse desde r-project) proveen la función de densidad y la función y el generador de vectores aleatorios de una distribución normal multivariante con media igual a mean y matriz de covarianzas igual a sigma, respectivamente.

Los argumentos de la función son:

```
dmvnorm(x, mean, sigma, log=FALSE) rmvnorm(n, mean, sigma,
method=c("svd", "chol"))
```

donde:

n: es el número de observaciones.

mean: es el vector de medias, por defecto es `rep(0, length = ncol(x))`.

sigma: matriz de covarianzas, por defecto es `diag(ncol(x))`.

log: función lógica; si `TRUE`, la densidades `d` son dadas como `log(d)`.

method: usa la descomposición matricial para determinar la matriz raíz de sigma, los métodos posibles son la descomposición en valor singular (`svd`, `default`) and la descomposición de Cholesky (`chol`).

```
Ejemplo: las instrucciones dmvnorm(x=c(0,0))
dmvnorm(x=c(0,0),mean=c(1,1)) sigma <- matrix(c(4,2,2,3), ncol=2)
x<-rmvnorm(n=500,mean=c(1,2), sigma=sigma) colMeans(x)
```

genera 500 observaciones de vector aleatorio cuya distribución es normal bivariado con vector de medias `mean` $= \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ y matriz de covarianzas `matrix` $= \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$.

Capítulo 4

Inferencia sobre la matriz de covarianzas

4.1 Introducción

En los capítulos precedentes se definió la matriz de covarianzas junto con algunas de sus propiedades, se hizo su estimación, se determinó su distribución bajo el supuesto de normalidad y se empleó en la inferencia sobre los vectores de medias. Este capítulo está dedicado a presentar la distribución de la *matriz de covarianzas* y la inferencia sobre ésta para una o varias poblaciones. Además, se muestra, de manera esquemática, su aplicación en el estudio de modelos de componentes de varianza y en algunos contrastes de independencia entre variables.

La matriz de covarianzas está ligada a varias formas cuadráticas del tipo $X'\mathbf{\Sigma}X$; por ejemplo: la distancia de Mahalanobis, la estadística T^2, las regiones de confianza para $\boldsymbol{\mu}$, algunas estadísticas para el análisis de varianza multivariado, entre otras. El elipsoide correspondiente a cada forma cuadrática tiene una representación que depende de la estructura de la matriz de covarianzas.

En la figura 4.1 se muestran algunos casos particulares de representaciones asociadas con la matriz de covarianzas de un vector bidimensional $X' = (X_1, X_2)$. La figura 4.1a corresponde a una forma cuadrática donde la matriz de covarianzas es diagonal, con elementos en la diagonal iguales; es decir, igualdad de varianzas (homocedasticidad) y no asociación lineal entre las variables. La figura 4.1b representa una forma cuadrática donde la matriz de covarianzas es diagonal, con elementos en la diagonal diferentes; esto es, varianzas distintas (heterocedasticidad) y no asociación lineal entre las variables. Las figuras 4.1c y 4.1d muestran las formas cuadráticas cuyas matrices de covarianzas contienen varianzas diferentes y covarian-

zas que señalan asociación lineal positiva y negativa entre las variables, respectivamente.

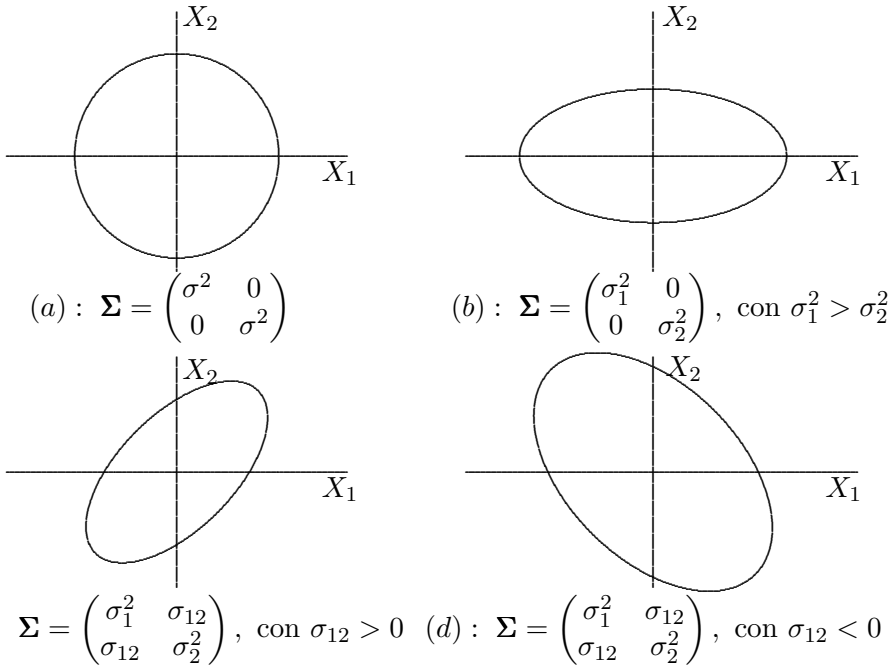


Figura 4.1 *Elipses asociadas con la matriz de covarianzas.*

4.2 Distribución muestral de la matriz de covarianzas

Dada una muestra aleatoria de vectores p–variantes de una población multinormal de media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$, el estimador máximo verosímil de $\boldsymbol{\Sigma}$ es (sección (3.2))

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{\alpha=1}^{n}(X_\alpha - \bar{\boldsymbol{X}})(X_\alpha - \bar{\boldsymbol{X}})' \\ &= \frac{1}{n}\boldsymbol{A} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{\boldsymbol{X}}_j)(X_{ik} - \bar{\boldsymbol{X}}_k)\right) \qquad \text{para } j,k = 1,\ldots,p. \end{aligned} \tag{4.1}$$

Se obtiene la distribución de

$$\boldsymbol{A} = \sum_{\alpha=1}^{n} (\boldsymbol{X}_\alpha - \bar{\boldsymbol{X}})(\boldsymbol{X}_\alpha - \bar{\boldsymbol{X}})',$$

paralelamente a como se procede en el caso univariado. Recuérdese que

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(n-1)}.$$

Téngase en cuenta que una variable aleatoria U tiene distribución ji-cuadrado, si su *fdp* es del tipo $\Gamma(U, \alpha = \frac{n}{2}, \beta = \frac{1}{2})$; es decir (ecuación B.10),

$$f_U(u) = \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{\frac{n}{2}} u^{\frac{n}{2}-1} e^{-\frac{1}{2}u}, \qquad u > 0.$$

Se afirma que la distribución ji–cuadrado es un caso especial de la distribución *gama*. En forma semejante, en el caso *p–variado* se define la función de densidad conjunta de *Wishart*, la cual está ligada a la función *gama multivariada*. A continuación se define la función gama multivariante y se muestra su relación con la distribución de Wishart, ésta es una definición alterna a la considerada en la sección (2.3.4).

Definición: la *función gama multivariante* está dada por

$$\Gamma_p(t) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\left[t - \frac{1}{2}(i-1)\right]. \tag{4.2}$$

Para el caso univariado, $p = 1$, se tiene la función gama que se muestra en la ecuación (B.11).

Definición: una matriz aleatoria, $\boldsymbol{A}$, de tamaño $(p \times p)$ tiene distribución de Wishart si su función de densidad se puede escribir de la forma

$$\mathcal{W}(\boldsymbol{A}|\boldsymbol{\Sigma}, n) = \frac{|\boldsymbol{A}|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}\operatorname{tra}(\boldsymbol{\Sigma}^{-1}\boldsymbol{A})}}{2^{\frac{1}{2}pn} |\boldsymbol{\Sigma}|^{\frac{1}{2}n} \Gamma_p(\frac{1}{2}n)}, \tag{4.3}$$

con $n \approx n - 1$.

Se nota $\boldsymbol{A} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n)$ para hacer referencia que la matriz $\boldsymbol{A}$ tiene una distribución asociada con la distribución Wishart, cuya matriz de escala es $\boldsymbol{\Sigma}$ y con grados de libertad iguales a n. Cuando $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, se dice que la distribución está en su forma *estándar*.

Algunas consecuencias de la definición anterior se resumen en las siguientes propiedades.

4.2.1 Propiedades de la matriz de covarianzas muestral

1. $\mathcal{E}(\boldsymbol{A}) = n\boldsymbol{\Sigma}$.

2. Si $\boldsymbol{B}$ es una matriz de tamaño $(k \times p)$, entonces $\boldsymbol{BAB'} \sim \mathcal{W}_k(\boldsymbol{B\Sigma B'}, n)$.

3. Si $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_q$ son matrices de tamaño $(p \times p)$, independientes y distribuidas conforme a una Wishart, es decir $\boldsymbol{A}_i \sim \mathcal{W}(\boldsymbol{\Sigma}, n_i - 1)$, entonces
$$\mathcal{A} = \sum_{i=1}^{q} \boldsymbol{A}_i$$
se distribuye $\mathcal{W}\Big(\boldsymbol{\Sigma}, \sum_{i=1}^{q}(n_i - 1)\Big)$.

4. Particionando $\boldsymbol{A}$ y $\boldsymbol{\Sigma}$ en q-filas y $(p-q)$-columnas
$$\boldsymbol{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$
si $\boldsymbol{A}$ se distribuye como $\mathcal{W}(\boldsymbol{\Sigma}, n)$, entonces A_{ii} se distribuye como $\mathcal{W}_\alpha(\boldsymbol{\Sigma}_{ii}, n)$ para $i = 1, 2$ y $\alpha = q,\ p - q$. Esta propiedad se puede hacer extensiva praa cualquier partición adecuada de las matrices $\boldsymbol{A}$ y $\boldsymbol{\Sigma}$.

▶ Distribución de $\widehat{\boldsymbol{\Sigma}}$

Suponga que $X_1, \ldots, X_n$ son n vectores aleatorios de tamaño $(p \times 1)$ que conforman una muestra aleatoria de una población normal p-variante; es decir, $X_\alpha \sim N_p(\boldsymbol{\mu}; \boldsymbol{\Sigma})$ para $\alpha = 1, \ldots, n$. De la definición y propiedades anteriores se puede concluir que

$$n\widehat{\boldsymbol{\Sigma}} = n\boldsymbol{S} = \boldsymbol{A} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n-1) \tag{4.4}$$

El análisis de varianza, como indica su nombre, consiste en particionar (sinónimo de analizar o descomponer) la variabilidad total de las variables consideradas en el modelo lineal propuesto. La variabilidad se expresa como una suma de cuadrados, con tales sumas de cuadrados se hace el contraste de las hipótesis respecto a los parámetros del modelo lineal, a través de una estadística F (cociente de variabilidad). La distribución de la estadística F se determina al considerar que las sumas de cuadrados están ligadas a un distribución ji-cuadrado y son independientes. Este requerimiento se tiene, en la mayoría de las aplicaciones del análisis de varianza en virtud

del *Teorema de Cochran*. Este teorema se presenta ahora en la versión multivariada.

▶ Teorema de Cochran

Sean $\mathbb{Y} = (Y_1, Y_2, \ldots, Y_n)$ una matriz $n \times p$, donde los Y_i son vectores aleatorios de tamaño $(p \times 1)$, independientes y distribuidos conforme a $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Supóngase que la matriz $\boldsymbol{C}_i = (c^i_{\alpha\beta})$, asociada con la forma cuadrática:

$$\boldsymbol{Q}_i = \mathbb{Y} C_i \mathbb{Y}' = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_n \end{pmatrix} \begin{pmatrix} c^i_{11} & c^i_{12} & \cdots & c^i_{1n} \\ c^i_{21} & c^i_{22} & \cdots & c^i_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c^i_{n1} & c^i_{n2} & \cdots & c^i_{nn} \end{pmatrix} \begin{pmatrix} Y'_1 \\ Y'_2 \\ \vdots \\ Y'_n \end{pmatrix} = \sum_{\alpha,\beta}^{n} c^i_{\alpha\beta} Y_\alpha Y'_\beta,$$

es una matriz simétrica de tamaño $(n \times n)$ de rango r_i, $r(\boldsymbol{C}_i) = r_i$, para $i = 1, \ldots k$, y

$$\boldsymbol{Q} = \sum_{i=1}^{k} \boldsymbol{Q}_i = \sum_{\alpha,\beta}^{n} Y_\alpha Y'_\alpha.$$

Así, $n = \sum_{i=1}^{k} r_i$ es condición necesaria y suficiente para que los

$\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_k$ sean independientes y distribuidos $\mathcal{W}(\boldsymbol{\Sigma}, r_i)$.

Este resultado es particularmente útil para el análisis de varianza (univariado o multivariado) cuando la descomposición de formas cuadráticas se expresa como sumas de otras formas cuadráticas. Es el caso de la descomposición dada en (3.73):

$$\underbrace{\sum_{i=1}^{q}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\boldsymbol{Y}}_{..})(Y_{ij} - \bar{\boldsymbol{Y}}_{..})'}_{\text{Covariabilidad total } (\boldsymbol{Q})} = \underbrace{\sum_{i=1}^{q} n_i(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})(\bar{\boldsymbol{Y}}_{i.} - \bar{\boldsymbol{Y}}_{..})'}_{\text{Covariabilidad entre } (\boldsymbol{Q}_1)} + \underbrace{\sum_{i=1}^{q}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\boldsymbol{Y}}_{i.})(Y_{ij} - \bar{\boldsymbol{Y}}_{i.})'}_{\text{Covariabilidad dentro } (\boldsymbol{Q}_2)}$$

donde *"la covariabilidad total"* es desagregada en *"la covariabilidad debida al modelo"* (entre) y *"la covariabilidad debida al error"* (dentro).

La versión univariada del *Teorema de Cochran* es la siguiente:

> Sean Z_i variables aleatorias independientes y distribuidas conforme a la $n(0, 1)$ para $i = 1, \ldots, n$ y si
>
> $$\sum_{i=1}^{n} Z_i^2 = Q_1 + \cdots + Q_k,$$
>
> donde $k \leq n$ y Q_i con n_i grados de libertad $(i = 1, \ldots, k)$. Entonces $Q_1, \ldots, Q_k$ son variables aleatorias independientes *ji-cuadrado* de $n_1, \ldots, n_k$ *grados de libertad*, respectivamente, si y sólo si,
>
> $$n = n_1 + \cdots + n_k.$$

4.3 Contraste de hipótesis sobre la matriz de covarianzas

4.3.1 Una población

Mediante una muestra aleatoria de n observaciones vectoriales $X_1, \ldots, X_n$, de una población $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con $\boldsymbol{\Sigma}$ definida positiva, se quiere contrastar el juego de hipótesis

$$H_0: \quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 \quad \text{frente a} \quad H_1: \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0. \tag{4.5}$$

La supuesta matriz de covarianzas $\boldsymbol{\Sigma}_0$, es una matriz sobre la cual se tiene un propósito específico respecto a sus valores, o puede ser una matriz resultante de experiencias anteriores.

La razón de máxima verosimilitud, suministra la estadística de prueba para (4.5). Los estimadores de máxima verosimilitud para los parámetros de la distribución normal multivariante, restringidos a H_0 son $\bar{\boldsymbol{X}}$ y $\boldsymbol{\Sigma}_0$ respectivamente; mientras que los estimadores en todo el espacio de parámetros son $\bar{\boldsymbol{X}}$ y S. La razón de verosimilitud es

$$\boldsymbol{\lambda} = \left[\left(\frac{n-1}{n}\right)^p \frac{|\boldsymbol{S}|}{|\boldsymbol{\Sigma}_0|}\right]^{\frac{n}{2}} \exp\Big\{-\frac{1}{2}\left[(n-1)\operatorname{tra}(\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}) - np\right]\Big\}.$$

Si se asume $n \approx n-1 = v$, después de aplicar logaritmos y efectuar las simplificaciones del caso, se consigue

$$\boldsymbol{\lambda}^* = v\Big(\ln|\boldsymbol{\Sigma}_0| - \ln|\boldsymbol{S}| + \operatorname{tra}\big(\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}\big) - p\Big) \tag{4.6}$$

La estadística (4.6) se expresa en términos de los valores propios $\lambda_1, \ldots, \lambda_p$ de la matriz $(\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1})$, a través de las siguientes expresiones

$$\operatorname{tra}\big(\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}\big) = \sum_{i=1}^{p} \lambda_i$$

$$\ln|\boldsymbol{\Sigma}_0| - \ln|\boldsymbol{S}| = -\ln|\boldsymbol{\Sigma}_0^{-1}| - \ln|\boldsymbol{S}| = -\ln|\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1}| = -\ln\left(\prod_{i=1}^{p}\lambda_i\right),$$

después de reemplazar en (4.6) las cantidades anteriores se obtiene

$$\begin{aligned}\boldsymbol{\lambda}^* &= v\left[-\ln\left(\prod_{i=1}^{p}\lambda_i\right) + \sum_{i=1}^{p}\lambda_i - p\right] \\ &= v\left[\sum_{i=1}^{p}(\lambda_i - \ln\lambda_i) - p\right].\end{aligned} \tag{4.6a}$$

Para v moderadamente grande (o para n) y bajo H_0, $\boldsymbol{\lambda}^*$ se distribuye *ji-cuadrado* con $p(p+1)/2$ grados de libertad. Bartlett (1954) propuso la estadística

$$\boldsymbol{\lambda}_1^* = \left\{1 - \frac{1}{6(n-1)}\left[2p+1-\frac{2}{p+1}\right]\right\}\boldsymbol{\lambda}^*, \tag{4.7}$$

cuya distribución se aproxima a la de una *ji–cuadrado*. Se rechaza H_0 cuando $\boldsymbol{\lambda}_1^* > \chi^2_{(\alpha,1/2p(p+1))}$. Se observa que los grados de libertad de la estadística χ^2 son $\frac{1}{2}p(p+1)$ y están ligados al número de parámetros distintos de la matriz de covarianzas $\boldsymbol{\Sigma}$.

Observación:

La hipótesis respecto a la independencia y homocedasticidad de las variables, asumida en la mayoría de los modelos de regresión lineal y en el análisis de varianza clásico, se expresa como $H_0: \boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, donde σ^2 es la varianza común y desconocida. De acuerdo con la figura 4.1a, esta hipótesis corresponde a la "*esfericidad*" de la forma cuadrática asociada con la matriz de covarianzas, de otra forma, la hipótesis se puede leer como variables ortogonales (de covarianza cero) y con varianza constante en cualquier dirección, es decir, varianza constante de manera "isotrópica". La razón de máxima verosimilitud para verificar $H_0: \boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$ es

$$\boldsymbol{\lambda}^* = \left[\frac{|\boldsymbol{S}|}{(\mathrm{tra}(\boldsymbol{S}/p))^p}\right]^{n/2},$$

como, para n grande, $-2\ln\boldsymbol{\lambda}^*$ tiene aproximadamente una distribución $\chi^2_{(v)}$ donde v es igual al número total de parámetros menos el número de parámetros estimados bajo la restricción impuesta por H_0. De esta forma, la razón de máxima verosimilitud se reduce a:

$$-2\ln\boldsymbol{\lambda}^* = -n\ln\left[\frac{|\boldsymbol{S}|}{(\mathrm{tra}\,\boldsymbol{S}/p)^p}\right] = -n\ln\boldsymbol{\lambda}_1^*,$$

donde

$$\boldsymbol{\lambda}_1^* = (\boldsymbol{\lambda}^*)^{2/n} = \frac{p^p\prod_{i=1}^p\lambda_i}{\left(\sum_{i=1}^p\lambda_i\right)^p},$$

con $\lambda_1,\ldots,\lambda_p$ los valores propios de la matriz $\boldsymbol{S}$. Una aproximación sobre $-n\ln\boldsymbol{\lambda}_1^*$ es

$$\boldsymbol{\lambda}_2^* = -\left(v - \frac{2p^2+p+2}{6p}\right)\ln\boldsymbol{\lambda}_1^*,$$

la cual tiene aproximadamente una distribución χ^2 con $(\frac{1}{2}p(p+1)-1)$ grados de libertad. Se rechaza H_0 si $\boldsymbol{\lambda}_2^* \geq \chi^2_{(\alpha,\frac{1}{2}p(p+1)-1)}$.

Ejemplo 4.1 Se tomaron 20 sujetos[1] y se les midió los tiempos de reacción ante un estímulo en centésimas de segundo. Los estímulos consisten en preparar al individuo mediante tres intervalos de tiempo con duración diferente. Los datos se asumen asociados a una distribución normal trivariada. Se quiere verificar la siguiente hipótesis

$$H_0: \quad \boldsymbol{\Sigma} = \begin{pmatrix} 4 & 3 & 2 \\ 3 & 6 & 5 \\ 2 & 5 & 10 \end{pmatrix},$$

la cual ha sido sugerida por observaciones anteriores.

De los datos muestrales, la matriz de covarianzas estimada es

$$\boldsymbol{S} = \begin{pmatrix} 3.42 & 2.60 & 1.89 \\ 2.60 & 8.00 & 6.51 \\ 1.89 & 6.51 & 9.62 \end{pmatrix}.$$

Las cantidades requeridas en (4.6) son

$$\boldsymbol{\Sigma}_0^{-1} = \begin{pmatrix} 0.4070 & -0.2326 & 0.0349 \\ -0.2326 & 0.4186 & -0.1628 \\ 0.0349 & -0.1628 & 0.1744 \end{pmatrix}$$

y

$$\boldsymbol{S}\boldsymbol{\Sigma}_0^{-1} = \begin{pmatrix} 0.8531 & -0.0147 & 0.0257 \\ -0.5752 & 1.6844 & -0.0761 \\ -0.4092 & 0.7195 & 0.6841 \end{pmatrix}.$$

de donde,

$$|\boldsymbol{\Sigma}_0| = 86, \quad |S| = 88.635538, \quad \text{tra}(S\boldsymbol{\Sigma}_0^{-1}) = 3.2216,$$
$$v = 19, \quad \boldsymbol{\lambda}^* = 3.65 \text{ y } \boldsymbol{\lambda}_1^* = 3.44.$$

Como $\boldsymbol{\lambda}_1^* < \chi^2_{(\alpha,6)}$, para valores de α incluso del 10%, entonces no se rechaza la hipótesis de que la matriz de covarianzas es la propuesta en H_0. ☑

[1] Morrison (1990, pág., 293).

4.3.2 Varias poblaciones

La igualdad de matrices de covarianza es un supuesto que se requiere para aplicar adecuadamente algunas técnicas tales como la comparación de medias en dos o más poblaciones (estadística T^2 y en el ANAVAMU), el análisis discriminante, entre otras.

• Para el caso univariado ($p = 1$) se han propuesto varios procedimientos, uno de ellos es la prueba de Bartlett (1937) para contrastar la homogeneidad de varianzas, la cual ha sido extendida a situaciones multivariadas. Para verificar la hipótesis

$$H_0 : \ \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_q^2,$$

se calcula

$$c = 1 + \frac{1}{3(q-1)}\left[\sum_{i=1}^{q}\frac{1}{v_i} - \frac{1}{\sum_{i=1}^{q} v_i}\right], \quad s_p^2 = \frac{\sum_{i=1}^{q} v_i s_i^2}{\sum_{i=1}^{q} v_i}$$

$$\text{y} \quad m = \Big(\sum_{i=1}^{q} v_i\Big)\ln s_p^2 - \sum_{i=1}^{q} v_i \ln s_i^2,$$

donde $s_1^2, \ldots, s_q^2$ son las varianzas muestrales y $v_1, \ldots, v_q$ los grados de libertad ($v_i = n_i - 1$) respectivos. La estadística

$$\frac{m}{c} \text{ se ditribuye aproximadamente como } \chi^2_{(q-1)}.$$

Se rechaza H_0 si $m/c > \chi^2_{(\alpha,q-1)}$.

• En el caso multivariado, se trata de contrastar la hipótesis sobre la igualdad de las matrices de covarianzas asociadas a varias poblaciones multinormales, mediante la información contenida en una muestra aleatoria de cada una de ellas.

Sea $X_{1_g}, \ldots, X_{n_g}$, con $g = 1, \ldots, q$, una muestra aleatoria de una población $N_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$; es decir, se dispone de q-muestras independientes de poblaciones multinormales. La hipótesis a contrastar es

$$H_0 : \quad \boldsymbol{\Sigma}_1 = \ \cdots \ = \boldsymbol{\Sigma}_q = \boldsymbol{\Sigma}. \tag{4.8}$$

De los datos muestrales se obtienen las matrices

$$\boldsymbol{A}_g = \sum_{\alpha=1}^{n_g} \left(X_{\alpha_g} - \bar{X}_g\right)\left(X_{\alpha_g} - \bar{X}_g\right)',$$

$$\boldsymbol{A} = \sum_{g=1}^{q} A_g,$$

$$\sum_{g=1}^{q} n_g = N, \quad \text{con} \quad g = 1, \ldots, q.$$

Mediante las matrices $\boldsymbol{A}_g$ y $\boldsymbol{A}$ se estiman $\boldsymbol{\Sigma}_g$ y $\boldsymbol{\Sigma}$, en el espacio de parámetros general y en el espacio de parámetros reducido por H_0, respectivamente. Así,

$$\widehat{\boldsymbol{\Sigma}}_g = \frac{1}{n_g}\boldsymbol{A}_g \quad \text{y} \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\boldsymbol{A}.$$

Considerando $v_g = (n_g - 1)$ y $v = \sum_{g=1}^{q} v_g = (N - q)$, se obtienen los estimadores insesgados para $\boldsymbol{\Sigma}_g$ y $\boldsymbol{\Sigma}$; éstos son respectivamente $\boldsymbol{S}_g$ y $\boldsymbol{S}_p$; es decir,

$$\boldsymbol{S}_g = \frac{1}{v_g}\boldsymbol{A}_g \quad \text{y} \quad \boldsymbol{S}_p = \frac{1}{v}\boldsymbol{A} = \frac{1}{v}\sum_{g=1}^{q} v_g \boldsymbol{S}_g. \tag{4.9}$$

La razón de máxima verosimilitud para verificar (4.8) es

$$\boldsymbol{\lambda}_1 = \frac{\prod_{g=1}^{q} |A_g|^{\frac{1}{2}n_g}}{|A|^{\frac{1}{2}N}} \frac{N^{\frac{1}{2}pN}}{\prod_{g=1}^{q} n_g^{\frac{1}{2}pn_g}}. \tag{4.10}$$

Se rechaza H_0 para valores pequeños de $\boldsymbol{\lambda}_1$ a un nivel de significación α; es decir, se rechaza H_0 para valores $\boldsymbol{\lambda}_1$ tales que $\boldsymbol{\lambda}_1 \leq \boldsymbol{\lambda}_1(\alpha)$.

Una modificación de (4.9) fue propuesta por Bartlett (1937) para el caso univariado ($p = 1$), donde se reemplazan los tamaños muestrales por los grados de libertad de $\boldsymbol{A}_g$ y de $\boldsymbol{A}$; esto es n_g por $v_g = (n_g - 1)$ y N por $v = \sum_{g=1}^{q} v_g = (N - q)$. La estadística correspondiente equivalente con (4.10)

$$\begin{aligned}\boldsymbol{\lambda}_1 &= \frac{\prod_{g=1}^{q} |\boldsymbol{A}_g|^{\frac{1}{2}v_g}}{|\boldsymbol{A}|^{\frac{1}{2}v}} \\ &= \left(\frac{|\boldsymbol{S}_1|}{|\boldsymbol{S}_p|}\right)^{v_1/2} \left(\frac{|\boldsymbol{S}_2|}{|\boldsymbol{S}_p|}\right)^{v_2/2} \cdots \left(\frac{|\boldsymbol{S}_g|}{|\boldsymbol{S}_p|}\right)^{v_g/2}.\end{aligned} \tag{4.11}$$

Para dos muestras, $q = 2$ y $p = 1$

$$A_1 = \sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 = v_1 s_1^2,$$
$$A_2 = \sum_{i=1}^{n_2}(x_{i2} - \bar{x}_2)^2 = v_2 s_2^2,$$
$$A = A_1 + A_2 = v_1 s_1^2 + v_2 s_2^2 = (v_1 + v_2) s_p^2,$$

las estadísticas s_1^2 y s_2^2 son los estimadores insesgados de σ_1^2 y σ_2^2. Al reemplazarlas en (4.11), resulta

$$\boldsymbol{\lambda}_1 = \frac{(v_1)^{\frac{1}{2}v_1}(v_2)^{\frac{1}{2}v_2}(s_1^2)^{\frac{1}{2}v_1}(s_2^2)^{\frac{1}{2}v_2}}{(v_1 s_1^2 + v_2 s_2^2)^{\frac{1}{2}(v_1+v_2)}}.$$

Recuérdese que la estadística s_1^2/s_2^2 tiene distribución F y se emplea para verificar la hipótesis $H_0 : \ \sigma_1^2 = \sigma_2^2$. Si se divide la última expresión por $(s_2^2)^{\frac{1}{2}(v_1+v_2)}$ se obtiene

$$\boldsymbol{\lambda}_1 = \frac{(v_1)^{\frac{1}{2}v_1}(v_2)^{\frac{1}{2}v_2}F^{\frac{1}{2}v_1}}{(v_1 F + v_2)^{\frac{1}{2}(v_1+v_2)}}.$$

La región crítica está dada por los valores muestrales tales que

$$\boldsymbol{\lambda}_1 \leq \boldsymbol{\lambda}_1(\alpha)$$

la cual es función de $F(n_1, n_2)$. La región crítica queda determinada por los valores de F tales que $F \leq F_1(\alpha)$ o $F \geq F_2(\alpha)$.

Anderson (1984, pág. 419) obtiene la distribución asintótica de $\boldsymbol{\lambda}_1$ al reemplazar n_g por v_g y N por v. Al aplicar logaritmos en los dos lados de la nueva expresión para $\boldsymbol{\lambda}_1$ y sustituir $\boldsymbol{A}_g$ por $n_g \boldsymbol{S}_g$ y $\boldsymbol{A}$ por $N\boldsymbol{S}_p$, se obtiene

$$-2\ln(\boldsymbol{\lambda}_{1_n}) = v \ln|\boldsymbol{S}_p| - \sum_{g=1}^{q} v_g \ln|\boldsymbol{S}_g|. \tag{4.12}$$

Box (1949) demuestra que si se introduce la cantidad $\boldsymbol{\rho}$ dada por

$$\boldsymbol{\rho} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(q-1)}\left(\sum_{g=1}^{q}\frac{1}{v_g} - \frac{1}{v}\right), \tag{4.13}$$

entonces

$$\boldsymbol{\varphi} = -2\boldsymbol{\rho}\ln(\boldsymbol{\lambda}_{1_n}), \tag{4.14}$$

se distribuye asintóticamente como ji-cuadrado con $p(p+1)(q-1)/2$ grados de libertad (el subíndice n resalta la distribución asintótica).

Ejemplo 4.2 La longitud del fémur dada en centímetros y el tiempo empleado para recorrer una distancia de 100 m. a "paso normal" fue medido en 26 personas que trabajan en oficinas, 23 trabajan como operadores de máquinas y 25 trabajan como conductores. Se desea verificar la hipótesis $H_0: \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3$.

Con los datos obtenidos, las estimaciones para cada una de las matrices de covarianzas son

$$\boldsymbol{S}_1 = \begin{pmatrix} 12.65 & -16.45 \\ -16.45 & 73.04 \end{pmatrix}, \quad \boldsymbol{S}_2 = \begin{pmatrix} 11.44 & -27.77 \\ -27.77 & 100.64 \end{pmatrix},$$

$$\boldsymbol{S}_3 = \begin{pmatrix} 14.46 & -31.26 \\ -31.26 & 101.03 \end{pmatrix}, \quad \boldsymbol{S}_p = \begin{pmatrix} 12.89 & -24.96 \\ -24.96 & 91.05 \end{pmatrix}.$$

En este caso $p = 2$, $q = 3$, $v_1 = (n_1 - 1) = 25$, $v_2 = (n_2 - 1) = 22$, $v_3 = (n_3 - 1) = 24$, $N = 74$ y $v = N - q = 71$. El valor de $\boldsymbol{\rho}$, de acuerdo con (4.11), es:

$$\boldsymbol{\rho} = 1 - \frac{13}{36}\left(\frac{1}{25} + \frac{1}{22} + \frac{1}{24} - \frac{1}{71}\right) = 0.9592.$$

A partir de (4.12) se calcula

$$\begin{aligned} -2\ln(\boldsymbol{\lambda}_{1_n}) &= 71\ln(550.2063) - [25\ln(653.3535) + 22\ln(380.1487) + 24\ln(483.7062)] \\ &= 6.9300. \end{aligned}$$

Como el valor de $\boldsymbol{\varphi} = 2\rho\ln(\boldsymbol{\lambda}_{1_n}) = 6.6472$ es menor que $\chi^2_{(5\%,6)} = 12.60$, se concluye que no hay evidencia suficiente para rechazar la hipótesis de igualdad en la variabilidad y covariabildad de las variables longitud del fémur y tiempo para recorrer 100 metros, respectivamente, para los tres tipos de actividad; es decir, las matrices de covarianzas asociadas con las medidas sobre personas de estos tres grupos no difieren de manera significativa. ☑

4.3.3 Dos poblaciones

Para dos poblaciones normales $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ con $i = 1, 2$ se desea verificar la hipótesis

$$H_0: \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$$

Para no perderse en la obtención de la prueba, la idea es emplear la estadística $\boldsymbol{\lambda}_1^*$ para el caso $q = 2$

$$\boldsymbol{\lambda}_1 = \frac{\prod_{g=1}^{q} |\boldsymbol{A}_g|^{\frac{1}{2}v_g}}{|\boldsymbol{A}|^{\frac{1}{2}v}} = v_1^{\frac{1}{2}pv_1} v_2^{\frac{1}{2}pv_2} \cdot \frac{|S_1|^{\frac{1}{2}v_1}|S_2|^{\frac{1}{2}v_2}}{|v_1S_1 + v_2S_2|^{\frac{1}{2}v}},$$

y obtener de ésta una estadística más sencilla mediante alguna transformación.

Para tal efecto, se busca una transformación de las X de manera que la prueba resulte invariante; es decir, que la región crítica de la prueba no cambie; en otras palabras, la decisión que se tome con los datos originales sea la misma que se tome con los datos transformados. La "ganancia" está en la simplicidad de la estadística que se obtenga con los datos transformados Una presentación más formal se encuentra en Arnold (1981, pág. 11-20).

Si los datos se transforman en

$$X_{(i)}^* = \boldsymbol{C}X_{(i)} + a,$$

con $\boldsymbol{C}$ matriz no singular y a vector de constantes, $i = 1, 2$, la prueba resultará invariante, pues esta transformación hace $\boldsymbol{\Sigma}_i^* = \boldsymbol{C}\boldsymbol{\Sigma}_i\boldsymbol{C}'$, y a $\boldsymbol{S}_i{}^* = \boldsymbol{C}\boldsymbol{S}_i\boldsymbol{C}'$. Las raices (valores propios) de

$$|\boldsymbol{\Sigma}_1 - \lambda\boldsymbol{\Sigma}_2| = 0$$

son invariantes bajo estas transformaciones, pues

$$\begin{aligned} |\boldsymbol{\Sigma}_1^* - \lambda\boldsymbol{\Sigma}_2^*| &= |\boldsymbol{C}\boldsymbol{\Sigma}_1 C' - \lambda\boldsymbol{C}\boldsymbol{\Sigma}_2 C'| \\ &= |\boldsymbol{C}\boldsymbol{C}'||\boldsymbol{\Sigma}_1 - \lambda\boldsymbol{\Sigma}_2| \\ &= |\boldsymbol{\Sigma}_1 - \lambda\boldsymbol{\Sigma}_2|. \end{aligned}$$

Los raices de la última ecuación son las únicas invariantes porque existe una matriz $\boldsymbol{C}$ no singular, tal que

$$\boldsymbol{C}\boldsymbol{\Sigma}_1\boldsymbol{C}' = \boldsymbol{\Lambda}, \text{ y } \boldsymbol{C}\boldsymbol{\Sigma}_2\boldsymbol{C}' = \boldsymbol{I},$$

la matriz $\boldsymbol{\Lambda}$ es una matriz diagonal $\text{Diag}(\lambda_i)$, con $\lambda_1 \geq \cdots \geq \lambda_p$. Una justificación semejante se tiene para las raices $l_1 \geq \cdots \geq l_p$ de

$$|\boldsymbol{S}_1 - l\boldsymbol{S}_2| = 0.$$

Las raices λ_i y l_i son los maximales invariantes para los $\boldsymbol{\Sigma}_i$ y $\boldsymbol{S}_i$ $i = 1, 2$; respectivamente (Arnold, pág. 13).

Con estos resultados se puede retornar a (4.15)

$$\boldsymbol{\lambda}_1^* = v_1^{\frac{1}{2}pv_1} v_2^{\frac{1}{2}pv_2} \cdot \frac{|\boldsymbol{S}_1|^{\frac{1}{2}v_1}|\boldsymbol{S}_2|^{\frac{1}{2}v_2}}{|v_1\boldsymbol{S}_1 + v_2\boldsymbol{S}_2|^{\frac{1}{2}v}},$$

multiplicando por $|\boldsymbol{C}\boldsymbol{C}'|^{\frac{1}{2}v}$ se obtiene

$$\begin{aligned}\boldsymbol{\lambda}_1^* &= v_1^{\frac{1}{2}pv_1} v_2^{\frac{1}{2}pv_2} \cdot \frac{|\boldsymbol{C}\boldsymbol{S}_1\boldsymbol{C}'|^{\frac{1}{2}v_1}|\boldsymbol{C}\boldsymbol{S}_2\boldsymbol{C}'|^{\frac{1}{2}v_2}}{|v_1\boldsymbol{C}\boldsymbol{S}_1\boldsymbol{C}' + v_2\boldsymbol{C}\boldsymbol{S}_2\boldsymbol{C}'|} = \frac{|\boldsymbol{L}|^{\frac{1}{2}v_1}|\boldsymbol{I}|^{\frac{1}{2}v_2}}{|v_1\boldsymbol{L} + v_2\boldsymbol{I}|^{\frac{1}{2}v}} \\ &= v_1^{\frac{1}{2}pv_1} v_2^{\frac{1}{2}pv_2} \cdot \prod_{i=1}^{p} \frac{l^{\frac{1}{2}v_1}}{(v_1 l_i + v_2)^{\frac{1}{2}v}}, \end{aligned} \tag{4.15}$$

notese que la matriz $\boldsymbol{L}$ es una matriz diagonal, Diag(l_i). De acuerdo con la última expresión, la regla de decisión es rechazar la hipótesis nula si las $\lambda_i \;\; i = 1, \ldots, p$, son, en extremo, pequeñas o grandes. Bajo H_0 sucede que $\lambda_i = 1$ para todo $i = 1, \ldots, p$. Una prueba invariante de la hipótesis nula tiene una región crítica en el espacio de los l_i que incluye los puntos que se apartan de $l_1 = \cdots = l_p = 1$.

Utilizando la aproximación de Box (1949) se obtiene que la distribución aproximada para $-2\boldsymbol{\rho} \ln \boldsymbol{\lambda}_1^*$, bajo la hipótesis nula, es $\chi^2_{(p(p+1)/2)}$, donde

$$\boldsymbol{\rho} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)} \left(\frac{1}{v_1} + \frac{1}{v_2} - \frac{1}{v} \right).$$

∘ *Una aplicación: modelos de componentes de varianza*

La comparación de dos matrices de covarianzas tiene aplicación en los *modelos de componentes de varianza*, pues estos modelos están asociados con diseños experimentales cuyos tratamientos son una muestra aleatoria de una población de tratamientos (considerada de tamaño infinito). El modelo de componentes de varianza de un factor se escribe

$$\boldsymbol{X}_{\alpha_{(g)}} = \boldsymbol{\mu} + \boldsymbol{\alpha}_g + \mathcal{E}_{\alpha_{(g)}} \text{ con } \;\; \alpha = 1, \ldots, a, \;\; g = 1, \ldots, q, \tag{4.16}$$

en el modelo, $\boldsymbol{\alpha}$ es una variable aleatoria con distribución $N(0, \boldsymbol{\Theta})$, de manera que el vector $\boldsymbol{X}$ tiene distribución $N(\boldsymbol{\mu}, \Theta + \boldsymbol{\Sigma})$, la estructura de la matriz de covarianzas de $\boldsymbol{X}$ justifica el calificativo de modelo de componentes de varianza, pues la matriz de covarianzas "total" se expresa como

suma de la matriz de covarianzas del "modelo" y la matriz de covarianzas del "error".

La hipótesis de no efecto de los tratamientos equivale a considerar que la variabilidad atribuible a ellos es nula; es decir,

$$H_0: \quad \boldsymbol{\Theta} = \mathbf{0}. \tag{4.17}$$

Similar al desarrollo seguido en la sección (3.5.3), se tiene que $\boldsymbol{E}$ y $\boldsymbol{H}$ de la estadística (3.55) corresponden a las matrices $\boldsymbol{A}_1$ y $\boldsymbol{A}_2$. Para el modelo de componentes de varianza presentado arriba, $\boldsymbol{E}$ tiene distribución $\mathcal{W}(\boldsymbol{\Sigma}, q(a-1))$ y $\boldsymbol{H}$ se distribuye $\mathcal{W}(\boldsymbol{\Sigma} + a\boldsymbol{\Theta}, q-1)$. La hipótesis nula anterior equivale a la igualdad de las matrices de covarianzas de las distribuciones de Wishart; es decir, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma} + a\boldsymbol{\Theta}$; mientras que la alternativa es la matriz $(\boldsymbol{\Sigma} + a\boldsymbol{\Theta}) - \boldsymbol{\Sigma}$ la cual es semidefinida positiva.

Sea $l_1 > \cdots > l_p$ las raices (valores propios) de

$$\left| \boldsymbol{H} - l\frac{1}{a-1}\boldsymbol{E} \right| = 0,$$

y sea

$$l_i^* = \begin{cases} l_i, & \text{si } l_i > 1 \\ 1, & \text{si } l_i \leq 1. \end{cases}$$

La razón de máxima verosimilitud para verificar la hipótesis $\Theta = \mathbf{0}$ frente a que $\boldsymbol{\Theta}$ es definida positiva y $\boldsymbol{\Theta} \neq \mathbf{0}$ es

$$a^{\frac{1}{2}qap} \prod_{i=1}^{p} \frac{l_i^{*\frac{1}{2}q}}{(l_i^* + a - 1)^{\frac{1}{2}qa}} = a^{\frac{1}{2}qap} \prod_{i=1}^{k} \frac{l_i^{\frac{1}{2}q}}{(l_i + a - 1)^{\frac{1}{2}qa}},$$

con k el número de raices mayores a 1. Para la distribución de esta estadística se puede aplicar como antes la aproximación a la ji-cuadrado.

4.3.4 Independencia entre variables

En el capítulo 2 se presentó el concepto de independencia entre vectores aleatorios. Bajo normalidad la independencia entre vectores aleatorios implica que la respectiva matriz de covarianzas es la matriz nula, y recíprocamente si la matriz de covarianzas es nula los vectores son independientes. Aunque en esta sección se desarrolla la independencia para el caso de dos "subvectores", para más de dos "subvectores" el tratamiento es semejante (Anderson 1984, pág. 376).

Sea $\boldsymbol{X}$ un vector aleatorio distribuido $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Particiónese $\boldsymbol{X}$ en los subvectores $\boldsymbol{X}_{(1)}$ y $\boldsymbol{X}_{(2)}$ de tamaño $(p_1 \times 1)$ y $(p_2 \times 2)$, respectivamente, donde $(p_1 + p_2 = p)$, de manera que

$$\begin{pmatrix} X_{(1)} \\ X_{(2)} \end{pmatrix} \sim N_{p_1+p_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right), \text{ con } \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} > 0.$$

El problema consiste en contrastar la hipótesis que $\boldsymbol{X}_{(1)}$ y $\boldsymbol{X}_{(2)}$ *son independientes*. Esto equivale a verificar que $\boldsymbol{\Sigma}_{12} = 0$.

Mediante la muestra aleatoria de tamaño n,

$$\begin{pmatrix} X_{(1)1} \\ X_{(2)1} \end{pmatrix}, \ldots, \begin{pmatrix} X_{(1)n} \\ X_{(2)n} \end{pmatrix}$$

una vez más, por el método de la razón de máxima verosimilitud se determina la prueba

$$\boldsymbol{\lambda} = \frac{|\boldsymbol{A}|^{\frac{1}{2}n}}{|\boldsymbol{A}_{11}|^{\frac{1}{2}n}|\boldsymbol{A}_{22}|^{\frac{1}{2}n}} = \frac{|\boldsymbol{S}_p|^{\frac{1}{2}n}}{|\boldsymbol{S}_{11}|^{\frac{1}{2}n}|\boldsymbol{S}_{22}|^{\frac{1}{2}n}}. \tag{4.19}$$

La hipótesis $\boldsymbol{\Sigma}_{12} = 0$ es equivalente a la hipótesis para el modelo lineal general multivariado, considerado en la sección (3.5.2), pues en un modelo de regresión $Y = X\beta + \epsilon$, la hipótesis $H_0 : \beta = 0$ es equivalente, con regresores aleatorios y bajo normalidad, a la independencia entre los regresores X y la variable respuesta Y.

La estadística $\boldsymbol{\lambda}$ equivale a la estadística que se obtiene al elevar (4.19) a la potencia $2/n$), ésta es:

$$\boldsymbol{\lambda}^* = \frac{|\boldsymbol{S}_p|}{|\boldsymbol{S}_{11}||\boldsymbol{S}_{22}|}, \tag{4.19a}$$

la cual se distribuye conforme a $\boldsymbol{\Lambda}_{(p_1,p_2,n-1-p_2)}$, o también como $\boldsymbol{\Lambda}_{(p_2,p_1,n-1-p_1)}$. Se rechaza la hipótesis de independencia entre los dos conjuntos de variables si $\boldsymbol{\lambda}^* \leq \boldsymbol{\Lambda}_{(p_1,p_2,n-1-p_2,\alpha)}$. Los valores críticos de esta distribución se encuentran en la tabla C.2.

Se demuestra también que:

$$\boldsymbol{\lambda} = \frac{|\boldsymbol{A}_{11.2}|^{\frac{1}{2}n}}{|\boldsymbol{A}_{11}|^{\frac{1}{2}n}} = \frac{|\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}|^{\frac{1}{2}n}}{|\boldsymbol{A}_{11}|^{\frac{1}{2}n}}.$$

Las raices $t_1 \geq \cdots \geq t_p$ de $|\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21} - t\boldsymbol{S}_{11}|$ son maximales e invariantes, para contrastar la hipótesis $\boldsymbol{\Sigma}_{12} = 0$. Al multiplicar por $|S_{11}^{-1}|$ resulta:

$$\boldsymbol{\lambda} = |I - S_{11}^{-1/2}S_{12}S_{22}^{-1}S_{21}S_{11}^{-1/2}|^{\frac{1}{2}n} = \prod_{i=1}^{p}(1-t_i)^{\frac{1}{2}n}. \tag{4.19b}$$

Las t_i corresponden a los valores propios de la matriz $\boldsymbol{S}_{11}^{-1/2}S_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1/2}$.

Una buena aproximación, bajo H_0, para la distribución de $\boldsymbol{\lambda}$ es $-2\boldsymbol{\rho}\ln\boldsymbol{\lambda}$, la cual se distribuye $\chi^2_{(p_1p_2)}$, donde

$$\boldsymbol{\rho} = 1 - \frac{p_1 + p_2 + 3}{2n}.$$

Ejemplo 4.3 Un investigador en cultivos perennes tomó 40 árboles de durazno, variedad "Rey Negro", de edades semejantes, midió el diámetro del tronco principal (X_1 en cm), el área foliar (X_2 en cm^2), tiempo para la maduración del fruto (X_3 en dias) y el peso en pulpa por fruto (X_4 en gm).

No sobra señalar que estas medidas son el promedio de algunas mediciones preliminares, es el caso del peso por fruto el cual corresponde al promedio del peso de frutos tomados aleatoriamente de la parte inferior, media y superior de cada árbol.

Con los datos recogidos (redondeados para facilitar cálculos) se estimó la matriz de covarianzas

$$\boldsymbol{S} = \begin{pmatrix} 2 & 5 & 1 & 1 \\ 5 & 15 & 1 & 2 \\ 1 & 1 & 5 & 3 \\ 1 & 2 & 3 & 2 \end{pmatrix}.$$

Con estos datos se pretende verificar la hipótesis que la contextura del árbol está relacionada con la calidad del fruto que produce; más técnicamente, que estas variables fisiológicas están asociadas con las variables morfológicas o de estructura del árbol. Particularmente, que las variables X_1 y X_2 se relacionan con las variables X_3 y X_4. Aquí, $p_1 = p_2$ y

$$\boldsymbol{S}_{11} = \begin{pmatrix} 2 & 5 \\ 5 & 15 \end{pmatrix} \quad \boldsymbol{S}_{12} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \quad \boldsymbol{S}_{21} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \quad S_{22} = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}$$

$$\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21} = \begin{pmatrix} 1 & 3 \\ 3 & 10 \end{pmatrix},$$

luego

$$\begin{aligned} |\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21} - t\boldsymbol{S}_{11}| &= 0 \\ 5t^2 - 5t + 1 &= 0 \\ t &= \frac{5 \pm \sqrt{5}}{10}. \end{aligned}$$

Los valores de $\boldsymbol{\rho}$ y de $\boldsymbol{\lambda}$ son, respectivamente

$$\boldsymbol{\rho} = 1 - \frac{p_1 + p_2 + 3}{2n} = 1 - \frac{7}{80} = 0.9125,$$

y

$$\boldsymbol{\lambda} = 20[\ln(1 - t_1) + \ln(1 - t_2)] = -32.18876,$$

como el valor $-2\boldsymbol{\rho} \ln \boldsymbol{\lambda} = 58.744484$ es muy superior a $\chi^2_{(1\%,4)} = 13.3$ (de la tabla C.7), no se rechaza la existencia de alguna clase de dependencia entre estos pares de variables. Es decir, la calidad del fruto está asociada con la estructura del árbol. ☑

4.3.5 Contraste sobre la igualdad de varias distribuciones normales

Una distribución normal multivariada queda determinada por el vector de medias y la matriz de covarianzas. En el capítulo 3 se presentarom las pruebas sobre la igualdad de los vectores de medias, asumiendo que las matrices de covarianzas son iguales; es decir,

$$H_{0_a}: \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_q, \quad \text{dado que} \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_q$$

La hipótesis sobre la igualdad de varias matrices de covarianzas, expresada como $H_{0_b}: \quad \boldsymbol{\Sigma}_1 = \quad \cdots \quad = \boldsymbol{\Sigma}_q = \boldsymbol{\Sigma}$, se desarrolló en la sección (4.3.2). La hipótesis a considerar ahora es una combinación de H_{0_a} y H_{0_b}. Ésta es

$$H_{0_c}: \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_q, \text{ y } \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_q. \tag{4.18}$$

Sean $\boldsymbol{L}_a$, $\boldsymbol{L}_b$ y $\boldsymbol{L}_c$ los máximos de la funciones de verosimilitud bajo cada una de las tres hipótesis y $\boldsymbol{L}$ el máximo de la función de verosimilitud sin restricción alguna. La hipótesis H_{0_a} es condicionada respecto a la hipótesis H_{0_b}, luego las respectivas razones de máxima verosimilitud son:

$$\boldsymbol{\lambda}_a = \frac{\boldsymbol{L}_a}{\boldsymbol{L}_b}, \ \boldsymbol{\lambda}_b = \frac{\boldsymbol{L}_b}{\boldsymbol{L}} \text{ y } \boldsymbol{\lambda}_c = \frac{\boldsymbol{L}_c}{\boldsymbol{L}}$$

como $\boldsymbol{L}_c$ y $\boldsymbol{L}_a$ son iguales, se concluye que

$$\boldsymbol{\lambda}_c = \boldsymbol{\lambda}_a \cdot \boldsymbol{\lambda}_b. \tag{4.19}$$

De la relación anterior se puede obtener $\boldsymbol{\lambda}_c$ a través de $\boldsymbol{\lambda}_a$ y $\boldsymbol{\lambda}_b$. La estadística

$$\begin{aligned} -2\ln \boldsymbol{\lambda}_c &= -2\ln \boldsymbol{\lambda}_a - 2\ln \boldsymbol{\lambda}_a \\ &= v\ln|\ln\frac{1}{v}\boldsymbol{W}| - \sum_i v_i \ln|\boldsymbol{S}_i|, \end{aligned} \tag{4.20}$$

la cual tiene asociada, de manera asintótica, una distribución *ji–cuadrado* con $\frac{1}{2}p(q-1)(p+3)$ grados de libertad.

Para el caso de una población ($q=1$), la hipótesis $H_{0_c}:\ \boldsymbol{\mu}=\boldsymbol{\mu}_0,$ y $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_0$, frente a la alternativa $H_{0_c1}:\ \boldsymbol{\mu}\neq\boldsymbol{\mu}_0,$ o $\boldsymbol{\Sigma}\neq\boldsymbol{\Sigma}_0$, se verifica mediante la estadística

$$\boldsymbol{\lambda}_c = \Big(\frac{e}{n}\Big)^{\frac{1}{2}pn}|\boldsymbol{H}\boldsymbol{\Sigma}_0^{-1}|^{\frac{1}{2}n}e^{-\frac{1}{2}\left[\operatorname{tra}\left(\boldsymbol{H}\boldsymbol{\Sigma}_0^{-1}\right)+n(\bar{X}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\bar{X}-\boldsymbol{\mu}_0)\right]}, \tag{4.21}$$

con $X_1,\cdots,X_n$, una muestra aleatoria de una población normal p-variante de media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$, $\boldsymbol{H}=\sum_{i=1}^n(X_i-\bar{X})(X_i-\bar{X})'$ y $\bar{X}=\sum_{i=1}^n X_i/n$.

Cuando la hipótesis nula es cierta, la estadística $-2\ln\boldsymbol{\lambda}_c$ tiene distribución asintótica *ji–cuadrado* con $\frac{1}{2}p(p+1)+p$ grados de libertad.

La distribución exacta para la estadística de razón de verosimilitud asociada a la hipótesis anterior fue desarrollada por Nagarsenker y Pillai (1974) a través de dos métodos: (a) series ji–cuadrado y (b) series beta. Estos autores elaboraron tablas para:

$$\alpha\ =\ 0.005,\ 0.01,\ 0.025,\ 0.05,\ 0.1,\ 0.25;\ p\ =\ 2(1)6,$$

que significa entre 2 y 6 variando de a 1; y $n\ =\ 4(1)\ 20(2)\ 40(5)\ 100$. No obstante, el avance paralelo de los métodos numéricos y la computación, hacen posible cada vez más los cálculos requeridos en estos procedimientos.

4.4 Rutina SAS para calcular la estadística de prueba sobre una matriz de covarianzas

```
TITLE1 'EJEMPLO 4.1';
TITLE2 'PRUEBA QUE SIGMA=SIGMA\_0';
```

```
PROCIML; /* invoca el procedimiento IML */
USE EJEMP4\_1;
SIGMA\_0={ 4 3 2, 3 6 5, 2 5 10};
```

/* matriz de covarianzas Σ_0 */

```
S={ 3.42 2.60 1.89, 2.60 8.00 6.51, 1.89 6.51 9.62 }; /* matriz de covarianzas
```

muestral S */

```
P = NROW(SIGMA\_0); /* número de filas de la matriz Σ0 */
GL=(1/2)*P*(P+1); /* grados de libertad */
N=20; /* tamaño de muestra */
ISIGMA\_0=INV(SIGMA\_0); /* matriz Σ0^-1 */
SISIG\_0=(S)*(ISIGMA\_0); /* producto entre S y Σ0^-1 */
D\_SIGMA\_0=DET(SIGMA\_0); /* determinante de Σ0 */
D\_S=DET(S); /* determinante de S */
E\_LAMBDA=(N-1)*(LOG(D\_SIGMA\_0)-LOG(D\_S)+TRACE(SISIG\_0)-P);
                                    /* cálculo de la estadística λ*, ec. (4.6)*/
E\_LAMBDA1=(1-(1/(6*(N-1)))*(2*P+1-2/(P+1)))*(E\_LAMBDA);
                                    /* cálculo de la estadística λ1* ec. (4.7) */
P\_VAL=1-PROBCHI(E\_LAMBDA1,GL); /* calcula el p valor */
RUN;
```

4.5 Procesamiento de datos con R

Se desarrolla el programa para la sección 4.4 en la función `sigma.test()`. El usuario entrega las matrices $\mathbf{\Sigma}_0$, $\mathbf{S}$ y el valor de n y la función regresa el valor de la estadística λ_1^* junto con el $p-$valor

```
sigma.test<-function(Sigma_0,S,n){

# numero de filas la matriz  Sigma
p<-nrow(S)
gl<-(1/2)*p*(p+1)   # grados de libertad

# producto entre S y la inversa de Sigma_0
Sisigma_0<-S%*%solve(Sigma_0)

#determinante de sigma_0
D_sigma_0<-det(Sigma_0)

#determinante de S
D_S<-det(S)

# estadística  lambda de Ecua 4.6
E_lambda<-(n-1)*(log(D_sigma_0)-log(D_S)+
sum(diag(Sisigma_0))-p)
```

```
# estadística  lambda de Ecua 4.7
E_lambda1<-(1-(1/(6*(n-1)))*(2*p+1-2/(p+1) ))* E_lambda
p_val<-pchisq(E_lambda1,gl) # pvalor
list(E_lambda1=E_lambda1, P_valor=p_val) }

# llamado de la función
Sigma_0<-matrix(c(4,3,2,3,6,5,2,5,10),nrow=3 )
S<-matrix(c(3.42,2.60,1.89,2.60,8,6.51,1.89,6.51,9.62),nrow=3)
sigma.test(Sigma_0,S,n=20)
```

La siguiente función desarrolla la prueba de igualdad de matrices de varianzas y covarianzas para dos o mas poblaciones, está basada en el estadístico λ_1 de la ecuación 4.11.

```
var.igual<-function(datos,grupos) {

# matriz de ceros para guardar la suma de v_i*S_i
St<-matrix(0,nrow=ncol(datos),ncol=ncol(datos))

# matrices de covarianzas.
Covs<-by(datos,grupos,cov)

# numero de obs en cada grupo
ni<-as.vector( by(datos,grupos,nrow ) )
vi<-ni-1
lev<-levels(grupos) # niveles del factor

# un vector para guardar los logaritmos de los determinantes.
LnSi<-numeric(1)

# este ciclo for calcula los logaritmos de los determinantes
# y la matriz para calcular Sp
for(i in 1:length(lev)){
Si<- Covs[[i]]
LnSi[i]<-log(det(Si))
St<-St+vi[i]*Si
}

# matriz Sp
Sp<- St/sum(vi)

# se implementa la ec 7.23 Rencher
LnM<-0.5*sum(vi* LnSi) - 0.5*sum(vi)*log(det(Sp))
p<-ncol(Sp)
k<-length(ni)

# numero de grupos
# ec. 7.22 Rencher
c1<-(sum(1/vi)-1/sum(vi))*( (2*p^2+3*p-1)/(6*(p+1)*(k-1)) )
```

```
# ec. 7.24 Rencher
c2<-(sum(1/vi^2)-1/sum(vi)^2)*( (p-1)*(p+2)/( 6*(k-1) )   )
a1<-0.5*(k-1)*p*(p+1)
a2<-(a1+2)/(abs(c2-c1^2))
b1<-(1-c1-a1/a2)/a1
b2<- (1-c1-2/a2)/a2

# ec. 7.21
u<- -2*(1-c1)*LnM
if(c2>=c1^2){
F<- -2*b1*LnM  # ec. 7.25 Rencher
pvalor<-pf(F,a1,a2,lower.tail=FALSE)
}
else{
F<- -( a2*b2*LnM ) /(a1*(1+2*b2*LnM)) # ec. 7.26 Rencher
pvalor<-pf(F,a1,a2,lower.tail=FALSE)
}
Resp<-matrix(c(LnM,u,c1,c2,a1,a2,b1,b2,F,pvalor,-2*LnM),
nrow=1)
colnames(Resp)<-c("LnM","u","c1","c2","a1","a2","b1","b2","F",
"pvalor","-2*LnM")
print(Resp)
}
```

Para verificar la hipótesis que las matrices de varianzas y covarianzas de los tratamientos del ejemplo 3.9 son iguales se usa el comando: (después de haber definido la función)

```
var.igual(Mdatos,ejemp3_9[,1] )
```

Parte II

Métodos

Capítulo 5

Análisis de componentes principales

5.1 Introducción

En el capítulo 1 se hizo una sinopsis de los diferentes métodos de análisis multivariado, éstos se presentan en dos clases: los que suministran información sobre la *interdependencia* entre las variables y los que dan información acerca de la *dependencia* entre una o varias variables respecto a otra u otras. En este capítulo se presenta el *análisis de componentes principales* (en adelante ACP), como uno de los métodos de interdependencia.

En el trabajo de recolección de la información sobre un campo determinado, uno de los problemas que enfrenta el investigador es la elección de las variables a medir. En un proceso de investigación, durante las etapas iniciales frecuentemente hay una escasa teoría sobre el campo a abordar; consecuentemente, el investigador recoge información sobre un número amplio de variables, que a su juicio son relevantes en el problema. En casos donde resultan muchas variables se presentan algunos problemas con la estimación de parámetros, así por ejemplo, con díez variables puede hacerse necesario estimar 45 correlaciones, con 20 se pueden estimar 190 coeficientes de correlación, y así, el número de correlaciones a estimar crece conforme aumenta el número de variables. Además del problema de estimación, está el de la comprensión, de tal forma que se hace necesario abocar alguna técnica que resuma la información contenida en las variables y facilite su análisis.

El ACP tiene como objetivo la estructuración de un conjunto de datos multivariado mediante la reducción del número de variables. Esta es una metodología de tipo matemático para la cual no es necesario asumir distribución probabilística alguna.

En esta sección se desarrolla la técnica del *análisis por componentes principales*, la cual es una metodología para la reducción de datos.

Para comenzar se puede decir que el análisis de componentes principales transforma el conjunto de variables originales en un conjunto más pequeño de variables, las cuales son combinaciones lineales de las primeras, que contienen la mayor parte de la variabilidad presente en el conjunto inicial.

El análisis por componentes principales tiene como objetivos, entre otros, los siguientes:

- Generar nuevas variables que expresen la información contenida en un conjunto de datos.
- Reducir la dimensión del espacio donde están inscritos los datos.
- Eliminar las variables (si es posible) que aporten poco al estudio del problema.
- Facilitar la interpretación de la información contenida en los datos.

El análisis por componentes principales tiene como propósito central la determinación de unos pocos factores (componentes principales) que retengan la mayor variabilidad contenida en los datos. Las nuevas variables poseen algunas características estadísticas "deseables", tales como independencia (bajo el supuesto de normalidad) y no correlación.

En el caso de la no correlación entre las variables originales, el ACP no tiene mucho que hacer, pues las componentes se correspoderían con cada variable por orden de magnitud en la varianza; es decir, la primera componente coincide con la variable de mayor varianza, la segunda componente con la variable de segunda mayor varianza, y así sucesivamente.

A continuación se presenta la interpretación geométrica, el concepto de componente principal, su generación y algunas de sus aplicaciones.

5.2 Interpretación geométrica de las componentes principales

Antes de entrar en la formalidad geométrica de esta técnica, se muestra un caso cuyas observaciones específicas se presentan en la tabla 5.1, ésta

contiene 12 observaciones y 2 variables (X_1 y X_2), junto con los datos corregidos por la media (X_1^* y X_2^*).

Tabla 5.1 Datos originales y centrados

Obs.	X_1	X_1^*	X_2	X_2*
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	-1
5	10	2	8	5
6	9	1	-1	-4
7	8	0	4	1
8	7	-1	6	3
9	5	-3	-3	-6
10	3	-5	-1	-4
11	2	-6	-3	-6
12	0	-8	0	-3
Media	8	0	3	0
Varianza	23.091	23.091	21.091	21.091

Las matrices de covarianzas $\boldsymbol{S}$ y de correlaciones muestral $\boldsymbol{R}$ son,

$$\boldsymbol{S} = \begin{pmatrix} 23.091 & 16.455 \\ 16.455 & 21.091 \end{pmatrix} \text{ y } \boldsymbol{R} = \begin{pmatrix} 1.000 & 0.746 \\ 0.746 & 1.000 \end{pmatrix}.$$

Las varianzas de X_1 y X_2 son 23.091 y 21.091, respectivamente, y la *varianza total* de las dos variables es $23.091 + 21.091 = 44.182$. Además, que las variables X_1 y X_2 están correlacionadas, con un coeficiente de correlación de 0.746. Los porcentajes de la variabilidad total retenida por X_1 y X_2 son, respectivamente, 52.26% y 47.74%.

La figura 5.1 muestra la ubicación de los 12 puntos corregidos por la media. Sea Y_1 un nuevo eje que forma un ángulo θ con el eje X_1. La proyección de las observaciones sobre el eje Y_1 da las coordenadas de las observaciones con respecto a Y_1. Estas coordenadas son una combinación lineal de las coordenadas originales. Por geometría elemental se tiene

$$y_1 = \cos\theta \times x_1^* + \operatorname{sen}\theta \times x_2^*,$$

donde y_1 es la coordenada de la observación con respecto a Y_1. x_1^* y x_2^* son, respectivamente, las coordenadas de la observación con respecto a X_1^* y X_2^*. Por ejemplo, para un valor $\theta = 10^o$, la ecuación para la combinación lineal es

$$\begin{aligned} y_1 &= 0.985x_1^* + 0.174x_2^* \\ &= 0.985(x_1 - 8) + 0.174(x_2 - 3) \\ &= -8.402 + 0.985x_1 + 0.174x_2, \end{aligned}$$

la cual se usa para obtener las coordenadas de las 12 observaciones respecto al eje Y_1. Nótese que las ecuaciones anteriores se pueden expresar en términos de las variables originales; de donde resulta que la respectiva coordenada es una combinación lineal de las variables originales más una constante. Por ejemplo, la coordenada para la primera observación es 8.747. Las coordenadas o proyecciones de las observaciones sobre Y_1 pueden considerarse como los valores y_1 de esta nueva variable. La figura 5.1 muestra los 12 puntos proyectados sobre el eje Y_1. La tabla 5.2 contiene la media y la varianza para los 12 valores de las variables X_1^*, X_2^* y Y_1, respectivamente. De esta tabla se observa que: (1) la nueva variable permanece corregida (con media igual a cero), y (2) la varianza de Y_1 es 28.659 y retiene el 64.87% (28.659/44.182) del total de la varianza de los datos. Nótese que la varianza retenida por Y_1 es mayor que la retenida por cualquiera de las variables originales.

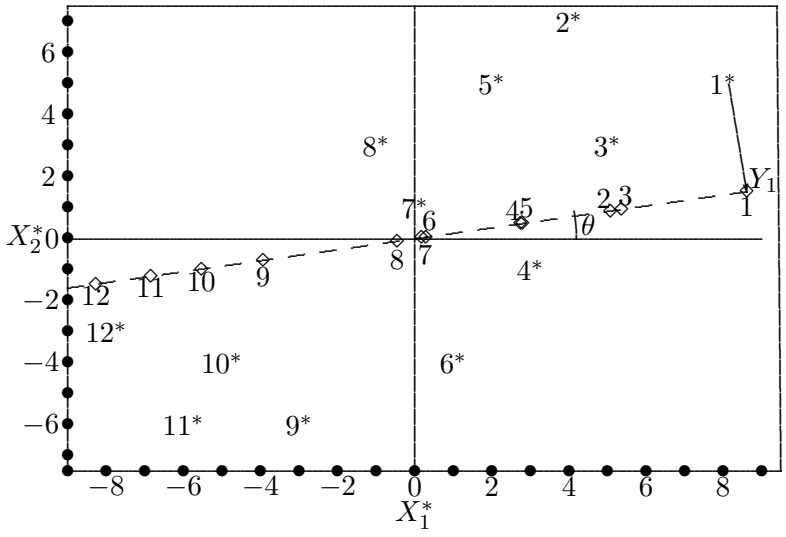


Figura 5.1 *Datos corregidos (∗) y proyectados sobre* Y_1 (⋄).

Tabla 5.2 Puntajes en la primera componente

Obs.	X_1^*	X_2^*	Y_1
1	8	5	8.747
2	4	7	5.155
3	5	3	5.445
4	3	-1	2.781
5	2	5	2.838
6	1	-4	0.290
7	0	1	0.174
8	-1	3	-0.464
9	-3	-6	-3.996
10	-5	-4	-5.619
11	-6	-6	-6.951
12	-8	-3	-8.399
Media	0.000	0.000	0.000
Varianza	23.091	21.091	28.659

Ahora, supóngase que el ángulo entre la variable Y_1 y la variable centrada X_1^* es 20^o en lugar de 10^o. De la misma manera, se obtiene la proyección de las observaciones sobre este nuevo eje. La tabla 5.3 contiene la varianza total, la varianza retenida por la proyección y el porcentaje de varianza retenida para diferentes ángulos que forma la variable Y_1 y la variable centrada X_1^*.

Tabla 5.3 Varianza retenida por el primer eje

Ángulo (θ)	Vza. Total	Vza. de Y_1	Porc. %
0	44.182	23.091	52.263
10	44.182	28.659	64.866
20	44.182	33.434	75.676
30	44.182	36.841	83.387
40	44.182	38.469	87.072
43.261	44.182	38.576	87.312
50	44.182	38.122	86.282
60	44.182	35.841	81.117
70	44.182	31.902	72.195
80	44.182	26.779	60.597
90	44.182	21.091	47.772

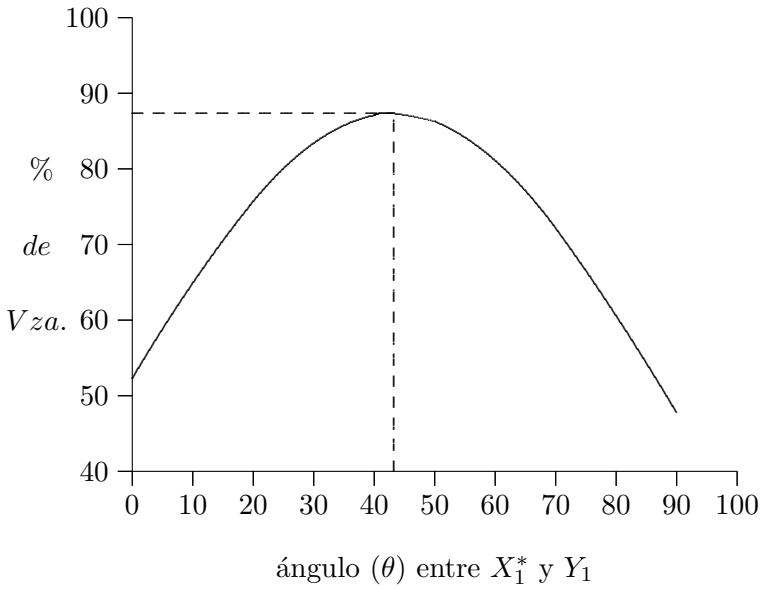


Figura 5.2 *Porcentaje de la varianza total retenida por* Y_1.

La figura 5.2 contiene una gráfica del porcentaje de varianza retenido por Y_1 y el ángulo formado entre Y_1 y la variable centrada X_1^* (primera y última columna de la tabla 5.3). La tabla y la figura permiten apreciar que el porcentaje de varianza explicado por Y_1 crece en tanto el ángulo θ crece, y que después de cierto valor máximo, la varianza reunida por Y_1 decrece. De tal forma, que hay *un único eje nuevo* y es una variable que retiene la máxima cantidad de la variabilidad contenida en los datos. Después de varios ensayos en busca del valor máximo, con la ayuda del gráfico, se advierte que el valor del ángulo óptimo del eje repecto a X_1^* es cercano a 43.261^o. La ecuación para calcular los valores de y_1, en términos de las variables originales, es

$$\begin{aligned} y_1 &= \cos 43.261 \times x_1^* + \operatorname{sen} 43.261 \times x_2^* \\ &= 0.728x_1^* + 0.685x_2^* \\ &= -7.879 + 0.728x_1 + 0.685x_2. \end{aligned}$$

Nótese que la variable Y_1 retiene el 87.31% (38.576/44.182) de la variabilidad total de los datos. Por tanto, es posible identificar un segundo eje, que corresponda a una segunda nueva variable, tal que reuna el máximo

de varianza no retenida por el primer eje Y_1. Sea Y_2 el nuevo segundo eje, el cual se considera ortogonal a Y_1. Así, como el ángulo entre Y_1 y X_1^* es θ entonces el ángulo entre Y_2 y X_2^* también es θ. De manera análoga, la combinación lineal para conformar y_2 es

$$y_2 = -\operatorname{sen}\theta \times x_1^* + \cos\theta \times x_2^*.$$

Para $\theta = 43.261$ la ecuación anterior es

$$\begin{aligned} y_2 &= -\operatorname{sen} 43.261 \times x_1^* + \cos 43.261 \times x_2^* \\ &= -0.685x_1^* + 0.728x_2^* \\ &= -3.296 + -0.685x_1 + 0.728x_2. \end{aligned}$$

La tabla 5.4 contiene los valores de los datos centrados X_1^* y X_2^* y las coordenadas de los 12 datos proyectados sobre los nuevos ejes ortogonales Y_1 y Y_2 (factoriales). En la tabla 5.4 también se reportan las medias y la varianza de las respectivas variables. Además, se han calculado la matriz de covarianzas y la matriz de correlación, $\boldsymbol{S}_Y$ y $\boldsymbol{R}_Y$, respectivamente, entre las nuevas variables. En la figura 5.3 se han graficado las observaciones centradas y los nuevos ejes. A continuación se presentan algunas conclusiones derivadas del desarrollo hecho hasta este punto.

1. La orientación y configuración de los puntos u observaciones no cambia en los dos espacios bidimensionales. Las observaciones pueden, entonces, ser representadas con relación a cualquiera de los dos sistemas: el "viejo" o el "nuevo".

2. La proyección de los puntos hacia los ejes originales reproducen los valores de las variables originales, y recíprocamente, las proyecciones de los puntos sobre los nuevos ejes dan los valores para las nuevas variables. Los nuevos ejes o variables se denominan *componentes principales* y los valores de las nuevas variables se llaman *puntajes de las componentes principales*.

3. Cada una de las nuevas variables es una combinación lineal de las variables originales y se conservan centradas (media cero).

4. La variabilidad total de las variables nuevas ($38.576 + 5.606 = 44.182$) es la misma que la variabilidad total contenida en las variables originales ($23.091 + 21.091 = 44.182$). Es decir, la variabilidad total de los datos no se altera por transformaciones ortogonales de éstos.

5. Los porcentajes de variabilidad retenida por las componentes principales Y_1 y Y_2 son, respectivamente, 87.31% (38.576/44.182) y 12.69% (5.606/44.182). La varianza reunida por la primera nueva variable, Y_1, es mayor que la reunida por cualquiera de la variables originales. La segunda nueva variable, Y_2, reúne la varianza que no ha sido reunida por la primera nueva variable. Las dos variables reúnen toda la variabilidad.

6. Las dos nuevas variables son incorrelacionadas; es decir, su correlación es cero.

Tabla 5.4 Coordenadas factoriales

Obs.	X_1^*	X_2^*	Y_1	Y_2
1	8	5	9.253	-1.841
2	4	7	7.710	2.356
3	5	3	5.697	-1.242
4	3	-1	1.499	-2.784
5	2	5	4.883	2.271
6	1	-4	-2.013	-3.598
7	0	1	0.685	0.728
8	-1	3	1.328	2.870
9	-3	-6	-6.297	-2.313
10	-5	-4	-6.382	0.514
11	-6	-6	-8.481	-0.257
12	-8	-3	-7.882	3.298
Media	0.000	0.000	0.000	0.000
Varianza			38.576	5.606

$$\boldsymbol{S}_Y = \begin{pmatrix} 38.576 & 0.000 \\ 0.000 & 5.606 \end{pmatrix} \quad \boldsymbol{R}_Y = \begin{pmatrix} 1.000 & 0.000 \\ 0.000 & 1.000 \end{pmatrix}.$$

La ilustración anterior de ACP puede extenderse fácilmente a más de dos variables. Con este propósito se muestra la técnica, manteniendo el punto de vista geométrico, para seguir de alguna manera la presentación hecha en la literatura de la escuela francesa. La figura 5.4 es útil para representar y leer las filas y las columnas de la matriz de datos $\mathbb{X}$ como elementos de espacios de dimensión p y n respectivamente. $X_{(1)}, \ldots, X_{(n)}$ indican cada uno de los ejes coordenados para $\mathbb{R}^n$ (variables) y $X^{(1)}, \ldots, X^{(p)}$ indican cada uno de los ejes coordenados para $\mathbb{R}^p$ (individuos).

La distancia entre los puntos fila o individuos tiene un significado. Así, dos puntos cercanos en $\mathbb{R}^p$ implican que las coordenadas de estos puntos deben tener valores similares. En cambio, una distancia pequeña entre dos columnas o variables (puntos de $\mathbb{R}^n$), registradas sobre el conjunto de individuos, significa que ellas miden casi lo mismo. Al decir que dos

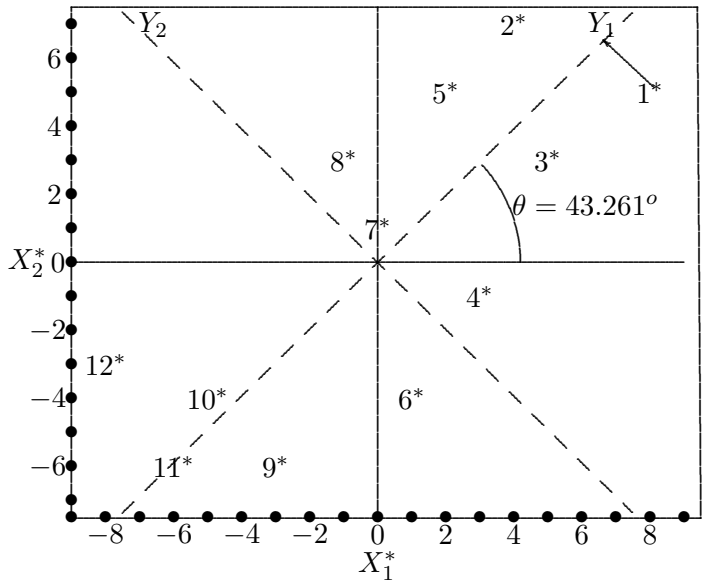


Figura 5.3 *Datos corregidos (∗) y nuevos ejes.*

puntos son cercanos, esto tan sólo significa que ellos tienen valores similares en algunas variables; no necesariamente en todas. El ACP considera y aprovecha este tipo de cercanía.

La técnica de componentes principales se puede comparar con la siguiente situación: a un grupo de personas se les debe tomar una fotografía, de tal manera que la cabeza de cada una sea equivalente a uno de los puntos anteriores en $\mathbb{R}^4$ (tres coordenadas para el espacio y una para el tiempo o la fecha). No es difícil imaginar como la ubicación de la cámara respecto al grupo, producirá fotografías diferentes del mismo grupo; de manera que individuos cercanos en una fotografía, aparecerán muy apartados en otra.

El ACP busca "tomar la mejor fotografía" sobre un conjunto de datos, en este caso una buena fotografía corresponde al *subespacio* de menor dimensión, p por ejemplo, el que provea un buen ajuste para las observaciones y las variables, de tal forma que las distancias entre los puntos en el subespacio suministren una buena representación de las distancias originales. Y para cerrar este paralelo, una fotografía no es otra cosa que la representación en dos dimensiones de un evento que ocurre en cuatro dimensiones (espacio–tiempo).

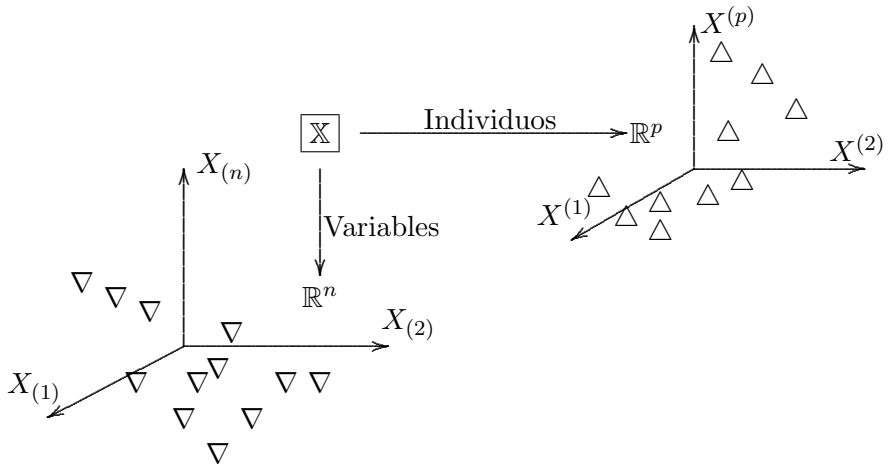


Figura 5.4 *Espacio fila y columna.* $\triangle$*: Individuo,* (∇)*: Variable.*

Para encontrar un subespacio de dimensión q $(q < p)$, tal que los n puntos (filas de $\mathbb{X}$) queden aproximadamente en éste, se empieza por hallar un subespacio de dimensión uno; es decir, una línea recta que contenga al origen, la cual se ajuste lo mejor posible a los datos (Lebart y colaboradores 1984, págs. 3-29). La figura 5.5 muestra el ajuste de los datos a la línea recta $\mathcal{CP}_1$.

La proyección de un vector cualquiera OQ_i (individuos) sobre la recta $\mathcal{CP}_1$ es el vector OP_i. Sea u un vector unitario del subespacio $\mathcal{CP}_1$, entonces la proyección OP_i es el producto escalar entre OQ_i y u. De esta forma, el producto de $\mathbb{X}$ y u es la proyección de cada una de las filas de $\mathbb{X}$ sobre $\mathcal{CP}_1$.

Uno de los criterios para encontrar el "mejor" ajuste es el de mínimos cuadrados. La figura 5.5 ilustra esta técnica. De la relación pitagórica entre los lados del triángulo OP_iQ_i resulta, al sumar sobre cada una de los n triángulos, determinados por sus proyecciones sobre el subespacio $\mathcal{CP}_1$,

$$\sum_{i=1}^{n}(Q_iP_i)^2 = \sum_{i=1}^{n}(OQ_i)^2 - \sum_{i=1}^{n}(OP_i)^2. \tag{5.1}$$

Las proyecciones OP_i reflejan la información recogida en el subespacio $\mathcal{CP}_1$ de cada punto. Se quiere maximizar esta cantidad de información. Como los puntos Q_i están a una distancia fija del origen O_i, maximizar (5.1) es

equivalente a minimizar las distancias OP_i dadas en la misma ecuación. La cantidad a maximizar en función de $\mathbb{X}$ es

$$\sum_{i=1}^{n}(OP_i)^2 = (\mathbb{X}u)'(\mathbb{X}u) = u'\mathbb{X}'\mathbb{X}u, \tag{5.2}$$

con la restricción $u'u = 1$.

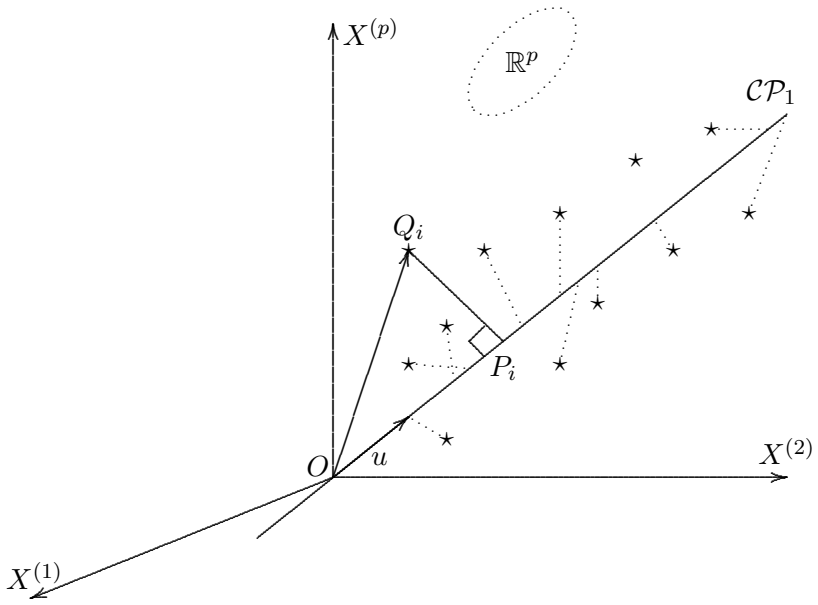


Figura 5.5 *Proyección sobre una línea recta.*

El mejor subespacio bidimensional que ajusta los n puntos es el generado por u_1 y u_2, donde u_2 es el segundo vector en la base para este subespacio ortogonal a u_1 que maximiza a $u_2'\mathbb{X}'\mathbb{X}u_2$.

En forma iterativa el subespacio q-dimensional ($q \leq p$), es el "mejor" en el sentido mínimo cuadrático, y se determina en forma semejante. Este subespacio es generado por los vectores propios $u_1, \ldots, u_q$ de la matriz $\mathbb{X}'\mathbb{X}$, los cuales corresponden a los q valores propios más grandes, como se muestra analíticamente en la sección (5.3).

Se justifica ahora el uso de la palabra el "mejor" respecto al ajuste del subespacio a los datos. Se trata de maximizar la expresión (5.2) con la restricción $u'u = 1$, lo cual es un problema de optimización con restricción

que se resuelve a través de multiplicadores de Lagrange. La expresión a maximizar es $u'\mathbb{X}'\mathbb{X}u - \lambda(u'u - 1)$, al derivar respecto a u e igualar a cero resulta

$$\begin{aligned} 2\mathbb{X}'\mathbb{X}u - 2\lambda us = 0, \text{ equivalentemente} \\ \mathbb{X}'\mathbb{X}u = \lambda u. \end{aligned}$$

con esto se observa que u es el vector propio de $\mathbb{X}'\mathbb{X}$ (A2.24). Ahora, como

$$\begin{aligned} u'\mathbb{X}'\mathbb{X}u &= u'(\mathbb{X}'\mathbb{X}u) = u'\lambda u \\ &= \lambda u'u = \lambda, \end{aligned}$$

se muestra que el máximo se consigue en el *valor propio más grande* de $\mathbb{X}'\mathbb{X}$. Así, se nota por u_1 al vector propio asociado con el valor propio más grande λ_1 de la matriz $\mathbb{X}'\mathbb{X}$. Con esto se concluye que u_1 genera el subespacio $\mathcal{CP}_1$, llamado *el primer componente*.

Si además se busca el subespacio de dos dimensiones con características similares de ajuste al anterior y que lo contenga, entonces, se debe buscar un segundo vector unitario u_2 que maximice $u_2'\mathbb{X}'\mathbb{X}u_2$ y sea ortogonal a u_1. La expresión a optimizar con las restricciones ($u_2'u_1 = 0$ y $u_2'u_2 = 1$) es

$$u_2'\mathbb{X}'\mathbb{X}u_2 - \lambda(u_2'u_2 - 1) - \psi u_2'u_1,$$

con λ y ψ los respectivos multiplicadores de Lagrange.

Al derivar respecto a u_2' e igualar a cero se consigue:

$$2\mathbb{X}'\mathbb{X}u_2 - 2\lambda u_2 - \psi u_1 = 0.$$

Premultiplicando en la igualdad anterior por u_1', y de la ortogonalidad entre u_1 y u_2, se obtiene que

$$2u_1'\mathbb{X}'\mathbb{X}u_2 - 2\lambda u_1'u_2 - \psi u_1'u_1 = 0,$$

como $u_1'\mathbb{X}'\mathbb{X} = \lambda u_1'$, entonces, reemplazando en la ecuación anterior se obtiene

$$2\lambda u_1'u_2' - \psi = 0,$$

de donde, nuevamente por la ortogonalidad entre u_1 y u_2, se concluye que $\psi = 0$. Así:

$$\mathbb{X}'\mathbb{X}u_2 = \lambda u_2';$$

se concluye que u_2 es el segundo vector propio correspondiente al segundo valor propio más grande λ_2 de $\mathbb{X}'\mathbb{X}$. En consecuencia, u_2 genera la recta

$\mathcal{CP}_2$, ortogonal a $\mathcal{CP}_1$, llamada la *segunda componente principal*, y además, $\{u_1, u_2\}$ generan el subespacio de dimensión dos que "mejor" ajusta a los datos.

Mediante un procedimiento análogo, se sigue hasta obtener un subespacio de dimensión $q \leq p$, generado por los q vectores propios ligados a los q-valores propios más grandes de $\mathbb{X}'\mathbb{X}$.

5.2.1 Relación entre los subespacios de $\mathbb{R}^p$ y de $\mathbb{R}^n$

Las p columnas de la matriz $\mathbb{X}$ pertenecen al espacio $\mathbb{R}^n$. Las proyecciones de estos p puntos sobre la línea recta de "mejor" ajuste corresponden a las coordenadas del vector $\mathbb{X}'v$, donde v es un vector unitario contenido en la recta. En forma semejante que en $\mathbb{R}^p$, se trata de maximizar la información contenida en la proyección; es decir, el cuadrado de la longitud del vector $\mathbb{X}'v$

$$v'\mathbb{X}\mathbb{X}'v,$$

con la restricción $v'v = 1$. Se consigue también que el máximo esté en la dirección del vector propio v_1, generado por el valor propio más grande de $\mathbb{X}\mathbb{X}'$. Iterativamente, se obtienen los vectores $v_2, \ldots, v_r$ generadores del subespacio que se ajusta en forma óptima a los datos.

La relación entre u_α y v_α es la siguiente: por definición del vector propio

$$\mathbb{X}\mathbb{X}'v_\alpha = \psi_\alpha v_\alpha,$$

donde v_α y ψ_α son el α-ésimo vector y valor propio de $\mathbb{X}\mathbb{X}'$, respectivamente.

Al premultiplicar por $\mathbb{X}'$ en la igualdad anterior, resulta

$$\begin{aligned}\mathbb{X}'\mathbb{X}(\mathbb{X}'v_\alpha) &= \psi_\alpha \mathbb{X}'v_\alpha \\ (\mathbb{X}'\mathbb{X})u_\alpha &= \psi_\alpha u_\alpha,\end{aligned}$$

con $u_\alpha = \mathbb{X}'v_\alpha$. Se concluye entonces que todo valor propio no nulo, de la matriz $\mathbb{X}\mathbb{X}'$, es un valor propio de la matriz $\mathbb{X}'\mathbb{X}$, y los vectores propios correspondientes se relacionan mediante

$$u_\alpha = k_\alpha \mathbb{X}'v_\alpha$$

con k_α una constante.

Premultiplicando por $\mathbb{X}$ los miembros de la ecuación $\mathbb{X}'\mathbb{X}u_\alpha = \lambda_\alpha u'_\alpha$, se obtiene

$$(\mathbb{X}\mathbb{X}')\mathbb{X}u_\alpha = \lambda_\alpha(\mathbb{X}u_\alpha).$$

Así, a todo vector propio u_α de $\mathbb{X}'\mathbb{X}$ le corresponde un vector propio $\mathbb{X}u_\alpha$ de $\mathbb{X}\mathbb{X}'$ con relación al mismo valor propio λ_α.

En conclusión

$$\lambda_\alpha = \psi_\alpha \quad \text{y} \quad v_\alpha = k'_\alpha \mathbb{X} u_\alpha,$$

y como $u'_\alpha u_\alpha = v'_\alpha v_\alpha = 1$, se puede establecer la relación $k_\alpha = k'_\alpha = 1/\sqrt{\lambda_\alpha}$. De lo anterior se derivan las siguientes relaciones, las cuales permiten obtener las coordenadas de un punto a partir de su representación en el otro espacio en forma recíproca (fila o columna)

$$\begin{aligned} u_\alpha &= \frac{1}{\sqrt{\lambda_\alpha}} \mathbb{X}' v_\alpha \\ v_\alpha &= \frac{1}{\sqrt{\lambda_\alpha}} \mathbb{X} u_\alpha. \end{aligned} \tag{5.3}$$

Las coordenadas de la nube de puntos sobre el eje α en $\mathbb{R}^p$ (o en $\mathbb{R}^n$) son las componentes de $\mathbb{X}u_\alpha$ (o de $\mathbb{X}'v_\alpha$). De manera que existe una relación de proporcionalidad entre las coordenadas sobre los respectivos ejes de los espacios fila o columna (Lebart y colaboradores 1985, pág. 282).

5.2.2 Reconstrucción de la matriz de datos

Volviendo al caso de la fotografía, el probrema ahora es: ¿Cómo quedó cada una de las personas del grupo dispuestas en la fotografía?. En general, habiendo proyectado un conjunto de puntos (filas o columnas de $\mathbb{X}$) sobre un subespacio de menor dimensión al inicial, ¿Cómo se ubican los objetos en este nuevo espacio?.

Siguiendo la construcción desarrollada anteriormente del "mejor" subespacio, la primera ubicación sería la proyección de los n puntos sobre el *primer componente principal* o *eje factorial* $\mathcal{CP}_1$ (figura 5.5), de esta forma el primer valor propio representa la cantidad de proyección recogida por este eje, como la suma de las proyecciones al cuadrado; es decir, $\lambda_1 = u'_1 \mathbb{X}' \mathbb{X} u_1$. En general, el subespacio de dimensión q posibilita reconstruir en forma adecuada la posición de los puntos si $\lambda_1 + \lambda_2 + \cdots + \lambda_q$ es una proporción alta de la traza de $\mathbb{X}'\mathbb{X}$; esto es, si

$$\lambda_1 + \lambda_2 + \cdots + \lambda_q \approx \text{tra}(\mathbb{X}'\mathbb{X}) = \sum_{\alpha=1}^{p} \lambda_\alpha.$$

De la relación mostrada en las ecuaciones (5.3), para el espacio $\mathbb{R}^p$ se escribe $\mathbb{X}u_\alpha = \sqrt{\lambda_\alpha} v_\alpha$. Postmultiplicando esta ecuación por u'_α se consigue

$$\begin{aligned}\mathbb{X}u_\alpha u'_\alpha &= \sqrt{\lambda_\alpha}v_\alpha u'_\alpha \\ &= \mathbb{X}\left\{\sum_{\alpha=1}^{p} u_\alpha u'_\alpha\right\} \\ &= \sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha}v_\alpha u'_\alpha.\end{aligned}$$

La cantidad $\{\sum_{\alpha=1}^{p} u_\alpha u'_\alpha\} = \boldsymbol{I}_p$, pues es el producto de vectores ortonormales, de donde resulta

$$\underbrace{\mathbb{X}}_{n\times p} = \sum_{\alpha=1}^{p}\sqrt{\lambda_\alpha}\Big(\underbrace{v_\alpha}_{n\times 1}\ \underbrace{u'_\alpha}_{1\times p}\Big).$$

Se consigue una reconstrucción aproximada de la matriz de datos $\mathbb{X}$, a través de la matriz $\hat{\mathbb{X}}$, la cual se obtiene a partir de los q primeros ejes principales, siempre que la proporción de traza no reunida por estos ejes ("ruido") sea pequeña. Así,

$$\mathbb{X} \approx \hat{\mathbb{X}} = \sum_{\alpha=1}^{q}\sqrt{\lambda_\alpha}\big(v_\alpha u'_\alpha\big). \tag{5.4}$$

Nótese que se están reemplazando $(n \times p)$ números de la matriz $\mathbb{X}$ por tan sólo $q \times (n + p)$ conformados por q–vectores $\sqrt{\lambda_\alpha}v_\alpha$ de tamaño $n \times 1$ y q–vectores u_α de tamaño $p \times 1$, respectivamente.

Así por ejemplo, si se tiene una matriz de tamaño (100×1000), un subespacio de dimensión 10 reduce los 100.000 datos a $10 \times (100 + 1000) = 11000$ datos.

Por comodidad y para efectos de aplicación, es una práctica común ubicar los datos, sean individuos o variables, en los dos primeros ejes. A este plano se le conoce con el nombre de *plano factorial*. Una representación en el plano factorial facilita la interpretación de los ejes, como también permite hacer algunas clasificaciones de los individuos, de las variables o ambos, la detección de posibles valores atípicos y el diagnóstico de la normalidad de los datos, entre otras aplicaciones.

5.3 Determinación de las componentes principales

En un estudio realizado sobre n-individuos mediante p-variables $X_1, \ldots, X_p$, es posible encontrar nuevas variables notadas por Y_k que sean combinaciones lineales de las variables originales X_j, y sujetas a ciertas condiciones.

El desarrollo del ACP es semejante a una regresión lineal del componente principal sobre las variables originales.

En tal sentido se determina la primera componente principal Y_1, la cual sintetiza la mayor cantidad de variabilidad total contenida en los datos. Así:

$$Y_1 = \gamma_{11}X_1 + \gamma_{12}X_2 + \ldots + \gamma_{1p}X_p, \tag{5.5}$$

donde las ponderaciones $\gamma_{11}, \ldots, \gamma_{1p}$ se escogen de tal forma que maximicen la razón de la varianza de Y_1 a la variación total; con la restricción: $\sum_{j=1}^{p} \gamma_{1j}^2 = 1$.

La segunda componente principal Y_2 es una combinación lineal ponderada de las variables observadas, la cual no está correlacionada con la primera componente principal y reune la máxima variabilidad restante de la variación total contenida en la primera componente principal Y_1. De manera general, la k-ésima componente es una combinación lineal de las variables observadas X_j, para $j = 1, \ldots, p$.

$$Y_k = \gamma_{k1}X_1 + \gamma_{k2}X_2 + \cdots + \gamma_{kp}X_p, \tag{5.6}$$

la cual tiene la varianza más grande entre todas las siguientes. De otra manera, los Y_k sintetizan en forma decreciente la varianza del conjunto original de datos.

A continuación se muestra cómo generar las componentes principales.

Supóngase que el vector aleatorio $X' = (X_1, \ldots, X_p)$ tiene matriz de varianzas y covarianzas $\mathbf{\Sigma}$. Sin pérdida de generalidad asúmase que la media de los X_i es cero, para todos los $i = 1, \ldots, p$; esto siempre es lícito, pues de otra manera sólo basta con centrar (restando la media) el vector X . Para encontrar la primera componente principal, se examina el vector de coeficientes $\Gamma' = (\gamma_{11}, \ldots, \gamma_{1p})$, tal que la varianza $\Gamma' X$ sea un máximo sobre la clase de todas las combinaciones lineales $\Gamma' X$, con la restricción $\Gamma'\Gamma = 1$.

De esta manera, se determina la combinación lineal

$$Y = \sum_{j=1}^{p} \gamma_{1j} X_j, \tag{5.7}$$

tal que

$$\text{var}(Y) = var\Big(\sum_{j=1}^{p} \gamma_{1j} X_j\Big), \tag{5.8}$$

sea máxima, donde $\sum\limits_{j=1}^{p} \gamma_{1j}^2 = 1$.

La restricción que Γ sea un vector unitario, se hace para evitar el incremento de la varianza de manera arbitraria; de lo contrario, ésta se incrementaría tan sólo aumentando cualquiera de las componentes de Γ. El problema ahora es maximizar $\text{var}(\Gamma' X) = \Gamma' \mathbf{\Sigma} \Gamma$ con respecto a Γ, sujeto a $\Gamma'\Gamma = 1$, lo cual, por multiplicadores de Lagrange, equivale a resolver

$$(\mathbf{\Sigma} - \lambda_1 I)\Gamma_1 = 0. \tag{5.9}$$

Para que la solución de (5.9) sea diferente de la trivial, el vector de Γ_1 debe ser escogido de tal manera que

$$|\mathbf{\Sigma} - \lambda_1 \boldsymbol{I}| = 0. \tag{5.10}$$

La ecuación (5.10) corresponde a la ecuación característica, su solución es el valor propio más grande de $\mathbf{\Sigma}$ y Γ_1 el correspondiente vector propio.

Así, la primera componente principal puede escribirse de la siguiente forma

$$Y_1 = \Gamma_1' X. \tag{5.11}$$

La segunda componente principal se determina encontrando un segundo vector normalizado Γ_2, ortogonal a Γ_1, tal que $Y_2 = \Gamma_2' X$ tenga la segunda varianza más grande entre todos los vectores que satisfacen:

$$\Gamma_1'\Gamma_2 = 0 \qquad y \qquad \Gamma_2'\Gamma_2 = 1.$$

Mediante este mismo razonamiento, se demuestra que Γ_2 es el vector propio correspondiente al segundo valor propio más grande de $\mathbf{\Sigma}$. El proceso se desarrolla hasta encontrar los p-vectores; donde Γ_r es ortonormal a $\Gamma_1, \ldots, \Gamma_{r-1}$ con $r = 2, \ldots, p$. En la mayor parte de los análisis se asume que las raíces de $\mathbf{\Sigma}$ son distintas, esto implica que sus vectores propios asociados son mutuamente ortogonales; si además, se asume que $\mathbf{\Sigma}$ es definida positiva, entonces todas las raíces son positivas. En este caso, el rango corresponde al número de valores propios no nulos. Por un camino matricial se pueden obtener también las componentes principales. Por definición, $\mathbf{\Sigma}$ es una matriz real simétrica con raíces diferentes de cero, entonces ésta se puede escribir, de acuerdo con la descomposición espectral (A2.27), como

$$\mathbf{\Sigma} = \mathbf{\Gamma \Lambda \Gamma'}, \tag{5.12}$$

donde $\mathbf{\Lambda}$ es una matriz diagonal cuyos elementos son $\lambda_1, \ldots, \lambda_p$ y $\mathbf{\Gamma}$ es una matriz ortogonal cuya j-ésima columna es el j-ésimo vector propio Γ_j

asociado a λ_j. Los elementos de $\mathbf{\Gamma}$ son los γ_{ij}, los cuales dan cuenta de la contribución de la i-ésima variable en la j-ésima componente lineal.

El vector de componentes principales, que resulta de la transformación lineal $\mathbf{\Gamma}$ aplicada sobre el vector X, es

$$Y' = \big(Y_1, \ldots, Y_p\big);$$

se escribe

$$Y = \mathbf{\Gamma}' X. \tag{5.13}$$

La misma transformación aplicada sobre los datos contenidos en la matriz de datos $\mathbb{X}$, produce

$$\mathbb{Y} = \mathbf{\Gamma}' X, \tag{5.13a}$$

la cual corresponde a la matriz que representa a los mismos individuos representados en la matriz $\mathbb{X}$, pero ahora referidos a los "nuevos" ejes principales.

La matriz de varianzas y covarianzas de Y está dada por

$$\mathrm{Cov}(Y) = \mathbf{\Gamma}'\mathbf{\Sigma}\mathbf{\Gamma}. \tag{5.14}$$

Sustituyendo $\mathbf{\Sigma}$ por la ecuación (5.12)

$$\mathrm{Cov}(Y) = \mathbf{\Gamma}'\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}'\mathbf{\Gamma} = \mathbf{\Lambda}. \tag{5.15}$$

Como $\mathbf{\Lambda}$ es una matriz diagonal, las componentes principales son incorrelacionadas y la varianza de la k-ésima componente principal es su respectivo valor propio:

$$\mathrm{var}(Y_k) = \lambda_k.$$

La traza de la matriz $\mathbf{\Sigma}$ es:

$$\mathrm{tra}(\mathbf{\Sigma}) = \sum_{k=1}^{p} \sigma_{kk}^2. \tag{5.16}$$

Nuevamente, por (5.12) resulta

$$\mathrm{tra}(\mathbf{\Sigma}) = \mathrm{tra}(\mathbf{\Gamma}'\mathbf{\Lambda}\mathbf{\Gamma});$$

por las propiedades de traza (A2.13)

$$\begin{aligned} \mathrm{tra}(\mathbf{\Sigma}) &= \mathrm{tra}(\mathbf{\Gamma}'\mathbf{\Gamma}\mathbf{\Lambda}) \\ &= \mathrm{tra}(\mathbf{\Lambda}) = \sum_{j=1}^{p} \lambda_j. \end{aligned}$$

La expresión (5.17) indica que la varianza total de las variables originales es igual a la suma de las varianzas en cada una de las componentes principales.

En resumen, la transformación lineal que sintetiza la máxima variabilidad contenida en los datos corresponde a la generada por el valor propio más grande de los λ_i. Es costumbre notar al valor propio más grande como λ_1, de tal manera que los valores propios

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p, \tag{5.17}$$

generan las componentes principales, que en orden descendente, sintetizan la variabilidad del conjunto de datos originales.

▶ ACP bajo multinormalidad

Aunque, como se afirmó inicialmente, la técnica del ACP no requiere el supuesto de normalidad, se presenta aquí la caracterización de las componentes generadas bajo el ambiente de normalidad; es decir, cuando los datos $\boldsymbol{X}$ se distribuyen conforme a una $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

La deducción matemática de las componentes principales no se altera cuando las observaciones $\mathbb{X}$ proceden de una distribución normal p-variante. Si $\boldsymbol{\Gamma}$ denota la matriz de vectores propios de $\boldsymbol{\Sigma}$, entonces las componentes principales pueden escribirse de la siguiente forma

$$\boldsymbol{Y} = \boldsymbol{\Gamma}'(\boldsymbol{X} - \boldsymbol{\mu}),$$

la cual es equivalente a la ecuación (5.13), donde, antes de la rotación ortogonal, se ha efectuado una translación del origen para hacer que $\boldsymbol{Y}$ tenga media cero.

Ahora la diferencia es que al conocer la distribución de $\boldsymbol{X}$, se puede encontrar la distribución de $\boldsymbol{Y}$. De acuerdo con las propiedades de la distribución normal (sección (2.2)), cada componente del vector $\boldsymbol{Y}$ tiene distribución normal por ser una combinación lineal de variables aleatorias con distribución normal. Se demuestra que la distribución de $\boldsymbol{Y}$, a través del teorema de la transformación dado por la ecuación (1.8) de la sección (1.4.4), es normal multivariada. así,

$$g(y) = f_y(x)|J|,$$

donde $|J|$ es el Jacobiano de la transformación $\boldsymbol{Y}$; como $\boldsymbol{\Gamma}$ es una matriz ortogonal, la transformación inversa es $\boldsymbol{X} = \boldsymbol{\Gamma}Y + \boldsymbol{\mu}$. El Jacobiano asociado con la transformación es $|J| = |\frac{\partial X}{\partial Y}| = |\boldsymbol{\Gamma}| = 1$, nuevamente, por ser la

matriz $\mathbf{\Gamma}$ ortogonal (sección (A.2.2)). La función de distribución conjunta de $\boldsymbol{X}$ es (ecuación (2.1))

$$f_X(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\Big\{-\frac{1}{2}(x-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(x-\boldsymbol{\mu})\Big\},$$

entonces por la descomposición espectral mostrada en la ecuación (5.12), la matriz de covarianzas se puede expresar como

$$\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}';$$

así,

$$|\mathbf{\Sigma}| = |\mathbf{\Gamma}| \cdot |\mathbf{\Lambda}| \cdot |\mathbf{\Gamma}| = |\mathbf{\Lambda}|,$$

y además

$$\mathbf{\Sigma}^{-1} = \mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}',$$

de donde se tiene

$$\begin{aligned} g(y) &= \frac{1}{(2\pi)^{p/2}|\mathbf{\Lambda}|^{1/2}} \exp\Big\{-\frac{1}{2}y'\mathbf{\Lambda}^{-1}y\Big\} \\ &= \frac{1}{(2\pi)^{p/2}\Big(\prod_{i=1}^{p}\lambda_i\Big)^{\frac{1}{2}}} \exp\Big\{-\frac{1}{2}\sum_{i=1}^{p} y_i^2/\lambda_i\Big\}, \end{aligned}$$

así, $g(y)$ es el producto de normales independientes, y esto implica la normalidad de $\boldsymbol{Y}$.

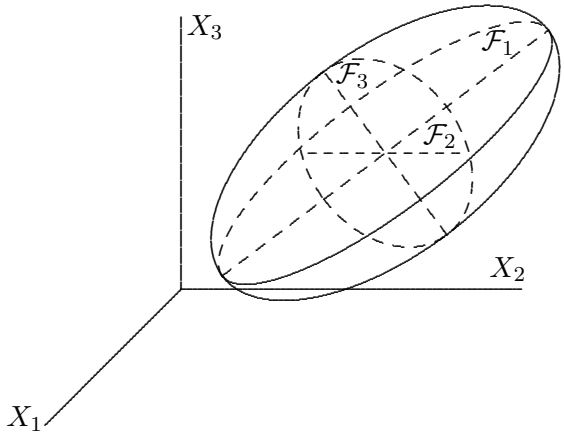


Figura 5.6 *Componentes principales bajo normalidad.*

De esta manera, en el caso normal multivariado, las componentes principales tienen una interpretación geométrica sencilla. Si la función de densidad conjunta de $\boldsymbol{X}$ es constante en un elipsoide del espacio $\mathbb{R}^p$, las componentes principales corresponden a los *ejes principales* del elipsoide. Como se muestra en las secciones (2.6) y (2.7), en el primer eje principal se encuentra el segmento de mayor longitud; cuyos extremos están en el elipsoide descrito por la ecuación $(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) = C$.

La figura 5.6 muestra los ejes principales $\mathcal{F}_1$, $\mathcal{F}_2$ y $\mathcal{F}_3$ para un conjunto de datos que proceden de una distribución normal trivariada.

5.4 Generación de las componentes principales

Aunque no es muy común ni adecuado hablar de estimación de las componentes principales, aquí se presenta como *la generación de componentes principales*, cuando la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ (o de correlación) no se conoce; esta debe ser estimada de la muestra. Dos son las formas más comunes de generar las componentes principales. La primera es a partir de la matriz de varianzas y covarianzas y la segunda a través de la matriz de correlación.

5.4.1 A partir de la matriz de varianzas y covarianzas

Supóngase que los valores de p-variables $X_1, \ldots, X_p$ se obtienen sobre una muestra de n-individuos, la matriz $\mathbb{X}$, de tamaño $(n \times p)$ representa tales datos.

Sean $\bar{X}_j$ la media muestral de la variable X_j; s_{kj} la covarianza muestral entre las variables X_j y X_k y la matriz $\boldsymbol{S} = (s_{jk})$, corresponde a la matriz de varianzas y covarianzas muestral de las p-variables.

Paralelamente a lo desarrollado en la última sección, se trata de encontrar la primera componente principal, que tenga máxima varianza muestral, a través de la combinación lineal

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p = \sum_{j=1}^{p} a_{1j}X_j.$$

La varianza muestral de esta combinación lineal es igual a $\boldsymbol{a}_1'\boldsymbol{S}\boldsymbol{a}_1$, donde el vector $\boldsymbol{a}_1' = (a_{11}, \ldots, a_{1p})$, es tal que $\|a_1\| = 1$. Las componentes de $\boldsymbol{a}_1$ deben satisfacer

$$[\boldsymbol{S} - l_1\boldsymbol{I}]\boldsymbol{a}_1 = 0, \tag{5.18}$$

con l_1 el multiplicador de Lagrange. Entonces l_1 es tal que

$$|\boldsymbol{S} - l_1\boldsymbol{I}| = 0; \tag{5.19}$$

la cantidad l_1 es el valor propio más grande de $\boldsymbol{S}$ y $\boldsymbol{a}_1$ su correspondiente vector propio.

El proceso para la extracción de las demás componentes principales es similar. Las m componentes principales ($m \leq p$) que contienen, de manera decreciente, fracciones de la varianza total, se generan a partir de los respectivos $\boldsymbol{a}_k$; con $l_1 \geq l_2 \geq \cdots \geq l_m \geq \cdots \geq l_p$ los valores propios de $\boldsymbol{S}$. El orden, en términos de la magnitud, de los valores propios (y variabilidad) con que se generan las componentes principales hace que algunos usuarios de esta metodología la califiquen como una *técnica de ordenamiento.*

La contribución de la i-ésima variable a la k-ésima componente principal está dada por la magnitud del coeficiente a_{ki}. La covarianza entre la variable X_i y la componente principal Y_k es :

$$\text{cov}(X_i, Y_k) = a_{ki}l_k. \tag{5.20}$$

La varianza muestral de las observaciones con respecto a la k-ésima componente principal es:

$$\text{var}(Y_k) = \boldsymbol{a}_k' S \boldsymbol{a}_k = l_k, \tag{5.21}$$

con l_k el k-ésimo valor propio ordenado descendentemente.

La varianza total de las p-variables es:

$$VT = \text{tra}(S) = \sum_{j=1}^{p} l_j. \tag{5.22}$$

Al dividir a (5.20) por la desviación estándar de X_i y Y_k respectivamente, se obtiene la correlación

$$r_{X_iY_k} = \frac{a_{ki}l_k}{(\sqrt{s_{ii}})(\sqrt{l_k})} = \frac{a_{ki}\sqrt{l_k}}{\sqrt{s_{ii}}}. \tag{5.23}$$

La expresión (5.23) suministra la *ponderación* (o grado de importancia) de la i-ésima variable sobre la k-ésima componente principal. Además, nótese que esta correlación depende en forma directa de las a_{ki}, pues s_{ii} y l_k son fijos. Ésta se convierte en una forma de leer e interpretar las componentes principales, ya que una observación de los valores a_{ki} auxilia la búsqueda del significado de las diferentes componentes principales.

En resumen, el procedimiento para obtener componentes principales, mediante la matriz de varianzas y covarianzas muestral, es el siguiente:

1. Estimar la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$; es decir, calcular la matriz $\boldsymbol{S}$.
2. Obtener los valores propios de la matriz $\boldsymbol{S}$; éstos corresponden a la varianza de cada componente principal.
3. Hallar la razón entre cada uno de los valores propios y la suma total de ellos (la traza de $\boldsymbol{S}$), e ir acumulando estas razones.
4. Los valores más altos obtenidos en (3) suministran un indicio del número de componentes relevantes.
5. Calcular las ponderaciones dadas en (5.23), las cuales indican el grado de asociación entre la variable y la componente principal respectiva.
6. Calcular los "nuevos" puntajes mediante la transformación $\boldsymbol{Y} = \boldsymbol{AX}$, donde $\boldsymbol{A}$ es la matriz ortogonal que define la rotación "rígida".
7. Interpretar los "nuevos" ejes. Ésta es, tal vez, la parte crucial de todo lo anterior, pues la carencia de una interpretación puede hacer que este trabajo se convierta en un ejercicio púramente numérico.

5.4.2 Mediante la matriz de correlaciones

Hasta ahora se han obtenido las componentes principales mediante los valores y vectores propios de la matriz de varianzas y covarianzas. Dado que la varianza de cualquier variable aleatoria no es invariante por cambios de escala, las componentes principales también sufren alguna variación. Este problema se puede obviar con la estandarización previa de las variables.

Sea $\boldsymbol{R}$ la matriz de correlaciones muestral, $\boldsymbol{R}$ se relaciona con la matriz de varianzas y covarianzas muestral $\boldsymbol{S}$, por medio de la siguiente expresión

$$\boldsymbol{R} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}, \tag{5.24}$$

donde $\boldsymbol{D}^{-\frac{1}{2}} = \text{Diag}(1/s_i)$.

El procedimiento para la consecusión de las componentes principales es igual al que se hizo a partir de la matriz de varianzas y covarianzas, sólo que aquí es necesario sustituir $\boldsymbol{S}$ por $\boldsymbol{R}$.

Existen algunas diferencias en la interpretación, la más relevantes son las siguientes:

- La suma de los valores propios es igual a p; es decir, la variabilidad total coincide con la dimensión de matriz $\boldsymbol{R}$.

- La proporción de la variabilidad total atribuible a cada componente principal es l_k/p.
- La ponderación de la variable i, en la k-ésima componente, está dada por $(a_{ki})(l_k)^{1/2}$.
- La matriz de transformación $\boldsymbol{A}$, en general, es diferente de la obtenida con $\boldsymbol{S}$. Esta característica de los valores y vectores propios hacen que el ACP sea sensible a cambios de escala. Por tal razón, se deben examinar cuidadosamente los datos originales en sus promedios y varianzas, con el ánimo de decidir sobre qué matriz conviene emplear y la interpretación que debe hacerse sobre los componentes generados.

En resumen, si se tiene una matriz de datos $\mathbb{X}$ de tamaño $(n \times p)$, los puntajes de las observaciones (filas de $\mathbb{X}$), respecto a las componentes principales, incluida una corrección por la media $\bar{X}$ o translación del origen a la media muestral, son dadas por la transformación

$$\mathbb{Y} = (\mathbb{X} - \bar{X})\boldsymbol{A}. \tag{5.25}$$

donde la matriz $\boldsymbol{A}$ está conformada por los vectores propios ortonormales, de $\boldsymbol{S}$, de $\boldsymbol{R}$ o de $\mathbb{X}'\mathbb{X}$.

Una aproximación a la matriz $\mathbb{X}$ se obtiene de (5.26) empleando tan sólo algunas $m < p$ componentes principales; así, asumiendo que la matriz $\mathbb{X}$ está centrada (medias iguales a cero), entonces la matriz $\hat{\mathbb{X}}$ aproxima a la matriz $\mathbb{X}$ mediante la siguiente expresión

$$\hat{\mathbb{X}} = \mathbb{Y}\boldsymbol{A}', \tag{5.26a}$$

donde $\mathbb{Y}$ es la matriz $(n \times m)$ de observaciones sobre las primeras m componentes principales y $\boldsymbol{A}'$ es la matriz $(p \times m)$ cuyas columnas son los primeros[1] m vectores propios de la matriz $\boldsymbol{S}$ (o de $\boldsymbol{R}$).

5.5 Selección del número de componentes principales

No hay criterios estrictamente formales para la determinación del número de componentes principales a mantener, excepto bajo normalidad como se muestra al final de esta sección. Los criterios sugeridos son de tipo empírico,

[1] El orden está asociado con la magnitud decreciente de los respectivos valores propios.

y se basan en la variabilidad (información) que en una situación particular se quiere mantener. Existen algunas ayudas gráficas con las cuales se decide acerca del número adecuado de componentes.

Se mostró (ecuación 5.23) que la suma de las varianzas originales, es la traza de $\boldsymbol{S}$ y es igual a la suma de los valores propios de $\boldsymbol{S}$. También, que la varianza de cada componente principal es igual al valor propio que la generó, ecuación (5.22). Es decir,

$$\sum_{j=1}^{p} s_{jj} = \sum_{k=1}^{p} l_k \quad \text{con } j, k = 1, \ldots, p; \tag{5.26}$$

cada componente principal explica una proporción de la variabilidad total, tal proporción se puede calcular mediante el cociente entre el valor propio y la traza de $\boldsymbol{S}$.

$$\frac{l_k}{\text{tra}(\boldsymbol{S})}. \tag{5.27}$$

El cociente (5.28) se denomina *la proporción de la variabilidad total explicada por el k-ésimo componente*. De acuerdo con la ecuación (5.28), un criterio consiste en tomar un número de componentes igual al número de valores propios que están por encima de la media; de esta manera, si las componentes han sido generadas desde la matriz de correlaciones, se seleccionan las componentes cuyos valores propios asociados sean mayores que 1.0 (pues $\sum_{k=1}^{p} l_k/p = 1.0$).

Para la construcción de cada componente principal, los valores propios se toman en orden decreciente (sección (5.4.1)), si de éstos se consideran los m-primeros, entonces la "eficiencia" será la proporción acumulada de variación total, explicada por ellos. Así,

$$\sum_{k=1}^{m} \frac{l_k}{\text{tra}(S)} \times 100\%, \tag{5.28}$$

es el porcentaje de variación total explicado por las m-primeras componentes principales; la variación no retenida por éstas ($\sum_{k=m+1}^{p} l_k/\text{tra}(\boldsymbol{S})$) se asume como *"ruido"* de los datos. En la figura 5.7 se muestra la proporción de la variabilidad total retenida hasta cada componente.

Una expresión equivalente a (5.29) es :

$$\frac{\sum_{k=1}^{m} l_k}{\sum_{k=1}^{p} l_k} \times 100\%; \; m < p. \tag{5.29}$$

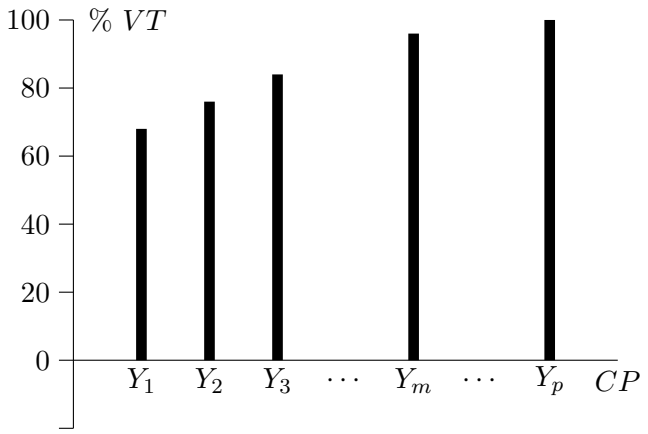


Figura 5.7 *Variación retenida hasta cada componente principal.*

Es inmediata la verificación de que si $m = p$ entonces (5.29) o (5.30) son iguales al 100%.

Una vez que se haya decidido por el porcentaje de variación explicado, que se considera satisfactorio; solo se debe escoger el número m que cumpla tal requerimiento; es decir, el número de componentes principales.

Dillon y Goldstein (1986, págs. 47-50) describen algunos métodos gráficos los cuales sirven como herramienta para la elección del número de componentes principales, suficientes para retener una proporción adecuada de la variabilidad total.

Los autores citan el procedimiento siguiente: en un diagrama cartesiano se ubican los puntos cuyas coordenadas son las componentes principales o factores (CP) y los valores propios, ordenados de forma descendente. Si a partir de algún punto (parte derecha) se puede trazar una línea recta de pendiente pequeña (a manera de ajuste), el número de componentes está dado por los puntos ubicados arriba de tal línea. La figura 5.8 representa una situación ideal. Obsérvese que los tres primeros factores son los candidatos para escoger las componentes principales que retienen una considerable cantidad de la variabilidad total.

Un segundo procedimiento, similar al anterior, consiste en elaborar un gráfico en donde se representa el porcentaje de variación explicado por cada componente o factor en las ordenadas y las componentes en orden decreciente en la abscisas (figura 5.9).

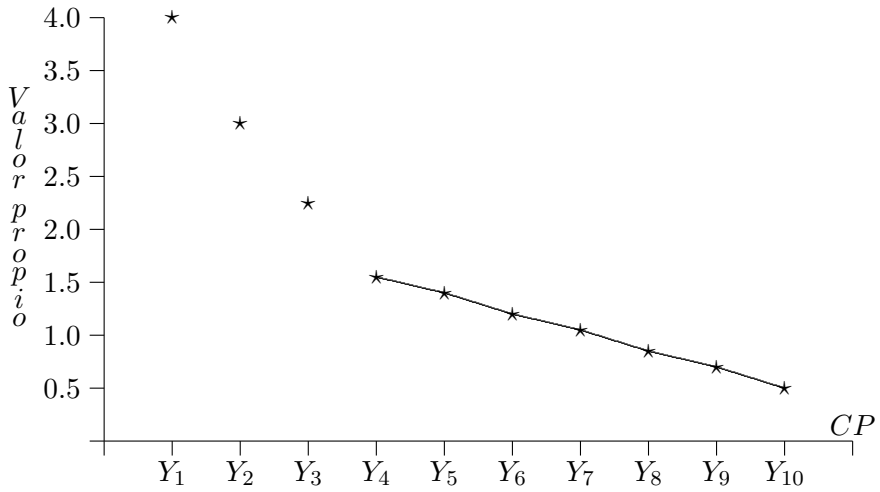


Figura 5.8 *Selección del número de componentes principales.*

La decisión es escoger los factores que retengan más variación. De acuerdo con la figura 5.9, para el caso (a) se escogen los tres primeros componentes, con los cuales se explica aproximadamente el 75% de la variabilidad; mientras que con el caso (b) sólo se tomariá el primer componente, pues éste recoge casi la misma variabilidad de los tres primeros de (a).

Hay dos posibles alternativas para decidir sobre el número de componentes a retener. La primera es ignorar las $(p-m)$ componentes si sus correspondientes valores propios son cero. Esto se tiene si el rango de la $\mathbf{\Sigma}$ es m, de donde el rango de $\boldsymbol{S}$ es también m. Esta situación se consigue trivialmente en la práctica.

La segunda es que la proporción de variabilidad explicada por los $(p-m)$ componentes, sea menor que cierto valor. De manera equivalente, que los $(p-m)$ últimos valores propios sean iguales; lo cual significa, geométricamente, una isotropía respecto a la variación.

Puede resultar útil hacer primero una prueba acerca de la independencia completa entre las variables, como se indica en la sección (4.3.1); es decir verificar la hipótesis $H_0: \mathbf{\Sigma} = \text{Diag}(\sigma_{ii})$, que equivale a verificar la hipótesis $H_0: \mathbf{\Sigma} = \sigma^2 \boldsymbol{I}$. Si los resultados indican que las variables son independientes, las variables por si mismas conforman cada una las componentes principales.

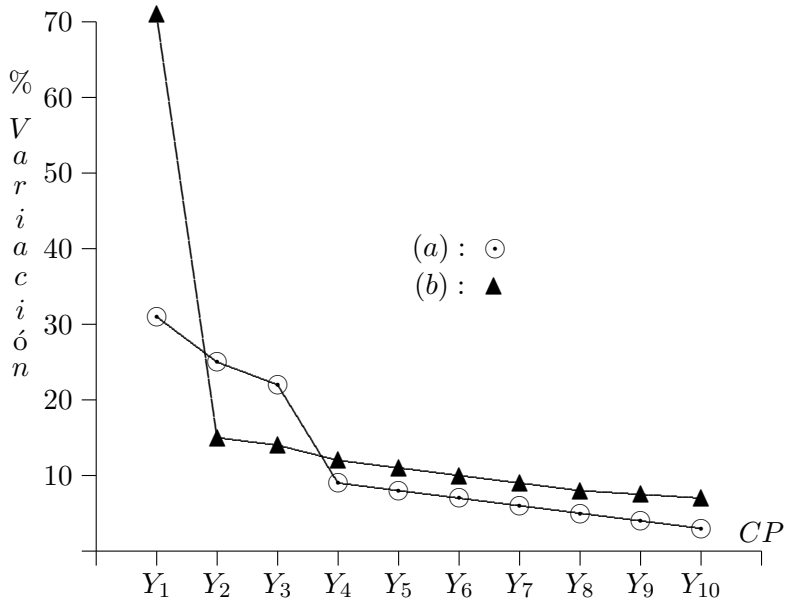


Figura 5.9 *Selección de componentes principales.*

Para desarrollar una prueba estadística acerca de la significancia de *"las m-componentes más grandes"*, se verifica la hipótesis de que los últimos k valores propios ($k = (p - m)$) son iguales y pequeños; es decir, la hipótesis nula $H_0 : \lambda_{p-k+1} = \lambda_{p-k+2} = \cdots = \lambda_p$, donde, $\lambda_1, \ldots, \lambda_p$, denotan los valores propios de la matriz $\mathbf{\Sigma}$. La implicación de la hipótesis anterior es que las primeras m componentes muestrales capturan las dimensiones esenciales, mientras que las últimas componentes reflejan "ruido". Si H_0 es cierta, los últimos $k = (p - m)$ valores propios tenderán a situarse sobre una línea recta casi horizontal, tal como se muestra en la figura 5.9.

Para probar $H_0 : \lambda_{p-m-1} = \lambda_{p-m+1} = \cdots = \lambda_p$, bajo el supuesto de multinormalidad, se emplea la estadística

$$\boldsymbol{u} = \left(n - \frac{2p+11}{6}\right)\left(k(\ln)\bar{\lambda} - \sum_{i=p-k+1}^{p} \ln \lambda_i\right), \tag{5.30}$$

la cual tiene aproximadamente una distribución χ^2_ν. Se rechaza la hipótesis H_0, si la estadística $\boldsymbol{u}$ es tal que $\boldsymbol{u} \geq \chi^2_{v,\alpha}$, donde los grados de libertad son dados por $\nu = \frac{1}{2}(k-1)(k+1)$.

Un desarrollo apropiado de este procedimiento empieza con verificar la hipótesis $H_{0(2)} : \lambda_{p-1} = \lambda_p$. Si no se rechaza esta hipótesis, se verifica

entonces la hipótesis $H_{0(3)}: \lambda_{p-2} = \lambda_{p-1} = \lambda_p$, y así, se continuan con las pruebas de esta forma, hasta que $H_{0(k)}$ sea rechazada para algún valor de k.

Todos los métodos presentados hasta ahora dependen únicamente de los valores propios. Sin embargo, los datos disponen de información adicional que puede emplearse para decidir sobre el número apropiado de componentes principales. Krzanowski (1994, págs. 81-83) desarrolla una estadística semejante al *PRESS* de regresión. Cada elemento x_{ij} de la matriz $\mathbb{X}$ se excluye y luego se estima a partir del *valor de la descomposición singular de rango reducido*. La precisión de la aproximación se basa en la suma de cuadrados de las diferencias entre el valor de x_{ij} y su estimado.

La estadística se define mediante la siguiente expresión

$$PRESS(m) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}(\hat{x}_{ij}^{(m)} - x_{ij})^2$$

donde $\hat{x}_{ij}^{(m)}$ es el estimador de x_{ij} basado en las primeras m componentes principales, omitiendo la observación x_{ij}. PRESS es la sigla que hace referencia a "PREdiction Sum of Squares". El número de componentes a retener es entonces determinado por el valor de

$$W_m = \frac{PRESS(m-1) - PRESS(m)}{PRESS(m)}\frac{p(n-1)}{(n+p-2m)},$$

donde W_m representa el incremento en la información predictiva suministrada por la m-ésima componente, dividida por el promedio de información predictiva en cada una de las componentes restantes. Si W_m es pequeño, la inclusión de la m-ésima componente tiene poco efecto sobre la aproximación. Si $W_m < 1$, la m-ésima componente principal lleva menos información que el promedio de componentes restantes, Krzanowski sugiere retener el número de componentes asociado con un $W_m = 0.9$.

5.6 Componentes principales en regresión

El análisis por componentes principales en regresión es una técnica alterna para encarar el problema de *multicolinealidad* en los regresores, lo mismo que la *regresión de borde (ridge)*. Mediante las componentes principales, como variables regresoras artificiales, se obtiene la estimación del modelo vía mínimos cuadrados.

Considérese la matriz de datos normalizados $\mathbb{X}^*$, de manera que $\mathbb{X}^{*'}\mathbb{X}^*$ es la matriz de correlación de los datos originales $\mathbb{X}$. Sean $\lambda_1, \ldots, \lambda_p$ los valores propios de la matriz de correlación y $\boldsymbol{\Lambda}$ la matriz diagonal de los respectivos valores propios, $\boldsymbol{PP'} = \boldsymbol{I}$ puesto que $\boldsymbol{P}$ es una matriz ortogonal. El modelo de regresión inicial se puede escribir en la forma

$$\begin{aligned} Y &= \beta_0 \boldsymbol{I} + \mathbb{X}^* \boldsymbol{\beta} + \boldsymbol{\epsilon} \\ Y &= \beta_0 \boldsymbol{I} + \mathbb{X}^* \boldsymbol{PP'} \boldsymbol{\beta} + \boldsymbol{\epsilon} = \beta_0 \boldsymbol{I} + \boldsymbol{Z\alpha} + \boldsymbol{\epsilon}, \end{aligned}$$

con $\boldsymbol{Z} = \mathbb{X}^* \boldsymbol{P}$ matriz de tamaño $(n \times p)$ y $\boldsymbol{\alpha} = \boldsymbol{P'\beta}$ vector $(p \times 1)$. Por la construcción hecha, las "nuevas" p-variables de las columnas de $\boldsymbol{Z}$ son ortogonales, pues ellas son las componentes principales. Entonces en forma similar a la ecuación (5.12)

$$\boldsymbol{Z'Z} = \boldsymbol{P'} \mathbb{X}^{*'} \mathbb{X}^* \boldsymbol{P} = \boldsymbol{\Lambda}.$$

Supóngase que de las p-componentes principales r son eliminados y s son incorporados al modelo de regresión, con $r + s = p$. Las matrices $\boldsymbol{P}$ y $\boldsymbol{\Lambda}$ se particionan acordemente, así

$$P = (P_r : P_s) \quad \text{y} \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_r & \vdots & 0 \\ \cdots & & \cdots \\ 0 & \vdots & \boldsymbol{\Lambda}_s \end{pmatrix},$$

las matrices $\boldsymbol{\Lambda}_r$ y $\boldsymbol{\Lambda}_s$ son submatrices diagonales de $\boldsymbol{\Lambda}$. El estimador mínimo cuadrático para $\boldsymbol{\alpha}$ es

$$\hat{\boldsymbol{\alpha}_s} = (\boldsymbol{Z'Z})^{-1} \boldsymbol{Z'}Y = \boldsymbol{\Lambda}_s^{-1} \boldsymbol{P'}_s \mathbb{X}^{*'} Y,$$

donde $\hat{\boldsymbol{\alpha}_s}$ es el estimador de los parámetros retenidos.

Ahora se observan algunas propiedades de estos estimadores. De la transformación de los βs dada por $\boldsymbol{P'\beta} = \boldsymbol{\alpha}$, se obtiene $\boldsymbol{\beta} = \boldsymbol{P\alpha}$. Si se nota por b_{cp} el estimador de $\boldsymbol{\beta}$, en el modelo que contiene s-componentes principales como regresores, entonces,

$$b_{cp} = \boldsymbol{P}_s \hat{\boldsymbol{\alpha}};$$

su valor esperado es

$$\begin{aligned} \mathcal{E}(b_{cp}) &= \boldsymbol{P}_s \boldsymbol{\alpha}_s \\ &= \boldsymbol{P}_s \boldsymbol{P'}_s \boldsymbol{\beta}, \quad \text{como } \boldsymbol{PP'} = \boldsymbol{P}_r \boldsymbol{P'}_r + \boldsymbol{P}_s \boldsymbol{P'}_s \text{ entonces,} \\ \mathcal{E}(b_{cp}) &= (I - \boldsymbol{P}_r \boldsymbol{P'}_r) \boldsymbol{\beta} \\ &= \boldsymbol{\beta} - \boldsymbol{P}_r \boldsymbol{\alpha}, \end{aligned}$$

de esta forma se prueba que los estimadores de los p-coeficientes de regresión son *sesgados*, el sesgo es $\boldsymbol{P}_r\boldsymbol{\alpha}_r$, y $\boldsymbol{\alpha}_r$ es el subvector de parámetros asociado con las componentes descartadas.

En el modelo de regresión lineal múltiple $\mathrm{Cov}(\hat{\beta}) = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$ para el caso de componentes principales como regresores se tiene que

$$\mathrm{Cov}(\hat{\alpha}_j) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = \sigma^2\boldsymbol{\Lambda}^{-1} = \sigma^2\,\mathrm{Diag}\Big(\frac{1}{\lambda_j}\Big).$$

Si todos las componentes principales son incorporadas al modelo de regresión, toda la variabilidad se mantiene, con esta regresión, lo que se consigue es una redistribución de ésta. Para situaciones de multicolinealidad extrema, se encontrará al menos un valor propio pequeño (una cuasi-singularidad de $\mathbb{X}^{*'}\mathbb{X}^{*}$). Al suprimir la componente ligada a tal valor propio, tal vez se reduzca la varianza total en el modelo, produciendo un mejoramiento en su predicción; en la sección (5.7.1) se hace una explicación más puntual sobre esto.

Ejemplo 5.1 En este aparte se desarrolla un caso particular, que intenta tomar los conceptos y procedimientos del ACP hasta aquí expuestos.

Los cálculos se desarrollan con la ayuda del procedimento PRINCOMP del paquete estadístico SAS (SAS User's Guide, 2000).

Los datos de la tabla 5.5 son algunas medidas corporales de pájaros. El objetivo es estudiar el efecto de la selección natural en tales aves. Las variables de interés son:

X_1**:** longitud total.

X_2**:** extensión de las alas.

X_3**:** longitud de pico y cabeza.

X_4**:** longitud del húmero.

X_5**:** longitud de la quilla (esternón o pecho).

Se midieron 49 pájaros moribundos después de una tempestad, 21 de los cuales sobrevivieron. Como cita Manly (1998), este trabajo está enmarcado en el estudio de la selección natural en aves.

La generación de las componentes principales se hace por medio de la matriz de correlación y de la matriz de covarianzas. A continuación se obtienen las componentes principales, primero mediante la matriz de correlación, y luego, mediante la matriz de covarianzas.

∘ *ACP mediante la matriz de correlación*

En el reporte del paquete SAS se tiene la matriz de varianzas y covarianzas y la matriz de correlaciones (señaladas como (A) y (B) respectivamente). Los valores propios para esta última se indican por (C).

La suma de los valores propios es igual a cinco (traza de R). En (F) están los vectores propios. Las componentes de los vectores propios suministran las ponderaciones o grados de importancia de cada variable con el respectivo componente principal.

El valor propio ligado con cada componente principal indica la cantidad de varianza retenida respecto a la varianza total. Así, con la primera componente se retiene:

$(3.615978/5.0) \times 100\% = 72.32\%$ de la variabilidad total; con la segunda el 10.63% y así sucesivamente (como se indica con (D) en la salida SAS). En la última línea, rotulada con (E), está la contribución acumulada hasta cada componente. Es inmediato, para estos datos, que la primera componente es más importante que las demás, pues como se observa en C ésta reune casi las tres cuartas partes (72.32%) de la variabilidad total.

De lo anterior, la primera componente con las variables normalizadas está dada por

$$Y_1 = 0.452X_1 + 0.462X_2 + 0.451X_3 + 0.471X_4 + 0.398X_5.$$

Y_1 es un indicador del tamaño de los pájaros. Nótese que los coeficientes de la combinación lineal que definen a Y_1 son todos positivos, y además, alrededor del 72.3% de la variación en los datos está relacionada con diferencias de tamaño; es decir, la primera componente reúne las variables que determinan el *tamaño* de las aves.

La segunda componente principal:

$$Y_2 = 0.051X_1 - 0.300X_2 - 0.325X_3 - 0.185X_4 + 0.877X_5.$$

Las variables X_2 (extensión de las alas), X_3 (longitud de pico y cabeza) y X_4 (longitud del húmero) contrastan con la variable X_5 (longitud de la quilla); es decir, al aumentar el primer grupo de medidas X_5 disminuye y viceversa. En consecuencia, Y_2 reune las variables que registran la *forma* de las aves. El valor tan bajo de X_1 en Y_2 significa que el tamaño de los pájaros afecta poco a Y_2. Similarmente se pueden hacer interpretaciones para Y_3, Y_4 y Y_5.

Tabla 5.5 Medidas corporales de gorriones

OBS.	X_1	X_2	X_3	X_4	X_5	OBS.	X_1	X_2	X_3	X_4	X_5
1	156	245	31.6	18.5	20.5	26	160	250	31.7	18.8	22.5
2	154	240	30.4	17.9	19.6	27	155	237	31.0	18.5	20.0
3	153	240	31.0	18.4	20.6	28	157	245	32.2	19.5	21.4
4	153	236	30.9	17.7	20.2	29	165	245	33.1	19.8	22.7
5	155	243	31.5	18.6	20.3	30	153	231	30.1	17.3	19.8
6	163	247	32.0	19.0	20.9	31	162	239	30.3	18.0	23.1
7	157	238	30.9	18.4	20.2	32	162	243	31.6	18.8	21.3
8	155	239	32.8	18.6	21.2	33	159	245	31.8	18.5	21.7
9	164	248	32.7	19.1	21.1	34	159	247	30.9	18.1	19.0
10	158	238	31.0	18.8	22.0	35	155	243	30.9	18.5	21.3
11	158	240	31.3	18.6	22.0	36	162	252	31.9	19.1	22.2
12	160	244	31.1	18.6	20.5	37	152	230	30.4	17.3	18.6
13	161	246	32.3	19.3	21.8	38	159	242	30.8	18.2	20.5
14	157	245	32.0	19.1	20.0	39	155	238	31.2	17.9	19.3
15	157	235	31.5	18.1	19.8	40	163	249	33.4	19.5	22.8
16	156	237	30.9	18.0	20.3	41	163	242	31.0	18.1	20.7
17	158	244	31.4	18.5	21.6	42	156	237	31.7	18.2	20.3
18	153	238	30.5	18.2	20.9	43	159	238	31.5	18.4	20.3
19	155	236	30.3	18.5	20.1	44	161	245	32.1	19.1	20.8
20	163	246	32.5	18.6	21.9	45	155	235	30.7	17.7	19.6
21	159	236	31.5	18.0	21.5	46	162	247	31.9	19.1	20.4
22	155	240	31.4	18.0	20.7	47	153	237	30.6	18.6	20.4
23	156	240	31.5	18.2	20.6	48	162	245	32.5	18.5	21.1
24	160	242	32.6	18.8	21.7	49	164	248	32.3	18.8	20.9
25	152	232	30.3	17.2	19.8						

del 1 al 21 sobrevivieron los demás no.
Fuente: Manly (1998, pág. 2).

Matriz de covarianzas (A)

	X_1	X_2	X_3	X_4	X_5
X_1	13.35374150	13.61096939	1.92206633	1.33061224	2.19221939
X_2	13.61096939	25.68282313	2.71360544	2.19770408	2.65782313
X_3	1.92206633	2.71360544	0.63163265	0.34226616	0.41464711
X_4	1.33061224	2.19770408	0.34226616	0.31841837	0.33937075
X_5	2.19221939	2.65782313	0.41464711	0.33937075	0.98282313

Media y desviación estándar

	X_1	X_2	X_3	X_4	X_5
MEDIA	157.979592	241.326531	31.4591837	18.4693878	20.8265306
DES. EST.	3.65427715	5.06782233	0.794753203	0.564285714	0.991374364

Matriz de Correlaciones (B)

	X_1	X_2	X_3	X_4	X_5
X_1	1.00000	0.73496	0.66181	0.64528	0.60512
X_2	0.73496	1.00000	0.67374	0.76851	0.52901
X_3	0.66181	0.67374	1.00000	0.76319	0.52627
X_4	0.64528	0.76851	0.76319	1.00000	0.60665
X_5	0.60512	0.52901	0.52627	0.60665	1.00000

Valores propios

		(1)	(2)	(3)	(4)	(5)
(**C**)	*Valor Propio*	3.61598	0.53150	0.38642	0.301574	0.16453
	Diferencia	3.08447	0.14508	0.08486	0.13704	
(**D**)	Proporción	0.72319	0.10630	0.07729	0.06031	0.03291
(**E**)	*Porc. Acum.*	0.72319	0.82950	0.90678	0.96709	1.0000

Vectores Propios (F)

Variable	$\mathcal{CP}_1$	$\mathcal{CP}_2$	$\mathcal{CP}_3$	$\mathcal{CP}_4$	$\mathcal{CP}_5$
X_1	0.45180	0.05072	−0.69047	0.42041	−0.37391
X_2	0.46168	−0.29956	−0.34045	−0.54786	0.53008
X_3	0.45054	−0.32457	0.45449	0.60630	0.34279
X_4	0.47073	−0.18468	0.41094	−0.38828	−0.65167
X_5	0.39768	0.87649	0.17846	−0.06887	0.19243

Se evalúa cada componente sobre cada uno de los 49 valores normalizados. Así, por ejemplo, para la primera observación,

$$x_{11} = 156,\ x_{12} = 245,\ x_{13} = 31.6,\ x_{14} = 28.5,\ \text{y}\ x_{15} = 20.5,$$

se les resta la media y se divide por la desviación estándar a cada componente; el resultado al reemplazar en la primera componente principal es

$$\begin{aligned} Y_1 &= 0.452 \times (-0.542) + 0.462 \times 0.725 + 0.451 \times 0.177 \\ &\quad + 0.471 \times 0.055 + 0.398 \times (-0.330) \\ &= 0.06429. \end{aligned}$$

Similarmente, el valor de la misma observación en la segunda componente es

$$\begin{aligned} Y_2 &= 0.051 \times (-0.542) - 0.300 \times 0.725 - 0.325 \times 0.177 \\ &\quad - 0.185 \times 0.055 + 0.877 \times (-0.330) \\ &= -0.60084. \end{aligned}$$

Para las demás componentes los cálculos son semejantes; éstos son los "nuevos" puntajes de los 49 pájaros con relación a las dos primeras componentes principales.

Los valores respecto a cada componente son las coordenadas de cada individuo (ave) respecto a los "nuevos" ejes. Es muy frecuente y cómodo ubicar las observaciones en el primer plano factorial, gráfico que puede sugerir alguna estructura de agrupamiento de los datos. De esta forma, por ejemplo, la primera ave se ubica en el punto de coordenadas $< 0.06429, -0.60084 >$, respecto a los dos primeros ejes principales (plano factorial).

En la tabla 5.6 se muestran las coordenadas de las 49 aves respecto a los cinco ejes factoriales, las cuales se pueden calcular como se hizo anteriormente para Y_1 y Y_2 o a través de la reconstrucción de la matriz de datos resumida en la ecuación (5.4).

En la figura 5.10 se han ubicado las 49 aves de acuerdo con sus coordenadas respecto a las dos primeras componentes $\mathcal{CP}_1$ y $\mathcal{CP}_2$.

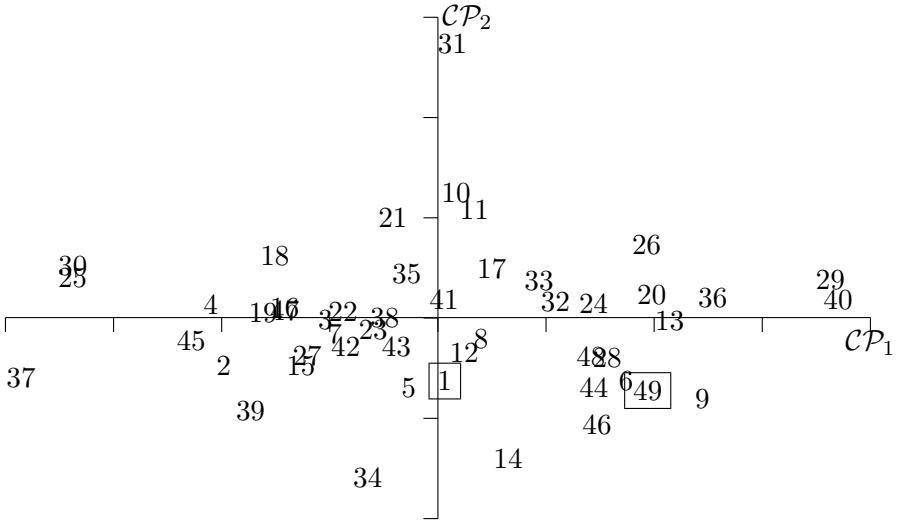


Figura 5.10 *Primer plano factorial.*

Tabla 5.6 Coordenadas factoriales de los gorriones

OBS.	$\mathcal{CP}_1$	$\mathcal{CP}_2$	$\mathcal{CP}_3$	$\mathcal{CP}_4$	$\mathcal{CP}_5$	OBS.	$\mathcal{CP}_1$	$\mathcal{CP}_2$	$\mathcal{CP}_3$	$\mathcal{CP}_4$	$\mathcal{CP}_5$
1	0.06	-0.60	0.17	-0.51	0.54	26	2.12	0.78	-0.28	-0.86	0.74
2	-2.18	-0.44	-0.40	-0.64	0.23	27	-1.32	-0.33	0.46	-0.18	-0.54
3	-1.14	0.01	0.67	-0.71	0.20	28	1.72	-0.36	1.21	-0.69	-0.27
4	-2.31	0.17	0.30	0.14	0.47	29	3.99	0.43	0.67	0.61	-0.79
5	-0.29	-0.66	0.47	-0.54	0.24	30	-3.71	0.57	-0.17	0.38	-0.01
6	1.91	-0.59	-0.62	0.00	-0.28	31	0.14	2.83	-1.19	-0.01	-0.17
7	-1.05	-0.11	-0.07	-0.08	-0.53	32	1.19	0.20	-0.46	0.12	-0.46
8	0.43	-0.16	1.64	0.81	0.56	33	1.02	0.42	-0.06	-0.10	0.56
9	2.69	-0.78	-0.36	0.46	-0.05	34	-0.71	-1.58	-1.49	-0.54	0.31
10	0.18	1.31	0.40	-0.29	-0.70	35	-0.31	0.49	0.23	-1.00	0.29
11	0.37	1.13	0.30	-0.14	-0.13	36	2.79	0.25	-0.51	-0.88	0.43
12	0.26	-0.31	-0.73	-0.39	-0.29	37	-4.24	-0.56	0.03	0.68	-0.11
13	2.35	0.01	0.37	-0.15	-0.22	38	-0.54	0.04	-0.86	-0.25	-0.07
14	0.71	-1.38	0.55	-0.47	-0.17	39	-1.90	-0.90	-0.05	0.31	0.20
15	-1.39	-0.44	0.17	0.92	-0.31	40	4.07	0.23	0.75	0.38	0.31
16	-1.55	0.14	-0.09	0.17	-0.05	41	0.06	0.22	-1.54	0.41	-0.23
17	0.54	0.54	-0.05	-0.40	0.37	42	-0.93	-0.24	0.51	0.64	0.06
18	-1.65	0.67	0.43	-0.76	0.07	43	-0.42	-0.24	-0.09	0.59	-0.45
19	-1.77	0.09	0.14	-0.62	-0.92	44	1.58	-0.66	0.00	0.01	-0.38
20	2.17	0.27	-0.37	0.70	0.48	45	-2.50	-0.18	-0.22	0.37	-0.03
21	-0.45	1.06	-0.03	1.00	0.02	46	1.61	-1.04	-0.50	-0.21	-0.43
22	-0.96	0.10	0.25	0.08	0.65	47	-1.55	0.11	0.75	-0.82	-0.54
23	-0.65	-0.07	0.24	0.14	0.34	48	1.55	-0.35	-0.33	0.81	0.43
24	1.58	0.18	0.62	0.74	0.14	49	2.13	-0.69	-0.85	0.38	0.07
25	-3.71	0.44	-0.01	0.38	0.40						

del 1 al 21 sobrevivieron los demás no.

De acuerdo con la interpretación que se le ha dado a los dos primeros factores, se puede afirmar que las aves sobrevivientes tienen un tamaño y forma cercano al origen de las coordenadas del primer plano factorial. Es interesante observar la ubicación de algunas aves. La número 31 tiene el valor más alto respecto a la quilla o esternón (pecho), nótese que la segunda componente está altamente influenciada por esta variable con una ponderación de 0.87649. Las aves numeradas como 30, 25 y 37 tienen los valores más bajos respecto a la variable longitud del húmero, X_4, la más importante en el primer componente con una ponderación de 0.47073, mientras que las aves numeradas con 29 y 40 tienen los valores más altos en esta misma variable. Finalmente, se puede apreciar que el ave "prototipo

es la número 41, sus valores respecto a las variables de más ponderación en cada uno de los dos factores principales están cerca de sus respectivos promedios.

La figura 5.11 permite apreciar las variables en el primer plano factorial. Se observa la influencia de cada una sobre estos ejes factoriales. Así por ejemplo, el primer eje está asociado con las variables X_4 y X_2, en tanto que el segundo está ligado con las variables X_5 y X_3. El ángulo formado entre las variables está en relación inversa con el grado de asociación entre ellas; recuérdese que el coeficiente de correlación de Pearson es igual al coseno del ángulo formado por las dos variables.

Como una estrategia para la interpretación se puede superponer el plano factorial de los individuos al de las variables, para apreciar una clasificación de las aves de acuerdo con su forma y tamaño. Sobre esto último se advierte acerca del cuidado que debe tenerse con la interpretación, ya que se trata de dos subespacios de espacios diferentes (individuos y variables).

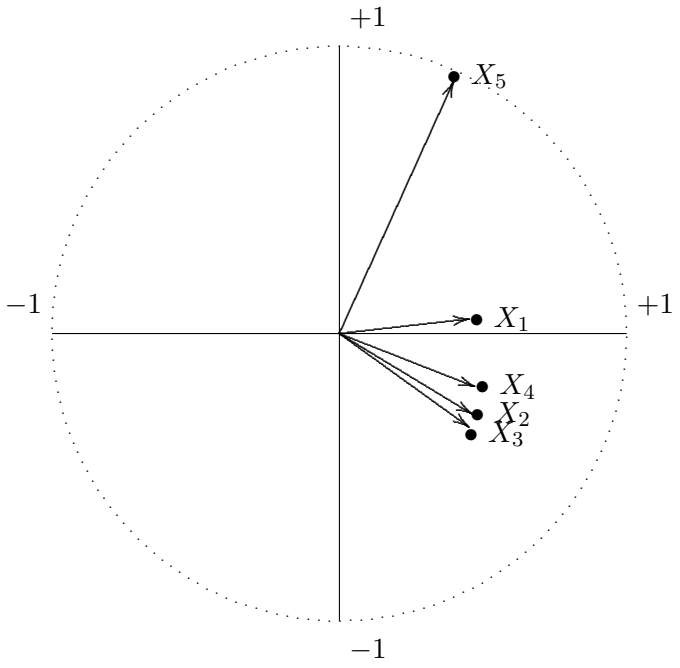


Figura 5.11 *Variables en el primer plano factorial.*

∘ *ACP mediante la matriz de covarianzas*

Con la matriz de covarianzas se generan los componentes principales de las variables que corresponden a las cinco características medidas sobre las aves. En esta parte se debe tener cautela con el uso de las componentes principales, pues éstas se afectan por los cambios en la escala. Nótese el comportamiento de la variable X_1, pues con las componentes principales generadas anteriormente desde la matriz de covarianzas, ésta no se muestra tan "importante" en la primera componente como cuando se obtienen desde la matriz de correlaciones. Esto se explica tanto por la magnitud de la varianza asociada a la variable X_1 (segunda en valor) como por la alta correlación observada con la variable X_2.

A continuación se muestran los valores y vectores propios junto con las respectivas fracciones de variabilidad recogida por cada una de las componentes principales. El lector puede observar que con tan solo la primera componente se reune el 86.2% de la variabilidad total, la cual a primera vista es superior a la variabilidad retenida por las dos primeras componentes principales (82.95%) generadas desde la matriz de correlación ; pero, como se afirmó anteriormente, esto no es más que una consecuencia de los altos valores para la media y la varianza de las variables X_1 y X_2 respecto a las demás.

Análisis de componentes principales

	(1)	(2)	(3)	(4)	(5)
$ValorPropio$	35.325757	4.622459	0.630918	0.312784	0.077521
$Diferencia$	30.703298	3.991542	0.318134	0.235263	
Proporción	0.8622	0.1128	0.0154	0.0076	0.0019
$P.Acumulada$	0.8622	0.9751	0.9905	0.9981	1.0000

La suma de los valores propios (traza de la matriz de covarianzas o varianza total) y el promedio de los valores propios de la matriz de covarianzas son, respectivamente,

$$VT = \text{tra}(\boldsymbol{S}) = 40.9694388 \text{ y}$$
$$\bar{l} = 8.19388776.$$

El siguiente recuadro contiene los vectores propios normalizados ($\mathcal{CP}_j$) de la matriz de covarianzas, las entradas de cada vector propio corresponden a cada una de las ponderaciones (a_{ij}) que definen cada componente principal.

Vectores propios (G)

$Variable$	$\mathcal{CP}_1$	$\mathcal{CP}_2$	$\mathcal{CP}_3$	$\mathcal{CP}_4$	$\mathcal{CP}_5$
X_1	0.53650	0.82810	−0.15649	−0.04021	0.01765
X_2	0.82902	−0.55051	−0.05774	−0.06902	−0.03964
X_3	0.09650	0.03356	0.23751	0.89762	−0.35695
X_4	0.07435	−0.01459	0.20324	0.30724	0.92658
X_5	0.10030	0.09923	0.93512	−0.30576	−0.11022

☑

De manera que la primera y segunda componente principal (redondeando los a_{ij}) son ahora:

$$Y_1 = 0.537X_1 + 0.829X_2 + 0.097X_3 + 0.074X_4 + 0.100X_5$$
$$Y_2 = 0.828X_1 - 0.551X_2 + 0.034X_3 - 0.015X_4 + 0.099X_5$$

5.7 Tópicos adicionales

Se presentan en esta última sección algunas aplicaciones e interpretaciones complementarias sobre los resultados del ACP.

5.7.1 Información de la última componente principal

Tradicionalmente las primeras componentes son consideradas útiles para resumir un conjunto de datos. Sin embargo, las últimas componentes principales pueden contener información que merece examinarse.

Se ha probado que los valores propios son las varianzas de las respectivas componentes principales. De esta manera, las últimas componentes son las que tienen la menor varianza. Si la varianza de un componente es cercana a cero, entonces la componente define una combinación lineal entre las variables que es aproximadamente constante sobre la muestra. Así, un valor propio extremadamente pequeño puede indicar una *colinealidad*, pues es una combinación lineal igual a cero; la cual el investigador puede pasar de manera inadvertida. Rencher (1998, pág. 352) sugiere que tales variables redundantes (las de mayor correlación con la componente) sean excluidas de manera que no distorsionen las primeras componentes principales. Supóngase, por ejemplo, que se tiene un vector de cinco variables $X' = (X_1, \cdots, X_5)$ y que $X_5 = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$. Entonces la matriz

de covarianzas $\boldsymbol{S}$ es singular, y, excepto un error de redondeo, λ_5 será cero. Así, $s_{Y_5} = 0$, y Y_5 es constante. Por lo tanto el valor de Y_5 es constante y su media es (cero)

$$Y_5 = \boldsymbol{a}_5 X = a_{51}X_1 + a_{52}X_2 + a_{53}X_3 + a_{54}X_4 + a_{55}X_5 = 0.$$

Como esto debe reflejar la dependencia de X_5 con X_1, X_2, X_3 y X_4, entonces $\boldsymbol{a}_5$ será proporcional a $(1, 1, 1, 1, -4)$, pues los X_i están centrados.

A continuación se describe una medida para cada observación, que puede utilizarse para indicar la "responsabilidad" de cada observación sobre el ajuste de las primeras componentes. En consecuencia, sirve como un instrumento para la detección de observaciones *atípicas*.

Se sabe que si los vectores propios de la matriz $\boldsymbol{S}$ se han normalizado (longitud 1), entonces

$$\boldsymbol{I} = \boldsymbol{a}_1\boldsymbol{a}_1' + \boldsymbol{a}_2\boldsymbol{a}_2' + \cdots + \boldsymbol{a}_p\boldsymbol{a}_p'.$$

Si se multiplica la igualdad anterior por el i-ésimo vector de observaciones x_i (fila i de la matriz de datos), se obtiene

$$\boldsymbol{I}x_i = \boldsymbol{a}_1(\boldsymbol{a}_1'x_i) + \boldsymbol{a}_2(\boldsymbol{a}_2'x_i) + \cdots + \boldsymbol{a}_p(\boldsymbol{a}_p'x_i)$$

Nótese que en el primer término del miembro derecho de la igualdad anterior $\boldsymbol{a}_1'x_i = y_{1i}$, es la primera componente principal evaluada para la observación x_i. Similarmente en el segundo término $\boldsymbol{a}_2'x_i = y_{2i}$, y así para los demás términos. De tal forma que

$$x_i = y_{1i}\boldsymbol{a}_1 + y_{2i}\boldsymbol{a}_2 + \cdots + y_{pi}\boldsymbol{a}_p, \qquad i = 1, 2, \ldots, n. \tag{5.31}$$

Se pueden emplear los últimos $(p-m)$ términos de (5.32), que corresponden a $y_{p-m,i}\boldsymbol{a}_{p-m} + \cdots + y_{pi}\boldsymbol{a}_p$, como una especie de "error" o "residual" para medir que tan bien ajustan ("reconstruyen") las m-primeras componentes la observación x_i. Como los $\boldsymbol{a}_i$ son ortonormales, el cuadrado de la longitud del vector de residuales $y_{p-m,i}\boldsymbol{a}_{p-m} + \cdots + y_{pi}\boldsymbol{a}_p$ es

$$d_i^2 = y_{p-m,i}^2 + \cdots + y_{pi}^2. \tag{5.32}$$

Se computa d_i^2 para cada una de las observaciones $x_1, x_2, \ldots, x_n$. Una observación con un valor demasiado extremo de d_i^2 indicará un ajuste "pobre" de las primeras $p - m - 1$ componentes principales, lo cual puede deberse a que la observación es *"aberrante"* o *atípica* con relación a la estructura de correlación.

Gnanadesikan (1997, págs. 294-297) muestra varias herramientas en las que se emplean las últimas componentes principales para la detección de observaciones atípicas. Algunas proyecciones de los datos sobre las últimas componentes útiles para la detección de observaciones atípicas son las siguientes:

1. Diagramas bivariados de las últimas componentes donde se proyectan las observaciones.

2. Gráficos de probabilidad de los valores dentro de cada una de las últimas filas de la matriz de puntajes; es decir, la matriz $\mathbb{Y} = \boldsymbol{A}(\mathbb{X} - \bar{X})$.

Por la linealidad de la transformación, se espera que estos valores tengan una distribución cercana a la normal, un gráfico de probabilidades será un buen punto de partida para el análisis. Este análisis puede ayudar a identificar coordenadas, en algunas de las últimas componentes principales, que parezcan no normales. La identificación de estas observaciones es semejante a como se verifica la normalidad de un conjunto de datos mediante los gráficos *cuantil-cuantil* (gráficos $Q \times Q$ de la sección (2.5))

3. Gráficos de los valores en cada una de las últimas filas de $\mathbb{Y}$ frente a ciertas distancias en el espacio de las primeras componentes principales.

Por ejemplo, si la mayor parte de la variabilidad de un conjunto de datos, de dimensión $p = 5$, está asociada con la variabilidad de las primeras dos componentes principales, puede ser informativo un gráfico de las proyecciones sobre cada uno de los tres ejes principales restantes, versus la distancia al centroide de cada uno de los puntos proyectados en el plano asociado a los dos primeros ejes principales. Así, se tiene un gráfico donde el eje horizontal corresponde a uno de los últimos ejes principales y el eje vertical a la distancia de la respectiva observación, en el plano determinado por los dos primeros ejes principales (primer plano factorial), al centroide en el mismo primer plano factorial.

Con las metodologías anteriores, si una observación es detectada como aberrante, se puede excluir de las estimaciones de $\boldsymbol{S}$ (o de $\boldsymbol{R}$) y entonces repetir el proceso de obtención y análisis de los residuales de las componentes principales. Una alternativa al problema de datos atípicos son los métodos de estimación robustos para la matriz de covarianzas (o correlaciones), Gnanadesikan (1977, sección (5.3.2)).

5.7.2 Selección de variables

En la motivación presentada al comienzo de este capítulo se consideró el ACP como una técnica útil para reducir el número de variables, por ejemplo,

preguntas en un formulario de encuesta. En esta sección se muestra como emplear el ACP para este propósito.

En técnicas tales como el análisis de regresión, el análisis de varianza multivariado y el análisis discriminante se presentan algunos criterios para seleccionar variables. Estos criterios se relacionan con separación de grupos, factores externos, tales como variables, o con una tasa de clasificación adecuada. En el contexto de las componentes principales, no se tienen variables dependientes, como en regresión, o grupos de observaciones, como en el análisis discriminante. Sin considerar influencias externas, se quiere encontrar un subconjunto de las variables originales que *mejor capturen* la variación interna (y la covariación) de las variables.

Un procedimiento consiste en asociar una variable a cada una de las primeras componentes y retener éstas, por ejemplo, de 50 variables seleccionar un subconjunto de 10. Otra aproximación consiste en asociar una variable a cada una de las últimas componentes y excluirlas; para el ejemplo nuevamente, asociar una variable a cada una de la últimas 40 componentes y excluir estas 40 variables. Para asociar una variable con una componente principal, se escoge la variable correspondiente al coeficiente de la componente más grande (en valor absoluto), siempre que la variable no se haya seleccionado previamente. El procedimiento es aplicable para componentes generadas desde $\boldsymbol{S}$ o desde $\boldsymbol{R}$.

5.7.3 Biplots

Mediante la expresión (5.26a) se muestra que una aproximación mínimo cuadrática de la matriz centrada de datos $\mathbb{X}$ de tamaño $(n \times p)$ se obtiene al reemplazar las p columnas de $\mathbb{X}$ por un número más pequeño de $m < p$ columnas derivadas desde las componentes principales. La aproximación matricial se denota por $\hat{\mathbb{X}} = \mathbb{Y}_1 \boldsymbol{A}_1'$, donde $\mathbb{Y}_1$ es una matriz de tamaño $(n \times m)$ que corresponde a las observaciones o puntajes sobre las primeras m componentes principales, y $\boldsymbol{A}_1$ es una matriz $p \times m$ que contiene las columnas consistentes de los m primeros vectores propios.

La *descomposición en valor singular* (ecuaciones A2.29 y A2.30) de la matriz centrada $\mathbb{X}$ es

$$\mathbb{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}', \tag{5.33}$$

donde las columnas de la matriz $\boldsymbol{U}$, de tamaño $(n \times p)$, son ortogonales, $\boldsymbol{D}$ es una matriz diagonal, de tamaño $(p \times p)$, que contiene los *valores singulares* de $\mathbb{X}$ (raices cuadradas de los valores propios positivos de $\mathbb{X}'\mathbb{X}$) y $\boldsymbol{V}$ es una matriz ortogonal $(p \times p)$. Los valores singulares de $\boldsymbol{D}$ están dispuestos en orden decreciente.

Una aproximación mínimo cuadrática a $\mathbb{X}$ a una dimensión $m < p$, es dada por la matriz $\hat{\mathbb{X}}$, donde

$$\hat{\mathbb{X}} = \sum_{i=1}^{m} d_i \boldsymbol{u}_i \boldsymbol{v}_i', \tag{5.34}$$

y por tanto $\mathbb{Z} = \hat{\mathbb{X}}$ minimiza la expresión $\text{tra}(\mathbb{X} - \mathbb{Z})(\mathbb{X} - \mathbb{Z})'$ bajo la restricción de que el rango de $\mathbb{Z}$ sea menor o igual que m.

La descomposición en el valor singular se puede relacionar con las componentes principales. Así, para una matriz $\mathbb{X}$, dada por (5.34), $\mathbb{X}'\mathbb{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}'$, por tanto, los vectores singulares a derecha de $\mathbb{X}$ son los vectores propios de $\mathbb{X}'\mathbb{X}$ y los valores propios de la matriz $\mathbb{X}'\mathbb{X}$ son los cuadrados de los valores singulares de $\mathbb{X}$. En forma similar, para la matriz $\mathbb{X}\mathbb{X}' = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}'$, se observa que los vectores singulares a izquierda de $\mathbb{X}$ son los vectores propios de $\mathbb{X}\mathbb{X}'$ y que los valores propios de $\mathbb{X}\mathbb{X}'$ son los cuadrados de los valores singulares de la matriz $\mathbb{X}$. En consecuencia los valores propios de $\mathbb{X}'\mathbb{X}$ y de $\mathbb{X}\mathbb{X}'$ son equivalentes.

Las componentes principales de $\mathbb{X}'\mathbb{X}$ están dadas por $\mathbb{Y} = \mathbb{X}\boldsymbol{A}$, y como $\mathbb{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$ y considerando que $\boldsymbol{A} = \boldsymbol{V}$, se tiene entonces que

$$\mathbb{Y} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D},$$

por tanto, las componentes principales de $\mathbb{X}'\mathbb{X}$ son simplemente una versión escalada de los vectores singulares a izquierda de $\mathbb{X}$. En consecuencia, la descomposición en el valor de $\mathbb{X}$ se puede expresar como $\mathbb{X} = \mathbb{Y}\boldsymbol{V}'$.

Simétricamente, se observa que en las componentes principales para $\mathbb{X}\mathbb{X}'$, denotadas por $\mathbb{Z} = \mathbb{X}'\boldsymbol{U}$, se tiene $\mathbb{Z} = \boldsymbol{V}\boldsymbol{D}$. Las componentes principales de $\mathbb{X}\mathbb{X}'$ se obtienen por el escalamiento de los valores singulares a derecha de $\mathbb{X}$, con vectores propios de $\mathbb{X}\mathbb{X}'$ iguales a los vectores singulares a izquierda de $\mathbb{X}$. Se concluye entonces que las componentes principales de $\mathbb{X}\mathbb{X}'$ están relacionadas con los vectores propios de $\mathbb{X}'\mathbb{X}$ y recíprocamente, que los vectores propios de $\mathbb{X}\mathbb{X}'$ están relacionados con las componentes principales de $\mathbb{X}'\mathbb{X}$.

Como se sugiere en la ecuación (5.34), la matriz $\mathbb{X}$ puede aproximarse mediante $\hat{\mathbb{X}} = \boldsymbol{U}_1\boldsymbol{D}_1\boldsymbol{V}_1'$, donde $\boldsymbol{U}_1$ es una matriz de tamaño $(n \times m)$, $\boldsymbol{D}_1$ es una matriz de tamaño $(m \times m)$ y $\boldsymbol{V}_1$ es una matriz de tamaño $(p \times m)$, las cuales representan las primeras m columnas tanto de $\boldsymbol{U}$ como de $\boldsymbol{V}_1$, y, los correspondientes valores singulares de $\boldsymbol{D}$. Las columnas de $\boldsymbol{V}_1$ suministran información sobre las primeras m columnas o variables de $\mathbb{X}$, y las columnas de $\boldsymbol{U}_1$ proveen información acerca de las m primeras filas u objetos de $\mathbb{X}$. El término *biplot* hace referencia a la consideración simultánea tanto del espacio columna como del espacio fila de la matriz de datos $\mathbb{X}$. La más

cercana traducción de este término es *"doble gráfico"*; pero una vez que se ha explicado su sentido, se conserva en este texto el término original.

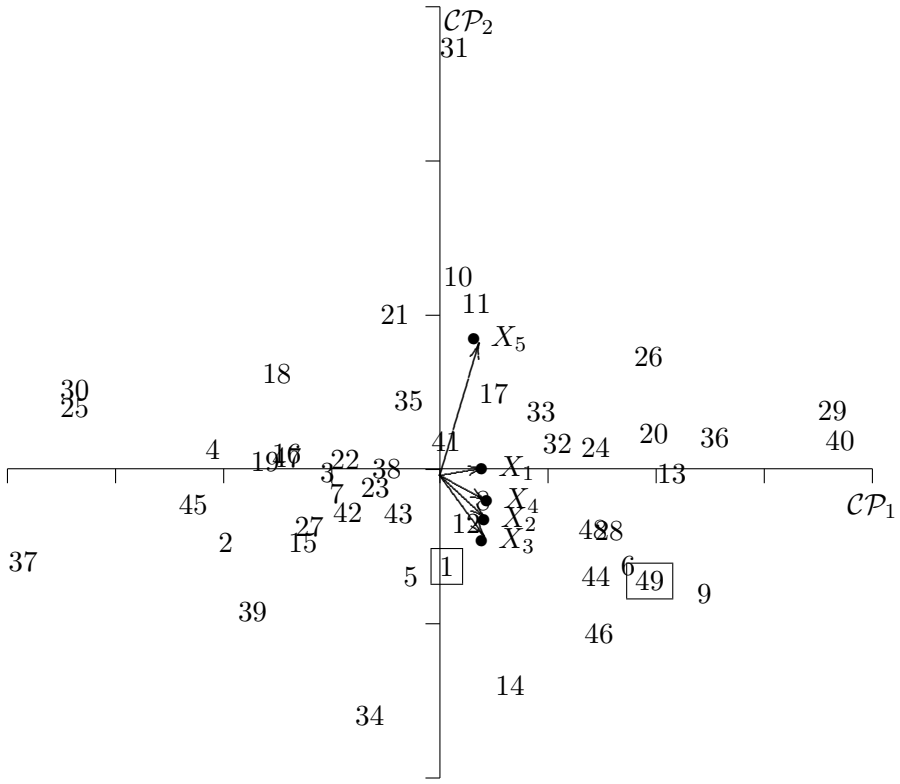


Figura 5.12 *Biplot para el ejemplo 5.1.*

Un biplot se usa para proveer una representación bidimensional de una matriz de datos $\mathbb{X}$. Se emplean únicamente dos dimensiones para hacer más fácil el gráfico. Se asume entonces, que una aproximación mediante la descomposición en valor singular de $\mathbb{X}$ basada en $m = 2$ es adecuada.

Una descomposición en valor singular de la matriz $\mathbb{X}$ basada en dos dimesiones está dada por $\hat{\mathbb{X}} = \boldsymbol{U}_1\boldsymbol{D}_1\boldsymbol{V}_1'$, donde las filas de la matriz $\boldsymbol{V}_1'$ $(2 \times p)$ son los vectores propios de $\mathbb{X}'\mathbb{X}$ y las columnas de $\boldsymbol{U}_1$ $(n \times 2)$ son los vectores propios de $\mathbb{X}\mathbb{X}'$.

En ACP una aproximación para $\mathbb{X}'\mathbb{X}$ es dada por

$$\hat{\mathbb{X}}'\hat{\mathbb{X}} = \boldsymbol{V}_1\boldsymbol{D}_1^2\boldsymbol{V}_1' = \boldsymbol{V}_1\boldsymbol{\Lambda}_1\boldsymbol{V}_1',$$

donde $\boldsymbol{\Lambda}_1$ denota una matriz diagonal de tamaño (2×2) con los dos primeros valores propios sobre la diagonal, y las dos columnas de $\boldsymbol{V}_1$ las pondera-

ciones correspondientes a las dos primeras componentes principales de $\mathbb{X}'\mathbb{X}$. En resumen, la aproximación $\hat{\mathbb{X}} = \boldsymbol{U}_1\boldsymbol{D}_1\boldsymbol{V}_1' = \mathbb{Y}_1\boldsymbol{V}_1'$ representa el producto de los puntajes de las componentes principales (dos columnas de $\mathbb{Z}_1$) y las correspondientes ponderaciones (dos filas de $\boldsymbol{V}_1$).

Anteriormente se mostró que los puntajes en las componentes principales de los n objetos se pueden graficar en el plano determinado por las dos primeras componentes principales. Además, se indicó que las ponderaciones de las componentes principales (vectores propios) pueden graficarse como rayos que salen desde el origen en un plano que contiene las dos primeras componentes principales como ejes. Un biplot de componentes principales simplemente reune los dos gráficos anteriores en uno solo. Así, las relaciones entre los objetos y las variables pueden apreciarse en este tipo de gráficos. En el ejemplo desarrollado para las aves, un biplot corresponde a superponer, a la manera de dos acetatos o transparencias, las gráficas de las figuras 5.10 y 5.11.

5.8 Rutina SAS para componentes principales

Mediante el siguiente procedimiemto computacional del PROC PRINCOMP del paquete SAS, se obtienen las componentes principales asociadas a un conjunto de datos cuya escala sea al menos de intervalo. Las componentes principales pueden ser generadas desde la matriz de covarianzas o desde la matriz de correlación. El procedimiento permite la creación de un archivo con las "nuevas" coordenadas (puntajes), como también la elaboración de planos factoriales para la proyección de observaciones. Para algunas opciones de más cálculo y computacionales consultar (SAS User's Guide, 2001).

```
TITLE 'Generaci\'on de componentes principales';
DATA EJEMP5\_2; /* archivo o matriz de datos */
INPUT X1 X2 X3 X4 X5 X6 X7 X8 ; /* variables X1 a X8 */
CARDS; /* ingreso de datos */
        insertar aquí los datos
; PROC PRINCOMP /* procedimiento para desarrollar componentes
                  principales. */
OUT= nombre SAS /* nombre SAS de un archivo de salida que contiene
       /* los datos originales y los puntajes de la comp.ppales. */
COV /* desarrolla comp. ppales. desde la matriz de covarianzas. */
    /* Si se omite COV, toma la matriz de correlación */
N= n /* especifica el número de comp. ppales. a computar, si no, */
     /* hace tantas como variables */
```

```
PREFIX= nombre; /* nombre a las comp. ppales., por defecto asigna */
               /* PRIN1, PRIN2,...  */
VAR lista de variables; /* variables para el ACP, por omisióon */
                       /* considera las numéricas */
PROC PLOT; /* para ubicar puntos en un plano */
PLOT PRIN2*PRIN1; /* ubica las observs. en el plano de eje vertical */
                /* PRIN2 y eje horizontal PRIN1 */
TITLE ' Primer plano factorial'; RUN;
```

5.9 Procesamiento de datos con R

Con el siguiente código R se realizan los cálculos del ejemplo 5.1

```
# lectura de datos datos<-scan()
 156.0 245.0 31.6 18.5 20.5
 154.0 240.0 30.4 17.9 19.6

 insertar aquí el resto de datos

 162.0 245.0 32.5 18.5 21.1
 164.0 248.0 32.3 18.8 20.9

 datos2<-matrix(datos,ncol=5,byrow=TRUE)
 ejemp5_1<-data.frame(datos2)

 # matriz de varianza covarianza
 cov(ejemp5_1)
 # vector de medias
 mean(ejemp5_1)
 # desviaciones estándar
 sqrt(diag(cov(ejemp5_1) ))
 # matriz de coreelaciones
 cor(ejemp5_1)

 # análisis de componentes principales desde la matriz
 #  de correlaciones
 acp<-princomp(ejemp5_1,cor=TRUE)
 summary(acp)
 # gráfico scree
 plot(acp)
 # la desviación estándar de cada componente principal
 # es decir la raiz de los valores propios de la matriz
 acp$sdev
 # Matriz con los vectores propios
 acp$loadings
 # la media de las variables originales con la que se corrigen
 # las obs
 acp$center
 # numero de observaciones
 acp$n.obs
```

```
# las coordenadas factoriales
acp$scores

#biplots
par(mfrow=c(2,2))
# primer plano factorial
biplot(acp)
#segundo plano factorial
biplot(acp,choices = c(1,3))
# tercer plano factorial
biplot(acp,choices = c(2,3))

# Acp desde la matriz de covarianzas
# ( opción por defecto )
acpCov<-princomp(ejemp5_1)
```

Capítulo 6

Análisis de factores comunes y únicos

6.1 Introducción

Uno de los propósitos de la actividad científica es condensar las relaciones observadas entre eventos, para explicar, predecir, controlar o hacer formulaciones teóricas sobre el campo donde se inscriben tales observaciones. Un procedimiento para alcanzar este objetivo consiste en tratar de incluir la máxima información contenida en las variables originales, en un número menor de variables derivadas[1], manteniendo en lo posible una solución de fácil interpretación. En tales casos el investigador, frecuentemente, acopia información sobre las variables que hacen visibles los conceptos puestos en consideración, para tratar de descubrir si las relaciones entre las variables observadas son consistentes con los conceptos asumidos y que ellas pretenden medir o si, por vía alterna, deben plantearse estructuras diferentes o más complejas.

En muchas áreas del conocimiento no siempre es posible medir directamente los conceptos sobre los que se tiene algún interés; por ejemplo, en psicología la *inteligencia*, en economía el *nivel de desarrollo* de un país. En tales casos el investigador acude a una serie de indicadores de los conceptos y trata de descubrir si las relaciones entre estas variables observadas son consistentes con lo que se quiere que ellas midan.

Así, el *análisis de factores comunes y únicos*, más conocido como *análisis factorial* (AF), persigue describir la relación de covariación entre múltiples variables, en términos de pocas variables aleatorias no observables, llamadas *factores*. El análisis factorial se basa en un modelo, el cual considera el vector de observaciones compuesto por una parte sistemática y por un error

[1] Principio de parsimonia.

no observable. La parte sistemática se asume como una combinación lineal de un número pequeño de "nuevas" variables no observables (latentes), llamadas *factores*, la parte no sistemática corresponde a los errores, los cuales se asumen incorrelacionados. De esta manera, el análisis se concentra en los efectos de los factores. Como en los modelos lineales, se desarrolla la estimación para la parte sistemática y se verifica su ajuste. La estimación se hace a través de algunos métodos tales como *el de la componente principal*, *el del factor principal* y *el de máxima verosimilitud*. En algunas circunstancias los factores conseguidos no muestran una asociación clara e interpretable con las variables, razón por la cual, mediante algunas rotaciones, y con la ayuda de los especialistas de cada campo, se facilita la interpretación.

6.2 El Modelo factorial

El análisis factorial se dirige a establecer si las covarianzas o correlaciones observadas sobre un conjunto de variables pueden ser explicados en términos de un número pequeño no observable de variables latentes.

De esta manera, considérese a X como un vector aleatorio de tamaño $(p\times 1)$ con media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$; se trata entonces de indagar acerca del siguiente modelo

$$X = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathfrak{f} + U, \tag{6.1}$$

escrito más explícitamente como

$$\begin{cases} X_1 &= \mu_1 + \lambda_{11}\mathfrak{f}_1 + \cdots + \lambda_{1k}\mathfrak{f}_k + u_1 \\ X_2 &= \mu_2 + \lambda_{21}\mathfrak{f}_1 + \cdots + \lambda_{2k}\mathfrak{f}_k + u_2 \\ & \vdots \\ X_p &= \mu_p + \lambda_{p1}\mathfrak{f}_1 + \cdots + \lambda_{pk}\mathfrak{f}_k + u_p, \end{cases}$$

donde $\boldsymbol{\Lambda}$ es una matriz de constantes (ponderaciones, cargas o pesos) de tamaño $(p \times k)$, $\mathfrak{f}$ es un vector columna de k componentes $(k \leq p)$ y U un vector aleatorio de tamaño $(p \times 1)$ con distribución independiente de $\mathfrak{f}$. Respectivamente:

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{22} & \cdots & \lambda_{pk} \end{pmatrix}, \ \mathfrak{f} = \begin{pmatrix} \mathfrak{f}_1 \\ \mathfrak{f}_2 \\ \vdots \\ \mathfrak{f}_k \end{pmatrix} \text{ y } U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_p \end{pmatrix}.$$

La escritura anterior señala que la información contenida por cada variable "engloba" varios aspectos (los ƒ's), compartidos en grado o intensidad distinta por las demás variables, y alguna información exclusiva de la variable. Los elementos de ƒ son llamados los *factores comunes* y los elementos de U *factores únicos o específicos.*

Para efectos de estimación se asume que

$$\mathcal{E}(\mathfrak{f}) = 0, \qquad \mathrm{Cov}(\mathfrak{f}) = I.$$

$$\mathcal{E}(U) = 0, \quad \mathrm{Cov}(U) = \mathcal{E}(UU') = \boldsymbol{\Psi} \quad \text{y} \ \mathrm{Cov}(\mathfrak{f}, U) = 0. \tag{6.2}$$

Por la incorrelación de los errores, la matriz $\boldsymbol{\Psi}$ debe ser una matriz diagonal.

Observación:

- Sin pérdida de generalidad se puede asumir $\boldsymbol{\mu} = 0$; ya que el procedimiento es invariante respecto a la localización de los datos; es decir; los resultados son equivalentes para datos a los cuales se les resta la media $\boldsymbol{\mu}$.
- Se pueden considerar dos tipos de modelos de acuerdo con la aleatoriedad o no de ƒ. Tomar a ƒ como un vector aleatorio es apropiado cuando diferentes muestras constan de diferentes individuos. En el caso de no aleatoriedad, una escritura más precisa es $X_\alpha = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathfrak{f}_\alpha + U$, donde el subíndice α señala a un individuo particular. El segundo modelo es apropiado cuando existe interés en un conjunto definido de individuos y no en la estructura de los factores.
- Considerar $\mathrm{Cov}(\mathfrak{f}) = \mathcal{E}(\mathfrak{f}\mathfrak{f}') = \boldsymbol{I}$, significa que los factores son *ortogonales*; de otra manera si $\mathrm{Cov}(\mathfrak{f}) = \boldsymbol{\Phi}$, entonces ésta es una matriz no diagonal y los factores se denominan *oblicuos.*

El modelo expresado en (6.1) muestra que

$$X_i = \mu_i + \sum_{j=1}^{k} \lambda_{ij}\mathfrak{f}_j + u_i, \quad i = 1, \ldots, p,$$

y por tanto,

$$\mathrm{var}(X_i) = \sigma_{ii} = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_{ii}.$$

De tal forma que la varianza de X_i puede ser descompuesta en dos partes; la primera

$$\lambda_{i1}^2 + \cdots + \lambda_{ik}^2 = \sum_{j=1}^{k} \lambda_{ij}^2 = h_i^2, \tag{6.3}$$

se denomina la *comunalidad* y representa la varianza de X_i, compartida con las otras variables a través de los factores comunes $\mathfrak{f}$. La segunda parte ψ_{ii}, representa la variabilidad exclusiva de X_i; es decir, la varianza que no es compartida con las otras variables, se llama la *especificidad* o la *varianza única*. La escritura matricial que resume los supuestos anteriores es

$$\text{Cov}(X) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{6.4}$$

Si los factores son ortogonales $\boldsymbol{\Phi} = \boldsymbol{I}$, y por lo tanto (6.4) se transforma en

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{6.4a}$$

Cuando las variables originales se han estandarizado, el análisis puede desarrollarse a partir de la matriz de correlación $\boldsymbol{R}$ y así (6.4a) se escribe

$$\boldsymbol{R} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{6.4b}$$

Aunque la escritura de las matrices $\boldsymbol{\Lambda}$ y $\boldsymbol{\Psi}$ que conforman las desagregaciones (6.4a) y (6.4b) es la misma, se advierte que estas matrices, en general, no coinciden en las dos descomposiciones.

La contribución del factor $\mathfrak{f}_j$ a la varianza total es

$$V_j = \sum_{i=1}^{p} \lambda_{ij}^2 = \boldsymbol{\lambda}_j'\boldsymbol{\lambda}_j,$$

donde $\boldsymbol{\lambda}_j$ denota la j-ésima columna de la matriz $\boldsymbol{\Lambda}$.

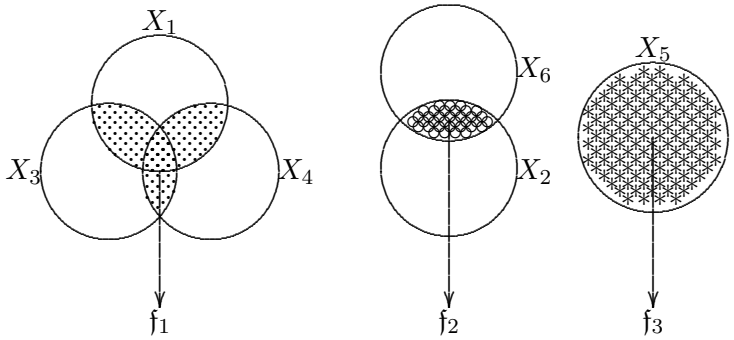


Figura 6.1 *Variables y factores.*

A manera de ilustración, la figura 6.1 muestra una situación donde, por ejemplo, las variables X_1, X_3 y X_4 comparten, con intensidades diferentes,

el factor $\mathfrak{f}_1$, las variables X_2 y X_6 comparten, con intensidades diferentes, el factor $\mathfrak{f}_2$ y la variable X_5 coincide con el factor $\mathfrak{f}_3$. La región sombreada se asocia con la *comunalidad* y la no sombreada con la *unicidad* para cada variable.

La variabilidad retenida por todos los factores comunes es la comunalidad total H, la cual se define como:

$$H = \sum_{j=1}^{k} V_j = \sum_{i=1}^{p} \sum_{j=1}^{k} \lambda_{ij}^2.$$

La participación del factor $\mathfrak{f}_j$, de la comunalidad total, está dada por:

$$V_c = \frac{V_j}{H}.$$

Bajo normalidad, toda la información acerca de la estructura factorial se obtiene de $\mathcal{E}(X) = \boldsymbol{\mu}$ y de $\text{Cov}(X) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi}$.

Un cambio de escala, en las variables aleatorias que conforman X, se obtiene mediante la transformación $Y = \boldsymbol{C}X$, donde $\boldsymbol{C}$ es una matriz diagonal. En el modelo (6.1), llámese $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_X$ y $\boldsymbol{\Psi} = \boldsymbol{\Psi}_X$, de aquí

$$\begin{aligned} X &= \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathfrak{f} + U, \quad \text{al premultiplicar por } \boldsymbol{C}, \text{ se obtiene} \\ \boldsymbol{C}X &= \boldsymbol{C\mu} + \boldsymbol{C\Lambda}\mathfrak{f} + \boldsymbol{C}U \\ Y &= \boldsymbol{C\mu} + \boldsymbol{C\Lambda}\mathfrak{f} + \boldsymbol{C}U, \end{aligned}$$

con

$$\begin{aligned} \text{Cov}(Y) &= \boldsymbol{C\Sigma C} \\ &= \boldsymbol{C\Lambda}_X\Lambda_X'\boldsymbol{C} + \boldsymbol{C\Psi}_X\boldsymbol{C}. \end{aligned}$$

Luego el modelo k–factorial no se afecta por un cambio de escala de las variables, pues la matriz de las ponderaciones factoriales resultante es igual a $\boldsymbol{\Lambda}_Y = \boldsymbol{C\Lambda}_X$ con varianzas específicas $\boldsymbol{\Psi}_Y = \boldsymbol{C\Psi}_X\boldsymbol{C} = \text{Diag}(c_i^2\psi_{ii})$.

Por ejemplo, la matriz $\boldsymbol{C}$ es aquella cuyos elementos sobre la diagonal son iguales a los recíprocos de las desviaciones estándar observables, es decir: $(c_{ii} = 1/\sqrt{\sigma_{ii}})$. La matriz de covarianzas del "nuevo" vector es $\text{Cov}(Y) = \boldsymbol{C\Sigma C}$, la cual coincide con la matriz de correlaciones.

El siguiente ejemplo[2] ilustra lo expuesto. Se desarrollaron pruebas sobre idioma Clásico (X_1), Francés (X_2) e Inglés (X_3) en jóvenes. La matriz de

[2]Mardia y colaboradores (1979, pag. 255).

correlación es

$$\begin{bmatrix} 1 & 0.83 & 0.78 \\ 0.83 & 1 & 0.67 \\ 0.78 & 0.67 & 1 \end{bmatrix},$$

la cual es una matriz no singular, las tres variables pueden expresarse de la siguiente manera:

$$X_1 = \lambda_1 \mathfrak{f} + u_1, \qquad X_2 = \lambda_2 \mathfrak{f} + u_2 \ \text{ y } \ X_3 = \lambda_3 \mathfrak{f} + u_3.$$

Para esta situación, $\mathfrak{f}$ representa el factor común y λ_1, λ_2 y λ_3 representan las *ponderaciones factoriales*. El factor común $\mathfrak{f}$ se interpreta como la "habilidad general" y la variación de los u_i representa el complemento de la habilidad general en cada idioma; es decir, la habilidad que no se contempla en la habilidad general, como por ejemplo, el error de medida sobre cada sujeto (o también lo exclusivo de ese individuo) que no es registrado por $\mathfrak{f}$.

En este caso, tres variables han sido representadas por una sola variable $\mathfrak{f}$, esto equivale a decir que la información contenida en un espacio de dimensión tres se ha representado en un espacio de dimensión uno. Otro problema es responder a la pregunta ¿Qué tan buena es esta representación?.

▶ No unicidad de las ponderaciones en los factores

Los coeficientes o ponderaciones en el modelo (6.1) o (6.1a) pueden multiplicarse por una matriz ortogonal sin que estos pierdan la capacidad de generar la matriz de covarianzas en la forma (6.4). Para mostrar esta propiedad, considérese la matriz ortogonal $\boldsymbol{T}$; es decir, $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{T}\boldsymbol{T}' = \boldsymbol{I}$, de tal forma que el modelo (6.1) puede escribirse como

$$X - \boldsymbol{\mu} = \boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{T}'\mathfrak{f} + U = (\boldsymbol{\Lambda}\boldsymbol{T})(\boldsymbol{T}'\mathfrak{f}) + U = \boldsymbol{\Lambda}^*\mathfrak{f}^* + U,$$

donde $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{T}$ y $\mathfrak{f}^* = \boldsymbol{T}'\mathfrak{f}$.

Si se reemplaza $\boldsymbol{\Lambda}$ por $\boldsymbol{\Lambda}^* = (\boldsymbol{\Lambda}\boldsymbol{T})$ en $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, se obtiene

$$\begin{aligned} \boldsymbol{\Sigma} &= (\boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\prime} + \boldsymbol{\Psi}) = \boldsymbol{\Lambda}\boldsymbol{T}(\boldsymbol{\Lambda}\boldsymbol{T})' + \boldsymbol{\Psi} \\ &= \boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{\Lambda}' + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \end{aligned}$$

de esta forma se muestra que *transformaciones ortogonales* de los factores reproducen la matriz de covarianzas; es decir,

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\prime} + \boldsymbol{\Psi}.$$

Los nuevos factores $\mathfrak{f}^*$ satisfacen los supuestos presentados en las ecuaciones (6.2) para el modelo de factores; es decir, $\mathcal{E}(\mathfrak{f}^*) = 0$, $\text{Cov}(\mathfrak{f}^*) = \boldsymbol{I}$ y $\text{Cov}(\mathfrak{f}^*, U) = 0$.

Las comunalidades $h_i^2 = \lambda_{i1}^2 + \cdots + \lambda_{ik}^2$, para $i = 1, \ldots, p$, resultan inalteradas por la transformación $\boldsymbol{\Lambda T}$, pues

$$h_i^2 = \boldsymbol{\lambda}_j^* \boldsymbol{\lambda}_j^{*\prime} = \boldsymbol{\lambda}_j' \boldsymbol{T} \boldsymbol{T}' \boldsymbol{\lambda}_j = \boldsymbol{\lambda}_j' \boldsymbol{\lambda}_j = h_i^2.$$

En resumen, la no unicidad de las ponderaciones en los factores se tiene por la rotación ortogonal de éstos, cada rotación ortogonal produce "nuevos" pesos de los factores que reproducen la misma estructura de la matriz de covarianzas.

6.3 Comunalidad

En el modelo propuesto para el análisis factorial se resaltan los componentes comunes y específicos de las variables. El interés se dirige a la cantidad de variabilidad que una variable comparte con las demás. La ecuación (6.4b) muestra que si a la matriz de correlación $\boldsymbol{R}$ se le cambian los elementos de la diagonal por las respectivas *comunalidades*, se obtiene la matriz de *correlación reducida* $\boldsymbol{R}^*$; pues $\boldsymbol{R} - \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$. Para estimar la matriz $\boldsymbol{\Psi}$ se deben estimar primero las comunalidades. Algunos de los procedimientos más conocidos se presentan, en forma resumida, enseguida.

1. La comunalidad de la i-ésima variable se estima mediante la correlación más alta, en valor absoluto, observada entre la variable X_i y las demás $(p-1)$ variables. Este valor se ubica en el respectivo sitio de la diagonal de la matriz de correlación.

2. Un método alternativo para estimar las comunalidades está dado por

$$h_i^2 = \frac{r_{ij} r_{ik}}{r_{jk}}, \tag{6.5a}$$

donde X_j y X_k son dos variables con la correlación más alta respecto a X_i.

3. También se puede estimar la comunalidad, mediante el promedio de las correlaciones de las respectivas variables, así

$$h_i^2 = \sum_{j=1}^{p} \frac{r_{ij}}{p-1}, \quad \text{con } i \neq j, \ i = 1, \ldots, p. \tag{6.5b}$$

Los valores extremos que las comunalidades pueden alcanzar son: de una parte 0.0 si las variables no tienen correlación, de otra parte, 1.0 si la varianza es perfectamente reunida por el conjunto de factores propuesto. Comunalidades negativas no tienen sentido y no ocurren excepto por errores de redondeo de comunalidades cercanas a cero. Cuando las comunalidades

se estiman con (6.5a) pueden resultar valores mayores que 1, tales casos Gorsuch (1983, pág 102) los denomina "casos Heywood" (Heywood, 1931) y sugiere igualarlos a 0.99 o a 1.0.

4. Otro procedimiento consiste en tomar la comunalidad de una variable X_i como el cuadrado de su *coeficiente de correlación múltiple* con las demás $(p-1)$ variables . El cuadrado de la correlación múltiple suministra el porcentaje de varianza que la variable tiene en común con todas las demás variables en la matriz de datos inicial, a manera de una regresión de la variable X_i sobre las demás $(p-1)$ variables.

5. Procedimientos iterativos han sido desarrollados gracias al empleo de la tecnología computacional. Se inicia con la matriz de correlación corregida en sus valores diagonales. La suma de cuadrados de los coeficientes factoriales, para un número predeterminado de factores, es empleada como las comunalidades. El procedimiento sigue con una nueva matriz de correlaciones. Las iteraciones se desarrollan hasta que las comunalidades se estabilicen, de acuerdo con una regla de convergencia establecida.

6.4 Métodos de estimación

Con los vectores observados $X_1, \ldots, X_n$ constituidos por p-variables aleatorias, se pretende responder a la pregunta: *¿Representa el modelo factorial (6.1), con un número pequeño de factores, adecuadamente los datos (ajuste del modelo)?*. Existen varios métodos para estimar las ponderaciones factoriales λ_{ij} y las varianzas ψ_{ij}; aquí se consideran los tres más comunes: *el método de la componente principal*, *el del factor principal* y el de *máxima verosimilitud*.

6.4.1 Método de la componente principal

El nombre de la técnica puede contribuir a la confusión entre análisis factorial y análisis de componentes principales. En el método de la componente principal para estimar las ponderaciones λ_{ij}, no se calcula componente principal alguna. Con el desarrollo de la metodología se despejará esta aparente ambigüedad.

A través de una muestra aleatoria $X_1, \ldots, X_n$, se obtiene la matriz de covarianza $\boldsymbol{S}$ y se pretende buscar un estimador $\hat{\boldsymbol{\Lambda}}$ que se aproxime a la expresión (6.4a) con $\boldsymbol{S}$ en lugar de $\boldsymbol{\Sigma}$; es decir,

$$\boldsymbol{S} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}. \tag{6.6}$$

En la aproximación mediante el componente principal, se considera la matriz $\hat{\boldsymbol{\Psi}}$ como insignificante, entonces, la matriz de covarianzas muestral se factoriza de la forma $\boldsymbol{S} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$. La descomposición espectral (expresión A2.27) de $\boldsymbol{S}$ es

$$\boldsymbol{S} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}'. \tag{6.7}$$

Así, la matriz $\boldsymbol{S}$ se puede escribir en la forma $\boldsymbol{S} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$, pero la matriz $\hat{\boldsymbol{\Lambda}}$ no se define como $\boldsymbol{P}\boldsymbol{D}^{1/2}$ porque $\boldsymbol{P}\boldsymbol{D}^{1/2}$ es de tamaño $(p \times p)$, mientras que $\hat{\boldsymbol{\Lambda}}$ es de tamaño $(p \times k)$, con $k < p$. De esta manera se debe definir la matriz $\boldsymbol{D}_1$ como aquella que contenga los k valores propios más grandes $\theta_1 > \theta_2 > \cdots > \theta_k$ y la matriz $\boldsymbol{P}_1$ conformada por los correspondientes vectores propios. La matriz $\boldsymbol{\Lambda}$ se estima por

$$\hat{\boldsymbol{\Lambda}} = \boldsymbol{P}_1 {\boldsymbol{D}_1}^{1/2}, \tag{6.8}$$

donde $\hat{\boldsymbol{\Lambda}}$ es de tamaño $(p \times k)$, $\boldsymbol{P}_1$ es de tamaño $(p \times k)$ y ${\boldsymbol{D}_1}^{\frac{1}{2}}$ es de tamaño $(k \times k)$.

Como una ilustración a la estructura de los λ_{ij} mostrados en las ecuaciones (6.9), considérese $p = 5$ y $k = 2$:

$$\begin{pmatrix} \hat{\lambda}_{11} & \hat{\lambda}_{12} \\ \hat{\lambda}_{21} & \hat{\lambda}_{22} \\ \hat{\lambda}_{31} & \hat{\lambda}_{32} \\ \hat{\lambda}_{41} & \hat{\lambda}_{42} \\ \hat{\lambda}_{51} & \hat{\lambda}_{52} \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ p_{31} & p_{32} \\ p_{41} & p_{42} \\ p_{51} & p_{52} \end{pmatrix} \begin{pmatrix} \sqrt{\theta_1} & 0 \\ 0 & \sqrt{\theta_2} \end{pmatrix} = \begin{pmatrix} \sqrt{\theta_1}p_{11} & \sqrt{\theta_2}p_{12} \\ \sqrt{\theta_1}p_{21} & \sqrt{\theta_2}p_{22} \\ \sqrt{\theta_1}p_{31} & \sqrt{\theta_2}p_{32} \\ \sqrt{\theta_1}p_{41} & \sqrt{\theta_2}p_{42} \\ \sqrt{\theta_1}p_{51} & \sqrt{\theta_2}p_{52} \end{pmatrix}.$$

En esta última expresión se encuentra la explicación del calificativo *método del componente principal*, pues se observa que las ponderaciones del j-ésimo factor (columnas de la matriz $\hat{\boldsymbol{\Lambda}}$) son proporcionales a los coeficientes (o ponderaciones) del j-ésimo componente principal.

Los elementos de la diagonal de la matriz $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$ corresponden a la suma de cuadrados de la respectiva fila de la matriz $\hat{\boldsymbol{\Lambda}}$, o $\hat{\boldsymbol{\lambda}}_{\boldsymbol{i}}\hat{\boldsymbol{\lambda}}_{\boldsymbol{i}}{}' = \sum_{j=1}^{k} \hat{\lambda}_{ij}^2$. Por tanto, para completar la aproximación de $\boldsymbol{S}$, conforme a (6.6), se define

$$\hat{\psi}_i = s_{ii} - \sum_{j=1}^{k} \hat{\lambda}_{ij}^2,$$

para poder escribir

$$\boldsymbol{S} \cong \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}},$$

donde $\hat{\boldsymbol{\Psi}} = \text{Diag}(\hat{\psi}_1, \ldots, \hat{\psi}_p)$. Se nota que las varianzas sobre la diagonal de la matriz $\boldsymbol{S}$ se tienen de manera exacta, pero las que están por fuera de la diagonal no.

En esta metodología de estimación, la suma de los cuadrados en las filas y las columnas de la matriz $\hat{\mathbf{\Lambda}}$ son iguales a las comunalidades y a los valores propios, respectivamente, así:

$$\hat{h}_i^2 = \sum_{j=1}^{k} \hat{\lambda}_{ij}^2, \qquad \sum_{i=1}^{p} \hat{\lambda}_{ij}^2 = \sum_{i=1}^{p} (\sqrt{\theta_j} p_{ij})^2 = \theta_j \sum_{i=1}^{p} p_{ij}^2 = \theta_j$$

Ejemplo 6.1 Para un estudio de referencia[3], se obtuvo una muestra aleatoria de consumidores a quienes se les indagó acerca de los siguientes atributos sobre un producto nuevo: gusto X_1, costo X_2, sabor X_3, tamaño por porción X_4 y calorías suministradas X_5. La matriz de correlación es la siguiente:

$$\mathbf{R} = \begin{pmatrix} 1.00 & .02 & \boxed{.96} & .42 & .01 \\ .02 & 1.00 & .13 & .71 & \boxed{.85} \\ .96 & .13 & 1.00 & .50 & .11 \\ .42 & .71 & .50 & 1.00 & \boxed{.79} \\ .01 & .85 & .11 & .79 & 1.00 \end{pmatrix}.$$

Las correlaciones enmarcadas indican que las respectivas variables se pueden agrupar para formar nuevas variables. Así, los grupos de variables son $\{X_1,\ X_3\}$, $\{X_2,\ X_5\}$, mientras que la variable X_4 está más cercana al segundo grupo que al primero.

Las relaciones lineales que se pueden derivar de estas correlaciones sugieren que la información representada por estas variables se puede sintetizar a través de dos o tres factores.

Los valores propios de la matriz de correlaciones $\mathbf{R}$ son: $\theta_1 = 2.853$, $\theta_2 = 1.806$, $\theta_3 = 0.203$, $\theta_4 = 0.102$ y $\theta_5 = 0.033$. Los vectores propios asociados con valores propios distintos son ortogonales, los cuales son normalizados para conformar la matriz $\boldsymbol{P}$; ésta viene dada por:

$$\boldsymbol{P} = \begin{pmatrix} 0.331 & 0.607 & 0.098 & 0.139 & 0.702 \\ 0.460 & -0.390 & 0.743 & -0.282 & 0.0717 \\ 0.382 & 0.557 & 0.168 & 0.117 & -0.709 \\ 0.556 & -0.078 & -0.602 & -0.568 & 0.002 \\ 0.473 & -0.404 & -0.221 & 0.751 & 0.009 \end{pmatrix}.$$

[3]Johnson y Wichern (1998, pág. 525)

La proporción de variabilidad acumulada hasta cada uno de los factores está indicada en el siguiente recuadro

$k:$	(1)	(2)	(3)	(4)	(5)
$PV_k:$	0.5706	$\boxed{\frac{2.853+1.806}{5}=0.9318}$	0.9726	0.9930	1.0000

Se nota que con dos factores ($k = 2$) se reune una buena proporción de la variabilidad total presente en los datos iniciales (93.18%). La matriz de ponderaciones factoriales se obtiene como se indica en la ecuación (6.8), tales ponderaciones para los datos en consideración son:

$$\hat{\mathbf{\Lambda}} = \begin{pmatrix} 0.559 & 0.816 \\ 0.777 & -0.524 \\ 0.645 & 0.748 \\ 0.939 & -0.105 \\ 0.798 & -0.543 \end{pmatrix}.$$

En el cuadro siguiente se resumen las ponderaciones factoriales, las comunalidades y las varianzas específicas.

Variable	Peso factorial estimado		Comunalidad	Varianza específica
	$\hat{\lambda}_{ij} = \sqrt{\theta_j} p_{ij}$		h_i^2	$\hat{\psi}_i^2 = 1 - h_i^2$
	f1	f2		
X_1	0.559	0.816	$\boxed{0.559^2 + 0.816^2 = 0.978}$	$\boxed{1 - 0.978 = 0.022}$
X_2	0.777	− 0.524	0.878	0.122
X_3	0.645	0.748	0.975	0.025
X_4	0.939	− 0.105	0.892	0.108
X_5	0.798	− 0.543	0.931	0.069

Sobre los resultados anteriores y de una manera descriptiva, se puede sugerir un modelo con dos factores para los datos. Se aplaza la interpretación de cada uno de estos factores hasta la sección de rotación de factores. ☑

6.4.2 Método del factor principal

Este método se llama también el *método del eje principal*, y se basa en una estimación inicial de $(\hat{\mathbf{\Psi}})$ y la factorización de $(\boldsymbol{S} - \hat{\mathbf{\Psi}})$ o $(\boldsymbol{R} - \hat{\mathbf{\Psi}})$ para obtener

$$\boldsymbol{S} - \hat{\mathbf{\Psi}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}', \quad \text{o} \quad \boldsymbol{R} - \hat{\mathbf{\Psi}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}', \tag{6.9}$$

donde $\hat{\boldsymbol{\Lambda}}$ es una matriz de tamaño $(p \times k)$ que se calcula a partir de la expresión (6.8) empleando los valores y vectores propios de $\boldsymbol{S} - \hat{\boldsymbol{\Psi}}$ o $\boldsymbol{R} - \hat{\boldsymbol{\Psi}}$.

Los elementos de la diagonal de la matriz $\boldsymbol{S} - \hat{\boldsymbol{\Psi}}$, por definición, son las comunalidades $\hat{h}_i^2 = s_{ii} - \hat{\psi}_i$ y los elementos de la diagonal de la matriz $\boldsymbol{R} - \hat{\boldsymbol{\Psi}}$ son las comunalidades $\hat{h}_i^2 = 1 - \hat{\psi}_i$. Naturalmente, tanto $\hat{h}_i^2$ como $\hat{\psi}_i$ tienen valores diferentes, dependiendo de si se emplea $\boldsymbol{S}$ o $\boldsymbol{R}$. Los valores en la diagonal de $(\boldsymbol{S} - \hat{\boldsymbol{\Psi}})$ y $(\boldsymbol{R} - \hat{\boldsymbol{\Psi}})$ son $\hat{h}_1^2, \ldots, \hat{h}_p^2$, respectivamente. Se insiste, en que a pesar de escribirse de la misma manera para las matrices $(\boldsymbol{S} - \hat{\boldsymbol{\Psi}})$ y $(\boldsymbol{R} - \hat{\boldsymbol{\Psi}})$, éstas no necesariamente son iguales.

$$\boldsymbol{S} - \hat{\boldsymbol{\Psi}} = \begin{pmatrix} \hat{h}_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & \hat{h}_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & \hat{h}_p^2 \end{pmatrix},$$

$$\boldsymbol{R} - \hat{\boldsymbol{\Psi}} = \begin{pmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & \hat{h}_p^2 \end{pmatrix}. \tag{6.10}$$

En las ecuaciones anteriores no se conoce el valor de las comunalidades $\hat{h}_i^2$. Un método usado para estimarlas en $\boldsymbol{R} - \hat{\boldsymbol{\Psi}}$ es

$$\hat{h}_i^2 = 1 - \frac{1}{r^{ii}}, \tag{6.11}$$

donde r^{ii} es el i-ésimo elemento sobre la diagonal de la matriz $\boldsymbol{R}^{-1}$. Se demuestra que $1 - 1/r^{ii} = R_i^2$, es el coeficiente de correlación múltiple entre la variable X_i y las demás $(p-1)$ variables (como se indicó en sección (6.3)). Nótese que si un factor está asociado con sólo una variable, por ejemplo X_i, el uso de $\hat{h}_i^2 = R_i^2$ mostrará pequeñas ponderaciones para X_i en todos los factores, incluyendo el factor asociado con X_i; esto se debe a que $\hat{h}_i^2 = R_i^2 = \hat{\lambda}_{i1}^2 + \cdots + \hat{\lambda}_{ik}^2$ y R_i^2 será pequeño debido a que X_i tiene poco en común con las demás $(p-1)$ variables.

Para $\boldsymbol{S} - \hat{\boldsymbol{\Psi}}$, un estimador inicial de la comunalidad, como en (6.11), es

$$\hat{h}_i^2 = s_{ii} - \frac{1}{s^{ii}}, \tag{6.12}$$

donde s_{ii} es el i-ésimo elemento sobre la diagonal de $\boldsymbol{S}$ y s^{ii} es el i-ésimo elemento sobre la diagonal de $\boldsymbol{S}^{-1}$. Se demuestra también que (6.12) es equivalente a

$$\hat{h}_i^2 = s_{ii} - \frac{1}{s^{ii}} = s_{ii} R_i^2, \tag{6.12a}$$

Después de estimar la comunalidad, se calculan los valores y vectores propios de $(\boldsymbol{S}-\hat{\boldsymbol{\Psi}})$ o $(\boldsymbol{R}-\hat{\boldsymbol{\Psi}})$, los cuales se utilizan para obtener estimadores de las ponderaciones $\hat{\lambda}_{ij}$, elementos de $\hat{\boldsymbol{\Lambda}}$. De esta manera, las columnas y filas de $\hat{\boldsymbol{\Lambda}}$ pueden emplearse para calcular nuevos valores propios (varianza explicada) y comunalidades, respectivamente. Así, la suma de los cuadrados de los elementos de la j-ésima columna de $\hat{\boldsymbol{\Lambda}}$ es el j-ésimo valor propio de $(\boldsymbol{S}-\hat{\boldsymbol{\Psi}})$ o de $(\boldsymbol{R}-\hat{\boldsymbol{\Psi}})$, y la suma de los cuadrados de la i-ésima fila de $\hat{\boldsymbol{\Lambda}}$ es la comunalidad de la variable X_i.

El procedimiento anterior puede desarrollarse de una manera iterativa "mejorando" la estimación de las comunalidades en cada etapa. Se inicia con los valores "ad hoc" de la comunalidad señalados en (6.11) o (6.12), con estascomunalidades se obtiene $\hat{\boldsymbol{\Lambda}}$ a partir de (6.9), de donde se pueden obtener nuevas comunalidades mediante la suma de cuadrados en cada fila; es decir, $\hat{h}_i^2 = \sum_{j=1}^{k} \hat{\lambda}_{ij}^2$. Estos valores de $\hat{h}_i^2$ son sustituidos en $(\boldsymbol{S}-\hat{\boldsymbol{\Psi}})$ o $(\boldsymbol{R}-\hat{\boldsymbol{\Psi}})$, con los cuales se pueden obtener nuevos valores para la matriz $\hat{\boldsymbol{\Lambda}}$. Este proceso continúa hasta que las comunalidades estimadas "se estabilicen" o converjan.

6.4.3 Método de máxima verosimilitud

Si los factores comunes ƒ y los errores U se pueden asumir con distribución normal, entonces, es procedente estimar, vía máxima verosimilitud, los coeficientes factoriales y las varianzas específicas.

El problema consiste en encontrar $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$ y $\boldsymbol{\Phi}$ que satisfagan

$$\mathrm{Cov}(X) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}.$$

Se imponen algunas restricciones para asegurar la existencia y unicidad de las soluciones (Anderson 1984, pág. 557). Se supone que $\boldsymbol{\Phi} = \boldsymbol{I}$; es decir los factores son independientes o incorrelacionados, además que la matriz $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} = \boldsymbol{\Gamma}$ es diagonal.

Sea $X_1, \ldots, X_n$ una muestra aleatoria de $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, la función de verosimilitud para esta muestra es:

$$\boldsymbol{L} = (2\pi)^{-\frac{1}{2}pn}|\boldsymbol{\Sigma}|^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{n}(X_\alpha - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(X_\alpha - \boldsymbol{\mu})\right\}. \qquad (6.13)$$

Sea $\boldsymbol{A} = \sum_{\alpha=1}^{n}(X_\alpha - \bar{\boldsymbol{X}})(X_\alpha - \bar{\boldsymbol{X}})'$.

Maximizar (6.13) es equivalente a maximizar su logaritmo. Con $\boldsymbol{\mu}$ reemplazado por $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ y de la igualdad

$$\sum_{\alpha=1}^{n}(X_\alpha - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(X_\alpha - \boldsymbol{\mu}) = \mathrm{tra}\Big(\boldsymbol{\Sigma}^{-1}\Big\{\sum_{\alpha=1}^{n}(X_\alpha - \bar{\boldsymbol{X}})'(X_\alpha - \bar{\boldsymbol{X}}) + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'(\bar{\boldsymbol{X}} - \boldsymbol{\mu})\Big\}\Big),$$

se obtiene

$$\ln \boldsymbol{L} = l = -\frac{1}{2}pn\ln(2\pi) - \frac{1}{2}n\ln|\boldsymbol{\Sigma}| - \mathrm{tra}(\boldsymbol{A}\boldsymbol{\Sigma}^{-1}).$$

Se demuestra que la maximización de la última expresión con respecto a los elementos de $\boldsymbol{\Gamma}$ y de $\boldsymbol{\Psi}$ conducen a las siguientes ecuaciones para los estimadores de máxima verosimilitud $\hat{\boldsymbol{\Gamma}}$ y $\hat{\boldsymbol{\Psi}}$ respectivamente, considerando $n \approx (n-1)$ y por tanto $\frac{1}{n}\boldsymbol{A} = \boldsymbol{S}$,

$$\begin{aligned} \boldsymbol{S}\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} &= \hat{\boldsymbol{\Lambda}}(\boldsymbol{I} + \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Gamma}}) \\ \mathrm{Diag}(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}} + \hat{\boldsymbol{\Psi}}) &= \mathrm{Diag}(\boldsymbol{S}). \end{aligned} \tag{6.14}$$

Las ecuaciones (6.14) deben resolverse en forma iterativa. Existen varios procedimientos para encontrar un máximo de la función de verosimilitud, tales como Newton-Raphson.

Jöreskog (1967) y Lawley (1967) desarrollaron un algoritmo para obtener el máximo de $l = \ln \mathbf{L}(\boldsymbol{\Gamma}, \boldsymbol{\Psi})$. El procedimiento empieza con un valor $\boldsymbol{\Psi}_0$ con el cual se calcula $\boldsymbol{\Gamma}_0$. La función $l_0 = \ln \mathbf{L}(\boldsymbol{\Gamma}_0, \boldsymbol{\Psi})$ se maximiza para obtener $\boldsymbol{\Psi}_1$ y el respectivo $\boldsymbol{\Gamma}_1$; $l_1 = \ln \mathbf{L}(\boldsymbol{\Gamma}_1, \boldsymbol{\Psi})$ se maximiza respecto a $\boldsymbol{\Psi}$ y así sucesivamente. La existencia del máximo de la función de verosimilitud, con la restricción que $\boldsymbol{\Psi} > 0$, no se puede garantizar, sea por la falta de ajuste de los datos al modelo normal supuesto o por problemas de muestreo. En estos casos, algunos elementos de $\boldsymbol{\Psi}$ se aproximan a cero o son negativos, en el proceso iterativo. Este inconveniente se corrige desarrollando la maximización dentro de una región R_δ para la cual $\psi_j^2 \geq \delta$, para todo j, con δ un número pequeño positivo y arbitrario. Para mayores detalles del proceso consultar a Morrison (1990, pág. 357-362). Actualmente este problema ha sido satisfactoriamente resuelto por los paquetes *SAS* y *SPSS*. El primero mediante el procedimiento *FACTOR* ha mostrado conseguir las mejores estimaciones (SAS/STAT, 2001).

6.5 Número de factores a seleccionar

Nuevamente, una de las preguntas que se le interponen al usuario del análisis factorial es *¿Qué tan bueno es el ajuste a los datos del modelo*

con un número particular de factores comunes?. Éste es casi el mismo problema tratado en el capítulo de componentes principales. Varios criterios se han propuesto para escoger el valor adecuado de k, el número de factores o variables latentes, algunos son similares a los empleados para componentes principales.

Muchas de las respuestas a esta pregunta se dan con procedimientos bastante informales, basados en la experiencia e intuición, más que en un modelo analítico o matemático. Por ejemplo, uno de los criterios más populares es considerar un número de factores igual al número de valores propios que sean mayores que la unidad, siempre que los factores hayan sido generados o estimados a partir de la matriz de correlación. Otro procedimiento informal consiste en graficar el número de orden de los valores propios frente a su magnitud, se escoge el número de factores correspondiente al punto donde los valores propios empiecen a conformar una línea recta, aproximadamente horizontal. Este procedimiento es descrito también en el capítulo de componentes principales.

A continuación se presentan algunos criterios para decidir sobre el número de factores a considerar, junto con la explicación y justificación pertinente.

• *Método 1.* Se aplica particularmente cuando se han obtenido estimadores a través del método de la componente principal. Como se deduce del desarrollo hecho en la sección (6.4.1), la proporción de la varianza muestral total debida al j-ésimo factor, obtenido con base en $\boldsymbol{S}$, es:

$$(\hat{\lambda}_{1j}^2 + \cdots + \hat{\lambda}_{pj}^2)/\operatorname{tra}(\boldsymbol{S}).$$

La proporción correspondiente con base en la matriz de correlación $\boldsymbol{R}$ es:

$$(\hat{\lambda}_{1j}^2 + \cdots + \hat{\lambda}_{pj}^2)/p.$$

La contribución de todos los k factores a $\operatorname{tra}(\boldsymbol{S})$ o a p, es por tanto igual a $\sum_{i=1}^{p}\sum_{j=1}^{k}\hat{\lambda}_{ij}^2$, que es la suma de los cuadrados de todos los elementos de $\hat{\boldsymbol{\Lambda}}$. Para el método de la componente principal, se observa que por la propiedad expuesta en la sección (6.4.1), la anterior suma es igual tanto a la suma de los primeros k-valores propios como a la suma de las p-comunalidades; es decir,

$$\sum_{i=1}^{p}\sum_{j=1}^{k}\hat{\lambda}_{ij}^2 = \sum_{i=1}^{p}\hat{h}_i^2 = \sum_{j=1}^{k}\theta_j^2.$$

De tal forma, que se debe escoger un k suficientemente grande tal que la suma de las comunalidades o la suma de los valores propios (varianza retenida) constituya una proporción suficiente de la $\operatorname{tra}(\boldsymbol{S})$.

Esta estrategia puede ser extendida al método del factor principal, donde estimaciones a priori de las comunalidades son usadas para formar $(\boldsymbol{S}-\hat{\boldsymbol{\Psi}})$ o $(\boldsymbol{R}-\hat{\boldsymbol{\Psi}})$. Sin embargo, como algunos valores propios de las matrices anteriores pueden ser negativos, entonces la proporción $\sum_{j=1}^{k}\theta_j/\sum_{j=1}^{p}\theta_j$, puede exceder a 1.0, de manera que para alcanzar un porcentaje determinado (por ejemplo 80%) se necesitan menos de k factores; en consecuencia el valor adecuado de k es el correspondiente al valor propio con el cual el 100% es excedido por primera vez.

• *Método 2.* Este método consiste en escoger el valor de k como el número de valores propios mayores que la media de ellos. Así, para factores estimados con base en la matriz $\boldsymbol{R}$ el promedio es 1;y, para factores estimados con base en la matriz $\boldsymbol{S}$ es $\sum_{j=1}^{p}\theta_j/p$. Este método se encuentra incluido como una opción que opera por defecto en varios paquetes estadísticos.

• *Método 3.* Con este método se pretende verificar la hipótesis de que k es un número adecuado de factores para ajustar la estructura de covarianza; es decir, $H_0:\ \boldsymbol{\Sigma}=\boldsymbol{\Lambda}\boldsymbol{\Lambda}'+\boldsymbol{\Psi}$, donde $\boldsymbol{\Lambda}$ es una matriz de tamaño $(p\times k)$.

De acuerdo con el procedimiento desarollado en el capítulo 4, la estadística adecuada para contrastar H_0 es

$$\left(n-\frac{2p+4k+11}{6}\right)\ln\left(\frac{|\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'|}{|\boldsymbol{S}|}\right), \tag{6.15}$$

la cual se distribuye aproximadamente conforme a una distribución ji-cuadrado con $\nu=\frac{1}{2}[(p-k)^2-p-m]$ grados de libertad y $\hat{\boldsymbol{\Lambda}}$ y $\hat{\boldsymbol{\Psi}}$ estimadores de máxima verosimilitud. Si se rechaza H_0, entonces $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'+\hat{\boldsymbol{\Psi}}$ no se ajusta adecuadamente a $\hat{\boldsymbol{\Sigma}}$, y debe ensayarse con un valor más grande que k factores.

6.6 Rotación de factores

El objetivo con el análisis factorial es la obtención de una estructura de factores o variables latentes simple, las cuales puedan ser identificadas por el investigador. Cuando los modelos para los factores estimados no revelen su significado, una rotación ortogonal u oblicua de éstos puede ayudar en tal sentido. En la rotación ortogonal los factores son rotados manteniendo la ortogonalidad entre éstos (rotación "rígida"), mientras que con la rotación oblicua no.

La interpretación de las ponderaciones o coeficientes factoriales es adecuada si cada variable pondera altamente sólo un factor determinado, y si cada uno

de éstos es positivo y grande o cercano a cero. Las variables se particionan en correspondencia con cada uno de los factores; las variables que se puedan asignar a más de un factor se dejan de lado. La interpretación de un factor es la característica común, media o genérica sobre las variables cuyo l_{ij} es grande.

6.6.1 Rotación ortogonal

Uno de los problemas en el análisis factorial es la asignación apropiada del nombre a cada uno de los factores. En ACP se mostró que una representación de las variables y los individuos en el primer plano factorial, puede ayudar a la interpretación de las componentes. En el análisis de factores comunes y únicos esta representación puede resultar insuficiente y ambigua para tal propósito, pues algunas variables pueden ubicarse cerca de las diagonales del plano factorial (simétricas respecto a alguno de los ejes). En estos casos es conveniente efectuar una rotación θ de los ejes factoriales. La figura 6.2 muestra este caso para una situación particular. El plano factorial $\mathfrak{f}_1 \times \mathfrak{f}_2$ se ha rotado un ángulo θ produciendo los "nuevos" ejes $\mathfrak{f}'_1$ y $\mathfrak{f}'_2$, los cuales generan el plano factorial $\mathfrak{f}'_1 \times \mathfrak{f}'_2$. Respecto a este último sistema de coordenadas, los factores $\mathfrak{f}'_1$ y $\mathfrak{f}'_2$ se podrán interpretar con la ayuda de las variables más próximas a cada uno de ellos. Después de la rotación, las variables X_6 y X_3 tienen ponderaciones más altas, mientras que con referencia al plano inicial $\mathfrak{f}_1 \times \mathfrak{f}_2$ estas variables tienen ponderaciones casi iguales respecto a $\mathfrak{f}_1$ y a $\mathfrak{f}_2$; esto dificulta la interpretación de los ejes. Algo semejante se puede decir con respecto al eje $\mathfrak{f}_2$ de las variables X_1, X_4 y X_5.

Se conoce que una transformación ortogonal corresponde a una rotación "rígida" de los ejes de coordenadas, por tal razón la matriz de pesos factoriales se rota mediante

$$\mathbf{\Delta} = \mathbf{\Gamma}\hat{\mathbf{\Lambda}}, \tag{6.16}$$

donde $\mathbf{\Gamma}$ es una matriz de tamaño $(k \times k)$ ortogonal, δ_{ij} denota la i-ésima respuesta del j-ésimo factor rotado. La rigidez de la rotación hace que las p comunalidades h_i^2 no cambien.

A continuación se describen algunas técnicas de rotación ortogonal.

• *Rotación Varimax*

El principal objetivo de esta rotación es tener una estructura de factores, en la cual cada variable pondere altamente a un único factor. Es decir, una variable deberá tener una ponderación alta para un factor y cercana a cero

para los demás. De esta forma, resulta una estructura donde cada factor representa un constructo (o concepto) diferente.

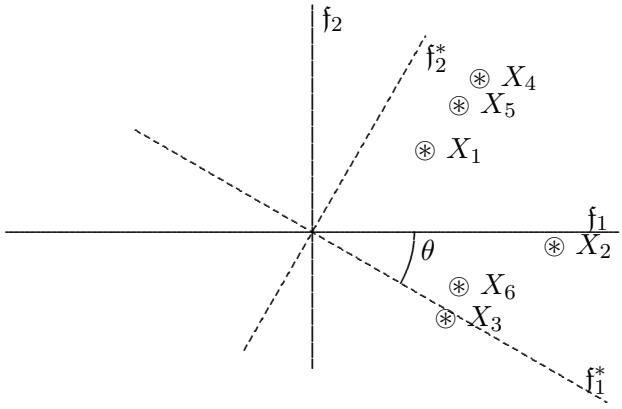


Figura 6.2 *Rotación de factores.*

De otra manera, el objetivo de la rotación *varimax* es determinar una matriz de transformación $\mathbf{\Gamma}$, tal que cualquier factor tenga algunas variables con ponderación alta y otras con ponderación baja. Esto se logra mediante la *maximización de la varianza asociada con los cuadrados de las ponderaciones ($\hat{\lambda}_{ij}$) sobre todas las variables*, con la restricción de que la comunalidad para cada variable no se altere; de aquí el nombre *varimax*. Esto se traduce en que, para un factor dado j,

$$\mathcal{V}_j = \frac{\sum_{i=1}^{p}\left(\hat{\lambda}_{ij}^2 - \bar{\hat{\lambda}}_j^2\right)^2}{p} = \frac{\sum_{i=1}^{p}\hat{\lambda}_{ij}^4 - \left(\sum_i^p \hat{\lambda}_{ij}^2\right)^2}{p^2},$$

donde $\mathcal{V}_j$ es la varianza de las comunalidades de las variables dentro del factor j y $\bar{\hat{\lambda}}_j^2 = \sum_{i=1}^{p}\hat{\lambda}_{ij}^2/p$, es el promedio de los cuadrados para las ponderaciones del factor j. La varianza para todos los factores está dada por:

$$\begin{aligned} \boldsymbol{V} &= \sum_{j=1}^{k}\mathcal{V}_j = \sum_{j=1}^{k}\left(\frac{\sum_{i=1}^{p}\hat{\lambda}_{ij}^4 - \left(\sum_i^p \hat{\lambda}_{ij}^2\right)^2}{p^2}\right) \\ &= \frac{\sum_{j=1}^{k}\sum_{i=1}^{p}\hat{\lambda}_{ij}^4}{p} - \frac{\sum_{j=1}^{k}\left(\sum_i^p \hat{\lambda}_{ij}^2\right)^2}{p^2}. \end{aligned} \tag{6.17}$$

Como el número de variables permanece constante, la maximización se hace sobre

$$p\boldsymbol{V} = \sum_{j=1}^{k}\sum_{i=1}^{p}\hat{\lambda}_{ij}^{4} - \frac{\sum_{j=1}^{k}\left(\sum_{i}^{p}\hat{\lambda}_{ij}^{2}\right)^{2}}{p}. \tag{6.18}$$

La matriz ortogonal, $\boldsymbol{\Gamma}$, se obtiene de tal forma que la ecuación (6.19) sea máxima, con la restricción de que la comunalidad para cada variable permanezca constante.

Kaiser (1958) demuestra que la rotación de los ejes un ángulo θ, implica satisfacer la siguiente ecuación

$$\tan 4\theta = \frac{2\left[2p\sum_i(\gamma_{ij}^2-\gamma_{ij'}^2)\gamma_{ij}^2\gamma_{ij'}^2 - \sum_i(\gamma_{ij}^2-\gamma_{ij'}^2)2\sum_i(\gamma_{ij}\gamma_{ij'})\right]}{p\sum_i[(\gamma_{ij}-\gamma_{ij'}^2)-(2\gamma_{ij}\gamma_{ij'})^2]-\{[\sum_i(\gamma_{ij}^2-\gamma_{ij'}^2)]^2-(2\sum_i\gamma_{ij}\gamma_{ij'})^2\}} \tag{6.19}$$

donde

$$\gamma = \frac{\hat{\lambda}_{ij}}{\sum_{j=1}^{k}\hat{\lambda}_{ij}^{2}}.$$

El ángulo 4θ se asigna, de acuerdo con el signo, al cuadrante correspondiente. El procedimiento iterativo es como sigue: rotar el primero y segundo factor de acuerdo con el ángulo solución de la ecuación (6.19), el primer nuevo factor se rota con el tercer factor original, y así sucesivamente hasta completar las $k(k-1)/2$ pares de rotaciones. Este procedimiento iterativo se desarrolla hasta cuando todos los ángulos sean menores que ϵ, de acuerdo con algún crterio de convergencia.

• *Rotación cuartimax*

El objetivo de la rotación *cuartimax* es identificar una estructura factorial en la que todos las variables tengan una fuerte ponderación con el mismo factor, y además, que cada variable, que pondere otro factor, tenga ponderaciones cercanas a cero en los demás factores. De esta forma, se persigue que las variables ponderen altamente los mismos factores y de manera relevante a otros. Este objetivo se logra por la maximización de la varianza de las ponderaciones a través de los factores, con la restricción de que la comunalidad de cada variable permanezca constante. Así, para una variable i, se define

$$Q_i = \frac{\sum_{j=1}^{k}\left(\hat{\lambda}_{ij}^{2} - \hat{\overline{\lambda}}_{i}^{2}\right)}{k}, \tag{6.20}$$

donde Q_i es la varianza de las comunalidades de la variable i (el cuadrado de las ponderaciones) y $\hat{\lambda}_{ij}^{2}$ es el cuadrado de la ponderación de la i-ésima

variable sobre el j-ésimo factor, además, $\hat{\bar{\lambda}}_i^2 = \sum_{j=1}^k \hat{\lambda}_{ij}^2/k$ es el promedio de los cuadrados de las ponderaciones en la i-ésima variable, donde k es el número de factores. La ecuación (6.20) puede escribirse en la forma siguiente:

$$Q_i = \frac{k\sum_{j=1}^k \lambda_{ij}^4 - \left(\sum_{j=1}^k \lambda_{ij}^2\right)^2}{k^2}. \tag{6.21a}$$

La varianza total sobre las p-variables está dada por:

$$= \sum_{i=1}^p Q_i = \sum_{i=1}^p \left[\frac{k\sum_{j=1}^k \lambda_{ij}^4 - \left(\sum_{j=1}^k \lambda_{ij}^2\right)^2}{k^2}\right]. \tag{6.21}$$

Como en el caso varimax, *la matriz de rotación* $\mathbf{\Gamma}$ se encuentra maximizando la función dada en (6.21), bajo la restricción de mantener constante la comunalidad para cada variable. Dado que el número de factores k se asume constante y que en la ecuación (6.21a) el término $\sum_{j=1}^k \lambda_{ij}^2$ es la comunalidad de la variable i, la cual también es constante, la maximización de (6.21) se reduce a maximizar la ecuación

$$\boldsymbol{Q} = \sum_{i=1}^p \sum_{j=1}^k \hat{\lambda}_{ij}^4. \tag{6.22}$$

• *Otras rotaciones ortogonales*

• Si se tienen tan sólo dos factores ($k = 2$), se puede emplear una rotación basada sobre una inspección visual de un *gráfico* en el que se ubican las ponderaciones, tal como lo muestra la figura 6.2. En el gráfico, los puntos corresponden a las filas de la matriz $\hat{\mathbf{\Lambda}}$; es decir, $(\hat{\lambda}_{i1}, \hat{\lambda}_{i2})$, $i = 1, \ldots, p$, para cada una de las variables $X_1, \ldots, X_p$. Se escoge un ángulo θ, a través del cual los ejes puedan ser rotados, hasta que se ubiquen cerca de la mayoría de los puntos. Las nuevas ponderaciones $(\hat{\lambda}_{i1}^*, \hat{\lambda}_{i2}^*)$ corresponden a la proyección ortogonal de cada punto sobre los "nuevos" ejes. Más formalmente, se hace una transformación ortogonal del tipo $\hat{\mathbf{\Lambda}}^* = \hat{\mathbf{\Lambda}}\mathbf{\Gamma}$, donde

$$\mathbf{\Gamma} = \begin{pmatrix} \cos\theta & \operatorname{sen}\theta \\ -\operatorname{sen}\theta & \cos\theta \end{pmatrix}.$$

La aplicación de esta estrategia requiere de una buena apreciación visual y un poco de paciencia.

Observación:

Es claro que el procedimiento de rotación varimax, maximiza la varianza total de las ponderaciones en el sentido de las columnas de $\hat{\mathbf{\Lambda}}$,

mientras que el procedimiento cuartimax lo hace en el sentido de las filas. En consecuencia, es posible desarrollar una rotación que maximice la suma ponderada de la varianza tanto en el sentido de filas como de columnas. Es decir, maximizar

$$\boldsymbol{Z} = \alpha \boldsymbol{Q} + \beta p \boldsymbol{V}, \tag{6.23}$$

donde $p\boldsymbol{V}$ está dado por (6.18) y $\boldsymbol{Q}$ por (6.22). Al reemplazar, por las respectivas expresiones, se obtiene

$$\begin{aligned} \boldsymbol{Z} &= \alpha \left(\sum_{i=1}^{p} \sum_{j=1}^{k} \hat{\lambda}_{ij}^4 \right) + \beta \left(\sum_{j=1}^{k} \sum_{i=1}^{p} \hat{\lambda}_{ij}^4 - \frac{\sum_{j=1}^{k} \left(\sum_{i}^{p} \hat{\lambda}_{ij}^2 \right)^2}{p} \right) \\ &= (\alpha + \beta) \sum_{i=1}^{p} \sum_{j=1}^{k} \hat{\lambda}_{ij}^4 - \beta \frac{\sum_{j=1}^{k} \left(\sum_{i}^{p} \hat{\lambda}_{ij}^2 \right)^2}{p}. \end{aligned} \tag{6.24}$$

Si se divide por $(\alpha + \beta)$ resulta

$$\boldsymbol{Z}^* = \sum_{i=1}^{p} \sum_{j=1}^{k} \hat{\lambda}_{ij}^4 - \gamma \frac{\sum_{j=1}^{k} \left(\sum_{i}^{p} \hat{\lambda}_{ij}^2 \right)^2}{p}, \tag{6.25}$$

con $\gamma = \beta/(\alpha + \beta)$.

Varios son los tipos de rotación que resultan de acuerdo con los valores de γ. Así, para $\gamma = 1$ ($\alpha = 0;\ \beta = 1$) la rotación es la tipo varimax; si $\gamma = 0$ ($\alpha = 1;\ \beta = 0$) la rotación corresponde a la cuartimax; si $\gamma = 1/2$ ($\alpha = 1;\ \beta = 1$) la rotación es la *bicuartimax*, y finalmemte, si $\gamma = k/2$ la rotación es del tipo *equimax*.

6.6.2 Rotación oblicua

Hasta ahora se ha presentado la rotación de los ejes factoriales conservando la perpendicularidad entre estos (no correlación o $\text{Cov}(\mathfrak{f}) = \boldsymbol{\Phi} = \boldsymbol{I}$). En algunos campos de investigación, como las ciencias sociales, los investigadores son renuentes a considerar la independencia entre factores (amparados por su marco conceptual), razón por la que permiten alguna correlación menor entre los factores. Con estas premisas se justifica la realización de una rotación[4] (*oblicua*) de los ejes factoriales. Se presenta una explicación de esta técnica, desde la óptica geométrica; para otros procedimientos más de tipo analítico se puede consultar a Gorsuch (1983, págs. 188-197) y Rencher (1998, págs. 389-390).

[4]Es más conveniente el término *transformación* oblicua.

Entre los procedimientos disponibles para la rotación oblicua está la *rotación visual*, en la cual los factores son rotados hasta una posición en la que permiten apreciar una estructura simple del conjunto de datos. Mediante una rotación oblicua se trata de expresar cada variable en términos de un número mínimo de factores; preferiblemente uno solo.

Una vez que se han conseguido los nuevos ejes factoriales, el patrón y la estructura de las ponderaciones cambia. Para obtener los "nuevos" pesos, se proyectan los puntos (cada fila de la matriz $\hat{\mathbf{\Lambda}}$) sobre los ejes oblicuos. Los dos procedimientos siguientes se emplean con frecuencia.

- El primero consiste en hacer la proyección de cada punto sobre un eje, en una dirección paralela al otro eje (figura 6.3a). Estas proyecciones suministran la configuración de los "nuevos" pesos ($\hat{\lambda}^*$). El cuadrado de la proyección da la única contribución que el factor hace sobre la varianza de la respectiva variable.

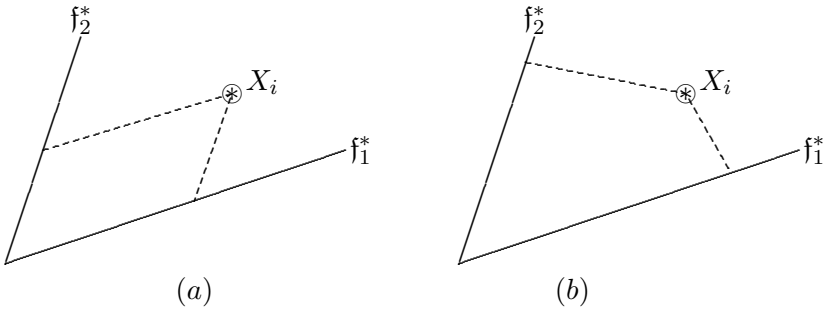


Figura 6.3 *Rotación oblícua de factores.*

- Mediante el segundo procedimiento, las proyecciones de cada punto se hacen trazando líneas perpendiculares a los "nuevos" ejes (figura 6.3b). Estas proyecciones suministran la estructura de los "nuevos" pesos ($\hat{\lambda}^*$). La estructura de las ponderaciones corresponde a la correlación simple entre la variables y los factores. El cuadrado de la proyección de una variable para cualquier factor, mide la contribución en varianza para la variable conjuntamente por el efecto del respectivo factor y los efectos de interacción del factor con otros factores. De manera que, la estructura de estas últimas ponderaciones no es muy útil para interpretar la estructura factorial, se recomienda observar la configuración conseguida en las ponderaciones para hacer una lectura adecuada de los factores.

Puede ser que el juego de las palabras "configuración" y "estructura" resulten ambiguos, se debe destacar que en éste contexto no lo son. La primera hace referencia a las proyecciones con dirección paralela a un eje y la otra a la proyección perpendicular.

El paquete *SAS* mediante la opción *ROTATE*, del procedimiento *FACTOR*, ofrece algunos métodos para la rotación de los ejes factoriales.

Ejemplo 6.2 (continuación) Con respecto al ejemplo sobre preferencia presentado anteriormente (ejemplo 6.1), se hace la rotación de factores vía varimax (SAS/STAT, 1998). La tabla 6.1 resume las coordenadas de las cinco variables respecto a los dos primeros factores, sin rotación y con ella, junto con las comunalidades. De la figura 6.4 se puede observar como las variables X_2, X_4 y X_5 están altamente ligadas con el primer factor mientras que el segundo factor lo está con las variables X_1 y X_3. Se puede calificar al primer factor $\mathfrak{f}_1^*$ con el nombre de *factor nutricional* y al segundo $\mathfrak{f}_2^*$ con el nombre de *factor gustativo*. En resumen, estas personas prefieren el producto de acuerdo con sus características nutricionales y gustativas (en este orden).

Tabla 6.1 Puntajes pre y post rotación

Variables	Coordenadas (l_{ij})		Coordenadas al rotar		Comunalidades
	$\mathfrak{f}_1$	$\mathfrak{f}_2$	$\mathfrak{f}_1'$	$\mathfrak{f}_2'$	h_i^2
Gusto X_1	0.56	0.82	0.02	0.99	0.98
Costo X_2	0.78	-0.52	0.94	-0.1	0.88
Sabor X_3	0.65	0.75	0.13	0.98	0.98
Tamaño X_4	0.94	-0.11	0.84	0.43	0.89
Calorias X_5	0.80	-0.54	0.97	-0.02	0.93

6.7 ¿Son apropiados los datos para un análisis de factores?

El análisis factorial tiene razón de ser cuando las variables están altamente correlacionadas; de otra manera, lo mismo que se muestra para componentes principales, la búsqueda de factores comunes no tendrá resultados satisfactorios. En esta dirección, la primera decisión que el usuario enfrenta es si los datos son o no apropiados para hacer sobre ellos un análisis factorial. La mayor parte de las medidas para este fin son de tipo heurístico o empíricas.

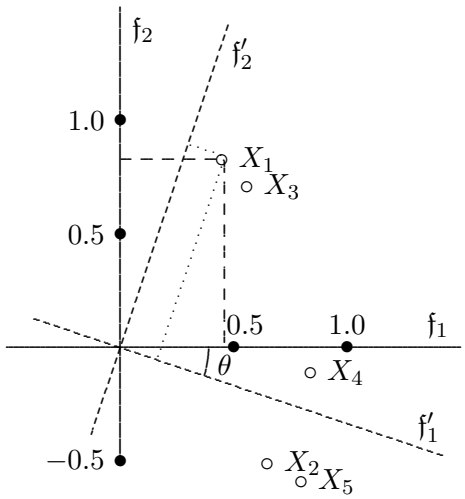


Figura 6.4 *Rotación de factores sobre preferencias.*

Una estrategia es el examen de la matriz de correlación, pues una correlación alta entre las variables indica que estas pueden ser agrupadas en conjuntos de variables. De manera que la búsqueda se dirige hacia aquellas características o atributos englobadas o agregadas en cada uno de estos conjuntos: a esto se le denomina *factores* o *variables latentes*. Una correlación baja entre las variables indica que las variables no tienen mucho en común. En el caso de disponer de un número grande de variables, la apreciación visual de la matriz de covarianzas puede tornarse pesada o de difícil lectura.

Por lo anterior, una primera inspección es sobre el determinante de la matriz de covarianzas; un valor bajo de éste señala baja correlación lineal entre las variables, pero no debe ser cero (matriz singular), caso en el cual se presentan algunas indeterminaciones en los cálculos, pues esto es un indicador de que algunas variables son linealmente dependientes. Pueden emplearse las técnicas que se tratan en el capítulo 4 (sección (4.3.1)) para hacer verificación de hipótesis acerca de la matriz de covarianzas, por ejemplo, verificar que la matriz de covarianzas es de la forma $\sigma^2 \boldsymbol{I}$, es decir una prueba de esfericidad.

Otra estrategia es observar la matriz de correlación parcial, donde para cada caso se controlan las demás variables. Se esperan correlaciones pequeñas para indicar que es adecuada una descomposición de la matriz de corre-

lación. El inconveniente aquí, como siempre en la toma de decisiones estadísticas, es el criterio para juzgar cuando una correlación es "pequeña".

Una tercera herramienta, consiste en examinar la *medida de adecuación de la muestra de Kaiser*, llamada de *Kaiser-Meyer-Olkin,* (KMO). Este indicador mide la adecuación de un conjunto de datos para el desarrollo de un análisis factorial sobre ellos. Se trata de una medida de la homogeneidad de las variables, (Kaiser, 1970). La fórmula para el cálculo del KMO es la siguiente:

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2} \tag{6.26}$$

donde r_{ij}^2 es el coeficiente de correlación simple entre las variables X_i y X_j, y, a_{ij}^2 es el coeficiente de correlación parcial entre las mismas variables X_i y X_j.

Este índice compara las magnitudes de los coeficientes de correlación simple r_{ij} con los coeficientes de correlación parcial a_{ij} observados. Si la suma de los cuadrados de los coeficientes de correlación parcial entre todos los pares de variables es pequeña en comparación con la suma de los cuadrados de los coeficientes de correlación, su valor es próximo a 1.0. Valores pequeños sugieren que el análisis factorial podría no ser conveniente, ya que las correlaciones entre pares de variables no pueden ser explicadas por las demás variables. Aunque no existe una estadística con la que se pueda probar la significancia de esta medida, en el siguiente cuadro se muestran algunos diagnósticos de adecuación de los datos, de acuerdo con el valor de la estadística de KMO.

Medida de KMO	Recomendación
≥ 0.90	Excelente
0.80^+	Meritorio
0.70^+	Bueno
0.60^+	Medio Bueno
0.50^+	Regular
< 0.50	No procedente

Obviamente es deseable, para los propósitos del análisis factorial, tener un valor alto de KMO. Se sugiere una medida mayor o igual que 0.80; aunque, una medida por encima de 0.60 es tolerable.

6.8 Componentes principales y análisis factorial

La semejanza de estos dos métodos está en que intentan explicar un conjunto de datos mediante un conjunto de variables, en un número menor que el inicial; es decir, ambas son técnicas de reducción de variables. De cualquier modo, existen algunas diferencias entre las dos metodologías, las cuales se resaltan a continuación.

1. El ACP es tan sólo una transformación de los datos. No se hace supuesto alguno sobre la forma de la matriz de covarianzas asociada a los datos. En cambio, el análisis factorial supone que los datos proceden de un modelo, como el definido en el modelo (6.1), con los supuestos considerados en las ecuaciones que se muestran en (6.2).

2. El análisis de componentes principales hace énfasis en explicar la varianza de los datos, mientras que el objetivo del análisis factorial es explicar la correlación entre variables.

3. En análisis de componentes principales las "nuevas" variables forman un índice. En el análisis factorial las "nuevas" variables son indicadores que reflejan la presencia de un atributo no manifiesto u observable (variable latente).

4. El ACP hace énfasis sobre la transformación de los valores observados a los componentes principales $Y = \mathbf{\Gamma} X$, mientras que el análisis factorial atiende a una transformación de los *factores comunes* ∫ a las variables observadas. Aunque, si la transformación por componentes principales es invertible, y se ha decidido mantener los k primeros componentes, entonces X puede aproximarse por tales componentes; es decir,
$$X = \mathbf{\Gamma} Y = \mathbf{\Gamma}_1 Y_1 + \mathbf{\Gamma}_2 Y_2 \doteq \mathbf{\Gamma}_1 Y_1,$$
ésta es una representación más elaborada que natural, pues se asume que las especifidades son nulas.

5. En el ACP se considera la variación total contenida en las variables, en tanto que en el análisis factorial la atención se dirige a la parte del total de varianza que es compartida por las variables.

6. En el ACP las variables se "agregan" adecuadamente para definir nuevas variables, mientras que en el análisis factorial, las variables se "desagregan" convenientemente en una serie de factores comunes desconocidos y en una parte propia de cada variable.

Éstas y otras diferencias serán más evidentes, en la medida que el investigador o usuario, las utilice conjugando los presupuestos estadísticos de la técnica y el marco conceptual donde se apliquen.

6.9 Rutina SAS para el cálculo de factores

Se muestra la sintaxis global del procedimiento FACTOR del paquete SAS. Se puede hacer cálculos a partir de la matriz de covarianzas o a partir de la matriz de correlación. Los métodos de estimación son algunos de los descritos aquí. El procedimiento tiene algunas opciones para la rotación de factores.

```
TITLE1 'Análisis factorial mediante el método';
TITLE2 'de la componente principal para el ejemplo 6.1 ';
DATA EJEMP6\_1 (TYPE=CORR); /* la declaración TYPE permite ingresar */
                           /* la matriz de correlación */
    _TYPE\_='CORR'; /* declaración obligatoria si se usa TYPE=CORR */
    INPUT X1 X2 X3 X4 X5; /* para declarar las variables */
CARDS; /* ingresar la matriz de correlación */
    1.00     0.02     0.96     0.42     0.01
    0.02     1.00     0.13     0.71     0.85
    0.96     0.13     1.00     0.50     0.11
    0.42     0.71     0.50     1.00     0.79
    0.01     0.85     0.11     0.79     1.00
;
PROC FACTOR
  METHOD=PRIN /* estimación mediante el método de la componente principal */
  CORR /* la estimación se hace desde la matriz de correlación */
  ROTATE=VARIMAX /* se hace rotación tipo VARIMAX */
  PREPLOT /* gráfica los factores antes de la rotación */
  PLOT    /* gráfica los factores después de la rotación */
  SCORE  /* imprime los  coeficientes de los factores   */
  SCREE /* imprime una gráfica descendente de los valores propios */
SIMPLE; /* imprime medias y desviaciones estándar; final de la instrucción */
VAR X1 X2 X3 X4 X5; /* variables a emplear para el análisis */
RUN;
```

6.10 Procesamiento de datos con R

Se realiza el ejemplo 6.1 usando el método de la componente principal y el método de máxima verosimilitud.

```
 # Matriz R del ejemplo 6.1
 R<-matrix(c(  1, .02, .96, .42, .01,
             .02,   1, .13, .71, .85,
```

```
            .96, .13,   1,  .5, .11,
            .42, .71, .50,   1, .79,
            .01, .85, .11, .79,   1  ),nrow=5)

# Valores y vectores propios
eig<-eigen(R)
# Matriz P (vectores propios)
P<-eig$vectors
P
# valores propios
eig$values
# ponderaciones factoriales
f<-P[,1:2]%*%D
f
# Comunalidad
comun<-matrix(rowSums(f^2))
comun
# varianza específica
1-comun

# Rotación varimax
varimax(f)

# Para realizar la rotación cuartimax se requiere instalar la
# librería GPArotation
library(GPArotation)
TV <- GPFoblq(f, method="quartimax", normalize=TRUE)
print(TV)
summary(TV)

#La libreria GPArotation realiza varios tipos de rotación
#para otros métodos pida ayuda mediante el comando
#help('GPFoblq')
#Análisis de factores usando lo función factanal()
#tenga en cuenta que esta función realiza el análisis de
#factores usando el método de maxima verosimilitud.

fac<-factanal(factors=2, covmat=R,rotation="none")
# los pesos se obtienen así:
fac$loadings
# Varianza específica
fac$uniquenesses
# Comunalidades
1-fac$uniquenesses
# Matriz de correlaciones usada
fac$correlation
# Análisis de factores con rotación varimax
fac<-factanal(factors=2, covmat=R,rotation="varimax")
# Como "varimax" es la opción por defecto para rotation,
# el siguiente comando produce el mismo resultado.
fac<-factanal(factors=2, covmat=R)
```

Capítulo 7

Análisis de conglomerados

7.1 Introducción

Conglomerado es un conjunto de objetos que poseen caracterí sticas similares. La palabra conglomerado es la traducción más cercana al término *"cluster"*, otros sinónimos son clases o grupos; incluso es muy frecuente el empleo directo de la palabra cluster. En la terminología del análisis de mercados se dice *segmento*, para denotar un grupo con determinado perfil; en biología se habla de familia o grupo para hacer referencia a un conjunto de plantas o animales que tienen ciertas características en común; en ciencias sociales se consideran *estratos* a los grupos humanos de condiciones socioeconómicas homogéneas. En este texto se usan los términos conglomerado, grupo y clase, indiferentemente, para aludir a un conjunto de objetos que comparten características comunes.

El análisis de conglomerados busca particionar un conjunto de objetos en grupos, de tal forma que los objetos de un mismo grupo sean similares y los objetos de grupos diferentes sean disímiles. Así, el análisis de conglomerados tiene como objetivo principal definir la estructura de los datos colocando las observaciones más parecidas en grupos.

Los propósitos más frecuentes para la construcción y análisis de conglomerados son los siguientes:

(i) La identificación de una estructura natural en los objetos; es decir, el desarrollo de una tipología o clasificación de los objetos.

(ii) La búsqueda de esquemas conceptuales útiles que expliquen el agrupamiento de algunos objetos.

(*iii*) La formulación de hipótesis mediante la descripción y exploración de los grupos conformados.

(*iv̈*) La verificación de hipótesis, o la confirmación de si estructuras definidas mediante otros procedimientos están realmente en los datos.

Los siguientes casos ejemplifican y motivan la utilidad y la necesidad del análisis de conglomerados.

- Un psicólogo clínico emplea una muestra de un determinado número de pacientes alcohólicos admitidos a un programa de rehabilitación, con el fin de construir una clasificación. Los datos generados sobre estos pacientes se obtienen a través de una prueba. La prueba contiene 566 preguntas de respuestas dicotómicas, las cuales se estandarizan y resumen en 13 escalas que dan un diagnóstico. Mediante una medida de similitud y la consideración de homogeneidad dentro y entre grupos, se conformaron cuatro grupos de alcohólicos: (1) emocionalmente inestables de personalidad, (2) psiconeuróticos con ansiedad-depresión,(3) de personalidad psicópata (4) alcohólico con abuso de drogas y características paranoicas.

- En taxonomía vegetal, el análisis de conglomerados se usa para identificar especies con base en algunas características morfológicas, fisiológicas, químicas, etológicas, ecológicas, geográficas y genéticas. Con esta información se encuentran algunos conglomerados de plantas, dentro de los cuales se comparten las características ya indicadas.

- El análisis de conglomerados puede emplearse con propósitos de muestreo. Así por ejemplo, un analista de mercados está interesado en probar las ventas de un producto nuevo en un alto número de ciudades, pero no dispone de los recursos ni del tiempo suficientes para observarlos todos. Si las ciudades pueden agruparse en conglomerados, un miembro de cada grupo podria usarse para la prueba de ventas; de otra parte, si se generan grupos no esperados esto puede sugerir alguna relación que deba investigarse.

Para alcanzar los propósitos anteriormente ilustrados se deben considerar los siguientes aspectos:

1. ¿Cómo se mide la similitud? Se requiere de un "dispositivo" que permita comparar los objetos en términos de las variables medidas sobre éllos. Tal dispositivo debe registrar la proximidad entre pares de objetos de tal forma que la distancia entre las observaciones (atributos del objeto) indique la similitud.

2. ¿Cómo se forman los conglomerados? Esta inquietud apunta a la arquitectura de los métodos; es decir, al procedimiento mediante el cual se agrupan las observaciones que son más similares dentro de un determinado conglomerado. Este procedimiento debe determinar la pertenencia al grupo de cada observación.

3. ¿Cuántos grupos se deben formar? Aunque se dispone de un amplio número de estrategias para decidir sobre el número de conglomerados a construir, el criterio decisivo es la homogeneidad "media" alcanzada dentro de los conglomerados. Una estructura simple debe corresponder a un número pequeño de conglomerados. No obstante, a medida que el número de conglomerados disminuye, la homogeneidad dentro de los conglomerados necesariamente disminuye. En consecuencia, se debe llegar a un punto de equilibrio entre el número de conglomerados y la homogeneidad de éstos. La comparación de las medias asociadas a los grupos o conglomerados construidos, desde el enfoque exploratorio, coadyuvan a la decisón acerca del número de éstos; el análisis discriminante es otra herramienta útil para tales propósitos. Aunque la decisión sobre el número de grupos a considerar, generalmente, es de la incumbencia del especialista asociado con el estudio en consideración.

Cualquiera que sea la estructura de clasificación conseguida, independientemente del método de clasificación seguido, no debe perderse de vista que se trata de un ejercicio exploratorio de los datos; de manera que se debe tener precaución con:

- la expansión o inferencia de resultados a la población a partir de la clasificación conseguida;

- la perpetuación o estatización, en el tiempo, en el espacio o en la población, de los grupos o clases conformados con una determinada metodología y sobre unos datos particulares.

Esta observación es pertinente, pues a pesar de que se garantice la calidad muestral (en términos de representatividad y probabilidad) de la información, los resultados descansan sobre los datos que participan en la conformación de los grupos. Aunque esta observación cabe para la mayoría de los procedimientos estadísticos, en el campo de la clasificación se ha visto que incluso una observación puede cambiar la estructura conseguida por las demás.

La técnica del análisis de conglomerados es otra técnica de reducción de datos. Se puede considerar la metodología de las componentes principales (capitulo 5) como un análisis de conglomerados, donde los objetos corresponden a las variables. Dos son los elementos requeridos en el análisis de conglomerados, el primero es la *medida* que señale el grado de *similitud* entre los objetos, el segundo es el *procedimiento* para la *formación* de los grupos o conglomerados.

7.2 Medidas de similitud

Reconocer objetos como similares o disimiles es fundamental para el proceso de clasificación. Aparte de su simplicidad, el concepto de *similitud* para aspectos cuantitativos se presenta ligado al concepto de métrica. Las medidas de similitud se pueden clasificar en dos tipos; en una parte están las que reunen las propiedades de *métrica*, como la distancia; en otra, se pueden ubicar los coeficientes de asociación, estos últimos empleados para datos en escala nominal.

Una métrica $d(;)$ es una función (o regla) que asigna un número a cada par de objetos de un conjunto $\mathbf{\Omega}$, es decir,

$$\begin{aligned} \mathbf{\Omega} \times \mathbf{\Omega} : &\xrightarrow{\quad d \quad} \mathbb{R} \\ (x, y) &\longrightarrow d(x, y), \end{aligned}$$

la cual satisface, sobre los objetos x, y y z de $\mathbf{\Omega}$, las siguientes condiciones:

1. *No negatividad.* $d(x, y) \geq 0$, y $d(x, y) = 0$, si y sólo si, $x = y$.

2. *Simetría.* Dados dos objetos x y y, la distancia, d, entre ellos satisface

$$d(x, y) = d(y, x).$$

3. *Desigualdad triangular.* Para tres objetos x, y y z las distancias entre ellos satisfacen la expresión

$$d(x, y) \leq d(x, z) + d(z, y).$$

   Esto, simplemente, quiere decir que la longitud de uno de los lados de un triángulo es menor o igual que la suma de las longitudes de los otros dos lados.

4. *Identificación de no Identidad.* Dados los objetos x y y

$$\text{si } d(x, y) \neq 0, \text{ entonces } x \neq y.$$

5. *Identidad.* Para dos elementos idénticos, x y x', se tiene que

$$d(x, x') = 0;$$

es decir, si los objetos son idénticos, la distancia entre ellos es cero.

Observación

Hay medidas que a cambio de la desigualdad triangular, propiedad (3), satisfacen

$$d(x, y) \leq \max\{d(x, z), d(z, y)\}, \text{ para todo } x, y \text{ y } z$$

a este tipo de distancia se le denomina *ultramétrica.* Esta distancia juega un papel importante en los métodos de clasificación automática.

Las medidas de similitud, de aplicación más frecuente, son las siguientes:

(1) Medidas de distancia.

(2) Coeficientes de correlación.

(3) Coeficientes de asociación.

(4) Medidas probabilísticas de similitud.

Antes de utilizar alguna de las medidas anteriores, se debe encontrar el conjunto de variables que mejor represente el concepto de similitud, bajo el estudio a desarrollar. Idealmente, las variables deben escogerse dentro del marco conceptual que explícitamente se usa para la clasificación. La teoría en cada campo, es la base racional para la selección de las variables a usar en el estudio.

La importancia de usar la teoría para la selección de las variables no debe subestimarse, pues resulta muy peligroso caer en un "empirismo ingenuo", por la facilidad con que los algoritmos nos forman grupos sin importar el número y el tipo de variables; ya que por la naturaleza heurística de las técnicas de agrupamiento se ha contaminado un poco su aplicación. Para la aplicación de esta técnica también se debe considerar la necesidad de estandarizar las variables, su transformación, o la asignación de un peso o ponderación para el cálculo de la medida de similitud y la conformación de los conglomerados (Aldenderfer y Blashfield 1984).

7.2.1 Medidas de distancia

En la sección (1.4.6) se presentaron algunas de estas medidas, las de uso más frecuente son:

- La distancia *euclidiana*, definida por

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2}.$$

- La distancia D^2 de *Mahalanobis*, también llamada la distancia generalizada

$$D^2 = d_{ij} = (X_i - X_j)'\mathbf{\Sigma}^{-1}(X_i - X_j)$$

  donde $\mathbf{\Sigma}$ es la matriz de varianzas y covarianzas de los datos, y X_i y X_j son los vectores de las mediciones que identifican los dos objetos i y j.

- Otra medida muy común es la de *Manhattan*, se define

$$d_{ij} = \sum_{k=1}^{p}|X_{ik} - X_{jk}|.$$

- Finalmente la medida de *Minkowski*

$$d_{ij} = \left(\sum_{k=1}^{p}|X_{ik} - X_{jk}|^r\right)^{1/r} \quad \text{con } r = 1, 2, \ldots$$

Ejemplo 7.1 Supóngase que se tienen cuatro personas cuya edad X_1 (en años), estatura X_2 (en metros), peso X_3 (en kilogramos) son los siguientes

Persona	Edad	Estatura	Peso
A	23	1.69	61
B	40	1.70	72
C	26	1.65	68
D	38	1.68	70

El vector de medias $\bar{\boldsymbol{X}}$, la matriz de covarianzas $\boldsymbol{S}$ y la matriz de correlación $\boldsymbol{R}$, manteniendo el orden de escritura anterior, son

$$\bar{\boldsymbol{X}} = \begin{pmatrix} 31.75 \\ 1.68 \\ 67.76 \end{pmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} 72.2500 & 0.0833 & 35.5833 \\ 0.0833 & 0.0004 & 0.0033 \\ 35.5833 & 0.0033 & 22.9166 \end{pmatrix}$$

y

$$\boldsymbol{R} = \begin{pmatrix} 1.0000 & 0.4538 & 0.8744 \\ 0.4538 & 1.0000 & 0.0322 \\ 0.8744 & 0.0322 & 1.0000 \end{pmatrix}.$$

La matriz de distancias euclidianas es

	A	B	C	D
A	0	20.25	7.62	17.49
B	20.25	0	14.56	2.83
C	7.62	14.56	0	12.17
D	17.49	2.83	12.17	0

donde la distancia entre A y B, por ejemplo, resulta del siguiente cálculo

$$d_{AB} = \sqrt{(23-40)^2 + (1.69-1.70)^2 + (61-72)^2} = 20.25$$

Se puede notar que los individuos más similares o cercanos son B y D, hecho que resalta fácilmente de los datos. Uno de los problemas de esta distancia es su sensibilidad a cambios de escala, dificultad que se supera mediante la distancia de Mahalanobis, la cual toma la distancia entre las variables estandarizadas; es decir, les "quita" el efecto de "la escala de medición" para calcular su similitud. La siguiente matriz resume las distancias de Mahalanobis entre las personas

	A	B	C	D
A	0	7.21	6.36	10.01
B	7.21	0	8.89	15.62
C	6.36	8.89	0	7.96
D	10.01	15.62	7.96	0

Un resultado, aparentemente curioso, es la distancia entre B y D, mientras que con la distancia euclidiana B y D son los más cercanos, con la distancia de Mahalanobis resultan ser los más lejanos, una situación inversa se tiene entre los objetos A y B con las dos distancias. Para una explicación a este resultado obsérvense las varianzas y las correlaciones entre las variables.

7.2.2 Coeficientes de correlación

Frecuentemente se les llama medidas angulares, por su interpretación geométrica. El más popular de ellos es el coeficiente producto momento de

Pearson, el cual determina el grado de correlación o asociación lineal entre casos. Está definido por:

$$r_{jk} = \frac{\sum_i (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{\sqrt{\sum_i (X_{ij} - \bar{X}_j)^2} \sqrt{\sum_i (X_{ik} - \bar{X}_k)^2}}, \quad \text{con } i = 1, \ldots, p$$

donde X_{ij} es el valor de la variable i para el caso j (objeto), y $\bar{X}_j$ es la media de todas las variables que definen el caso j. Esta medida se emplea para variables en escala al menos de intervalo; para el caso de variables binarias, éstas se transforman al conocido coeficiente φ. El coeficiente toma valores entre 1 y -1, un valor de cero significa no similitud entre los casos. Frecuentemente se le considera como una medida de *forma*, la cual es insensible a las diferencias en magnitud de las variables que intervienen en su cálculo.

El coeficiente de producto momento es sensible a la forma, esto significa que dos perfiles pueden tener correlación de +1.0, y no ser idénticos; gráficamente corresponde a líneas poligonales paralelas con alturas diferentes. La figura 7.1 muestra, un caso idealizado, de dos perfiles con base en seis variables con coeficiente de correlación $r = 1.0$. Sobre el eje horizontal se han ubicado las variables (el orden no es importante) y en el eje vertical se representan sus respetivos valores.

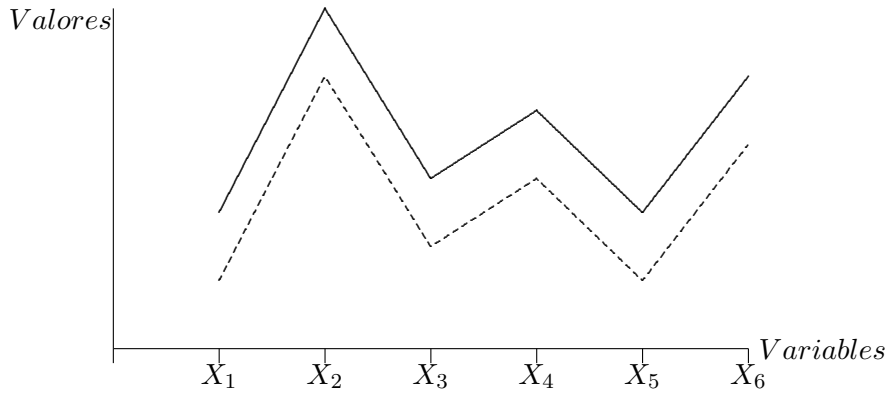


Figura 7.1 *Perfiles con coeficiente de correlación* $r = 1.0$.

Una limitación del coeficiente de correlación es que no siempre satisface la desigualdad triangular, y esto puede limitar la comparación entre perfiles. Otra limitación es su cálculo, pues debe obtenerse la media a través de diferentes tipos variables, y no a través de casos como corresponde a su definición estadística; de cualquier modo, el coeficiente demuestra ser bueno

frente a otros coeficientes de similitud en el análisis de conglomerados, por cuanto reduce el número de clasificaciones incorrectas.

7.2.3 Coeficientes de asociación

Son apropiados cuando los datos están en escala nominal. Cada variable toma los valores de 0 (de ausencia) y 1 (de presencia) de un atributo; una tabla de doble entrada resume toda la información (a manera de una matriz de diseño). Por ejemplo, la siguiente tabla contiene la información de dos OTU (Operational Taxonomic Unit) A y B con relación a 10 caracteres del tipo presencia/ausencia,

OTU	1	2	3	4	5	6	7	8	9	10
A	0	1	1	0	1	0	1	0	1	0
B	1	1	0	0	1	1	0	0	1	1

Al comparar estos dos objetos se tienen cuatro posibilidades (Crisci y López, 1983 págs. 42-49):

1. Que ambos tengan presente el carácter comparado $(1, 1)$.
2. Que ambos tengan ausente el carácter comparado $(0, 0)$.
3. Que el primero tenga el carácter presente y el segundo ausente $(1, 0)$.
4. Que el primero de ellos tenga el carácter ausente y el segundo presente $(0, 1)$.

La frecuencia con que se presentan estas cuatro características se resume en la siguiente tabla

	Objeto B	
Objeto A	1	0
1	(a)	(b)
0	(c)	(d)

El valor (a) es el número de atributos en los cuales el mismo estado es poseido por los dos objetos, (b) es la frecuencia de caracteres en los cuales el primer objeto lo posee y el segundo no, (c) es el número de caracteres en los que un estado está ausente en el primer objeto pero no en el segundo y

(d) es el número de caracteres en los cuales el mismo estado está ausente en ambos objetos.

Para el ejemplo de las OTU, la tabla de comparación de un mismo carácter es

	Objeto B	
Objeto A	1	0
1	(3)	(2)
0	(3)	(2)

- *Coeficiente de asociación simple* ($\mathcal{S}$): Es la medida de similitud más sencilla, entre los objetos i y j, se calcula mediante la siguiente fórmula
$$\mathcal{S}_{(i,j)} = \frac{a+d}{a+b+c+d}$$
sus valores están entre 0 y 1. Este coeficiente toma en cuenta la ausencia de una variable para los dos objetos en consideración.

- El coeficiente de *Jaccard* ($\mathcal{J}$), definido como
$$J_{(i,j)} = \frac{a}{a+b+c},$$
resuelve el problema de las ausencias conjuntas de una variable en el cálculo de la similaridad. Los biólogos anotan que con el empleo del coeficiente de asociación simple, algunos casos aparecerán como muy similares por el hecho de no poseer algún atributo en común; es algo así como decir, que una guayaba se parece a una naranja porque con ninguna de las dos se puede hacer jugo de mango.

- *Rogers y Tanimoto* ($\mathcal{RT}$): le da prelación a las diferencias, como en el caso de los dos anteriores coeficientes donde sus valores oscilan entre 0 y 1; es decir, valores de mínima y máxima similitud, respectivamente. Su cálculo se hace mediante la siguiente expresión:
$$RT_{(i,j)} = \frac{a+d}{a+(2b)+(2c)+d}.$$

- *Sϕrensen* o *Dice* ($\mathcal{SD}$): este coeficiente le confiere mayor importancia a las coincidencias en estado de presencia, se expresa como
$$SD_{(i,j)} = \frac{2a}{2a+b+c}.$$

Los valores de este coeficiente varían entre 0 y 1; y representan valores de mínima y máxima similitud, respectivamente.

- *Sokal y Sneath* ($\mathcal{SS}$): éste tiene más en cuenta las coincidencias, tanto por presencia como por ausencia de los atributos. Sus valores se obtienen calculando

$$SS_{(i,j)} = \frac{2(a+d)}{2(a+d)+b+c},$$

y toma valores entre 0 y 1 que equivalen a la mínima y máxima semejanza, respectivamente.

- *Hamann* ($\mathcal{H}$): considera importante las diferencias entre coincidencias y no coincidencias. Los valores de similitud están en el rango de -1 a 1, mínima y máxima similitud, respectivamente. Se expresa así

$$H_{(i,j)} = \frac{(a+d)-(b+c)}{a+b+c+d}$$

Aunque en la literatura de taxonomía numérica se encuentran otros coeficientes, con los anteriores se brinda la idea general de esta estrategia para medir similitud entre objetos.

Los valores de cada uno de estos coeficientes, para el ejemplo de las OTU, son los siguientes:

$$S_{(A,B)} = \frac{3+2}{3+2+3+2} = 0.5$$
$$J_{(A,B)} = \frac{3}{3+2+3} = 0.375$$
$$RT_{(A,B)} = \frac{3+2}{3+(2\times 2)+(2\times 3)+2} = 0.33$$

$$SD_{(A,B)} = \frac{2\times 3}{(2\times 3)+2+3} = 0.54$$
$$SS_{(A,B)} = \frac{2(3+2)}{2(3+2)+2+3} = 0.67$$
$$H_{(A,B)} = \frac{(3+2)-(2+3)}{3+2+3+2} = 0.$$

Una objeción que se le puede hacer a los coeficientes de asociación, es su aplicación solo a respuestas dicotómicas; aunque, los datos continuos se pueden transformar a valores de tipo 0 y 1, el problema se reduce a decidir a que valores se les asigna como 0 y a cuales como 1, esta transformación

hace que se pierda información; pues no tiene en cuenta la intensidad de los atributos.

Ejemplo 7.2

Otro caso semejante al de las OTU se puede construir para comparar viviendas Supóngase que se observan las variables:

- *pisos acabados* (X_1),
- *servicio de teléfono* (X_2),
- *servicio agua y luz* (X_3),
- *paredes en ladrillo y acabadas* (X_4),
- *cuatro o más alcobas* (X_5),
- *área superior a* $70m^2$ (X_6),
- *tres o más personas por alcoba* (X_7) y
- *cuatro o más electrodomésticos diferentes* (X_8).

Los datos, tomados en la forma presencia/ausencia, sobre 6 viviendas escogidas aleatoriamente de 6 zonas diferentes, son los siguientes:

	Variables							
Zona	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
A	1	0	0	1	0	0	0	0
B	0	0	1	0	0	0	1	0
C	0	1	0	1	1	0	0	0
D	1	0	0	0	1	0	1	0
E	1	1	0	1	1	0	1	1
F	1	0	0	0	1	1	1	0

Las frecuencias de coincidencia entre la zona A y la zona B se muestran en la siguiente tabla:

	Zona B	
Zona A	1	0
1	(0)	(2)
0	(2)	(4)

Los coeficientes de asociación simple y Jaccard toman los valores

$$S_{(A,B)} = \frac{4}{8} = 0.500 \text{ y } J_{(A,B)} = \frac{0}{4} = 0.000,$$

respectivamente. Mientras que el coeficiente simple indica una buena similitud entre éstas dos viviendas, el coeficiente de Jaccard señala que la asociación es débil.

	A	B	C	D	E	F
A	1	0.000	0.250	0.250	0.333	0.200
B	0.000	1	0.000	0.250	0.143	0.200
C	0.250	0.000	1	0.200	0.500	0.167
D	0.250	0.250	0.200	1	0.500	0.750
E	0.333	0.143	0.500	0.500	1	0.429
F	0.200	0.200	0.167	0.750	0.429	1

Los coeficientes de Jaccard para las seis viviendas, están contenidos en la matriz que se muestra a continuación. Allí se sugiere que las viviendas D, E y F son bastante similares, y que las viviendas A y B son totalmente disímiles. La presencia de muchos "empates" en los pares de casos resulta ser un problema para la conformación de los conglomerados; para el presente ejemplo hay tres pares de casos con $J = 0.250$, tres pares con $J = 0.200$ y dos con $J = 0.500$.

7.2.4 Coeficientes de probabilidad

Son bastante diferentes a los anteriores, este tipo de medida trabaja directamente sobre los datos originales. Al construir conglomerados, se considera la ganancia de información al combinar dos casos ; se fusionan los dos casos que suministren la menor ganancia de información. Una limitación de estas medidas probabilísticas es su utilización únicamente para variables dicotómicas. Puesto que estos coeficientes son muy utilizados en taxonomía numérica, se sugiere consultar a Clifford-Stephenson (1975).

7.3 Una revisión de los métodos de agrupamiento

Aunque no hay una definición universal de *conglomerado*, se toma la definición dada por Everitt (1980), quien dice que los conglomerados son *"regiones continuas de un espacio que contienen una densidad relativamente alta de*

puntos, las cuales están separadas por regiones que contienen una densidad relativamente baja de puntos".

Varios son los algoritmos propuestos para la conformación de conglomerados, se desarrollan, de una manera muy esquemática los métodos *jerárquicos*, los métodos de *partición o división*, *nubes dinámicas*, *clasificación difusa* y algunas herramientas *gráficas*. Cada uno de estos métodos representa una perspectiva diferente para la formación de los conglomerados, con resultados generalmente distintos cuando las diferentes metodologías se aplican sobre el mismo conjunto de datos. Para obviar en parte esta dificultad, se debe emplear un procedimiento concordante con la naturaleza de la tipología esperada, con las variables a considerar y la medida de similitud usada.

7.3.1 Métodos jerárquicos

Estos métodos empiezan con el cálculo de la matriz de distancias entre los objetos. Se forman grupos de manera *aglomerativa* o por un proceso de *división*. Una de las características de esta técnica es la localización irremovible de cada uno de los objetos en cada etapa del mismo. Con los procedimientos aglomerativos cada uno de los objetos empieza formando un conglomerado (grupos unitarios). Grupos cercanos se mezclan sucesivamente hasta que todos los objetos quedan dentro de un mismo conglomerado. Los métodos de división inician con todos los objetos dentro de un mismo conglomerado, éste es dividido luego en dos grupos, éstos en otros dos hasta que cada objeto llega a ser un conglomerado. Ambos procedimientos se resumen en un diagrama de árbol que ilustra la conformación de los distintos grupos, de acuerdo con el estado, de fusión o división, jerárquico implicado por la matriz de similaridades; este diagrama se conoce con el nombre de *dendrograma*[1]. Por su amplia aplicación, se explican solo los métodos aglomerativos; los procedimientos de división pueden consultarse en Dillon y Goldstein (1984, pag. 178), Krzanowski y Marriott (1995, págs. 61-94).

▶ Métodos aglomerativos

Son los más frecuentemente utilizados. Una primera característica de estos métodos es que buscan una matriz de similaridades de tamaño $(n \times n)$, (n número de objetos), desde la cual, secuencialmente, se mezclan los *casos más cercanos*; aunque cada uno tiene su propia forma de medir las distancias entre grupos o clases. Un segundo aspecto es que cada paso o etapa en la conformación de grupos puede representarse visualmente por

[1] Del griego *dendron*, que significa árbol.

un dendrograma. En tercer lugar, se requieren $(n-1)$ pasos para la conformación de los conglomerados de acuerdo con la matriz de similaridades. En el primer paso cada objeto es tratado como un grupo; es decir, se inicia con n conglomerados, y, en el paso final, se tienen todos los objetos en un solo conglomerado. Finalmente, los métodos jerárquicos aglomerativos son conceptualmente simples.

Aparte de las características y bondades anotadas, estos métodos adolecen de algunas fallas; por una parte, los cálculos requeridos en los algoritmos son muy numerosos, aunque aritméticamente simples, por ejemplo con 500 casos se requieren cerca de 125.000 valores en la matriz de similaridades, situación que demanda el uso de una buena máquina de cómputo; otra falla es que pasan sólo una vez a través de los datos; así, una partición pobre de los datos es irreversible en las etapas posteriores. A excepción del método de asociación simple, los demás métodos tienen el inconveniente de que generan diferentes soluciones al reordenar los datos en la matriz de similitud; por último, estos métodos son muy inestables cuando se extraen casos del análisis; en consecuencia son bastante sensibles a la presencia de observaciones atípicas.

◦ *Enlace simple o del "vecino más cercano"*

Después de iniciar con tantos grupos como objetos haya disponibles, se juntan los dos casos que estén a la menor distancia o dentro de un límite de similitud dispuesto. Ellos conforman el primer conglomerado. En la siguiente etapa puede ocurrir que un tercer objeto se junte a los dos ya conformados o que se una con otro más cercano a él, para formar un segundo conglomerado. La decisión se basa en establecer si la distancia entre el tercer objeto y el primer conglomerado es menor a la distancia entre éste y otro de los no agrupados. El proceso se desarrolla hasta que todos los objetos queden dentro de un mismo conglomerado. La distancia entre el conglomerado $\mathcal{A}$ y el conglomerado $\mathcal{B}$ se define mediante

$$d_{\mathcal{AB}} = \min_{\substack{i\in\mathcal{A}\\ j\in\mathcal{B}}}\{d_{ij}\}. \tag{7.1}$$

Así, la distancia entre dos conglomerados cualesquiera es la *menor distancia* observada desde un punto de un conglomerado a un punto del otro conglomerado.

Para ilustrar este procedimiento de agrupación, supóngase que cinco objetos se encuentran a las siguientes distancias.

	O_1	O_2	O_3	O_4	O_5
O_1	0	3	7	11	10
O_2	3	0	6	10	9
O_3	7	6	0	5	6
O_4	11	10	5	0	4
O_5	10	9	6	4	0

A una distancia cero, los cinco objetos conforman cada uno un grupo. La distancia más pequeña, de acuerdo con la matriz anterior, es 3, que corresponde entre O_1 y O_2. Así, a esta distancia se tienen cuatro grupos $\{O_1, O_2\}$, $\{O_3\}$, $\{O_4\}$ y $\{O_5\}$ Las distancias entre estos grupos se obtienen a través de (7.1); así, la distancia entre el conglomerado $\{O_1, O_2\}$ y los demás es

$$d_{\{O_1,O_2\}\{O_3\}} = \min\{d_{O_1O_3}, d_{O_2O_3}\} = \min\{7, 6\} = 6$$
$$d_{\{O_1,O_2\}\{O_4\}} = \min\{d_{O_1O_4}, d_{O_2O_4}\} = 10$$
$$d_{\{O_1,O_2\}\{O_5\}} = \min\{d_{O_1O_5}, d_{O_2O_5}\} = 9$$

Las distancias $d_{\{O_3\}\{O_4\}}$, $d_{\{O_3\}\{O_5\}}$ y $d_{\{O_4\}\{O_5\}}$, están contenidas en la matriz de distancias inicial. Así, la matriz de distancias entre los "nuevos" conglomerados, calculadas de acuerdo con la expresión(7.1), es

	$\{O_1, O_2\}$	$\{O_3\}$	$\{O_4\}$	$\{O_5\}$
$\{O_1, O_2\}$	0	6	10	9
$\{O_3\}$	6	0	5	6
$\{O_4\}$	10	5	0	4
$\{O_5\}$	9	6	4	0

De la matriz de distancias anterior, la siguiente distancia más pequeña es 4 y está entre los grupos $\{O_4\}$ y $\{O_5\}$; por tanto, a una distancia 4 se conforman los conglomerados: $\{O_1, O_2\}$, $\{O_3\}$ y $\{O_4, O_5\}$. La matriz de distancias entre éstos, calculadas mediante la fórmula (7.1), es

	$\{O_1, O_2\}$	$\{O_3\}$	$\{O_4, O_5\}$
$\{O_1, O_2\}$	0	6	9
$\{O_3\}$	6	0	5
$\{O_4, O_5\}$	9	5	0

La siguiente menor distancia es 5; corresponde a los grupos $\{O_3\}$ y $\{O_4, O_5\}$, la distancia entre éstos es: $\min\{d_{\{O_3\}\{O_4\}}, d_{\{O_3\}\{O_5\}}\} = 5$.

Quedan en esta etapa dos grupos $\{O_1, O_2\}$ y $\{O_3, O_4, O_5\}$. La matriz de distancias entre éstos, calculadas mediante la fórmula (7.1) es

	$\{O_1, O_2\}$	$\{O_3, O_4, O_5\}$
$\{O_1, O_2\}$	0	6
$\{O_3, O_4, O_5\}$	6	0

Por último, la siguiente distancia más pequeña es 6, corresponde a O_2 y O_3 y a O_3 y O_5. En este punto todos los objetos se pueden mezclar en el conglomerado $\{O_1, O_2, O_3, O_4, O_5\}$. La tabla siguiente resume el proceso.

Distancia	Conglomerado
0	$\{O_1\}, \{O_2\}, \{O_3\}, \{O_4\}, \{O_5\}$
3	$\{O_1, O_2\},\ \{O_3\},\ \{O_4\}, \{O_5\}$
4	$\{O_1, O_2\},\ \{O_3\},\ \{O_4, O_5\}$
5	$\{O_1, O_2\},\ \{O_3, O_4, O_5\}$
6	$\{O_1, O_2, O_3, O_4, O_5\}$

El dendrograma de la Figura 7.2 muestra la disposición de los objetos en cada uno de los conglomerados. El eje vertical contiene los niveles de distancia bajo los cuales se conforman los grupos; así, para una distancia de 4.5 se tienen tres grupos (bajo la línea punteada), estos son: $\{O_1, O_2\}$, $\{O_3\}$ y $\{O_4, O_5\}$.

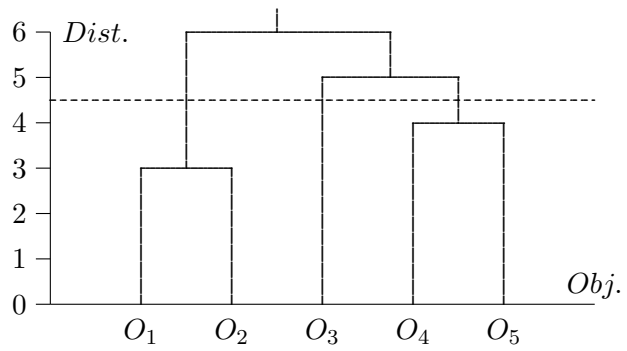


Figura 7.2 *Dendrograma: método del vecino más próximo.*

Las principales ventajas de este método son la invarianza respecto a transformaciones monótonas de la matriz de similaridades y su no afectación por la presencia de empates. La primera propiedad significa que la técnica no altera sus resultados cuando la transformación de los datos conserva el orden de los mismos.

◦ *Enlace completo o del "vecino más lejano"*

Este método es el opuesto lógico al de unión simple, la regla establece que cualquier candidato a incluirse en un grupo existente, debe estar dentro de un determinado nivel de similitud con todos los miembros de ese grupo; de otra manera, dos grupos son mezclados solo si los miembros más distantes de los dos grupos están suficientemente cerca de manera conjunta; el "suficientemente cerca" es dado por el nivel de similitud impuesto en cada etapa del algoritmo.

Para este procedimiento la distancia entre el conglomerado $\mathcal{A}$ y el conglomerado $\mathcal{B}$ está dado por

$$d_{\mathcal{AB}} = \max_{\substack{i\in\mathcal{A}\\ j\in\mathcal{B}}}\{d_{ij}\}. \tag{7.2}$$

En el ejemplo actual, en una primera etapa se fusionan los objetos O_1 y O_2 en un conglomerado. Las distancias entre los conglomerados resultantes se calculan a través de (7.2), por ejemplo las distancias entre el conglomerado $\{O_1, O_2\}$ y los demás son:

$$\begin{aligned} d_{\{O_1,O_2\}\{O_3\}} &= \max\{d_{O_1O_3}, d_{O_2O_3}\} = \max\{7,6\} = 7,\\ d_{\{O_1,O_2\}\{O_4\}} &= \max\{d_{O_1O_4}, d_{O_2O_4}\} = 11,\\ d_{\{O_1,O_2\}\{O_5\}} &= \max\{d_{O_1O_5}, d_{O_2O_5}\} = 10. \end{aligned}$$

La siguiente matriz contiene las distancias, tipo (7.2), entre los conglomerados obtenidos hasta ahora:

	$\{O_1, O_2\}$	$\{O_3\}$	$\{O_4\}$	$\{O_5\}$
$\{O_1, O_2\}$	0	7	11	10
$\{O_3\}$	7	0	10	6
$\{O_4\}$	11	10	0	4
$\{O_5\}$	10	6	4	0

En la matriz de distancias anterior, se observa que los objetos O_4 y O_5 pueden fusionarse, pues son los grupos más cercanos. La matriz de distan-

cias entre los conglomerados $\{O_1, O_2\}$, $\{O_3\}$ y $\{O_4, O_5\}$, aplicando nuevamente la expresión (7.2) es:

	$\{O_1, O_2\}$	$\{O_3\}$	$\{O_4, O_5\}$
$\{O_1, O_2\}$	0	7	11
$\{O_3\}$	7	0	6
$\{O_4, O_5\}$	11	6	0

El objeto O_3 se debe fusionar con el grupo constituido por los objetos O_4 y O_5, pues la distancia entre éste y los otros dos conglomerados, de acuerdo con las fórmula (7.2), es

$$d_{\{O_3\}\{O_1,O_2\}} = \max\{d_{O_3O_1}, d_{O_3O_2}\} = \max\{7, 6\} = 7,$$
$$d_{\{O_3\}\{O_4,O_5\}} = \max\{d_{O_3O_4}, d_{O_3O_5}\} = \max\{5, 6\} = 6.$$

Nótese que aunque O_3 dista de O_2 en 6 unidades, no está dentro de este nivel con O_1 (distan 7 unidades); es decir, no está conjuntamente cerca a este conglomerado. Hasta esta etapa se tienen los grupos o clases $\{O_1, O_2\}$, $\{O_3\}$ y $\{O_4, O_5\}$. En una última etapa los objetos conforman una sola clase.

La tabla siguiente muestra el algoritmo

Distancia	Conglomerado
0	$\{O_1\}, \{O_2\}, \{O_3\}, \{O_4\}, \{O_5\}$
3	$\{O_1, O_2\},\ \{O_3\},\ \{O_4\}, \{O_5\}$
4	$\{O_1, O_2\},\ \{O_3\},\ \{O_4, O_5\}$
5	$\{O_1, O_2\},\ \{O_3, O_4, O_5\}$
11	$\{O_1, O_2, O_3, O_4, O_5\}$

El respectivo dendrograma se exhibe en la figura 7.3. Es evidente que la determinación de los grupos en un nivel específico es ahora más clara que en el caso anterior. Se ilustran los conglomerados obtenidos al tomar una distancia de 5 y 7 unidades respectivamente.

∘ *Unión mediante el promedio*

Fue propuesto por Sokal y Michener (1958); es una salida a los extremos de los dos métodos anteriores. La distancia entre dos conglomerados $\mathcal{A}$ y $\mathcal{B}$ se define como el promedio de las distancias entre todos los pares de objetos,

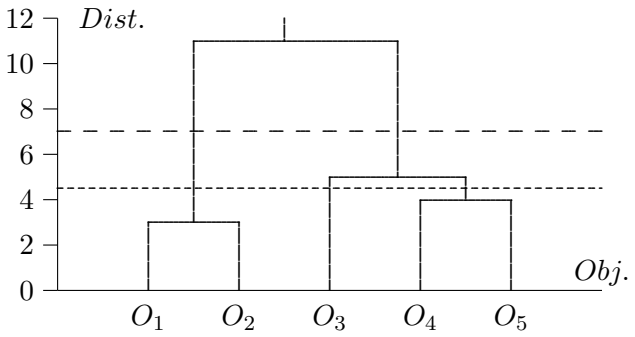


Figura 7.3 *Dendrograma: método del vecino más lejano.*

uno del conglomerado $\mathcal{A}$ y otro del conglomerado $\mathcal{B}$; es decir,

$$d_{\mathcal{AB}} = \frac{1}{n_{\mathcal{A}} n_{\mathcal{B}}} \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{B}} d_{ij}. \tag{7.3}$$

Se une el caso u objeto al conglomerado si se logra un determinado nivel de similitud con el valor promedio. El promedio más común es la media aritmética de las similaridades entre los objetos.

Con el ejemplo tratado, la tabla que resume el algoritmo y el dendrograma (figura 7.4) respectivo se presentan enseguida:

Distancia	Conglomerado
0	$\{O_1\}, \{O_2\}, \{O_3\}, \{O_4\}, \{O_5\}$
3	$\{O_1, O_2\},\ \{O_3\},\ \{O_4\}, \{O_5\}$
4	$\{O_1, O_2\},\ \{O_3\},\ \{O_4, O_5\}$
5.5	$\{O_1, O_2\},\ \{O_3, O_4, O_5\}$
8.8	$\{O_1, O_2, O_3, O_4, O_5\}$

Las distancias entre las clases $\{O_1, O_2\}$, $\{O_3\}$ y $\{O_4, O_5\}$ se calculan desde la expresión (7.3) como sigue:

$$d_{\{O_1,O_2\}\{O_3\}} = \frac{1}{2 \times 1}(d_{13} + d_{23}) = \frac{1}{2}(7 + 6) = 6.5.$$

$$d_{\{O_1,O_2\}\{O_4,O_5\}} = \frac{1}{2 \times 2}(d_{14} + d_{15} + d_{24} + d_{25}) = \frac{1}{4}(11 + 10 + 10 + 9) = 10.$$

$$d_{\{O_3\}\{O_4,O_5\}} = \frac{1}{1 \times 2}(d_{34} + d_{35}) = \frac{1}{2}(5 + 6) = 5.5.$$

Para el cuarto paso, por ejemplo, el caso O_3 está a una distancia en promedio del grupo $\{O_1, O_2\}$ de 6.5 y a 5.5 del grupo $\{O_4, O_5\}$, por eso se junta con éste último.

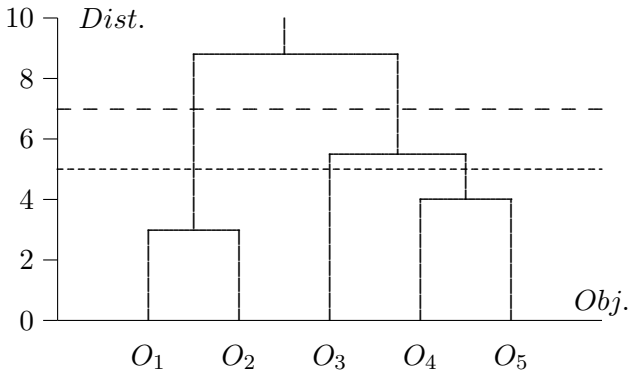


Figura 7.4 *Dendrograma: método del promedio.*

◦ *Método de Ward*

Con este método se busca la mínima variabilidad dentro de los conglomerados, se trata entonces de un problema de optimización. Ward (1963) basa su método sobre la pérdida de información resultante al agrupar casos en grupos, medida por la suma total del cuadrado de las desviaciones de cada caso al centroide del grupo al cual pertenece. La suma de cuadrados se calcula mediante

$$SCW = \frac{1}{(1/n_h + 1/n_k)} \|\bar{X}_h - \bar{X}_k\|^2, \tag{7.4}$$

con $\bar{X}_h$ y $\bar{X}_k$ los centroides, n_h y n_k los tamaños de los conglomerados h y k respectivamente.

Para un único atributo, la suma de cuadrados se obtiene de

$$SCW = \sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} X_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{n_j} X_{ij} \right)^2 \right), \tag{7.5}$$

donde X_{ij} es el valor del atributo para el i-ésimo individuo en el j-ésimo conglomerado, k es el número del conglomerado en cada etapa y n_j es el número de individuos para el j-ésimo conglomerado.

Se empieza con n grupos, un caso por grupo, aquí la suma de cuadrados de Ward (SCW) es cero. En el segundo paso se buscan los dos casos que

produzcan el menor incremento en la suma de cuadrados, dentro de todas las posibles combinaciones de a dos objetos. En la tercera etapa se toman los $(n-1)$ grupos conformados, se calcula la SCW y se juntan aquellos que produzcan el menor incremento en la variabilidad. El proceso continúa hasta obtener un grupo de n objetos o casos.

Para facilitar la comprensión del algoritmo se desarrolla el caso con cinco individuos sobre los cuales se mide un atributo.

Individuo	Atributo
A	3
B	7
C	8
D	11
E	14

El procedimiento en cada una de sus etapas es el siguiente;

• *Primera etapa*

La SCW para cada uno de los individuos es cero. Los grupos iniciales son

$$\{A\},\ \ \{B\},\ \ \{C\},\ \ \{D\},\ \text{y}\ \{E\}$$

• *Segunda etapa*

Los $\binom{5}{2} = 10$ posibles grupos o conglomerados de a dos individuos cada uno, producen la siguientes sumas de cuadrados

$$SCW_{\{A,B\}} = \left(3^2 + 7^2 - \frac{1}{2}(3+7)^2\right) = 8 \qquad SCW_{\{A,C\}} = 12.5$$
$$SCW_{\{A,D\}} = 32 \qquad SCW_{\{A,E\}} = 60.5$$
$$SCW_{\{B,C\}} = 0.5^{\bigstar} \qquad SCW_{\{B,D\}} = 12$$
$$SCW_{\{B,E\}} = 24.5 \qquad SCW_{\{C,D\}} = 5.5$$
$$SCW_{\{C,E\}} = 18 \qquad SCW_{\{D,E\}} = 4.5$$

Los individuos B y C son fusionados, pues producen la menor SCW. Los conglomerados resultantes son

$$\{A\},\ \ \{B,C\},\ \ \{D\}\ \text{y}\ \{E\}$$

• *Tercera etapa*

Se calcula la SCW para cada uno de los posibles agrupamientos ($\binom{4}{2} = 6$), entre los cuatro grupos encontrados en el paso anterior; resulta

$$SCW_{\{A\}\{B,C\}} = (3^2 + 7^2 + 8^2) - \frac{1}{3}(3 + 7 + 8)^2 = 14$$

$$SCW_{\{A,D\}} = 32 \qquad SCW_{\{A,E\}} = 60.5$$

$$SCW_{\{D\}\{B,C\}} = 8.67 \qquad SCW_{\{E\}\{B,C\}} = 28.67$$

$$SCW_{\{D,E\}} = 4.5^{\bigstar}$$

El grupo que registra la mayor homogeneidad es el conformado por D y E, ya que la fusión de estos dos objetos produce la menor variabilidad. Los grupos que se han formado hasta aquí son:

$$\{A\}, \ \{B, C\}, \text{ y } \{D, E\}.$$

• *Cuarta etapa*

Con los tres grupos anteriores se hacen los posibles reagrupamientos de a dos conglomerados, y luego se determina la SCW para cada una de las $\binom{3}{2} = 3$ "nuevos" arreglos. Los resultados se resumen en seguida

$$SCW_{\{A\}\{B,C\}} = 14^{\bigstar}, \quad SCW_{\{A\}\{D,E\}} = 64.67 \text{ y } SCW_{\{B,C\}\{D,E\}} = 30;$$

el grupo que muestra la mayor homogeneidad, en términos de la menor suma de cuadrados de Ward, lo constituyen A, B y C; de donde resultan los siguientes conglomerados: $\{A, B, C\}$ y $\{D, E\}$.

• *Quinta etapa*

El último conglomerado está constituido por A, B, C, D y E; con

$$SCW_{\{A,B,C\}\{D,E\}} = (3^2+7^2+8^2+11^2+14^2) - \frac{1}{5}(3+7+8+11+14)^2 = 60.2.$$

La Figura 7.5 contiene el dendrograma que ilustra el proceso de aglomeración jerárquica mediante la suma de cuadrados de Ward, para el ejemplo desarrollado.

El método de Ward tiende a formar conglomerados con pocas observaciones y tiende a conformar grupos con el mismo número de observaciones. Por basarse en promedios es muy sensible a la presencia de valores atípicos (outliers).

Para el caso de variables cualitativas, Pardo (1992) propone un procedimiento basándose en el método de Ward, para variables binarias y de tres categorías.

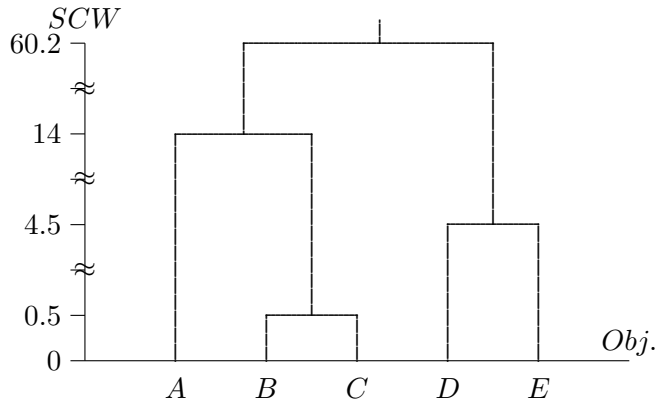


Figura 7.5 *Dendrograma: método de la SC de Ward.*

Finalmente, Gordon (1987) hace una revisión de los métodos jerárquicos de clasificación para la obtención de diagramas de árbol o dendrogramas y la validación de la clasificación obtenida.

7.3.2 Métodos de partición

A diferencia de los métodos de clasificación jerárquica, los métodos de *partición o no jerárquicos* no han sido muy empleados o examinados; razón por la que se aplican e interpretan, a veces, de una manera poco correcta. Se resumen estas técnicas de clasificación con las siguientes características:

1. Empiezan con una partición del conjunto de objetos en algún número especifíco de grupos; a cada uno de estos grupos se le calcula el centroide.

2. Ubican cada caso u objeto en el conglomerado cuyo centroide esté más cercano a éste.

3. Calculan el nuevo centroide de los conglomerados; éstos no son actualizados hasta tanto no se comparen sus centroides con todos los casos.

4. Continúan con los pasos (2) y (3) hasta que los casos resulten irremovibles.

Otra diferencia de las técnicas de partición con las jerárquicas, es que la ubicación de un objeto en un grupo no es definitiva.

∘ *Método de las K-medias*

Se asume que entre los individuos se puede establecer una distancia euclidiana. La idea central de estos métodos es la selección de alguna partición inicial de los objetos para luego modificar su configuración hasta obtener la "mejor" partición en términos de una función objetivo. Varios algoritmos propuestos para estos procedimientos difieren respecto al criterio de optimización (la "mejor" partición). Estos algoritmos son semejantes al de optimización, conocido como *el mayor descenso*, los cuales empiezan con un punto inicial y generan una serie de movimientos desde un punto a otro, calculando en cada paso el valor de una función objetivo, hasta que se encuentra un óptimo local.

El procedimiento de agrupamiento de *K-medias* consiste en particionar un conjunto de n individuos en k grupos, se nota la partición por $\mathcal{P}(n,k)$, con el siguiente criterio: primero se escogen los centroides de los grupos que minimicen la distancia de cada individuo a ellos, luego se asigna cada individuo al grupo cuyo centroide esté más cercano a dicho centroide.

Más formalmente, denótese por $X_{i,j}$ el valor del i-ésimo individuo sobre la j-ésima variable; con $i = 1,\ldots,n$ y $j = 1,\ldots,p$. La media de la j-ésima variable en el l-ésimo grupo se nota por $\bar{X}_{(l)j}$, $l = 1,\ldots,k$ y $n_{(l)}$ el número de individuos en el l-ésimo conglomerado. La distancia de un individuo a un conglomerado es

$$D_{(i,l)} = \left(\sum_{j=1}^{p} \left(X_{i,j} - \bar{X}_{(l)j} \right)^2 \right)^{1/2} . \tag{7.6}$$

Se define el componente de error de la partición por

$$\mathcal{E}\{\mathcal{P}(n,K)\} = \sum_{i=1}^{n} [D(i,l(i))]^2, \tag{7.7}$$

donde $l(i)$ es el grupo que contiene al i-ésimo individuo, y $D(i,l(i))$ es la distancia euclidiana entre el individuo i y el centroide del grupo que contiene al individuo. El procedimiento consiste en encontrar la partición con el error $\mathcal{E}$ más pequeño, moviendo individuos de un conglomerado a otro hasta que se estabilice la reducción de $\mathcal{E}$. En resumen, se trata de reubicar los individuos, de manera que se consigan grupos con la menor variabilidad posible.

Parte del problema está en la conformación de los K grupos iniciales. En la literatura sobre ésta técnica se sugieren, entre otras, las siguientes estrategias:

1. Escoger los primeros K objetos de la muestra como los K grupos iniciales de vectores de medias.

2. Escoger los K objetos más distantes.

3. Empezar con un valor de K tan grande como sea necesario, y proceder a formar centroides de los grupos espaciados a un múltiplo de desviación estándar sobre cada variable.

4. Rotular los objetos de 1 a n y escoger los que resulten marcados con los números $n/k, 2n/k, \ldots, (k-1)n/k$ y n.

5. Escoger K y la configuración inicial de los grupos por el conocimiento previo del problema.

◦ *Métodos basados en la traza*

Siguiendo la metodología del diseño experimental, se persigue minimizar la varianza dentro de los grupos, para detectar las diferencias entre ellos. Sea T la matriz de variación total, $\boldsymbol{E}$ la matriz de covariación dentro de los grupos, y $\boldsymbol{H}$ la matriz de covariación entre grupos; como en la sección (3.5), se tiene la igualdad

$$T = \boldsymbol{E} + \boldsymbol{H}. \tag{7.8}$$

Si se asumen K grupos , $\boldsymbol{E} = \sum_{i=1}^{k} \boldsymbol{E}_i$. En cualquier conjunto de datos T es fijo, entonces el criterio para la formación de conglomerados recae sobre $\boldsymbol{E}$ o $\boldsymbol{H}$. Algunos criterios son los siguientes:

- **La traza de $\boldsymbol{E}$**. Se trata de minimizar la traza de la matriz combinada de sumas de cuadrados y productos cruzados, por la identidad (7.8), minimizar la traza de $\boldsymbol{E}$ equivale a maximizar la traza de $\boldsymbol{H}$.

- **Determinante de $\boldsymbol{E}$**. La minimización del determinante de $\boldsymbol{E}$ es un criterio para la partición de grupos. Minimizar $|\boldsymbol{E}|$ equivale a maximizar $|T|/|\boldsymbol{E}|$.

- **Traza de $\boldsymbol{H}\boldsymbol{E}^{-1}$**. De manera análoga al análisis de varianza multivariado (sección (3.5)) se pretende maximizar $\boldsymbol{H}\boldsymbol{E}^{-1}$, esto puede ser expresado en términos de los valores propios $\lambda_1, \ldots, \lambda_p$ asociados con la matriz $\boldsymbol{H}\boldsymbol{E}^{-1}$, porque $\text{tra}(\boldsymbol{H}\boldsymbol{E}^{-1}) = \sum_i \lambda_i$.

◦ *Nubes dinámicas*

Los procedimientos de clasificación conocidos con el nombre de *nubes dinámicas* comienzan con una partición del conjunto de individuos, con el propósito de mejorarla u optimizarla respecto a una regla. La optimización

se consigue a través de procedimientos iterativos de cálculo, generalmente mediante los llamados métodos numéricos. Estos procedimientos requieren un criterio que permita comparar las calidades de dos particiones o clasificaciones que tienen el mismo número de clases o grupos. El procedimiento se termina cuando no se pueda mejorar la calidad de tal partición.

El algoritmo de las *nubes dinámicas de Diday (1972, 1974)* trata de optimizar el criterio llamado *función de agregación-separación* que expresa la adecuación entre una partición de un conjunto de individuos y una manera de representar las clases de esa partición. El algoritmo requiere definir la manera de representar los subconjuntos o clases de la partición; tal representación se llama *núcleo* y puede ser:

- el centro de gravedad de la clase (centroide),
- un grupo de individuos,
- una recta, un plano, etc.

Como se aprecia, tales núcleos no necesariamente son el "centro de gravedad" (centroide) sino que también pueden ser algunos de los individuos a clasificar, los cuales se consideran como un "prototipo" o "patrón" en cada grupo con un alto poder descriptivo. Este criterio de nucleización es orientado por el experto en cada campo, llámese biólogo, economista, psicólogo, ingeniero, médico, entre otros. La gráfica 7.6 ilustra algunos núcleos.

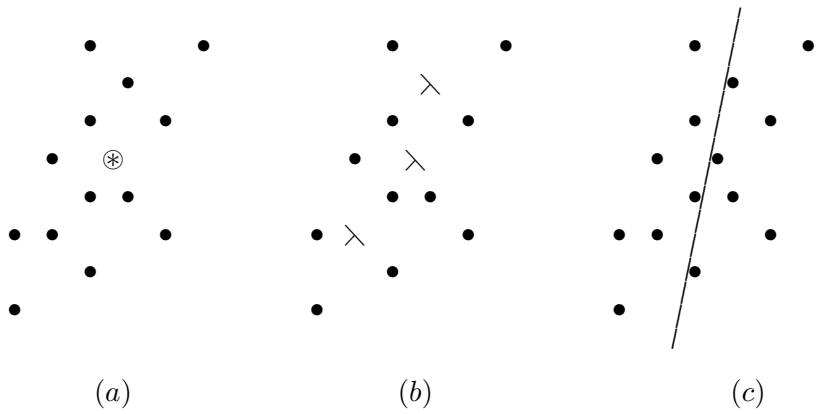


Figura 7.6 *Núcleos:* (*a*) *Centroides,* (*b*) *Individuos y* (*c*) *Recta.*

El algoritmo se desarrolla de la siguiente manera:

1. Se parte de k núcleos, seleccionados entre una familia de L núcleos. Éstos centros $\{L_1^0, \ldots, L_k^0\}$ inducen una partición del conjunto de objetos en k clases $\{C_1^0, \ldots, C_k^0\}$. El i-ésimo objeto es asignado a la clase cuyo núcleo esté más cercano a este objeto.

2. Se determinan los k "nuevos" núcleos $\{L_1^1, \ldots, L_k^1\}$ de las clases asociadas a la partición obtenida $\{C_1^0, \ldots, C_k^0\}$. Estos nuevos centros inducen otra partición, la cual se construye con la misma regla anterior, es decir, con la que se obtienen las clases $\{C_1^1, \ldots, C_k^1\}$.

3. El proceso se desarrolla hasta la m-ésima etapa, donde se encuentran k nuevos núcleos. Se empieza el proceso con los "nuevos" núcleos $\{L_1^m, \ldots, L_k^m\}$, los cuales corresponden al centro de gravedad de cada una de las clases $\{C_1^{m-1}, \ldots, C_k^{m-1}\}$. Con estos últimos núcleos se genera una nueva partición, cuyas clases son
   $\{C_1^m, \ldots, C_k^m\}$.

El criterio para "frenar" el proceso anterior es un problema de cálculo numérico y depende de la *función de agregación-separación* asumida (mínima varianza o inercia dentro de cada clase, distancia entre núcleos, número de iteraciones definido, etc.).

De manera esquemática, los principales aspectos del método son los siguientes:

- Se particiona $\mathbf{\Omega}$, el conjunto sobre el cual se quiere desarrollar una clasificación, en la forma $\mathcal{C} = \{C_1^0, \ldots, C_k^0\}$.
- Sea $\mathcal{L}^j = \{L_1^j, \ldots, L_k^j\}$ los núcleos de las clases en una etapa j.
- Sea $\mathcal{C}^j = \{C_1^{j-1}, \ldots, C_k^{j-1}\}$ una clasificación en una etapa j.
- La construcción de un criterio de adecuación global, de la forma

$$\boldsymbol{W}(\mathcal{C}, \mathcal{L}) = \sum_{l=1}^{k} D(C_l, L_l),$$

  donde D es una medida de adecuación (ajuste) del núcleo L_l a la clase C_l. Un valor pequeño de D muestra un buen ajuste entre L_l y C_l. Así, en cada iteración j, el decrecimiento del criterio muestra un

aumento del ajuste global entre las clases y los respectivos núcleos. Formalmente, el criterio es una aplicación de la forma:

$$\boldsymbol{W} : \mathcal{C} \times \mathcal{L} \longrightarrow \mathbb{R}^{+}.$$

7.3.3 Métodos gráficos

Haría falta más espacio para terminar la revisión de todas las técnicas de agrupamiento existentes hasta hoy. Finalmente se pueden citar algunas técnicas tales como, los *"glyphs"*, *estrellas*, *los rostros de Chernoff* y los gráficos de Fourier.

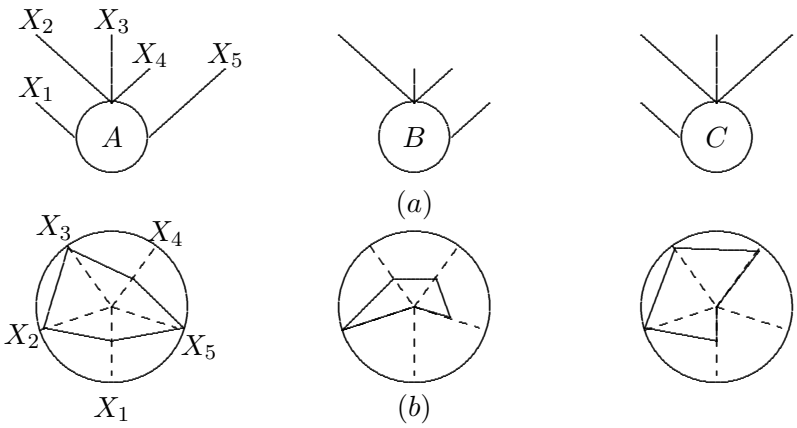


Figura 7.7 *Representación de tres individuos 5-dimensionales.*

• Un *glyph* consta de un círculo de radio r con p rayos que salen de él. La posición y longitud de cada rayo refleja el valor de la coordenada asociada con cada una de las p variables; las cuales pueden ser cualitativas (alto, medio y bajo, por ejemplo) o cuantitativas. En la figura 7.7a se representan 3 individuos A, B, y C a los cuales se les han registrado los atributos X_1, X_2, X_3, X_4 y X_5. En ella, los individuos B y C no aparecen con los rayos ligados a las variables X_1 y X_5, respectivamente; esto significa que el individuo B toma el valor cero (o nivel más bajo) en X_1 y el C toma el valor cero en X_5. Para conformar grupos, tan sólo es necesario buscar los *glyphs* (individuos) que más se parezcan respecto a las variables de interés, así por ejemplo, con relación a las variables X_2, y X_4 los individuos A y B son bastante semejantes, pero muy diferentes respecto a las demás variables.

• Una variación de los diagramas anteriores son los denominados de *estrellas*, en los cuales las variables se ubican sobre los radios de una estrella regular. La magnitud (o nivel) si es cualitativa, de cada variable se ubica sobre cada radio, así un valor máximo se representa en los extremos y un valor nulo (o bajo) en el centro de la circunferencia; el polígono que une los puntos ubicados sobre cada radio determina a un individuo. En la figura 7.7b, mediante una representación alterna, se muestran los gráficos de los atributos obtenidos por los mismos tres individuos A, B y C.

• Los *rostros de Chernoff* se basan en la representación de un vector de observaciones mediante características faciales como por ejemplo; la cabeza, la boca, la nariz, los ojos, las cejas y las orejas. Chernoff (1973) propuso hasta 18 dimensiones (variables) ligadas a 18 características faciales. Para un problema particular se asigna a cada una de las variables un rasgo facial determinado; por ejemplo, en un país al cual se le registra su producto interno bruto (X_1), población, (X_2), ingreso percápita (X_3), tasa de natalidad (X_4), tasa de mortalidad (X_5) y desempleo (X_6) se pueden identificar respectivamente estas variables con la longitud de la nariz, el ancho de la nariz, la distancia entre los ojos, la excentricidad de los ojos, el ángulo de las cejas y la curvatura de la boca. La figura 7.8 muestra nueve rostros, cada uno de los cuales representa, hipotéticamente, a un país.

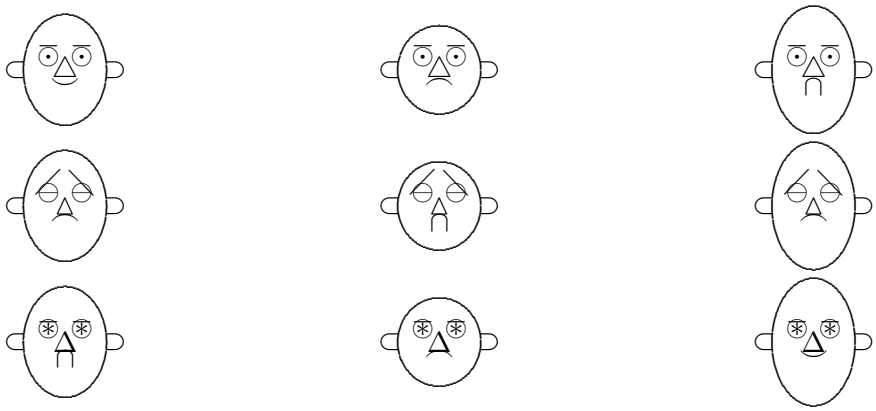


Figura 7.8 *Rostros de Chernoff.*

• Los *gráficos de Fourier* es otra técnica para la conformación de conglomerados. Andrews (1972) propone transformar los vectores de respuestas

p-dimensionales $X' = (X_1, \ldots, X_p)$ por series de Fourier de la forma

$$f_X(t) = \frac{X_1}{\sqrt{2}} + X_2 \operatorname{sen} t + X_3 \cos t + X_4 \operatorname{sen} 2t + X_5 \cos 2t + \cdots, \text{ con } -\pi \leq t \leq \pi.$$

Con n individuos se generan n curvas, una curva por individuo. La función f preserva las medias y las varianzas; las distancias se calculan a través de

$$\|f_{X_i} - f_{X_i'}\| = \int_{-\pi}^{\pi} [f_{X_i} - f_{X_i'}]^2 dt,$$

Para un valor específico de t_0, $f(t_0)$ es proporcional a la longitud de la proyección del vector $(X_1, \ldots, X_p)$ sobre el vector $\left(\frac{1}{\sqrt{2}}, \operatorname{sen} t_0, \cos t_0, \operatorname{sen} 2t_0, \cos 2t_0, \ldots\right)$. Esta proyección revela los grupos o conglomerados, a manera de bandas que contienen ondas "paralelas".

En la figura 7.9 se han clasificado los seis individuos en los grupos $\{A, B, D\}$ y $\{C, E, F\}$.

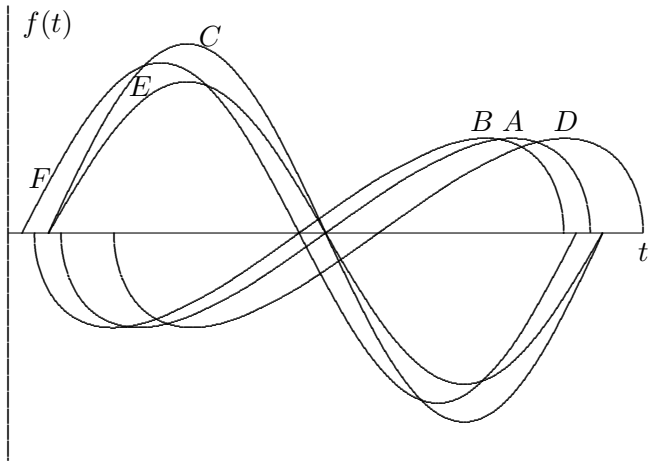


Figura 7.9 *Curvas de Andrews para clasificar seis objetos.*

7.3.4 Conglomerados difusos ("fuzzy")

El concepto de conjuntos difusos fue introducido por Zadeh (1965). Un conjunto *difuso* (borroso) es una clase de objetos con algún grado de perte-

nencia a éste. Hay casos en los que la relación de pertenencia de un objeto a un conjunto no está claramente definida, por ejemplo:

- las bacterias, los virus, la estrella de mar, tienen una situación ambigua con relación a la clase de los animales o de las plantas,
- la misma relación de ambigüedad se presenta entre el número 10 y la clase de números "mucho más grandes" que el número 1,
- la clase de las "mujeres bonitas",
- la clase de los "hombres altos";

estos grupos de objetos no constituyen clases o conjuntos en el sentido del término matemático usual. En esta parte se trata de mostrar la clasificación que se puede lograr con este tipo de objetos y conjuntos.

Más formalmente, sea $\mathcal{X}$ una colección de objetos, con un elemento genérico notado por x, así, se puede escribir $\mathcal{X} = \{x\}$. Un *conjunto difuso* $\mathcal{A}$ de $\mathcal{X}$ es caraterizado por una función de pertenencia (característica) $f_{\mathcal{A}}(x)$ la cual asocia a cada punto de $\mathcal{X}$ un número real en el intervalo $[0, 1]$. Con el valor de $f_{\mathcal{A}}(x)$ se representa "el grado de pertenencia" de x a $\mathcal{A}$. Un valor de $f_{\mathcal{A}}(x)$ cercano a 1 corresponde a un alto grado de pertenencia de x en $\mathcal{A}$. Cuando $\mathcal{A}$ es un conjunto en el sentido clásico, su función de pertenencia toma únicamente los valores 1 o 0, de acuerdo con la pertenencia o la no pertenencia de x a $\mathcal{A}$ (Yager y colaboradores, 1984).

Una clasificación difusa implica optimizar un criterio que involucra coeficientes de membresía. Esto permite la conformación de conglomerados a través de los procedimientos clásicos.

∘ *Similitud difusa*

Una relación difusa binaria $\mathcal{R}$, se define como una colección de pares ordenados, es decir, si $\mathcal{X} = \{x\}$ y $\mathcal{Y} = \{y\}$, son colecciones de objetos, entonces, una *relación difusa* de $\mathcal{X}$ en $\mathcal{Y}$ es un subconjunto $\mathcal{R}$ de $\mathcal{X} \times \mathcal{Y}$ caracterizado por la función de pertenencia $\mu_{\mathcal{R}}$, la cual asocia a cada par (x, y) de $\mathcal{X} \times \mathcal{Y}$ su grado de "pertenencia" $\mu_{\mathcal{R}}$ a $\mathbb{R}$. Se asume por simplicidad que el rango de $\mu_{\mathcal{R}}$ es el intervalo $[0, 1]$. El número $\mu_{\mathcal{R}}(x, y)$ se considera como la "fuerza" o el grado de la relación que hay entre x y y.

Una *relación de similitud difusa en* $\mathcal{X}$ es una relación de *similitud* $\mathcal{S}$ en $\mathcal{X}$, la cual satisface las siguientes propiedades:

(a) *Reflexiva*; es decir, $\mu_{\mathcal{S}}(x, x) = 1$ para todo x en el dominio de $\mathcal{S}$,

(b) *Simétrica*: $\mu_{\mathcal{S}}(x, y) = \mu_{\mathcal{S}}(y, x)$ para todo x, y en el dominio de $\mathcal{S}$, y

(c) *Transitiva*: $\mu_{\mathcal{S}}(x, z) \geq \sup_y\{\mu_{\mathcal{S}}(y, x) \wedge \mu_{\mathcal{S}}(y, z)\}$, para todo x, y y z en el dominio de $\mathcal{S}^2$.

Aquií el símbolo "$\wedge$" nota el máximo entre las funciones de pertenencia.

El complemento de la relación de similitud $\mathcal{S}$ es interpretado como una relación de disimilaridad $\mathcal{D}$ o una *función de distancia*, donde

$$\mu_{\mathcal{D}}(x, y) = 1 - \mu_{\mathcal{S}}(x, y) = d(x, y)$$

Considérese el conjunto de pares cuyo grado de similitud es mayor o igual que una cantidad α; es decir,

$$\mathcal{S}_\alpha = \{(x, y)\ en\ \mathcal{X} \times \mathcal{X} :\ \mu_{\mathcal{S}}(x, y) \geq \alpha\},\ \text{con } 0 \leq \alpha \leq 1.$$

La relación $\mathcal{S}_\alpha$ cumple las tres propiedades (a), (b) y (c) anteriores; luego induce una partición sobre el conjunto $\mathcal{X}$.

Se nota por Π_α a la partición en $\mathcal{X}$ inducida por $\mathcal{S}_\alpha$, con $0 \leq \alpha \leq 1$. Claramente, $\Pi_{\alpha'}$ es un *refinamiento* de Π_α si $\alpha' \geq \alpha$. Dos elementos x, y de $\mathcal{X}$ están en el mismo conglomerado (clase) de la partición $\Pi_{\alpha'}$, si y sólo si, $\mu_{\mathcal{S}}(x, y) \geq \alpha'$. Esto implica que $\mu_{\mathcal{S}}(x, y) \geq \alpha$ y por tanto que x y y están en el mismo conglomerado de Π_α.

Una sucesión de particiones $\Pi_{\alpha_1}, \ldots, \Pi_{\alpha_k}$ se puede representar mediante un *árbol* o *dendrograma*. El árbol está asociado con la matriz $\boldsymbol{\mu}_{\mathcal{S}}$ que contiene las similaridades; los objetos x_i y x_j pertenecen al mismo conglomerado de Π_α, si y sólo si, $\mu_{\mathcal{S}}(x_i, x_j) \geq \alpha$.

Ejemplo 7.3 La matriz siguiente contiene las similaridades difusas entre los objetos del conjunto $\mathcal{X} = \{O_1, O_2, O_3, O_4, O_5, O_6\}$

$$\boldsymbol{\mu}_{\mathcal{S}} = \begin{pmatrix} & O_1 & O_2 & O_3 & O_4 & O_5 & O_6 \\ O_1 & 1 & 0.2 & 1 & 0.6 & 0.2 & 0.6 \\ O_2 & 0.2 & 1 & 0.2 & 0.2 & 0.8 & 0.2 \\ O_3 & 1 & 0.2 & 1 & 0.6 & 0.2 & 0.6 \\ O_4 & 0.6 & 0.2 & 0.6 & 1 & 0.2 & 0.8 \\ O_5 & 0.2 & 0.8 & 0.2 & 0.2 & 1 & 0.2 \\ O_6 & 0.6 & 0.2 & 0.6 & 0.8 & 0.2 & 1 \end{pmatrix}$$

En la figura 7.10 se muestra la partición que se obtiene para cada uno de los valores $\alpha = 0.2,\ 0.6,\ 0.8$ y 1.0, respectivamente.

[2] En el sentido clásico $\mu_{\mathcal{S}}(;)$ es una relación de equivalencia.

Se observa por ejemplo, que a un grado de similitud $\alpha = 0.7$ se conforman los conglomerados $\{O_1, O_3, O_4, O_6\}$ y $\{O_2, O_5\}$; y una similitud $\alpha = 0.9$ se conforman los conglomerados $\{O_1, O_3\}$, $\{O_4, O_6\}$ y $\{O_2, O_5\}$.

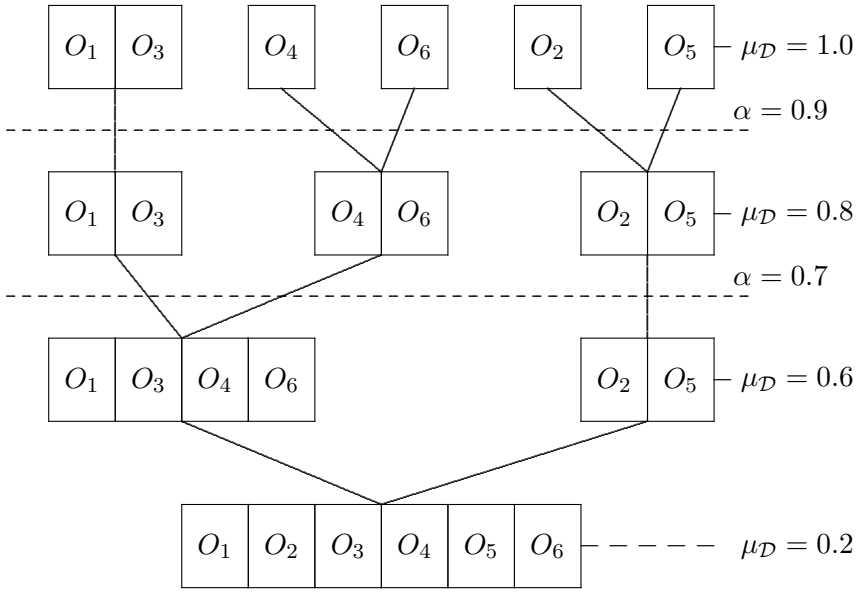


Figura 7.10 *Árbol para la relación de similitud difusa.* $\boldsymbol{\mu}_{\mathcal{S}}$

El algoritmo de *K-medias difuso*, referido en Krzanowski (1995, pág. 88), tiene como objetivo minimizar el criterio

$$\sum_{k=1}^{q} \sum_{i=1}^{n} f_{ik}^2 \|x_i - \mu_k\|,$$

donde f_{ik} es el coeficiente de pertenencia del i-ésimo objeto al k-ésimo conglomerado y $\|x_i - \mu_k\|$ es una medida de distancia, usualmente el cuadrado de la distancia euclidiana, entre el x_i y μ_k. Los centros de cada grupo μ_k son estimados de acuerdo con la siguiente expresión

$$\mathfrak{v}_{ik} = \frac{\sum_i f_{ik} x_{ij}}{\sum_h f_{hk}}, \text{ con } j = 1, \ldots, p;\ h = 1, \ldots, q.$$

A estas alturas el lector puede estar inquieto por la técnica que "mejor" clasifique un conjunto de datos, muy a pesar de que son más los métodos

omitidos que los considerados en este texto, la respuesta es desalentadora: no hay un método o algoritmo de clasificación que sea el "óptimo". La recomendación salomónica es hacer una especie de *panel* de metodologías clasificatorias sobre el conjunto de datos por agrupar para observar la confluencia de los métodos en términos de la tipología de clasificación obtenida[3], sin perder de vista el marco conceptual que cincunscribe los datos.

7.4 Determinación del número de conglomerados

Una de las inquietudes al emplear el análisis de conglomerados, es la decisión acerca del número apropiado de ellos. Los dendrogramas sugieren el número de conglomerados en cada paso, la pregunta sigue siendo ¿dónde cortar el árbol para obtener un número óptimo de grupos?. Esta pregunta no ha sido enteramente resuelta hasta hoy, aunque cada uno de los campos de aplicación le da una importancia diferente. Para las ciencias biológicas, por ejemplo, el problema de definir el número de grupos no es muy importante, simplemente porque el objetivo del análisis es la exploración de un patrón general de las relaciones entre los objetos, lo cual se logra a través de un árbol.

Procedimientos heurísticos son los más usados comunmente, en el caso más simple, un árbol jerárquico es cortado por inspección subjetiva en diferentes niveles. Este procedimiento es bastante satisfactorio porque generalmente es guiado por las necesidades y opiniones del investigador acerca de la adecuada estructura de los datos.

Otro método consiste en graficar el número de conglomerados de un árbol jerárquico en función del *coeficiente de fusión*, que corresponde al valor numérico bajo el cual varios casos se mezclan para formar un grupo. Los valores del coeficiente de fusión se ubican sobre el eje "Y" en el diagrama de árbol. Se traza la línea que une los puntos de coordenadas el coeficiente de fusión y el número de conglomerados; el punto desde donde la línea trazada se hace horizontal sugiere el número de conglomerados adecuado. La figura 7.11 muestra una situación hipotética. La línea se hace casi horizontal a partir del cuarto grupo, así que cuatro o tres conglomerados están presentes en los datos (semejante a la sección (5.5) para ACP).

Un procedimiento alterno consiste en examinar los valores del coeficiente de fusión para encontrar puntos donde el "salto" en el valor del coeficiente sea notorio. Un cambio brusco significa la mezcla de dos grupos dispares;

[3]No importa que se tome como una "perogrullada" estadística.

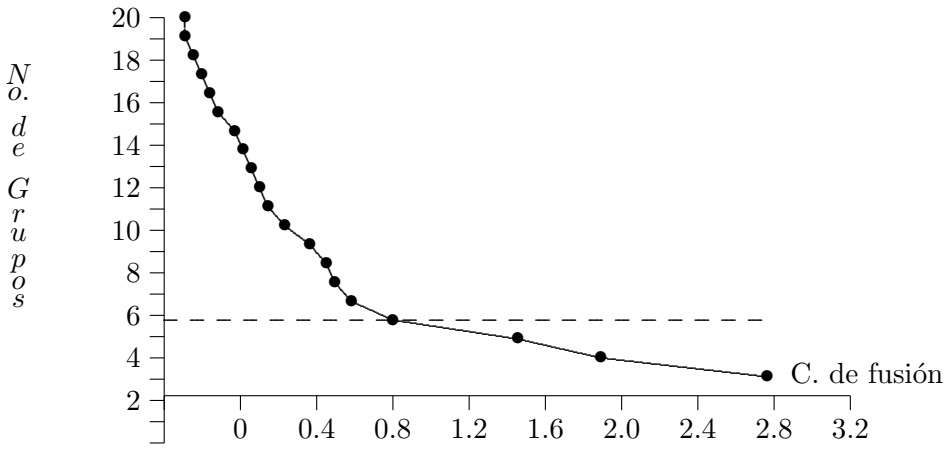


Figura 7.11 *Número de grupos vs coeficiente de fusión.*

es decir, que el número de conglomerados previos al punto de salto es el adecuado. Los datos siguientes corresponden al coeficiente de fusión asociado con el número de conglomerados, para un conjunto de datos

No. de conglomerados	10	9	8	7	6
Coeficiente de fusión	0.234	0.267	0.289	0.305	0.332
No. de conglomerados	5	4	3	2	1
Coeficiente de fusión	0.362	0.388	0.591	0.684	0.725

Un cambio brusco en la sucesión de valores del coeficiente de fusión se observa del cuarto al tercer conglomerado. Los valores del coeficiente para un número de grupos entre 10 y 4 se incrementan máximo en 3 centésimas; del grupo cuarto al tercero el incremento es alrededor de 2 décimas; así el número adecuado de conglomerados es cuatro. La dificultad de este procedimiento está en que muchos saltos de poca intensidad pueden presentarse, situación que hace difícil señalar el número de grupos apropiado.

Aunque no se han desarrollado formalmente pruebas estadísticas, algunas tienen una aceptación relativamente amplia. Lee (1979) considera algunas pruebas para la hipótesis de que los datos proceden de una población normal p-variada, en oposición a la alternativa de que provienen de dos poblaciones multinormales de diferente media. La prueba se basa en la razón de verosimilitud (sección (3.5)), y la siguiente ecuación

$$C_p = \max\{|T|/|\boldsymbol{E}|\}; \tag{7.9}$$

la maximización se hace sobre todas las posibles particiones de los datos en dos grupos. La distribución teórica de C_p es bastante complicada, sin embargo, es un punto de partida para determinar la posible diferencia entre grupos. El uso de esta prueba es limitada, pues es aplicable únicamente en el caso univariado.

Milligan y Cooper (1985) describen y proponen pruebas para identificar el número apropiado de grupos en un proceso de aglomeración jerárquica. Peck, Fisher y Ness (1989) encuentran un intervalo de confianza para el *número de conglomerados*, a través de un procedimiento "bootstrap". El procedimiento consiste en definir una función criterio que dependa de dos tipos de costos, un costo asociado con el número de conglomerados, y un costo asociado con la descripción de un individuo por su respectivo conglomerado (homogeneidad del conglomerado); se busca entonces un intervalo de confianza para k, el número de conglomerados, que minimice la función criterio.

En resumen, la técnica del análisis de conglomerados es otra técnica de reducción de datos. Se puede considerar la metodología de las componentes principales (capítulo 5) como un análisis de conglomerados, donde los objetos corresponden a las variables.

El análisis de conglomerados no tiene pretenciones inferenciales hacia una población a partir de una muestra, se emplea fundamentalmente como una técnica exploratoria. Las soluciones no son únicas; y además, siempre es posible conformar conglomerados, no obstante que los datos tengan una estructura "real o natural". Las tipologías encontradas en un análisis de conglomerados son fuertemente dependientes tanto de las variables relevantes como de las observaciones intervinientes en la construcción; así: una nueva variable, un nuevo individuo o los dos, pueden alterar cualquier estructura conseguida anteriormente. En consecuencia, se advierte sobre el cuidado que se debe tener con el uso de esta técnica en la toma de decisiones.

7.5 Rutina SAS para conformar conglomerados

◦ PROC CLUSTER: este procedimiento agrupa en forma jerárquica las observaciones en conglomerados o clases. Procedimientos tales como ACECLUS, FASTCLUS, MODECLUS, TREE y VARCLUS se encuentran disponibles en el paquete SAS (SAS User's Guide, 2001).

```
DATA   nombre SAS de los datos;
INPUT  escribir las variables;
CARDS; /* para entrar a continuación los datos */
```

```
;
PROC CLUSTER METHOD=
   AVERAGE  CENTROID  COMPLETE
   DENSITY  SINGLE    WARD; /* se debe elegir alguno de estos */
    /* métodos para desarrollar el análisis de conglomerados */
VAR  lista de variables numéricas para el análisis;
RUN;
```

○ FASTCLUS: sirve para la conformación de conglomerados disjuntos cuando el número de observaciones es grande (entre 100 y 100.000). Se especifica el número de conglomerados, y eventualmente, el radio mínimo de éstos.

```
PROC FASTCLUS ; VAR lista de variables numéricas para el análisis;
ID  variable, nominal o cuántica, que identifica las observaciones
     pedidas por la opción LIST en la declaración anterior;
BY  se usa para obtener análissis FASTCLUS separados sobre
    observaciones en los grupos definidos por el BY, se requiere un
    ordenamiento de las observaciones con el PROC SORT sobre la
     misma variable indicada en el BY;
RUN; para desarrollar la rutina
```

7.6 Procesamiento de datos con R

Se ilustra la conformación de conglomerados con R usando la matriz de distancias que aparece en la sección 7.3.1 asociada con la distancia 7.1. Se realiza el análisis usando varios métodos.

```
x<-matrix(c(0,   3, 7, 11, 10,
            3,   0, 6, 10,  9,
            7,   6, 0,  5,  6,
            11, 10, 5,  0,  4,
            10,  9, 6,  4,  0),ncol=5 )

dimnames(x)<-list(paste("O",1:5,sep=""),
                  paste("O",1:5,sep="") )
y<-as.dist(x)

# Enlace completo (la vecina más lejana): opción por defecto
cl<-hclust(y)
plot(cl,hang = -1)
abline(h=4.5,lty=2)

# Enlace simple (la vecina más cercana)
cl<-hclust(y, method = "single")
plot(cl,hang = -1)
abline(h=4.5,lty=2)
```

```
# Union mediante el promedio
cl<-hclust(y, method="average")
plot(cl,hang = -1)
abline(h=4.5,lty=2)

# método de ward
cl<-hclust(y, method="ward")
plot(cl,hang = -1)
abline(h=4.5,lty=2)
```

Si se tienen los datos, en lugar de la matriz de distancias, se puede calcular ésta mediante la función `dist()`, a continuación se ilustra el procedimiento.

```
x<-c(1,2,4,5,3,3)
y<-c(2,1,1,4,5,3)
datos<-data.frame(x=x,y=y,row.names=LETTERS[1:length(x)])
# distancia máxima entre dos componentes de X y Y
dd<-dist(datos,method="maximum")
# distancia de manhattan
dd<-dist(datos,method="manhattan")
# distancia de canberra
dd<-dist(datos,method="canberra")
# distancia de minkowski
dd<-dist(datos,method="minkowski")
# distancia euclidiana
dd<-dist(datos)
# distancia euclidiana
dd<-dist(datos,method="euclidean")

# conformación de los conglomerados
cl<-hclust(dd)
plot(cl,hang = -1)
```

Capítulo 8

Análisis discriminante

8.1 Introducción

Dos son los objetivos principales abordados por el *análisis discriminante*, de una parte está la *separación o discriminación de grupos*, y de otra, la *predicción* o *asignación* de un objeto en uno de entre varios grupos previamente definidos, con base en los valores de las variables que lo identifican. El primer objetivo es de carácter descriptivo, trata de encontrar las diferencias entre dos o más grupos a través de una *función discriminante*. Estas funciones se presentan en las secciones (3.4) y (3.5), en donde se comparan dos o más poblaciones con relación a sus centroides. En este capítulo se trata sobre el *análisis de clasificación*, el cual se orienta a "ubicar" un objeto o unidad muestral, en uno de varios grupos de acuerdo con una *regla de clasificación* (o *regla de localización*). Sin embargo, frecuentemente la mejor función para separar grupos provee también la mejor regla de localización de observaciones futuras; de tal forma que estos dos términos generalmente se emplean indistintamente.

Las siguientes son tan sólo algunos situaciones en las que se requeriría de un análisis discriminante:

- Una persona que aspira a ocupar un cargo en una empresa, es sometida a una serie de pruebas; de acuerdo con su puntaje se sugiere ubicarlo en alguno de los departamentos de la empresa.
- Un biólogo quiere clasificar una "nueva" planta en una de varias especies conocidas (taxonomía numérica).
- Un arqueólogo debe ubicar a un antepasado en uno de cuatro períodos históricos.

- En medicina forense, se debe determinar el género (sexo) de una persona con base en algunas medidas sobre determinados huesos de su cuerpo.

- De acuerdo con el registro de calificaciones que un estudiante históricamente ha mostrado, se quiere predecir si llegará a graduarse o no, en una determinada institución educativa.

Éstos son algunos casos típicos del análisis discriminante, pues de acuerdo con un conjunto de variables, se quiere obtener una función con la cual se pueda decidir sobre la asignación de un caso a una de varias poblaciones mutuamente excluyentes.

En el análisis discriminante se obtiene una función que separa entre varios grupos definidos a priori, esta función es una combinación, generalmente lineal, de las variables de identificación, la cual minimiza los errores de clasificación. El problema de la discriminación es entonces comprobar si tales variables permiten diferenciar las clases definidas previamente y precisar como se puede hacer.

Cabe resaltar que el problema es identificar la clase a la que se debe asignar un individuo, de quien se sabe que pertenece a una de las clases definidas de antemano, y para el cual sólo se conocen los valores de las variables "explicativas". Se sigue entonces una tarea de discriminación descriptiva en primer lugar, con la que se asignan individuos a las clases, más no se agrupan, puesto que no se trata de construir grupos sino de asignar individuos a éstos. La última característica diferencia la técnica de discriminación con la de clasificación, presentada en el capítulo 7, otra cosa es el empleo de estas técnicas para complementar o confrontar los resultados de una clasificación vía análisis de conglomerados, por ejemplo.

Para poblaciones multinormales con matriz de covarianzas iguales, las reglas de clasificación son en cierto sentido óptimas. En muchas aplicaciones, ya sea por desconocimiento, por descuido o por simple exploración de los datos, no se consideran los supuestos anteriores; más adelante se comentará acerca de la robustez de la técnica a la normalidad y a la igualdad de la matriz de covarianzas. Cuando se puede suponer que las poblaciones tienen probabilidades a priori, se incorpora esta información al análisis discriminante mediante una regla de discriminación bayesiana; en la sección (8.2.2) se esquematiza este caso. Al final del capítulo, sección (8.5), se consideran otras técnicas de discriminación de tipo no paramétrico. También se puede considerar el análisis discriminante para dos o para más de dos grupos; así se aborda en este capítulo.

8.2 Reglas de discriminación para dos grupos

La mayor parte de la literatura sobre análisis discriminante trata el problema para dos poblaciones. Con base en un vector X de variables medidas sobre una unidad de observación, que en adelante se indicará como la observación X, se quiere clasificar esta unidad en una de dos poblaciones.

A continuación se enuncia el resultado debido a Welch (1939), citado por Rencher (1998), a partir del cual se obtienen algunas reglas de clasificación o discriminación.

> *Sean $f(X|G_1)$ la función de densidad para X en G_1 y $f(X|G_2)$ la función de densidad para X en G_2, con G_1 y G_2 las dos poblaciones. (La notación $f(X|G_1)$ no representa una distribución condicional en el sentido usual). Sean p_1 y p_2 las probabilidades a priori, donde $p_1 + p_2 = 1$, entonces, la regla de discriminación óptima, es decir, la regla que minimiza la probabilidad total de clasificación incorrecta, es:*
>
> - asignar la observación X a G_1 si $p_1 f(X|G_1) > p_2 f(X|G_2)$,
> - o asignar a G_2, en otro caso.

8.2.1 Clasificación vía la máxima verosimilitud

Aunque esta situación, en la práctica, es muy poco frecuente, supóngase que se conocen las distribuciones de las dos poblaciones. Sean $f_1(X)$ y $f_2(X)$ las *fdp* de cada una de las poblaciones, con X vector de observaciones de tamaño $(p \times 1)$ (un caso). La regla de discriminación *máximo verosímil* para localizar el caso caracterizado por X en alguna de dos poblaciones, consiste en ubicarlo en la población para la cual X maximiza la verosimilitud o probabilidad.

En símbolos, si G_1 y G_2 son las dos poblaciones, entonces se localiza a X en G_i si

$$L_i(X) = \max_j \{L_j\}, \text{ con } i, j = 1, 2 \tag{8.1}$$

La regla dada en (8.1) es extendible a cualquier número de poblaciones. En caso de empates, X se asigna a cualquiera de las poblaciones.

▶ Clasificación en poblaciones con matrices de covarianzas iguales

Supóngase que las poblaciones G_i se distribuyen $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, con $i = 1, 2$, de manera que la verosimilitud de la i-ésima población es

$$L_i(X) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big\{-\frac{1}{2}(X - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}_i)\Big\}. \tag{8.2}$$

Maximizar (8.2) equivale a obtener el mínimo de $(X - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}_i)$, el cual es la distancia de Mahalanobis de X a $\boldsymbol{\mu}_i$. Se asigna el individuo, representado por X, a la población más cercana en términos de esta distancia; es decir, se asigna el caso X al grupo G_1 si

$$(X - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}_1) \leq (X - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}_2), \tag{8.3}$$

o al grupo G_2 si

$$(X - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}_1) > (X - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}_2), \tag{8.4}$$

Al desarrollar (8.3) y simplificar algunos términos, se obtiene que se asigna X a G_1 si

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}X - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0, \tag{8.5}$$

o a G_2 en caso contrario. El primer término de (8.5) es la *función discriminante lineal*, si se llama $b = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, entonces la función discriminante es de la forma $Y = b'X$; la cual es una combinación lineal de las medidas asociadas con las variables para un objeto o individuo particular.

Las reglas de ubicación, equivalentes con (8.3) y (8.4), son entonces

$$\begin{aligned} &\text{si } b'(X - \boldsymbol{\mu}_c) \geq 0, \text{ entonces } X \text{ se asigna a } G_1; \text{ o} \\ &\quad \text{si } b'(X - \boldsymbol{\mu}_c) < 0, \text{ entonces } X \text{ se asigna a } G_2 \end{aligned} \tag{8.5a}$$

donde $\boldsymbol{\mu}_c = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$.

Observación:

> La combinación lineal contenida en $b'X$, fue sugerida por R. A. Fisher (1936), de tal forma que la razón de las diferencias en las medias de las combinaciones lineales a su varianza sea mínima. Esto es, la combinación lineal es de la forma $Y = b'X$, y se quiere encontrar el vector de ponderaciones b que maximice la separación entre los dos grupos
>
> $$\frac{(b'\boldsymbol{\mu}_1 - b'\boldsymbol{\mu}_2)^2}{b'\boldsymbol{\Sigma}b}, \tag{8.6}$$

manteniendo constante la varianza de la combinación lineal $b'X$; es decir, $\text{var}(b'X) = b'\boldsymbol{\Sigma}b$, por multiplicadores de Lagrange se concluye que b es proporcional a $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

Hasta ahora se ha asumido que las dos poblaciones se conocen a través de su distribución, en la práctica los parámetros que las determinan e identifican se estiman e infieren desde muestras aleatorias independientes.

Supóngase que se extrae la muestra $X_{1(i)}, \ldots, X_{n_i(i)}$ de una población

$N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ para $i = 1, 2$. Con base en esta información se pretende asignar la observación X a G_1 o a G_2. Los estimadores para $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}$ son, respectivamente

$$\bar{X}_i = \sum_{j=1}^{n_i} X_{j(i)}/n_i, \ i = 1, 2,$$

$$\boldsymbol{S} = \frac{1}{n_1 + n_2 - 2}\left[\sum_{j=1}^{n_1}(X_{j(1)} - \bar{X}_{(1)})(X_{j(1)} - \bar{X}_{(1)})' + \sum_{j=1}^{n_2}(X_{j(2)} - \bar{X}_{(2)})(X_{j(2)} - \bar{X}_{(2)})'\right].$$

Al sustituir estas estimaciones en (8.5), la función discriminante muestral toma la forma $\hat{Y} = \hat{b}'X$. Se usan los mismos criterios dados en (8.5a). Con los datos muestrales, los criterios son:

$$\text{si } \hat{b}'X \geq \hat{b}'\bar{X}_c, \ X \text{ se asigna a } G_1 \text{ o,} \tag{8.7}$$

$$\text{si } \hat{b}'X < \hat{b}'\bar{X}_c, \ X \text{ se asigna a } G_2, \tag{8.8}$$

con $\bar{X}_c = \frac{1}{2}(\bar{X}_1 + \bar{X}_2)$ y $\hat{b} = \boldsymbol{S}^{-1}(\bar{X}_1 - \bar{X}_2)$.

Igual que en regresión, el centroide de los datos $(\bar{X}_1, \bar{Y}_1)$ y $(\bar{X}_2, \bar{Y}_2)$ satisface la ecuación $Y = b'X$; es decir, $\bar{Y}_1 = b'\bar{X}_1$ y $\bar{Y}_2 = b'\bar{X}_2$; de manera que $Y_c = b'X_c$. Las decisiones contempladas en (8.7) y (8.8) son equivalentes a:

$$\text{si } \hat{Y} \geq \bar{Y}_c = \frac{\bar{Y}_1 + \bar{Y}_2}{2}, \ X \text{ se asigna a } G_1, \text{ o} \tag{8.8a}$$

$$\text{si } \hat{Y} < \bar{Y}_c = \frac{\bar{Y}_1 + \bar{Y}_2}{2}, \ X \text{ se asigna a } G_2, \tag{8.8b}$$

con

$$\bar{Y}_c = \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2)$$

y

$$\begin{aligned}\hat{Y} &= \hat{b}X \\ &= \boldsymbol{S}^{-1}(\bar{X}_1 - \bar{X}_2)X.\end{aligned}$$

La figura 8.1 ilustra la discriminación entre dos grupos que tienen distribución normal bivariada, a través de la función discriminante lineal estimada $\hat{Y} = \hat{b}'X$. Por la forma y escala que se observa en las gráficas, se puede asumir que las matrices de covarianzas son casi iguales. Cuando la función se aplica en un punto $X_i = (X_{i1}, X_{i2})'$, se obtiene la combinación lineal $Y_i = b_1 X_{i1} + b_{i2}$, Y_i que corresponde a la proyección del punto X_i sobre la línea de separación óptima entre los dos grupos. Como las dos variables X_1 y X_2 tienen distribución normal (pues X es normal bivariada), la combinación lineal de éstas tiene también una distribución normal.

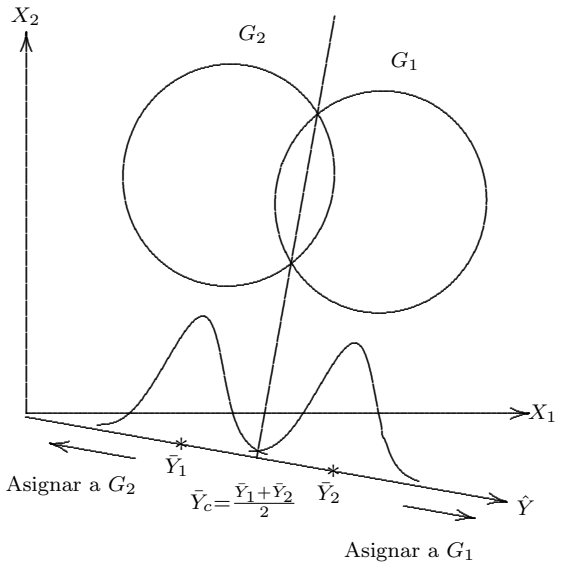


Figura 8.1 *Discriminación lineal.*

La magnitud del valor de la función de clasificación, calculada en el punto X, respecto al punto $\bar{Y}_c$, define la asignación de la observación X a uno de los dos grupos.

Ejemplo 8.1 Un grupo de 49 personas, de edad avanzada, que participaron en un estudio, fueron clasificadas mediante una evaluación psiquiátrica en una de las dos categorías: *senil* o *no senil*.

Los resultados de una prueba de inteligencia adulta, independientemente administrada a cada una de las personas, revela grandes diferencias entre los dos grupos en algunas partes de la prueba; razón por la que se decidió

considerar algunas partes de la prueba (subpruebas) con el fin de encontrar una regla de discriminación.

Las medias de estas subpruebas se resumen en la tabla 8.1.

Tabla 8.1 Evaluación psiquiátrica

Variable	Subprueba	No Senil ($n_1 = 37$)	Senil ($n_2 = 12$)
X_1	Información	12.57	8.75
X_2	Similaridades	9.57	5.33
X_3	Aritmética	11.49	8.50
X_4	Habil. artist.	7.97	4.75

Fuente: Morrison (1990, pág. 143)

Se asume que los datos de cada grupo (senil y no senil) siguen una distribución normal 4-variante con la misma matriz de covarianzas. La matriz de covarianzas muestral es:

$$\boldsymbol{S} = \begin{pmatrix} 11.2553 & 9.4042 & 7.1489 & 3.3830 \\ 9.4042 & 13.5318 & 7.3830 & 2.5532 \\ 7.1489 & 7.3830 & 11.5744 & 2.6170 \\ 3.3830 & 2.5532 & 2.6170 & 5.8085 \end{pmatrix}.$$

El valor de la función de discriminación la observación $X' = (X_1, X_2, X_3, X_4)$ viene dada por

$$\begin{aligned} \hat{Y} &= \hat{b}'X = (\bar{X}_{(1)} - \bar{X}_{(2)})'\boldsymbol{S}^{-1}X \\ &= \begin{pmatrix} 3.82 & 4.24 & 2.99 & 3.22 \end{pmatrix} \begin{pmatrix} 11.2553 & 9.4042 & 7.1489 & 3.3830 \\ 9.4042 & 13.5318 & 7.3830 & 2.5532 \\ 7.1489 & 7.3830 & 11.5744 & 2.6170 \\ 3.3830 & 2.5532 & 2.6170 & 5.8085 \end{pmatrix}^{-1} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \\ &= 0.030X_1 + 0.204X_2 + 0.010X_3 + 0.443X_4. \end{aligned}$$

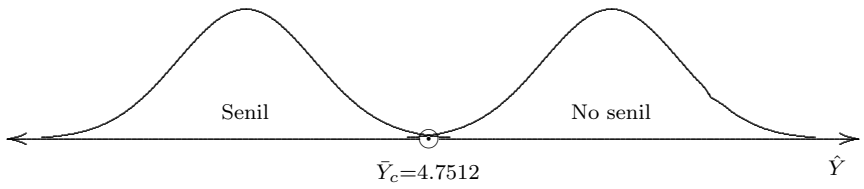


Figura 8.2 *Discriminación en senil o no senil.*

Para ubicar a un individuo en alguno de los dos grupos (senil o no senil) se utilizan los criterios expuestos en (8.7) y (8.8),

$$\bar{X}_c' = \frac{1}{2}(\bar{X}_1 + \bar{X}_2) = (10.66, 7.45, 9.99, 6.36),$$

como

$$\hat{b}'X_c = (0.030, 0.204, 0.010, 0.443)(10.66, 7.45, 9.99, 6.36)' = 4.7512.$$

Se asigna un individuo al grupo *no senil*, si la función de discriminación estimada $\hat{Y}_i \geq 4.7512$, y a la categoría *senil* si $\hat{Y}_i < 4.7512$ (figura 8.2). Supóngase que un individuo obtuvo los puntajes contenidos en el vector $X_0 = (10, 8, 7, 5)$, el valor de la función de discriminación en este caso es $\hat{Y} = \hat{b}'X_0 = 4.2115$ dado que este valor es menor que 4.7512, el individuo debe ser considerado como perteneciente al grupo senil. ☑

Observación:

- Se nota alguna semejanza entre el modelo de regresión lineal y la función discriminante. Aunque en algunos cálculos son parecidos, estas técnicas tienen algunas diferencias estructurales como las siguientes:

- En primer lugar, en el análisis de regresión se asume que la variable dependiente se distribuye normalmente y los regresores se consideran fijos. En análisis discriminante la situación es al revés, las variables independientes se asumen distribuidas normalmente y la variable respuesta se asume fija, la cual toma los valores cero o uno, según la ubicación del objeto en alguno de los dos grupos.

- En segundo término, el objetivo principal del análisis de regresión es predecir la respuesta media con base en el conocimiento de algunos valores fijos de un conjunto de variables explicativas; en cambio, el análisis discriminante pretende encontrar una combinación lineal de variables independientes, que minimicen la probabilidad de clasificar incorrectamente objetos en sus respectivos grupos.

- Finalmente, el análisis de regresión propone un modelo formal, sobre el que se hacen ciertos supuestos, con el fin de generar estimadores de los parámetros que tengan algunas propiedades deseables. El análisis discriminante busca un procedimiento para asignar o clasificar casos a grupos.

▶ Clasificación en poblaciones con matrices de covarianzas distintas

Si las dos poblaciones G_1 y G_2 tienen distribución normal p-variante con matrices de covarianzas distintas $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, el logaritmo de la razón de la verosimilitud para una observación particular X es el siguiente

$$\begin{aligned} Q(X) &= \ln\left(\frac{L_1(X)}{L_2(X)}\right) \\ &= \ln\left(\frac{(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_1|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(X-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(X-\boldsymbol{\mu}_1)\}}{(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_2|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(X-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(X-\boldsymbol{\mu}_2)\}}\right) \\ &= \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) - \frac{1}{2}(X-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1{}^{-1}(X-\boldsymbol{\mu}_1) + \frac{1}{2}(X-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2{}^{-1}(X-\boldsymbol{\mu}_2) \\ &= \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) - \frac{1}{2}\big(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1{}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2{}^{-1}\boldsymbol{\mu}_2\big) \\ &\quad + \big(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1{}^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2{}^{-1}\big)X - \frac{1}{2}X'\big(\boldsymbol{\Sigma}_1{}^{-1} - \boldsymbol{\Sigma}_2{}^{-1}\big)X. \end{aligned} \tag{8.9}$$

De acuerdo con este desarrollo, $Q(X)$ se puede escribir como:

$$Q(X) = \beta + \gamma X + X\Lambda X \tag{8.9a}$$

la cual expresa la forma cuadrática de la regla de clasificación contenida en (8.9), con:

$$\begin{aligned} \beta &= \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) - \frac{1}{2}\big(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1{}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2{}^{-1}\boldsymbol{\mu}_2\big) \\ \gamma &= \big(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1{}^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2{}^{-1}\big) \\ \Gamma &= \big(\boldsymbol{\Sigma}_1{}^{-1} - \boldsymbol{\Sigma}_2{}^{-1}\big). \end{aligned}$$

En la expresión $Q(X)$ el último término, $X'\big(\boldsymbol{\Sigma}_1{}^{-1} - \boldsymbol{\Sigma}_2{}^{-1}\big)X$, corresponde a los cuadrados y productos cruzados de las componentes del vector X, $Q(X)$ se denomina *función de discriminación cuadrática*. Nótese que si $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ entonces $Q(X)$ coincide con la función de discriminación lineal.

El criterio para clasificar una observación X es el siguiente:

si $Q(X) \geq 0$, entonces X se asigna a G_1; o

si $Q(X) < 0$, entonces X se asigna a G_2

En términos muestrales, si se obtiene una muestra de la población G_1 y una de la población G_2, se calcula un valor muestral de $\hat{Q}(X)$ al reemplazar $\boldsymbol{\mu}_i$ por $\bar{X}_i$ y $\boldsymbol{\Sigma}_i$ por $\boldsymbol{S}_i$ (i-ésimo grupo), ésta es:

$$\begin{aligned} \hat{Q}(X) =& \frac{1}{2}\ln\left(\frac{|\boldsymbol{S}_2|}{|\boldsymbol{S}_1|}\right) - \frac{1}{2}\big(\bar{X}_1'\boldsymbol{S}_1{}^{-1}\bar{X}_1 - \bar{X}_2'\boldsymbol{S}_2{}^{-1}\bar{X}_2\big) + \big(\bar{X}_1'\boldsymbol{S}_1{}^{-1} - \bar{X}_2'\boldsymbol{S}_2{}^{-1}\big)X \\ & - \frac{1}{2}X'\big(\boldsymbol{S}_1{}^{-1} - \boldsymbol{S}_2{}^{-1}\big)X. \end{aligned} \tag{8.10}$$

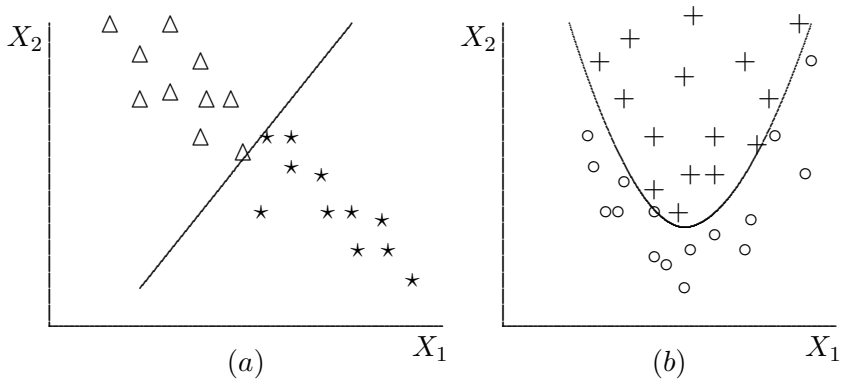


Figura 8.3 *Discriminación: (a) lineal, (b) cuadrática.*

Se observa que $\hat{Q}(X)$ tiene forma cuadrática, la cual se expresa en forma general como:

$$\hat{Q}(X) = b + c'X - X'\boldsymbol{A}X.$$

La regla para clasificar una observación muestral X es similar al caso poblacional como se indica en el recuadro anterior; es decir, se asigna la observación o individuo X al grupo G_1 si $\hat{Q}(X) \geq 0$; y al grupo G_2 en caso contrario.

Cuando $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, la función de clasificación cuadrática $\hat{Q}(X)$ es óptima de manera asintótica; aunque para muestras de tamaño pequeño $\boldsymbol{S}_i$ no es un estimador estable de $\boldsymbol{\Sigma}_i$, es decir, $\boldsymbol{S}_i$ varía bastante en muestras de la misma población o grupo. En tales casos Rencher (1998, pág. 233) recomienda emplear la regla de discriminación lineal. Para muestras de tamaño grande y con amplias diferencias entre $\boldsymbol{\Sigma}_1$ y $\boldsymbol{\Sigma}_2$, la función de discriminación cuadrática es la más recomendable.

En las figuras 8.3a y 8.3b se muestra la ubicación de las variables en el plano $X_1 \times X_2$. Para los datos de la figura 8.3a, donde $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ ya que la forma de las nubes de puntos es similar, es conveniente una regla de discriminación lineal; mientras que para los datos de la figura 8.3b, donde $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, la discriminación lineal no es conveniente, pues las observaciones están superpuestas, en este caso la discriminación de tipo cuadrático resulta más apropiada.

8.2.2 Regla de discriminación bayesiana

Hay situaciones en las que se pueden considerar probabilidades a priori para las poblaciones. Para dos poblaciones, sea p_i la probabilidad de que una observación provenga de la población G_i, $i = 1, 2$, con $p_1 + p_2 = 1$. Por ejemplo, con base en un diagnóstico clínico, se puede considerar a la gripe con más probabilidad de ocurrencia que el polio para un grupo humano determinado.

La regla de discriminación de Bayes localiza una observación X en la población con más alta probabilidad condicional, así por la regla ligada al teorema de Bayes, la observación X se asigna a la población G_1 si

$$\frac{p_1 f_1(X)}{p_1 f_1(X) + p_2 f_2(X)} \geq \frac{p_2 f_2(X)}{p_1 f_1(X) + p_2 f_2(X)}, \tag{8.11}$$

en caso contrario se asigna a G_2.

De la desigualdad (8.11), una regla equivalente es

$$\begin{aligned} &\text{Asignar } X \text{ a } G_1 \text{ si: } p_1 f_1(X) \geq p_2 f_2(X) \\ &\text{Asignar } X \text{ a } G_2 \text{ si: } p_1 f_1(X) < p_2 f_2(X). \end{aligned} \tag{8.11a}$$

Para dos o más poblaciones, como se muestra en la sección (8.3), la observación X se ubica en la población para la cual se maximiza

$$p_i L_i(X), \text{ con } i = 1, \ldots, k. \tag{8.12}$$

Nótese que el criterio dado en (8.1) es un caso especial de la regla de discriminación de Bayes cuando las probabilidades a priori son iguales.

Hasta aquí no se ha tenido en cuenta el problema de la clasificación incorrecta, ni los costos que implicarían clasificaciones erróneas. La tabla siguiente, ilustra los diferentes casos en la asignación de una observación X, las celdas indican el costo correspondiente.

	Decisión estadística	
Población	G_1	G_2
G_1	0	$C(2\|1)$
G_2	$C(1\|2)$	0

La regla para la ubicación de una observación X, que considere los costos de una clasificación incorrecta se obtienen de (Anderson, 1984 pág. 201):

$$\begin{aligned} &\text{Asignar } X \text{ en el grupo } G_1 \text{ si: } \frac{f_1(X)}{f_2(X)} \geq \frac{C(1|2)p_2}{C(2|1)p_1};\ o \\ &\text{Asignar } X \text{ en el grupo } G_2 \text{ si: } \frac{f_1(X)}{f_2(X)} < \frac{C(1|2)p_2}{C(2|1)p_1}. \end{aligned} \tag{8.13}$$

Si las dos poblaciones son multinormales, con la misma matriz de covarianzas, la decisión de asignación se toma de acuerdo con la siguiente regla (que es una consecuencia de la regla (8.13)):

Asignar X a G_1 si:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}X - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq \ln \frac{C(1|2)p_2}{C(2|1)p_1} = \ln k, \tag{8.14}$$

en caso contrario se asigna a G_2.

Los casos anteriores son situaciones particulares de éste último, allí se consideran poblaciones equiprobables con costos de clasificación incorrecta iguales; es decir, $k = 1$ y por consiguiente $\ln k = 0$.

8.3 Reglas de discriminación para varios grupos

Hasta ahora se ha considerado la clasificación de observaciones en el caso de sólo dos poblaciones. La práctica enfrenta al investigador con la clasificación de observaciones en varias poblaciones, por ejemplo, una entidad financiera puede estar interesada en clasificar a los solicitantes de tarjetas de crédito en varias categorías de riesgo. A las personas se les podría asignar, con base en el perfil e historial crediticio, en una de varias categorías de riesgo. Así, a un grupo de solicitantes no se les ofrece tarjetas de crédito, a un segundo grupo se le asigna tarjeta de crédito con un límite de 1000 unidades monetarias, a un tercer grupo se le asigna tarjeta de crédito con un límite de 3000 unidades monetarias, a un cuarto grupo de 6000 unidades monetarias, etc.

Se considera ahora el caso de muestras obtenidas a partir de k grupos independientes $G_1, G_2, \ldots, G_k$. Se desarrollan reglas de discriminación para el caso de varias poblaciones que tienen matrices de covarianzas igual o distinta, respectivamente.

8.3.1 Grupos con matrices de covarianzas iguales

Cuando se muestrea, varias poblaciones normales con matrices de covarianzas iguales, las funciones de discriminación óptima son lineales. Estas funciones de clasificación se obtienen aquí.

Si $p_1, p_2, \ldots, p_k$ son las probabilidades a priori de que una observación X proceda de la población $G_1, G_2, \ldots, G_k$, respectivamente; la regla de clasificación óptima, conociendo las funciones de densidad es la siguiente:

Asignar X a la población G_i si $p_i f_i(X) \geq p_j f_j(X)$, para todo $j = 1, \ldots, k$;

es decir, como en la ecuación (8.10), $p_i f_i(X) = max_j\{p_j f_j(X)\}$. Maximizar $p_i f_i(X)$ es equivalente a maximizar $\ln(p_i f_i(X))$. Si $X \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, se obtiene

$$\ln(p_i f_i(X)) = \ln(p_i) - \frac{1}{2} p \ln(2\pi) - \frac{1}{2} \ln|\boldsymbol{\Sigma}| - \frac{1}{2}(X - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1}(X - \boldsymbol{\mu}_i),$$

donde $\boldsymbol{\Sigma}$ es la varianza común a las k poblaciones. Nótese que cuando no hay información a priori sobre las p_i, se opta por asumir que éstas son iguales (distribución no informativa), y estas cantidades p_i desaparecen de la regla de clasificación; la regla de discriminación es entonces la de *máxima verosimilitud*. Además, se debe advertir que p es el número de variables, mientras que los p_i son las probabilidades a priori. Al desarrollar los cálculos algebraicos sobre la expresión anterior, se obtiene

$$\ln(p_i) + \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} X - \frac{1}{2} \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i.$$

Para observaciones muestrales (n_i por grupo), se asigna la observación X al grupo para el cual se maximice

$$\mathcal{D}_i = \ln(p_i) - \frac{1}{2}(X - \bar{X})'_i \boldsymbol{S}_p{}^{-1}(X - \bar{X}), \tag{8.15}$$

donde

$$\boldsymbol{S}_p = \frac{\sum_{i=1}^{k}(n_i - 1)\boldsymbol{S}_i}{\sum_{i=1}^{k}(n_i - 1)};$$

es una expresión semejante a la presentada en la sección (3.5.3) para estimar la matriz de covarianzas común a las k poblaciones.

Puesto que maximizar una función exponencial equivale a minimizar el exponente con signo negativo de la función, una expresión equivalente a (8.15) es:

$$\mathcal{D}_i^* = \frac{1}{2}(X - \bar{X})'_i \boldsymbol{S}_p{}^{-1}(X - \bar{X}) - \ln(p_i). \tag{8.15a}$$

Si se asume igual probabilidad a priori (p_i), entonces la observación X se asigna al grupo G_i que produzca el mayor valor $\mathcal{D}_i$. Alternativamente, se puede definir $\mathcal{D}_{ij} = \mathcal{D}_i - \mathcal{D}_j$, tal que la regla de asignación, por ejemplo para $i = 1, 2, 3$, sea:

- Asignar a G_1 si $\mathcal{D}_{12} > 0$ y $\mathcal{D}_{13} > 0$ (región $\mathcal{R}_1$).
- Asignar a G_2 si $\mathcal{D}_{12} < 0$ y $\mathcal{D}_{23} > 0$ (región $\mathcal{R}_2$).
- Asignar a G_3 si $\mathcal{D}_{13} < 0$ y $\mathcal{D}_{23} < 0$ (región $\mathcal{R}_3$).

De esta manera, el espacio de los individuos es dividido en tres regiones de discriminación, cuyas fronteras vienen dadas por las reglas de asignación $\mathcal{D}_{ij}$. En la figura 8.4 se muestran las regiones $\mathcal{R}_1$, $\mathcal{R}_2$ y $\mathcal{R}_3$ de discriminación para el caso de dos variables, X_1 y X_2 ($p = 2$) y tres grupos ($k = 3$).

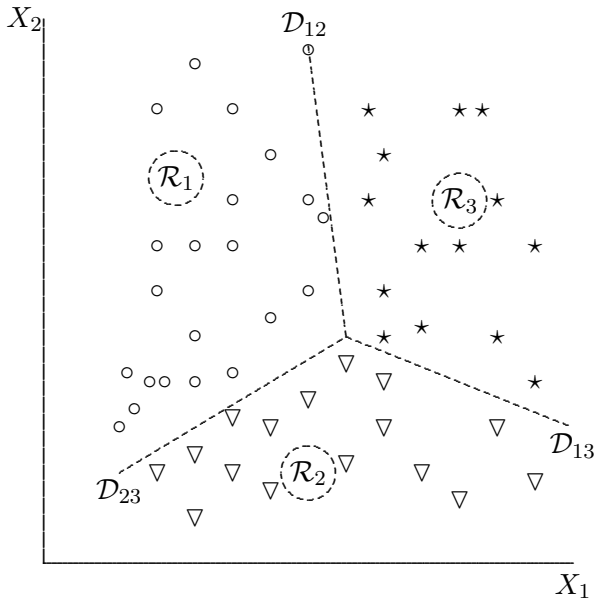


Figura 8.4 *Regiones de discriminación para tres grupos.*

La regla de discriminación bayesiana presentada en la sección (8.2.2) para dos grupos puede extenderse a varios grupos. La probabilidad a posteriori del i-ésimo grupo, dada la observación X, es:

$$P(G_i|X) = \frac{p_i f_i(X|G_i)}{\sum_{i=1}^{k} p_i f_i(X|G_i)}. \tag{8.16}$$

Algunos paquetes estadísticos, tales como SAS o SPSS, suministran la probabilidad a posteriori para cada observación X_{ij} (i-ésima observación del j-ésimo grupo). Éstas se calculan sustituyendo $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}$ por sus respectivos estimadores en (8.16). Así, para el caso de distribuciones multinormales $f(X|G_i) = N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$,

$$P(G_i|X) = \frac{p_i \exp\{-\frac{1}{2}D_i^2\}}{\sum_{i=1}^{k} p_i \exp\{-\frac{1}{2}D_i^2\}}, \tag{8.17}$$

donde $D_i = (X - \bar{X}_i)' \boldsymbol{S}_p{}^{-1}(X - \bar{X}_i)$ es la distancia de Mahalanobis de la observación X al centroide del i-ésimo grupo. Se asigna X al grupo con mayor probabilidad a posteriori.

8.3.2 Grupos con matrices de covarianzas distintas

Si se emplea la función de discriminación lineal para grupos con matrices de covarianzas distintas, las observaciones tienden a ser clasificadas en los grupos que tienen varianzas altas. De cualquier forma, la regla de clasificación puede modificarse conservando de manera óptima la clasificación, en términos de los errores de clasificación.

Considerando k-poblaciones de p-variables cada una, distribuidas $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, y cada una con probabilidades a priori $p_1, \ldots, p_k$, respectivamente, se tiene:

$$\ln[p_i f(X|G_i)] = \ln(p_i) - \frac{1}{2}p\ln(2\pi) - \frac{1}{2}|\boldsymbol{\Sigma}_i| - \frac{1}{2}(X - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(X - \boldsymbol{\mu}_i).$$

Para una muestra se emplea el vector de medias muestral $\bar{X}_i$ y la matriz de covarianzas muestral $\boldsymbol{S}_i$, para cada uno de los k-grupos. Omitiendo el término constante $-(p/2)\ln(2\pi)$, se obtiene la función de discriminación cuadrática,

$$Q_i(X) = \ln(p_i) - \frac{1}{2}|\boldsymbol{S}_i| - \frac{1}{2}(X - \bar{X}_i)'\boldsymbol{S}_i^{-1}(X - \bar{X}_i). \tag{8.18}$$

La regla de clasificación es: asignar la observación X al grupo para el cual $Q_i(X)$ *sea la más grande*. Una regla equivalente a (8.18) es considerar $-Q_i(X) = Q_i(X)^*$, la cual se escribe como

$$Q_i^*(X) = \frac{1}{2}(X - \bar{X}_i)'\boldsymbol{S}_i^{-1}(X - \bar{X}_i) + \frac{1}{2}\ln|\boldsymbol{S}_i| - \ln(p_i). \tag{8.18a}$$

Si las probabilidades a priori p_i son iguales o no se conocen, el término $\ln(p_i)$ puede descartarse de la función de discriminación. Nótese que para

que exista $\boldsymbol{S}_i$, se debe satisfacer que $n_i > p$, con $i = 1, \ldots, k$; es decir que el número de observaciones en cada grupo debe ser mayor que el número de variables.

Para poblaciones multinormales con matrices de covarianzas desiguales $\boldsymbol{\Sigma}_i$, la probabilidad a posteriori (bayesiana), empleando los estimadores de $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}_i$, está dada por:

$$P(G_i|X) = \frac{p_i|\boldsymbol{S}_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2}D_i^2\}}{\sum_{i=1}^{k} p_i|\boldsymbol{S}_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2}D_i^2\}}, \tag{8.19}$$

donde $D_i^2 = (X - \bar{X}_i)'\boldsymbol{S}_i^{-1}(X - \bar{X}_i)$. Aunque en la mayoría de las aplicaciones los valores de p_i no se tienen, algunos paquetes estadísticos los estiman como una proporción de los tamaños de muestra n_i; este procedimiento no es muy recomendado, a menos que las proporciones muestrales representen las proporciones poblacionales.

Para tamaños muestrales grandes, la función de discriminación cuadrática clasifica mejor que las lineales. Para muestras de tamaño pequeño, los resultados desde la discriminación cuadrática son menos estables en muestreos secuenciales o repetitivos que los resultados de la discriminación lineal; pues se deben estimar más parámetros en $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_k$ que en $\boldsymbol{S}_p$ y porque cada $\boldsymbol{S}_i$ tiene asociado algunos pocos grados de libertad de $\boldsymbol{S}_p$.

La sensibilidad a la no multinormalidad se observa también en la regla de discriminación cuadrática. Velilla y Barrio (1994) sugieren una transformación de los datos para aplicar la regla de discriminación lineal o cuadrática.

8.4 Tasas de error de clasificación

Una vez que se ha obtenido una regla de clasificación, la inquietud natural es acerca de *qué tan buena* es la clasificación generada a través de esta regla. Es decir, se quiere saber la *tasa de clasificación correcta*, referida como la probabilidad de clasificar una observación en el grupo al que verdaderamente pertenece. De manera complementaria, se tienen las *tasas de error* por clasificación incorrecta. El interés está en la probabilidad de que la regla de discriminación disponible clasifique incorrectamente una futura observación; de otra forma, se quiere evaluar la capacidad de la regla para predecir el grupo a que pertenece una observación.

La siguiente tabla ilustra la calidad de las posibles decisiones que se podrían tomar, con relación a la clasificación de objetos en uno de dos grupos.

	Decisión estadística	
Grupo	Asignar a G_1	Asignar a G_2
G_1 (n_1)	Decisión correcta (n_{11})	Error (n_{12})
G_2 (n_2)	Error (n_{21})	Decisión correcta (n_{22})

8.4.1 Estimación de las tasas de error

Un estimador simple de la tasa de error se obtiene al tratar de clasificar los objetos del mismo conjunto que se empleó para la construcción de la regla de clasificación. Este método se conoce como *resustitución*. A cada observación X_i se le aplica la función de clasificación y se asigna a uno de los grupos. Se cuentan entonces el número de clasificaciones correctas y el número de clasificaciones incorrectas conseguidas con la regla. La proporción de clasificaciones incorrectas se denomina la *tasa de error aparente*. Los resultados se disponen en una tabla como la siguiente.

Entre las n_1 observaciones de G_1, n_{11} son clasificadas correctamente en G_1 y n_{12} son clasificadas incorrectamente en G_2, con $n_1 = n_{11} + n_{12}$. Análogamente, de las n_2 observaciones de G_2, n_{21} son asignadas incorrectamente a G_1 y n_{22} son correctamente asignadas a G_2, con $n_2 = n_{21} + n_{22}$. De esta forma, la tasa de error aparente es

$$\textit{Tasa de error aparente} = \frac{n_{12} + n_{21}}{n_1 + n_2} = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}. \qquad (8.20)$$

El método de resustitución puede extenderse al caso de varios grupos, la tasa de error aparente es fácil de calcular, aunque la mayoría de los paquetes estadísticos la suministran. Esta tasa es un estimador de la probabilidad de que la función de clasificación, encontrada a partir de los datos, clasifique incorrectamente una observación. Tal probabilidad se denomina *tasa actual de error* (TAE). Si p_1 y p_2 son las probabilidades a priori para los grupos G_1 y G_2, respectivamente, la tasa actual de error es:

$$TAE = p_1 P(\textit{Asignar a } G_1 | G_2) + p_2 P(\textit{Asignar a } G_2 | G_1), \qquad (8.21)$$

donde $P(\textit{Asignar a } G_1|G_2)$ significa la probabilidad de clasificar X en el grupo G_1 cuando realmente procede del grupo G_2; una definición análoga se tiene para $P(\textit{Asignar a } G_2|G_1)$.

La definición de tasa actual de error se estima para procedimientos de clasificación basados en una muestra. Aunque se puede estar interesado en calcular la tasa actual esperada de error (TAEE) basados sobre todas las posibles muestras, es decir,

$$TAEE = p_1\mathcal{E}[P(Asignar\ a\ G_1|G_2)] + p_2\mathcal{E}[P(Asignar\ a\ G_2|G_1)]. \quad (8.22)$$

En el cálculo de (8.21) o de (8.22) se necesita conocer los parámetros poblacionales y asumir una distribución particular de los datos. Pero en la mayoría de los casos los parámetros poblacionales son desconocidos; y por tanto se requiere de algunos estimadores de las tasas de error. McLachlan (1992, págs. 337-377) suministra éstos y otros estimadores.

8.4.2 Corrección del sesgo de las estimaciones para las tasas de error aparente

Para muestras de tamaño grande la tasa de error aparente, como estimador de la tasa de error actual, tiene un sesgo pequeño. Para muestras de tamaño pequeño la situación, respecto a la disminución del sesgo, no es muy alagüeña. Se revisan a continuación algunas técnicas que permiten reducir el sesgo en la estimación de la tasa de error aparente.

▶ Partición de la muestra

Una forma de controlar el sesgo es mediante la división de la muestra en dos partes. Una de ellas (muestra de *ensayo*) se emplea para construir la regla de clasificación, mientras que la otra (muestra de *validación*) se utiliza para evaluar la bondad de la regla calculada. La regla de clasificación se evalúa en cada una de las observaciones de la muestra de validación. Como estas observaciones no se emplearon en la construcción de la regla de clasificación, entonces la tasa de error resultante es insesgada. Una forma de mejorar la estimación de las tasas de error es mediante el intercambio del papel de las dos muestras, de tal modo que la regla de clasificación se obtiene a partir de la muestra de validación, y la validación es hecha a partir de la muestra de ensayo; la tasa de error estimada se obtiene entonces como el promedio de las dos tasas de error calculadas.

Este procedimiento tiene fundamentalmente dos desventajas:

1. Se requiere de muestras de tamaño grande, las cuales pueden no ser alcanzables.

2. No evalúa la función de clasificación sobre la muestra completa, en consecuencia, las tasas de error tendrán varianzas más grandes que

las obtenidas con la muestra completa. Es decir, se debe decidir entre estimadores con sesgo pequeño o estimadores con varianza pequeña.

▶ Validación cruzada

Este procedimiento se puede considerar como un caso especial del anterior, pues se toman $(n-1)$ observaciones para construir la regla de clasificación y luego con ella se clasifica la observación omitida. Este procedimiento se repite una vez por cada observación (en total n veces). Aunque algunos califican a este método como de tipo *jacknife*, Seber (1984, pág. 289) dice que tal calificación es incorrecta. Rencher (1998, pág. 244) se refiere a la ventaja mostrada por este método usando el procedimiento de Monte Carlo.

▶ Estimación "Bootstrap"

El estimador de la tasa de error vía *bootstrap* es esencialmente una corrección del sesgo para la tasa error aparente, basados sobre un remuestreo de la muestra original (Efron y Tibshirani, 1993). Se describe este procedimiento para el caso de dos grupos con muestras de tamaños n_1 y n_2. En la primera muestra se toma una muestra aleatoria de tamaño n_1 con reemplazamiento. Se puede presentar que algunas observaciones de la muestra original no aparezcan en la nueva muestra, mientras que otras aparecerán más de una vez. De manera similar se remuestrea el segundo grupo. Con las dos "nuevas" muestras se *recalculan* las funciones de clasificación y con ésta se clasifican tanto las muestras originales como las nuevas. Las tasas de error en la clasificación para cada grupo se calculan con

$$d_i = \frac{e_{i.orig.} - e_{i.nva.}}{n-i}, \quad i = 1, 2; \tag{8.23}$$

donde $e_{i.orig.}$ es el número de observaciones del i-ésimo grupo original incorrectamente clasificadas y $e_{i.nva.}$, es el número de observaciones de la i-ésima muestra nueva que fueron mal clasificadas. Este procedimiento se desarrolla un buen número de veces (se sugieren entre 100 y 200 repeticiones) y se emplea el promedio de los d_i como corrector del término de sesgo, así:

$$Tasa\ de\ error\ bootstrap = tasa\ de\ error\ aparente + \bar{d}_1 + \bar{d}_2. \tag{8.24}$$

Ejemplo 8.2 Se quiere encontrar una regla para discriminar entre cuatro grupos de semillas de trigo.[1] Los grupos se definen de acuerdo con el sitio

[1] Tomado de Johnson (2000, págs. 235-243)

de cultivo y con la variedad del trigo. Así, los grupos 1 y 2 se corresponden con dos variedades (ARKAN y ARTHUR) cultivadas en un primer sitio (MAS0), mientras que los grupos 3 y 4 se corresponden con las mismas variedades cultivadas en un segundo sitio (VLAD12).

La investigación apunta a encontrar una manera (regla) de identificar las semillas de trigo con base en medidas físicas tales como: área, perímetro, longitud y ancho de cada grano.

Cada grano tiene un pliegue, de manera que se optó por tomar medidas tanto con el pliegue a la derecha como con el pliegue hacia abajo. Las variables que se midieron sobre el grano cuando el pliegue estaba hacia abajo son la raíz cuadrada del área Ar_1, perímetro Pe_1, longitud Lo_1 y el ancho An_1. Las variables Ar_2, Pe_2, Lo_2 y An_2 se definen de manera análoga, excepto que el grano se midió con el pliegue a la derecha. Las mediciones se obtuvieron mediante un analizador de imágenes. La tabla 8.2 esquematiza, con dos observaciones por grupo, la base de datos, aunque la base completa contiene 36 observaciones en cada uno de los grupos 1 y 2 y de a 50 para los grupos 3 y 4.

Tabla 8.2 Medidas sobre granos de trigo

Obs	Sitio	Variedad	Grupo	Ar_1	Pe_1	Lo_1	An_1	Ar_2	Pe_2	Lo_2	An_2
1	MAS0	ARKAN	1	54.418	219	89	43	56.6039	226	89	47
2	MAS0	ARKAN	1	55.1453	221	91	46	56.2583	224	91	46
37	MAS0	ARTHUR	2	50.4975	205	85	41	50.8724	215	86	42
38	MAS0	ARTHUR	2	52.4118	212	89	42	54.0185	217	91	44
73	VLAD12	ARKAN	3	52.4690	217	92	40	54.7175	221	93	44
74	VLAD12	ARKAN	3	56.7803	234	89	45	56.7891	230	95	46
123	VLAD12	ARTHUR	4	53.9907	220	100	43	50.1996	218	93	37
124	VLAD12	ARTHUR	4	56.6480	220	88	48	53.4509	213	87	44

Fuente: Johnson (2000, págs. 235-243)

Los cálculos para el ejemplo se desarrollan con el apoyo del procedimiento *DISCRIM* del paquete SAS. Una primera tarea es la verificación de la hipótesis sobre la igualdad de las cuatro matrices de covarianzas (sección (4.3.2)). Mediante la opción *POOL=TEST* del procedimiento DISCRIM se hacen los cálculos para decidir sobre el rechazo o no rechazo de la hipótesis

$$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3 = \mathbf{\Sigma}_4.$$

Con la opción *PRIORS PROP* se asignan probabilidades a priori para cada grupo, las cuales corresponden a la razón entre el número de observaciones

por grupo y el total de observaciones $(n_i/\sum_{i=1}^{k} n_i)$. Con la opción *CROSS-VALIDATE* se obtienen las estimaciones de las probabilidades de una clasificación incorrecta, mientras que con la instrucción *CROSSLIST* se produce una lista que indica el grupo en el que podría clasificarse cada una de los vectores de observaciones por el método de calibración cruzada.

Al final de este capítulo, en la sección (8.7), se escribe la sintaxis para el procedimiento *DISCRIM*.

Si no se rechaza la hipótesis de igualdad de las cuatro matrices de covarianzas, entonces la observación X se asigna, de acuerdo con (8.15a), al grupo para el cual $\mathcal{D}_i^*$ sea mínimo. En caso de rechazar la hipótesis de igualdad de las cuatro matrices de covarianzas el valor mínimo de $Q_i^*(X)$, expresión (8.18a), sugiere el grupo al cual se debe asignar la observación X.

De acuerdo con los datos las probabilidades a priori son:

$$p_1 = \frac{n_1}{n} = \frac{36}{172} = 0.209302 = p_2$$
$$p_3 = \frac{n_3}{n} = \frac{50}{172} = 0.290698 = p_4.$$

Para la verificación de la hipótesis $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_4$, de acuerdo con las expresiones (4.11) a (4.14) y con la salida del procedimiento DISCRIM, se tiene:

$$\boldsymbol{\varphi} = -2\boldsymbol{\rho}\ln(\boldsymbol{\lambda}_{1_n}) = 457.642902$$

que para una distribución ji-cuadrado con $p(p+1)(q-1)/2 = 108$ grados de libertad tiene un $p-valor$ igual a 0.0001, con lo cual se rechaza la hipótesis de igualdad de las matrices de covarianzas. En consecuencia la regla de clasificación adecuada es la contenida en la expresión (8.18a).

Las tablas 8.3 y 8.4 contienen las frecuencias y las tasas de clasificación incorrecta, de las semillas de trigo, de acuerdo con el método de *resustitución* y *clasificación cruzada*, respectivamente.

En la tabla 8.3 se muestra cómo serían clasificadas las semillas de los grupos mediante el método de resustitución. Se puede apreciar que la regla de discriminación clasifica de manera correcta a 66.67% las observaciones del grupo 1, 82% de las observaciones del grupo 3 y 94% de las observaciones del grupo 4, mientras que sólo el 22.22% de las del grupo 2 son correctamente asignadas.

Tabla 8.3 Número de observaciones y tasas de clasificación por resustitución

	Clasificación al grupo				
Del grupo	1	2	3	4	*Total*
1	24	1	9	2	36
	66.67	2.78	25.00	5.56	100.00
2	2	8	2	24	36
	5.56	22.22	5.56	66.67	100.00
3	5	1	41	3	50
	10.00	2.00	82.00	6.00	100.00
4	0	2	1	47	50
	0.00	4.00	2.00	94.00	100.00
Total	31	12	53	76	172
Porc.	18.02	6.98	30.81	44.19	100.00
Pr. a priori.	0.2093	0.2093	0.2907	0.2907	

Tabla 8.4 Número de observaciones y tasas de clasificación cruzada

	Clasificación al grupo				
Del grupo	1	2	3	4	*Total*
1	18	4	12	2	36
	50.00	11.11	33.33	5.56	100.00
2	2	7	2	25	36
	5.56	19.44	5.56	69.44	100.00
3	8	2	35	5	50
	16.00	4.00	70.00	10.00	100.00
4	0	4	3	43	50
	0.00	8.00	6.00	86.00	100.00
Total	28	17	52	75	172
Porc.	16.28	9.88	30.23	43.60	100.00
Pr. a priori.	0.2093	0.2093	0.2907	0.2907	

Debe tenerse en cuenta que posiblemente estas sean estimaciones sesgadas (por exceso) de las probabilidades verdaderas de asignación correcta, puesto que se obtienen de aplicar la regla sobre los mismos datos con que ésta fue construida.

Téngase presente que los grupos 1 y 3 corresponden a la misma variedad lo mismo que los grupos 2 y 4, de manera que la clasificación incorrecta se puede atribuir a los lugares y no tanto a las variedades.

Si se consideran las tasas de clasificación correcta por variedad, en el grupo 1 se clasifica 91.67% de las veces en la variedad correcta (pues 91.67 =

$66.67 + 25$), en el grupo 2 se clasifica el 88.89% de las veces, en el grupo 3 el 92% de las veces y en el grupo 4 el 98% de las veces.

Las estimaciones de las tasas verdaderas de clasificación correcta que se muestran en la tabla 8.4 son casi insesgadas, y por tanto mejores estimaciones que las obtenidas mediante el método de resustitución. En esta tabla se observa que el grupo 1 clasifica en la variedad correcta (grupos 1 y 3) el $50\% + 33.33\% = 83.33\%$ de las veces, el grupo 2 clasifica de esta forma el $19.44\% + 69.44\% = 88.88\%$ de las veces, el grupo 3 clasifica de la manera correcta el $16\% + 70\% = 86\%$ de las veces y el grupo 4 lo hace el $8\% + 86\% = 94\%$ de las veces. ☑

8.5 Otras técnicas de discriminación

8.5.1 Modelo de discriminación logística para dos grupos

Cuando las variables son discretas o son una mezcla de discretas y continuas, la discriminación a través del modelo logístico puede resultar adecuada.

Para distribuciones multinormales con $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, el logaritmo de la razón de densidades es

$$\begin{aligned} \ln \frac{f(X|G_1)}{f(X|G_2)} &= -\frac{1}{2} \underbrace{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}_{\boldsymbol{\alpha}} + \underbrace{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}}_{\boldsymbol{\beta}'} X \\ &= \boldsymbol{\alpha} + \boldsymbol{\beta}' X, \end{aligned} \tag{8.25}$$

la cual es una función lineal del vector observado X. Además de la normal multivariada, otras distribuciones multivariadas satisfacen (8.25), algunas de las cuales involucran vectores aleatorios discretos o mezcla de variables discretas y continuas. El modelo mostrado en la ecuación (8.25) se conoce como el *modelo logístico*, la regla para ubicar una observación X es: Asignar al grupo G_1 si

$$\boldsymbol{\alpha} + \boldsymbol{\beta}' X > \ln \frac{p_1}{p_2}, \tag{8.26}$$

y a G_2 en otro caso. Cuando las probabilidades a priori, p_1 y p_2, se pueden asumir iguales, el miembro izquierdo de la desigualdad (8.26) se compara contra el número cero. La *clasificación logística* es también referida como la *discriminación logística*.

La probabilidad a posteriori (sección (8.2.2)) en términos del modelo logístico, que señala la probabilidad de pertenencia de una observación X un grupo,

por ejemplo G_1, de acuerdo con el teorema de Bayes es:

$$\begin{aligned} P(G_1|X) &= \frac{p_1 f(X|G_1)}{p_1 f(X|G_1) + p_2 f(X|G_2)} \\ &= \frac{e^{\ln(p_1/p_2)+\alpha+\beta' X}}{1 + e^{\ln(p_1/p_2)+\alpha+\beta' X}} \\ &= \frac{e^{\alpha_0+\beta' X}}{1 + e^{\alpha_0+\beta' X}} = \frac{1}{1 + e^{-(\alpha_0+\beta' X)}}, \end{aligned} \tag{8.27}$$

donde $\alpha_0 = \ln(p_1/p_2) + \alpha$. De la expresión anterior se obtiene

$$P(G_2|X) = 1 - P(G_1|X) = \frac{1}{1 + e^{\alpha_0+\beta' X}}. \tag{8.28}$$

La estimación de $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$, se hace a través del método de mínimos cuadrados ponderados o mediante máxima verosimilitud para regresión logística (Seber 1984, págs. 312-315). La estimación conlleva a resolver sistemas de ecuaciones no lineales, cuya solución aproximada puede encontrarse con métodos numéricos tales como la técnica de "Newton-Raphson" o el método de "cuasi-Newton"; procedimientos incorporados en paquetes estadísticos como el SAS o el SPSS.

La figura 8.5 representa la función logística. Aquí se asigna la observación X al grupo G_1, si $P(G_1|X) \geq P(G_2|X)$ o al grupo G_2 en caso contrario. En general, para dos grupos, de acuerdo con la propiedad expresada en (8.28), se asigna la observación X al grupo G_i si $P(G_i|X) \geq 0.5$, $i = 1, 2$.

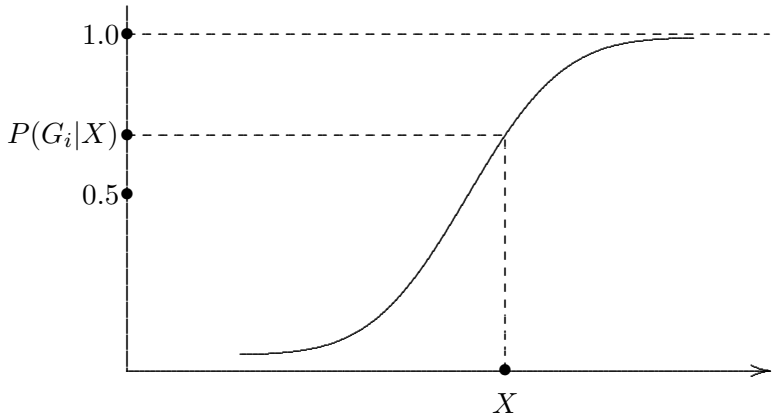


Figura 8.5 *Función logística.*

▶ Datos multinormales con $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$

Para estos datos la clasificación lineal es superior a la logística, sin embargo, para datos binarios, estos supuestos usualmente no se tienen y la clasificación a través de un modelo logístico resulta ser mejor. Ruiz-Velazco (1991) comparan la eficiencia del modelo logístico sobre el modelo de clasificación lineal.

Los modelos de discriminación logística lineal pueden emplearse en situaciones donde:

1. Las funciones de densidad sean multinormales con matrices de covarianzas iguales.
2. Las mediciones sean variables independientes tipo Bernoulli.
3. Las variables tipo Bernoulli sigan un modelo log-lineal con efectos de segundo orden o más iguales.
4. Situaciones 1 a 3 mezcladas.

Ejemplo 8.3 La clasificación logística se aplica, con buenos resultados, en investigación médica[2]. El objetivo, de esta ilustración, es predecir una trombosis postoperatoria de venas profundas, una condición es que se debe tratar a estos pacientes con anticoagulantes antes de la cirugía. Sin embargo, estos tratamientos producen problemas hemorrágicos en algunos pacientes, de donde resulta importante la identificación de los pacientes con más alto riesgo de trombosis.

De 124 pacientes en estudio, ninguno mostró evidencia preoperatoria de trombosis en venas profundas. Después de la intervención, 20 pacientes desarrollaron la condición (grupo G_1) y los 104 restantes no (grupo G_2). En el modelo logístico resultante se consideran, finalmente, cuatro variables continuas (X_1, X_2, X_3, X_4) y una variable discreta X_5. El modelo (8.25) estimado es

$$\begin{aligned}\omega &= \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}}' X \\ &= -11.3 + 0.009X_1 + 0.22X_2 + 0.085X_3 + 0.043X_4 + 2.19X_5.\end{aligned}$$

El valor de ω se calculó para cada uno de los 124 pacientes, reemplazando por los respectivos valores de X_1 a X_5. Si se aplica la regla de clasificación (8.25), con $p_1 = p_2$, los pacientes con $\omega > 0$ se asignan al grupo de trombosis de venas profundas (G_1). Con este procedimiento, 11 de los 124 pacientes

[2]Rencher (1998, págs. 255-256)

se clasificaron incorrectamente, es decir, con una tasa de error aparente de 9% (11/129). Sin embargo, usando el criterio de $\omega > 0$, se clasificaría incorrectamente a pacientes con alto riesgo (pues $p_1 > p_2$, para estos casos). Por tanto, se recomienda suministrar anticoagulante, antes de la cirugía, a los pacientes con $\omega > -2.5$. ☑

▶ Grupos con distribuciones multinormales donde $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Para estos datos la función logística no es lineal en los X, el logaritmo de la razón de densidades es

$$\begin{aligned} \ln \frac{f(X|G_1)}{f(X|G_2)} &= c_0 + (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1})X + \frac{1}{2}X'(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})X \\ &= c_0 + \boldsymbol{\delta}'X + X'\boldsymbol{\Delta}X, \end{aligned} \tag{8.29}$$

con

$$\begin{aligned} c_0 &= \frac{1}{2}\ln(|\boldsymbol{\Sigma}_2^{-1}|/|\boldsymbol{\Sigma}_1^{-1}|) - \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2), \\ \boldsymbol{\delta} &= (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2'\boldsymbol{\Sigma}_2^{-1}), \text{ y} \\ \boldsymbol{\Delta} &= (\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}). \end{aligned}$$

Aunque la función dada en (‘) no es lineal en los X, es lineal en los parámetros. Ésta se le conoce como la función *logística cuadrática*. Los parámetros se estiman mediante los mismos métodos iterativos citados anteriormente.

La función logística puede extenderse a varios grupos, puede emplearse para clasificar observaciones en varias poblaciones

8.5.2 Modelo de discriminación Probit

En algunos casos los grupos son definidos a través de un criterio *cuantitativo* en lugar de cualitativo. Por ejemplo, se puede particionar un grupo de estudiantes en dos grupos, con base en su promedio de rendimiento académico; que en un grupo se ubican los de rendimiento “alto” y en el otro los de rendimiento “bajo”. Con base en un vector X de puntajes y medidas, obtenidos para esta clase de estudiantes, se quiere predecir su pertenencia a uno de estos grupos.

A continuación se presentan los rasgos generales de la metodología. Sea Z una variable aleatoria continua, si t es un valor “umbral” o “límite”, entonces un individuo es asignado al grupo G_1 si $Z > t$ (por ejemplo, alto rendimiento) y si $Z \leq t$ se asigna al grupo G_2.

Para empezar se asume que el vector $(Z, X)'$ se distribuye $N_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, donde

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_Z \\ \boldsymbol{\mu}_X \end{pmatrix} \text{ y } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_Z^2 & \sigma_{ZX} \\ \sigma_{XZ} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}.$$

Por la propiedad 6 de la sección (2.2) (ecuaciones 2.2a y 2.2b), la distribución condicional de Z dado el vector X es normal con

$$\begin{aligned} \mathcal{E}(Z|X) &= \mu_{Z|X} = \mu_Z + \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}(X - \boldsymbol{\mu}_X), \\ \text{var}(Z|X) &= \sigma_{Z|X} = \sigma_Z^2 - \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XZ}. \end{aligned}$$

Por tanto,

$$\begin{aligned} P(G_1|X) &= P(Z > t|X) \\ &= P\left(\frac{Z - \mu_{Z|X}}{\sigma_{Z|X}} > \frac{t - \mu_{Z|X}}{\sigma_{Z|X}}\right) \\ &= 1 - \Phi\left(\frac{t - \mu_{Z|X}}{\sigma_{Z|X}}\right) \\ &= \Phi\left(\frac{-t + \mu_{Z|X}}{\sigma_{Z|X}}\right), \end{aligned}$$

donde $\Phi(\cdot)$ es la función de distribución normal estándar. De esta forma, reemplazando por las expresiones anteriores $\mu_{Z|X}$ y $\sigma_{Z|X}$, la probabilidad de que la observación X sea del grupo G_1 es

$$P(G_1|X) = \Phi\left[\frac{-t + \mu_Z + \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}(X - \boldsymbol{\mu}_X)}{\sqrt{\sigma_Z^2 - \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XZ}}}\right] = \Phi(\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 X), \quad (8.30)$$

donde

$$\begin{aligned} \boldsymbol{\gamma}_0 &= -(t - \mu_Z + \sigma'_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}(X - \boldsymbol{\mu}_X)/\sqrt{\sigma_Z^2 - \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XZ}}, \text{ y} \\ \boldsymbol{\gamma}_1 &= \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}/\sqrt{\sigma_Z - \boldsymbol{\sigma}_{ZX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\sigma}_{XZ}}. \end{aligned}$$

La regla de clasificación asigna la observación X al grupo G_1 si

$$P(Z > t|X) \geq P(Z < t|X);$$

es decir, si $P(G_1|X) \geq P(G_2|X)$, y al grupo G_2 en otro caso. De acuerdo con la expresión (8.30) la regla es:

Asignar la observación X al grupo G_1 si $\Phi(\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 X) \geq 1 - \Phi(\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 X)$,

lo cual equivale a que $\Phi(\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 X) \geq \frac{1}{2}$. En términos de $\gamma_0 + \gamma_1 X$, la regla puede expresarse como: asignar X la grupo G_1 si

$$\boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 X \geq 0, \tag{8.31}$$

y al grupo G_2 en el otro caso (figura 8.6). Los parámetros $\boldsymbol{\gamma}_0$ y $\boldsymbol{\gamma}_1$ se estiman a través del método de máxima verosimilitud (con soluciones iterativas), empleando una dicotomización del tipo: $\omega = 0$ si $Z \leq t$ y $\omega = 1$ si $Z > t$. No se requiere que X tenga una distribución multinormal, únicamente que la distribución condicional de Z dado X sea normal. Esto posibilita la inclusión en X de variables aleatorias discretas.

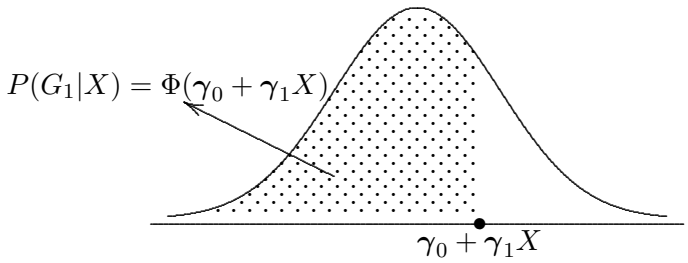


Figura 8.6 *Discriminación probit.*

8.5.3 Discriminación con datos multinomiales

La mayoría de los datos procedentes de encuestas corresponden a variables de tipo categórico. Las combinaciones de las categorías constituyen un resultado (valor) de una variable aleatoria multinomial. Por ejemplo, considérense las siguientes cuatro variables categóricas: género (masculino o femenino), credo político (liberal, conservador e independiente), tamaño de la ciudad de residencia (menos de 10.000 habitantes, entre 10.000 y 100.000 y más de 100.000) y nivel de escolaridad (primaria, media, universitaria y de posgrado). El número de posibles valores que toma esta variable multinomial es el producto del número de modalidades de cada una de las variables: $2 \times 3 \times 3 \times 4 = 72$. Para este caso, supóngase que se desea predecir si una persona votará en las próximas elecciones, después de habérsele observado alguna de las 72 categorías descritas anteriormente. De esta manera se tienen dos grupos: el grupo G_1 constituido por los *votantes* y el grupo G_2 por los *no votantes*.

De acuerdo con la regla de Welch (sección (8.2)), se asigna la observación X a G_1 si

$$\frac{f(X|G_1)}{f(X|G_2)} > \frac{p_2}{p_1}, \tag{8.32}$$

y a G_2 en caso contrario. En este ejemplo la expresión $f(X|G_1)$ se representa por q_{1i}, $i = 1, \ldots, 72$, y $f(X|G_2)$ por q_{2i}, $i = 1, \ldots, 72$, donde q_{1i} es la probabilidad de que una persona del grupo de votantes (G_1) quede en la categoría i, la definición es análoga para q_{2i}. La regla de clasificación (8.32), en términos de las probabilidades multinomiales, es: asignar a la persona identificada con el vector de observaciones X a la población G_1 si

$$\frac{q_{1i}}{q_{2i}} > \frac{p_2}{p_1}, \tag{8.33}$$

y a G_2 en el otro caso.

Si las probabilidades q_{1i} y q_{2i} se conocen, se reemplazan en la expresión (8.33) para cada una de las categorías $i = 1, \ldots, 72$; de tal forma que las 72 categorías se particionan en dos clases, una de las cuales se corresponde con individuos del grupo G_1 y la otra con individuos del grupo G_2.

En la práctica los valores para las probabilidades q_{1i} y q_{2i} no se conocen, éstos deben estimarse desde los datos muestrales; mientras que los valores de p_1 y p_2 se deben conocer a priori, en caso contrario se asumen iguales ($p_1 = p_2 = 0.5$). Supóngase que el número de indivuos de la i-ésima categoría en los grupos G_1 y G_2 es, respectivamente, n_{1i} y n_{2i}. Se estiman q_{1i} y q_{2i} mediante

$$\hat{q_{1i}} = \frac{n_{1i}}{N_1} \text{ y } \hat{q_{2i}} = \frac{n_{2i}}{N_2}, \tag{8.34}$$

donde $N_1 = \sum_i n_{1i}$ y $N_2 = \sum_i n_{2i}$ son el número de individuos en cada uno de los dos grupos.

Hay situaciones en donde las categorías o modalidades de las variables individuales admiten un orden. Si todas las variables tienen categorías ordenadas, entonces se les asigna un rango (puesto) a cada categoría, y de esta forma se trabaja de manera directa con los rangos y las reglas usuales de clasificación. Para el caso tratado, el tamaño de la ciudad y el grado de escolaridad son variables de este tipo, así por ejemplo, a las categorías de la variable escolaridad se les asignan los números 1, 2, 3 y 4 respectivamente. Se ha demostrado que las funciones de discriminación lineal se desempeñan aceptablemente bien sobre datos ordinales.

Para variables cuyas modalidades no admiten un ordenamiento, por ejemplo el *credo político* de un individuo, el tratamiento debe ser diferente. Así, para una variable con k modalidades no ordenables, éstas pueden ser reemplazadas por $(k-1)$ variables "ficticias" (dummy) y emplear sobre estas

la discriminación lineal. Para el caso, las tres categorías de la variable *credo político* pueden convertirse en variables ficticias como se muestra a continuación

$$Y_1 = \begin{cases} 1, & \text{si es liberal.} \\ 0, & \text{en otro caso.} \end{cases} \qquad Y_2 = \begin{cases} 1, & \text{si es conservador.} \\ 0, & \text{en otro caso.} \end{cases}$$

Así, el par de variables (Y_1, Y_2) toman los valores $(1, 0)$ para un liberal, $(0, 1)$ para un conservador y $(0, 0)$ para un independiente.

8.5.4 Clasificación mediante funciones de densidad

Las reglas de clasificación presentadas en las secciones (8.2) y (8.3) se basan en el supuesto de multinormalidad de los datos. Además, estas reglas se obtienen del principio de asignación óptima de Welch, con el cual una observación X se asigna al grupo para el que $p_i f(X|G_i)$ sea máxima. Si la forma de $f(X|G_i)$ no es normal o es desconocida, la función de densidad puede estimarse directamente desde los datos; este procedimiento se conoce como estimación *"kernel"* (*núcleo*). En este texto se mantendrán los dos términos de manera indistinta. De manera que el propósito es desarrollar una metodología que no requiera postular modelos para la distribución condicionada a cada grupo, en este sentido se puede considerar este tipo de clasificación como de "distribución libre" o no paramétrico; aunque en estricto sentido un procedimiento de clasificación siempre requerirá una distribución.

A continuación se describe el procedimiento kernel para una variable aleatoria continua y unidimensional X. Supóngase que X tiene función de densidad $f(x)$, la cual se quiere estimar mediante una muestra $x_1, \ldots, x_n$. Un estimador de $f(x_0)$ para un punto arbitrario x_0 se basa en la proporción de puntos contenidos en el intervalo $(x_0 - h, x_0 + h)$. Si se nota por $N(x_0)$ el número de puntos en el intervalo, entonces la proporción de $N(x_0)/n$ es un estimador de $P(x_0 - h < X < x_0 + h)$, la cual es aproximadamente igual al área del rectángulo inscrito en el recinto delimitado por el intervalo $(x_0 - h, x_0 + h)$ y la función f; es decir, $2hf(x_0)$. Así, $f(x_0)$ se estima por

$$\hat{f}(x_0) = \frac{N(x_0)}{2hn}. \tag{8.35}$$

Se expresa a $\hat{f}(x_0)$ como una función de los x_i muestrales definiendo

$$\mathcal{K}(u) = \begin{cases} \frac{1}{2}, & \text{para } |u| \leq 1, \\ 0, & \text{para } |u| > 1. \end{cases} \tag{8.36}$$

Dado que $(x_0 - x_i) \leq h$, la función definida en (8.36) se calcula por medio de $\mathcal{K}[(x_0 - x_i)/h]$, de esta forma $N(x_0) = 2\sum_{i=1}^{n} \mathcal{K}[(x_0 - x_i)/h]$, y el estimador (8.35) de f es ahora

$$\hat{f}(x_0) = \frac{1}{hn} \sum_{i=1}^{n} \mathcal{K}\Big(\frac{x_0 - x_i}{h}\Big). \tag{8.37}$$

La función $\mathcal{K}(\cdot)$ se llama el "*kernel*". Por su propia definición, la función de densidad estimada vía kernel es robusta al efecto de datos atípicos o "outliers". Esto porque, en general, la catidad $K[(x_0 - x_i)/h]$ se hace pequeña cuando x_i se aleja de x_0.

En la estimación dada por (8.37), $\mathcal{K}[(x_0 - x_i)/h]$ toma el valor $\frac{1}{2}$ para los x_i dentro del intervalo $(x_0 - h, x_0 + h)$ y cero para los puntos que estén fuera. De esta forma, cada punto del intervalo contribuye con $1/(2hn)$ a $\hat{f}(x_0)$ y con cero para los puntos fuera de éste. La gráfica de $\hat{f}(x_0)$ en función de x_0 es la correspondiente a una función de paso (escalonada), puesto que habrá un salto (o caida), siempre que x_0 esté a una distancia máxima h con alguno de los x_i. Nótese que los promedios móviles tienen esta propiedad.

Para un estimador "suave" de $f(x)$, se debe escoger un núcleo suave. Se presentan las siguientes dos opciones, entre otras,

$$\mathcal{K}(u) = \frac{1}{\pi}\frac{\operatorname{sen}^2 u}{u^2}, \text{ o } \mathcal{K}(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \tag{8.38}$$

las cuales tienen la propiedad de que todos los n puntos muestrales $x_1, \ldots, x_n$ contribuyen a $\hat{f}(x_0)$ con ponderaciones altas para los puntos cercanos. Aunque el segundo núcleo suave de (8.38) tiene la forma de una distribución normal, esto no significa supuesto alguno sobre la forma de la densidad $f(x)$. Se ha usado este tipo de función dado que es simétrica y unimodal, aunque se puede emplear cualquier otro tipo de funciones como núcleo; se prefieren las simétricas y unimodales.

Para funciones de densidad multivariadas, si $x_0' = (x_{01}, \ldots, x_{0p})$ es un punto arbitrario cuya densidad se quiere estimar, una extensión de (8.37) es

$$\hat{f}(x_0) = \frac{1}{nh_1h_2, \cdots, h_p} \sum_{i=1}^{n} \mathcal{K}\Big(\frac{x_{01} - x_{i1}}{h_1}, \ldots, \frac{x_{0p} - x_{ip}}{h_p}\Big). \tag{8.39}$$

Un estimador basado sobre un núcleo normal multivariado está dado por

$$\hat{f}(x_0) = \frac{1}{nh^p |\boldsymbol{S}_p|^{\frac{1}{2}}} \sum_{i=1}^{n} e^{(x_0 - x_i)' \boldsymbol{S}_p{}^{-1} (x_0 - x_i)/2h^2}, \tag{8.40}$$

donde los h_i son iguales y $\boldsymbol{S}_p$ es la matriz de covarianzas calculada a partir de los k grupos muestrales.

La selección del parámetro de suavizamiento h es clave para el uso de estimadores de densidad tipo kernel. El tamaño de h determina la cantidad de contribución de cada x_i a $\hat{f}(x_0)$. Si h es demasiado pequeño, $\hat{f}(x_0)$ presenta "picos" en cada x_i, y si h es grande, $\hat{f}(x_0)$ es casi uniforme. En consecuencia, los valores de h dependen del tamaño de la muestra n, los cuales tienen una relación inversa con éste; a mayor tamaño de muestra menor será el valor de h y recíprocamente. En la práctica se debe intentar con varios valores de h y evaluarlos en términos de los errores de clasificación obtenidos con cada uno de ellos.

Para emplear las estimaciones hechas sobre las funciones de densidad, a través de núcleos, en análisis discriminante, se aplica la densidad estimada en cada grupo y se obtiene $\hat{f}(x_0|G_1), \ldots, \hat{f}(x_0|G_k)$, donde x_0 es el vector de medidas de un individuo.

La regla de clasificación es: asignar x_0 al grupo G_i para el cual la cantidad

$$p_i \hat{f}(x_0|G_i) \tag{8.41}$$

tome el valor máximo.

Ejemplo 8.4 Se quiere establecer la posible relación existente entre el diseño de un casco para fútbol (americano) y las lesiones en el cuello[3].

Para esto se tomaron 6 mediciones sobre cada uno de 90 deportistas, los cuales estaban divididos en grupos de a 30 en cada una de las siguientes tres clases: Futbolistas universitarios (grupo 1), futbolistas de educación media (grupo 2), y deportistas no futbolistas (grupo 3).

Las seis variables son:

X_1: ancho máximo de la cabeza.

X_2: circunferencia de la cabeza.

X_3: distancia entre la frente y la nuca a la altura de los ojos.

X_4: distancia de la parte superior de la cabeza a los ojos.

X_5: distancia de la parte superior de la cabeza a las orejas.

X_6: ancho de quijada.

[3]Rencher (1995, pág. 346)

Se emplea como núcleo la distribución normal multivariada en (8.40). Con $h = 2$ se obtiene $\hat{f}(x_0|G_i)$, para los tres grupos ($i = 1, 2, 3$). Asumiendo que $p_1 = p_2 = p_3$, la regla de clasificación de acuerdo con (8.41) es: asignar x_0 a al grupo para el cual $\hat{f}(x_0|G_i)$ sea la más grande. La tabla 8.5 muestra los resultados de la clasificación de los 90 individuos junto con la tasa de error aparente.

Tabla 8.5 Clasificación de los futbolistas

Grupo actual	Número de obs.	Grupo asignado		
		G_1	G_2	G_3
1	30	25	1	4
2	30	0	12	18
3	30	0	3	27

La tasa aparente de clasificación correcta es: $(25 + 12 + 27)/90 = 0.711$.

La tasa de error aparente en la clasificación es: $1 - 0.711 = 0.289$. ☑

8.5.5 Clasificación mediante la técnica de "el vecino más cercano"

El método de clasificación llamado *"el vecino más cercano"* se considera como una técnica de tipo no paramétrico. Para el procedimiento se determina la distancia de Mahalanobis de una observación X_i respecto a las demás observaciones X_j, mediante

$$D_{ij} = (X_i - X_j)' \boldsymbol{S}_p^{-1} (X_i - X_j), \quad i \neq j. \tag{8.42}$$

Para clasificar la observación X_i en uno de dos grupos, se examinan los k puntos más cercanos a X_i, si la mayoría de estos k puntos pertenecen al grupo G_1, se asigna la observación X_i a G_1, en otro caso se asigna a G_2. Si se nota el número de individuos (objetos) de G_1 por k_1 y a los restantes por k_2 en G_2, con $k = k_1 + k_2$, entonces la regla se expresa también como: asignar X_i a G_1 si

$$k_1 > k_2, \tag{8.43}$$

y G_2 en otro caso. Si los tamaños muestrales de cada grupo son n_1 y n_2 respectivamente, la decisión es: asignar X_i a G_1 si

$$\frac{k_1}{n_1} > \frac{k_2}{n_2}. \tag{8.44}$$

De una manera coloquial, una observación X_i se asigna al grupo donde se "inclinen" la mayoría de sus vecinos; es decir, por votación la mayoría decide el grupo donde se debe ubicar cada observación.

Además, si se consideran las probabilidades a priori: asignar x_i a G_1 si

$$\frac{k_1/n_1}{k_2/n_2} > \frac{p_2}{p_1}. \tag{8.45}$$

Estas reglas se pueden extender a más de dos grupos. Así, en (8.44): se asigna la observación al grupo que tenga la más alta proporción k_j/n_j, donde k_j es el número de observaciones en el grupo G_j entre las k observaciones más cercanas a X_i.

Respecto al valor k, se sugiere tomar un valor cercano a $\sqrt{n_i}$ para algún n_i típico. En la práctica se puede ensayar con varios valores de k y usar el que menor tasa de error provoque.

8.5.6 Clasificación mediante redes neuronales

Se ha observado que muchos problemas en patrones de reconocimiento han sido resueltos más "fácilmente" por humanos que por computadores, tal vez por la arquitectura básica y el funcionamiento de su cerebro. Las *redes neuronales* (RN) son diseñadas mediante emulaciones, hasta ahora incompletas, con el cerebro humano para imitar el trabajo humano y tal vez su inteligencia. El término red neuronal artificial es usado para referirse a algoritmos de cómputo que usan las estructuras básicas de las neuronas biológicas.

Una neurona recibe impulsos de otras neuronas a través de las dendritas. Los impulsos que llegan son enviados por los terminales de los axones a las otras neuronas. La transmisión de una señal de una neurona a otra se hace a través de una conexión (sinapsis) con las dentritas de las neuronas vecinas. La sinapsis es un proceso físico-químico complejo, el cual genera una inversión de potencial en la célula receptora; si el potencial alcanza cierto umbral, la célula envía una señal a través de su axón y en consecuencia se establece una comunicación con las que se le conecten directa o indirectamente.

Una neurona artificial (en adelante simplemente neurona) en computación consta de: unas entradas o estímulos, una caja de procesamiento y una respuesta. El modelo más simple de neurona artificial es el modelo de McCulloch y Pits (Torres y otros 1993, págs. 2–7). Supóngase que la atención está sobre la neurona k, esta neurona recibe una serie de entradas

Y_{ik}, cada una de las cuales puede ser la salida de la i-ésima neurona vecina. La neurona desarrolla una suma ponderada de las entradas y produce como salida un cero o un uno dependiendo de si la suma supera un valor *umbral* μ_k asignado a la neurona. La figura 8.7 ilustra este modelo de neurona.

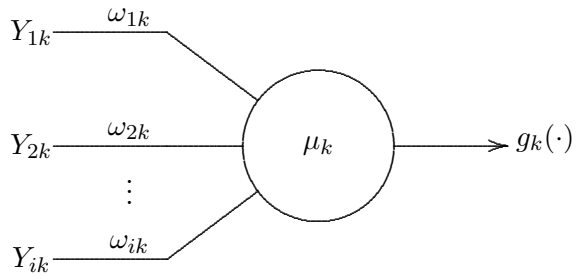


Figura 8.7 *Modelo de neurona simple.*

Las entradas $Y_{1k}, \ldots, Y_{ik}$ corresponden a las salidas de las neuronas conectadas con la neurona k.

Las cantidades $\omega_{1k}, \ldots, \omega_{ik}$ son las ponderaciones de conexión entre la salida de la j-ésima neurona y la entrada a la k-ésima neurona.

μ_k es el umbral de la señal de la k-ésima neurona.

$g_k(\cdot)$ es la función de *salida, respuesta* o *transferencia* de la k-ésima neurona

La ecuación de nodo es

$$Z_k = g_k\Big(\sum_j \omega_{jk} Y_{jk} - \mu_k\Big) = \begin{cases} 1, & \text{si } \sum_j \omega_{jk} Y_{jk} \geq -\mu_k, \\ 0, & \text{si } \sum_j \omega_{jk} Y_{jk} < \mu_k. \end{cases}$$

Otras funciones de transferencia son las siguientes:

Función rampa

$$g(x) = \begin{cases} 0, & si \;\; x < 0, \\ x, & si\; 0 \leq x \leq 1\,, \\ 1, & si\; x > 1. \end{cases}$$

Función logística

$$g(x) = \tfrac{1}{1+e^{-x}},$$

Función signo

$$g(x) = \begin{cases} -1, & si\; x < 0, \\ 1, & si\; x \geq 0. \end{cases}$$

Una red consiste en un conjunto de neuronas o unidades de cómputo. Cada neurona en una red desarrolla un cálculo simple. Tres son los elementos básicos de una red neuronal: las *neuronas, nodos* o *unidades de cómputo*; *la arquitectura (topología)* de la red, la cual describe las conexiones entre los nodos; y *el algoritmo de "entrenamiento"* usado para encontrar los valores particulares de los parámetros, con los cuales la red desarrolla eficientemente una tarea particular.

Un *perceptrón* es una red neuronal, que está conformado por varias neuronas que desarrollan un trabajo específico. Un perceptrón multicapa está constituido por varias capas de neuronas interconectadas con alguna arquitectura específica. Este tipo de modelos es el que más atención ha recibido para clasificación.

Rosenblant (1962), citado por Krzanowski y Marriott (1995), demuestra que si dos conjuntos de datos se separan por un hiperplano, entonces mediante el modelo tipo perceptrón se determina un plano que los separe.

La asignación de un individuo determinado por el vector $X' = (X_1, \ldots, X_p)$ a uno de q-grupos $G_1, \ldots, G_q$, puede verse como un proceso matemático que transforma las p entradas $X_1, \ldots, X_p$ en q unidades de salida $Z_1, \ldots, Z_q$, las cuales definen la localización de un individuo en un grupo; es decir, $Z_i = 1$ y $Z_j = 0$, para todo $i \neq j$ si el individuo es localizado en el grupo G_i. El perceptrón multicapa lleva a cabo, la tarea de transformación tratando a los X_i como valores de p-unidades en la *capa de entrada*, los Z_j son los valores de las q-unidades en la *capa de salida*; además entre estas dos capas hay algunas *capas escondidas* (intermedias) de nodos o neuronas. Usualmente cada unidad en una capa está conectada a todas las unidades de la capa adyacente y no a otras (aunque algunas redes permiten conectar unidades de capas no contiguas). La *arquitectura* o *topología* de una red es determinada por el número de capas, el número de unidades en cada capa y las conexiones entre unidades.

La figura 8.8 muestra una red de tres capas que contiene cuatro unidades en la capa de entrada, tres unidades en una capa escondida y dos unidades en la capa de salida; una conexión completa se establece entre capas vecinas. Para cada conexión entre la j-ésima unidad, en la i-ésima capa y la k-ésima unidad en la $i+1$-ésima capa se asocia una ponderación $\omega_{i(jk)}$. El valor para cualquier unidad X_j, en la i-ésima capa, se transfiere a la k-ésima unidad en la $(i+1)$-ésima capa transformado por $f_i(x_i)$ y multiplicado por la respectiva ponderación. De esta manera, a la unidad k de la capa $i+1$ "llegan" las contribuciones de las unidades ubicadas en la capa anterior, éstas se combinan aditivamente y se adiciona una constante α_{ik}, para producir el valor $y_k = \alpha_{ik} + \sum_j \omega_{i(jk)} f_i(x_j)$ para esta unidad. Este proceso se continúa

de manera sucesiva entre una capa y otra hasta que hayan sido asignados valores a todas las unidades de la red.

De acuerdo con los tres elementos básicos de una red descritos anteriormente, para el perceptrón presentatado, tan sólo se han desarrollado los dos primeros (los nodos y la arquitectura). El último está relacionado con el *entrenamiento* de la red, y consiste en encontrar los mejores valores de las ponderaciones $\omega_{i(jk)}$ y las constantes α_k. El término "mejores" hace referencia a los valores con los cuales la red predice en forma óptima (mínimo error de clasificación). Lo anterior implica la optimización de alguna función objetivo, la cual compara lo observado con los valores producidos por cada una de las unidades de la red sobre todos los datos de los n individuos de entrenamiento. La función más común es la suma de cuadrados de los residuales, aunque existen otros criterios como la verosimilitud (Krzanowski y Marriott 1995, págs. 50-52).

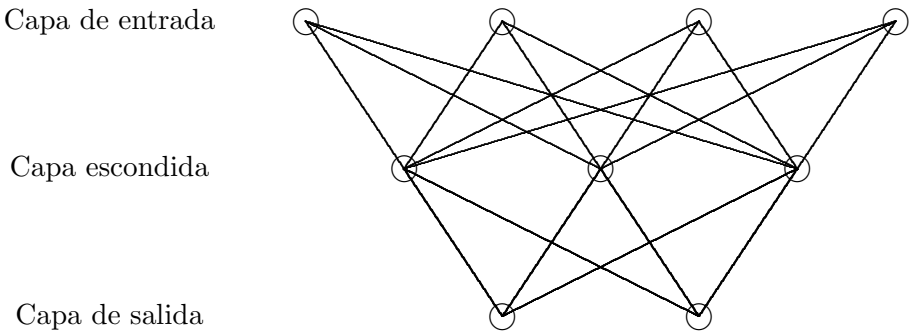


Figura 8.8 *Perceptrón multicapa.*

Se presenta, de manera condensada, la optimización con el criterio de mínimos cuadrados. Supóngase que se tienen datos de ensayo para n individuos, el i-ésimo de los cuales está caracterizado mediante el vector $X_i = (X_{i1}, \ldots, X_{ip})'$. Para simplificar la notación se ignora la presencia de capas y se centra la atención sobre las unidades o nodos. Así, se nota ω_{jk} para indicar la ponderación entre las unidades j y k. Se escribe I_{ij} para señalar el valor de entrada recibido por la unidad j correspondiente al individuo i y O_{ij} expresa el valor de salida emanado desde la misma unidad. De esta forma, $I_{ij} = X_{ij}$ si j es una unidad de entrada e $I_{ij} = \sum_k O_{ik}\omega_{kj}$ en otro caso, la suma se hace sobre todas las unidades de la capa anterior conectadas con la unidad j. Similarmente, $O_{ij} = I_{ij}$ para una unidad de

entrada, mientras que $O_{ij} = f(I_{ij})$ en otro caso (funciones apropiadas f se presentan al comienzo de esta sección). Si se escribe el valor objetivo de salida como T_{ij} de la unidad j en el individuo i, la función objetivo a optimizar es

$$\mathcal{E} = \sum_{i=1}^{n} \mathcal{E}_i = \sum_{i=1}^{n} \Big[\frac{1}{2} \sum_{k} (O_{ik} - T_{ik})^2 \Big]. \tag{8.46}$$

La minimización de (8.46) se logra de manera iterativa con el empleo de aproximaciones tales como el "menor descenso", en cada iteración las ponderaciones se actualizan de acuerdo con el punto correspondiente al menor decrecimiento de $\mathcal{E}$. Este proceso iterativo es conocido como el *algoritmo de propagación hacia atrás.*

El problema es decidir cuando parar el proceso. Una estrategia es considerar la tasa de clasificación incorrecta, de manera que el proceso se frena cuando ésta sea suficientemente cercana a cero.

Ejemplo 8.5 Para ilustrar como se constuye una red neuronal con el fin de emplearla en la clasificación de objetos, se considera el caso (hipotético) de clasificar gatos de acuerdo con el color del pelo (caracterización fenotípica). Los gatos considerados tienen una representación del tipo (X_1, X_2) con $X_1, X_2 = 0,\ 1$, las cuales corresponden a la siguiente caracterización alélica de los gatos:

$0\ 0 \Longrightarrow$ "Blanco"

$1\ 0 \Longrightarrow$ "Gris"

$0\ 1 \Longrightarrow$ "Pardo"

$1\ 1 \Longrightarrow$ "Negro"

Ésta obedece a los genes que determinan la pigmentación del pelo, los cuales determinan su color.

Después de cubrir las fases de entrenamiento y aprendizaje, se propone la red neuronal cuyas capas, conexiones y ponderaciones (arquitectura) se muestran en la figura 8.9.

Los números 1.5 y 0.5 corresponden a los valores umbral μ_k; de manera que la salida, en cada una de ellas, es 1.0 o 0.0 si la suma ponderada que entra en ella es superior a estos valores. De manera más explícita, un gato pardo se identifica con $(0, 1)$, a la neurona de la capa media ingresan los valores $(1) \times 0 + (1) \times 1 = 1$, el cual como es menor que 1.5 produce una salida de 0.0, a la última neurona ingresa la cantidad $(1) \times 0 + (-2) \times 0 + (1) \times 1 = 1$,

que por ser mayor que 0.5 hace que esta neurona produzca como salida el 1. De esta manera un gato de color pardo lo identifica mediante el 1, algo similar ocurre con un gato gris. La tabla 8.6 contiene el proceso y resultado de la clasificación.

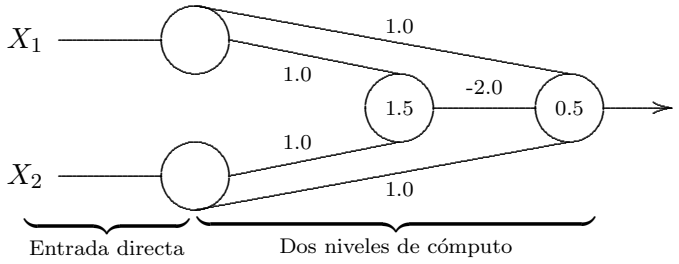


Figura 8.9 *Clasificación mediante una red neuronal.*

Tabla 8.6 Clasificación mediante una red neuronal

Entrada		Neurona interna		Neurona Final	
X_1	X_2	Entra	Sale	Entra	Sale
0	0	0+0	0	0+0+0	0
1	0	1+0	0	1+0+0	1
0	1	0+1	0	0+0+1	1
1	1	1+1	1	1-2+1	0

Se observa que a los gatos blancos y negros los identifica con el 0, mientras que a los otros con el 1, en genética se habla de homocigotos y heterocigotos, respectivamente. ☑

8.6 Selección de variables

La selección de variables en el análisis discriminante está asociada con el uso que se pretenda dar a la metodología. De acuerdo con los dos objetivos presentados al comienzo de este capítulo, uno corresponde a la *separación* de grupos y el otro a la *localización o clasificación* de observaciones o casos. Las metodologías empleadas para la separación de grupos se relacionan con las estadísticas parciales T^2 o Lambda de Wilks (Λ), con las cuales se verifica la influencia de un subconjunto de variables en la separación (diferencia de medias) de dos o más grupos (capítulo 3). En esta parte se comentan algunas metodologías para el segundo propósito.

Es importante advertir sobre el cuidado que se debe tener al intercambiar el uso de metodologías cuyos propósitos son la separación de grupos o la localización de observaciones, respectivamente.

El problema sobre la contribución de cada variable en la discriminación, tal como se procede en el análisis de regresión, está ligado a la búsqueda de la función de predicción con las variables que mejor contribuyan a la discriminación. Naturalmente, se procura incorporar al modelo el menor número variables predictoras (principio de parsimonia). Uno de los criterios de selección de variables es escoger el subconjunto que produzca la menor tasa de error.

A continución se comentan los procedimientos más empleados, los cuales están incorporados en la mayoría de los paquetes estadísticos.

Para el caso de dos grupos se recomiendan dos procedimientos:

(1) Las estadísticas F parciales con niveles de significancia nominal entre 0.10 y 0.25. Con estas estadísticas se observa el aporte "extra" que cada variable hace al modelo, una vez que han ingresado las demás, se incorparan aquellas que tengan el mayor valor F.

(2) Un estimador de la probabilidad de clasificación correcta basado en la distancia de Mahalanobis entre dos grupos (McLachlan 1992, págs. 366-367).

Un mecanismo formal para la selección del "mejor" subconjunto de variables en cualquier problema de modelamiento requiere un criterio que evalue la bondad del ajuste, de un procedimiento para el cálculo (generalmente computacional), y tal vez, de una regla necesaria para "frenar" el proceso (Krzanowski 1995, pág. 41). Dentro de los procedimientos para el cálculo de la bondad del ajuste en la selección de variables se cuentan la selección *hacia adelante* (forward), la eliminación hacia *atrás* (backward) y la selección *"stepwise" (selección paso a paso).*

En la selección hacia adelante ("forward") la función de clasificación se inicia con la variable que bajo algún criterio sea la más apropiada (generalmente a través de la estadística F). En una segunda etapa se adiciona, entre las restantes $(p-1)$ variables, la que mejor desempeño muestre en la regla de clasificación, luego se agrega a estas dos variables una entre las $(p-2)$ restantes la de mejor desempeño, y así sucesivamente.

La eliminación hacia atrás (backward) trabaja en sentido opuesto a la técnica anterior. Se empieza la función con todas las p variables, se remueve en cada etapa la variable que menos afecte el "buen desempeño" de la función de clasificación.

La estrategia de selección basada en el método "stepwise" trabaja en forma parecida al procedimiento de selección hacia adelante, la diferencia es que en cada etapa una de las variables ya incorporadas al modelo puede ser removida sin que menoscabe el desempeño de la función de clasificación.

La tres estrategias anteriores requieren una regla para finalizar el proceso, en términos de mejoramiento o deterioro. La regla natural es terminar el proceso cuando la adición de nuevas variables no incremente significativamente el buen desempeño de la función, o cuando la exclusión de cualquiera de las variables ya incorporadas al modelo no deteriore su desempeño. El término "desempeño" puede ser juzgado a través de la tasa de clasificación, de la estadística Lambda de Wilks (Λ) para un subconjunto de variables, o de algún incremento en términos de suma de cuadrados tal como se hace en análisis de regresión.

Otro procedimiento consiste en combinar el procedimiento "stepwise" con el criterio de estimación del error mediante validación cruzada. En este procedimiento cada observación es excluida, un subconjunto de variables es seleccionado para construir la regla de clasificación, y luego la observación excluida es clasificada empleando reglas de clasificación lineal computadas desde las variables seleccionadas. Las tasas de error resultantes son usadas para escoger la variable que en cada etapa debe incorporarse al modelo.

Se puede emplear también el análisis de componentes principales (capítulo 5) para seleccionar variables, o utilizar los mismos componentes como predictores en la función de discriminación (Biscay, Valdes y Pascual (1990)).

8.7 Rutina SAS para hacer análisis discriminante

Para un conjunto de observaciones que contienen variables cuantitativas y una variable de clasificación, que define el grupo de cada observación, el PROC DISCRIM desarrolla un criterio de discriminación para asignar cada observación en uno de los grupos. SAS también tiene el procedimiento STEPDISC, el cual desa-rrolla análisis discriminante con selección de variables (tipo "stepwise", "forward" y "backward").

Al frente (o debajo) de cada instrucción se explica su propósito dentro de los símbolos $/*$ y $*/$.

```
/* Análisis discriminante */
DATA  nombre SAS; /* nombre del archivo de datos */
INPUT variables; /* variables, incluyendo la de clasificación */
```

```
CARDS; /* ingreso de datos */
      /* escribir aquí los datos */
;
PROC DISCRIM CROSSVALIDATE POOL=YES CROSSLIST;
/* desarrolla discriminación asumiendo igualdad de las matrices */
/* de covarianzas e imprime la  validación cruzada por observación */
CLASS variable;  /* se indica la variable que define los grupos */
VAR  lista de variables; /* se escriben las variables cuantitativas */
                                                /* para el análisis */
PRIORS EQUAL$|$PROP$|$probabilidades;
 /* (EQUAL) toma iguales las probabilidades a priori para cada grupo */
 /* (PROP) hace las probabilidades proporcionales a los tamaños de grupo */
 /* también se puede dar las probabilidades a priori para cada grupo */
 /* Ejemplo, para tres grupos 1, 2 y 3, se escribe  PRIORS '1'=0.25 */
                                                /* '2'=0.35 '3'=0.40; */
 /* Por defecto se considera la opción EQUAL */
RUN;
```

8.8 Procesamiento de datos con R

Se presenta el ejemplo 8.2, como son muchos datos (172 filas) la lectura se hace desde un archivo externo usando la función (read.table())

```
# lectura de los datos
ejemp8_2<-read.table("ejemplo8_2.txt",header=TRUE)
# transformación mediante raíz cuadrada
ejemp8_2$DOWN_A<-sqrt(ejemp8_2$DOWN_A)
ejemp8_2$RIGHT_A<-sqrt(ejemp8_2$RIGHT_A)
# definición del factor
ejemp8_2$GRP<-factor(ejemp8_2$GRP)
head(ejemp8_2)
# requiere la librería MASS
library(MASS)
# análisis discriminante lineal
z<-lda(GRP ~.,ejemp8_2,prior =c(36/172,36/172,50/172,50/172))
# clasificación de las observaciones por medio de la regla
# obtenida.
clasif<-predict(z,ejemp8_2[,-9])$class
clasif
# tabla de clasificación
addmargins(table(ejemp8_2$GRP,clasif))
```

Suponga que se tiene la observación

Ar_1	Pe_1	Lo_1	An_1	Ar_2	Pe_2	Lo_2	An_2
54.36	220.07	90.08	44.84	53.86	220.17	90.27	43.53

y queremos aplicar la regla de discriminación sobre ella, en R procedemos así:

```
nuevo<-data.frame(Ar_1=54.36,Pe_1=220.07,Lo_1=90.08,
                  An_1=44.84,Ar_2=53.86,Pe_2=220.17,
                   Lo_2=90.27,An_2=43.53)
predict(z,nuevo)$class
```

Clasificación usando la función de discriminación cuadrática

```
zq<-qda(GRP ~.,ejemp8_2,prior =c(36/172,36/172,50/172,50/172))
clasifq<-predict(zq,ejemp8_2[,-9])$class
clasifq
tabla<-table(ejemp8_2$GRP,clasif)
addmargins(tabla)
```

El código R para generar la tabla 8.4 es el siguiente:

```
# Estimación de la probabilidad de clasificación
# errónea por validación cruzada
clasifq<-numeric(nrow(ejemp8_2))
for(i in 1:nrow( ejemp8_2)){
zq<-qda(GRP~.,ejemp8_2[-i,],prior=c(36,36,50,50)/172)
clasifq[i]<-as.numeric(predict(zq,ejemp8_2[i,-9])$class)
}
tabla8_3<-table(ejemp8_2$GRP,clasifq)
addmargins(tabla8_3)
```

Capítulo 9

Análisis de correlación canónica

9.1 Introducción

Hay situaciones en las que un conjunto de variables se debe dividir en dos grupos para estudiar la relación existente entre las variables de éstos. El llamado *análisis de correlación canónica (ACC)* o simplemente *análisis canónico*, es una de las herramientas desarrolladas para tales propósitos. En el análisis de *regresión múltiple* se mide la relación entre un conjunto de variables llamadas *regresoras* y *una* variable *respuesta* o *dependiente*, se puede considerar entonces, al ACC como una generalización del modelo de regresión múltiple; el cual busca establecer la relación entre un conjunto de variables predictoras y un conjunto de variables respuesta; se puede advertir lo difícil y complejo que resultaría desarrollar un análisis de regresión para cada una de las variables respuesta. El ACC se propone determinar la correlación entre una *combinación lineal* de las variables de un conjunto y una *combinación lineal* de las variables del otro conjunto. Nótese que la estrategia consiste en volver al caso clásico, donde se encuentra la correlación entre pares de variables; cada una de las cuales es una combinación lineal de las variables de los respectivos conjuntos. Una vez que se tienen estas correlaciones, el problema es encontrar el par de combinaciones lineales con la *mayor correlación*; éste nuevamente es un problema de reducción del espacio de las variables.

Por interés histórico e ilustrativo, se presenta el siguiente ejemplo desarrollado por Hotelling (1936), creador de esta técnica, citado en Manly (2000, pág. 146). Se midió la velocidad de lectura (X_1), la capacididad de lectura

(X_2), la velocidad aritmética (Y_1) y la capacidad aritmética (Y_2) en un grupo de 140 estudiantes de séptimo grado. La intención era determinar si la habilidad en lectura (medida por X_1 y X_2) se relaciona con la habilidad aritmética (medida por Y_1 y Y_2). Con el análisis canónico se busca una combinación lineal U de X_1 y X_2 y otra V de Y_1 y Y_2,

$$\begin{cases} U = a_1X_1 + a_2X_2 \\ V = b_1Y_1 + b_2Y_2, \end{cases}$$

tal que la correlación entre U y V sea tan grande como se pueda. El procedimiento de optimización es similar al de componentes principales, excepto que aquí se maximiza la correlación en lugar de la varianza. Las variables U y V reciben el nombre de *variables canónicas*.

Hotelling encontró, de acuerdo con sus datos, que las "mejores" selecciones para U y V son

$$\begin{cases} U = -2.78X_1 + 2.27X_2 \\ V = -2.44Y_1 + 1.00Y_2, \end{cases}$$

con una correlación de 0.62. Es fácil observar que U mide la diferencia entre la capacidad y la velocidad de lectura, mientras que V mide la diferencia entre la capacidad y la velocidad aritmética. El valor de la correlación (0.62) indica que una diferencia grande entre X_1 y X_2 va acompañada de una diferencia alta entre Y_1 y Y_2. En resumen la lectura y aritmética están altamente correlacionados en los estudiantes de séptimo grado.

En diferentes campos del conocimiento aparece la necesidad de buscar la relación entre dos conjuntos de variables; por ejemplo, en:

- economía puede haber interés en establecer la relación entre las variables consumo agregado (C), producto interno bruto (PIB) inversión bruta (I) y las variables gasto público (G), oferta monetaria (M) e interés a corto plazo (R);

- médicina el interés se dirige a determinar si ciertos estilos de vida y hábitos de alimentación individual tienen algún efecto sobre la salud de un grupo de pacientes; la salud se mide mediante algunas variables asociadas tales como hipertensión, peso, ansiedad, y niveles de tensión.

- mercadeo, se busca la relación entre las variables tamaño, precio por marca, punto de venta (distancia al consumidor) y las variables volumen de ventas por tamaño, frecuencia de compra por marca;

- ecología hay interés por indagar acerca de la relación existente entre algunas variables ambientales (temperatura, precipitación anual, altitud y densidad vegetal) con algunas variables morfológicas (medidas corporales), sobre especies animales o vegetales;
- psicología, a un grupo de estudiantes se les registran logros (habilidades y destrezas) para observar su relación con un conjunto de variables de personalidad y actitudes.

En resumen, el ACC tiene como objetivo encontrar el par de combinaciones lineales, una en cada conjunto, que tengan la correlación más alta entre ellas para determinar si existe algún grado de asociación entre los dos conjuntos de variables. Si sobre bases teóricas o por interés en el estudio, se hace que uno de los conjuntos sea un conjunto de variables predictoras o independientes y el otro de variables dependientes o respuesta, entonces el objetivo del ACC es determinar si el conjunto de variables predictoras afecta o explica el conjunto de variables respuesta.

9.2 Geometría de la correlación canónica

Considérese un conjunto de datos asociado con las variables $\boldsymbol{X} = \{X_1, X_2\}$ y con las variables $\boldsymbol{Y} = \{Y_1, Y_2\}$, predictoras y respuesta, respectivamente (Sharma 1996, pág. 392). La tabla 9.1 contiene los datos de dos conjuntos de variables $\boldsymbol{X}$ y $\boldsymbol{Y}$ corregidos por su media. Los datos se deben representar en un espacio de dimensión cuatro, como esto no es posible hacerlo en un plano como el que se dispone para dibujar (esta hoja de papel), se procede a hacer una representación geométrica de los datos para la variables X y Y en forma separada.

Las figuras 9.1a y 9.1b muestran los dispersogramas de los conjuntos de variables $\boldsymbol{X}$ y $\boldsymbol{Y}$, respectivamente. Supóngase que en el primer conjunto se identifica un nuevo eje, U_1, el cual forma un ángulo $\theta_1 = 10^o$ con el eje X_1. La proyección de los 24 puntos sobre el "nuevo" eje corresponde a una combinación lineal de las variables del conjunto $\boldsymbol{X}$. Por geometría elemental, el valor de la nueva variable U_1 se calcula mediante la siguiente ecuación:

$$U_1 = \cos 10^o X_1 + \operatorname{sen} 10^o X_2 = 0.985X_1 + 0.174X_2.$$

En la tabla 9.1 se muestran los valores de U_1 en cada una de las 24 observaciones en las X. Así por ejemplo, para el primer punto $(1.051, -0.435)$,

$$U_1 = 0.985(1.051) + 0.174(-0.435) = 0.959.$$

Tabla 9.1 Datos hipotéticos

Observación	X_1	X_2	Y_1	Y_2	U_1	V_1
1	1.051	-0.435	0.083	0.538	0.959	0.262
2	-0.419	-1.335	-1.347	-0.723	-0.645	-1.513
3	1.201	0.445	1.093	-0.112	1.260	0.989
4	0.661	0.415	0.673	-0.353	0.723	0.512
5	-1.819	-0.945	-0.817	-1.323	-1.956	-1.220
6	-0.899	0.375	-0.297	-0.433	-0.820	-0.427
7	3.001	1.495	1.723	2.418	3.215	2.446
8	-0.069	-2.625	-2.287	-1.063	-0.524	-2.513
9	-0.919	0.385	-0.547	0.808	-0.838	-0.238
10	-0.369	-0.265	-0.447	-0.543	-0.410	-0.606
11	-0.009	-0.515	0.943	-0.633	-0.098	0.670
12	0.841	1.915	1.743	1.198	1.161	2.047
13	0.781	1.845	1.043	2.048	1.089	1.680
14	0.631	-0.495	0.413	-0.543	0.535	0.202
15	-1.679	-0.615	-1.567	-0.643	-1.760	-1.692
16	-0.229	-0.525	-0.777	-0.252	-0.317	-0.817
17	-0.709	-0.975	0.523	-0.713	-0.868	0.248
18	-0.519	0.055	-0.357	0.078	-0.502	-0.309
19	0.051	0.715	0.133	0.328	0.174	0.237
20	0.221	0.245	0.403	0.238	0.260	0.460
21	-1.399	-0.645	-0.817	-1.133	-1.490	-1.155
22	0.651	0.385	1.063	-0.633	0.708	0.782
23	-0.469	-0.125	-0.557	-0.393	-0.484	-0.658
24	0.421	1.215	-0.017	1.838	0.625	0.612
Media	0.000	0.000	0.000	0.000	0.000	0.000
Desv. Est.	1.052	1.033	1.018	1.011	1.109	1.140

En la figura 9.1b se identifica un "nuevo" eje V_1 para el segundo conjunto, que forma un ángulo $\theta_2 = 20^o$ respecto al eje Y_1. Similar al caso de las variables $\boldsymbol{X}$, la proyección de los puntos hacia V_1 se consigue mediante una combinación lineal de las variables del conjunto $\boldsymbol{Y}$. Los valores de esta nueva variable se obtienen de:

$$V_1 = \cos 20^o Y_1 + \operatorname{sen} 20^o Y_2 = 0.940 Y_1 + 0.342 Y_2.$$

En la última columna de la tabla 9.1 se presenta la proyección sobre V_1 de los puntos con coordenadas en las Y.

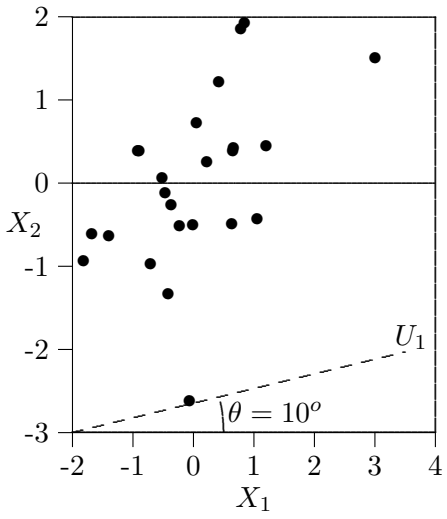


Figura 9.1a *Conjunto* $\boldsymbol{X}$.

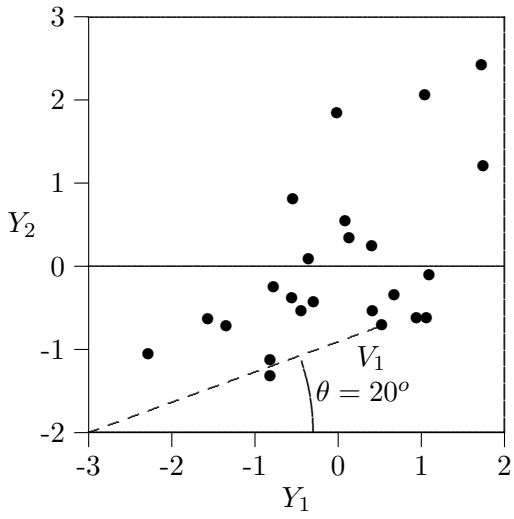


Figura 9.1b *Conjunto* $\boldsymbol{Y}$.

La correlación simple entre el nuevo par de variables U_1 y V_1 es igual a 0.831. Recuérdese que este valor corresponde al coseno del ángulo formado por estos dos vectores; para este caso $\theta_{UV} = 33.8^o$.

En la tabla 9.2 se consigna la correlación entre las nuevas variables U_1 y V_1, generadas desde diferentes ángulos θ_1 y θ_2 respectivamente. Se observa que la correlación más grande entre las dos nuevas variables es 0.961, la cual se tiene cuando el ángulo formado entre U_1 y X_1 es $\theta_1 = 57.6^o$ y el ángulo formado entre V_1 y Y_1 es $\theta_2 = 47.2^o$. Las figuras 9.1a y 9.1b, muestran los dos nuevos ejes U_1 y V_1, respectivamente.

Tabla 9.2 Correlación entre variables canónicas

Ángulo entre U_1 y X_1 (θ_1)	Ángulo entre V_1 y Y_1 (θ_2)	Correlación entre U_1 y V_1
10	20	0.830
20	10	0.846
10	30	0.843
40	40	0.946
57.6	47.2	0.961
30	10	0.872
20	40	0.894
40	20	0.919
60	70	0.937

Las proyecciones de los 24 puntos hacia los dos nuevos ejes U_1 y V_1, de donde resultan las "nuevas" variables, se obtienen mediante la transformación contenida en las siguientes ecuaciones:

$$\begin{cases} U_1 = \cos 57.6^o X_1 + \operatorname{sen} 57.6^o X_2 = 0.536X_1 + 0.844X_2 \\ V_1 = \cos 47.2^o Y_1 + \operatorname{sen} 47.2^o Y_2 = 0.679Y_1 + 0.734Y_2. \end{cases}$$

Así, las nuevas variables con una alta correlación, generadas desde cada conjunto $\boldsymbol{X}$ y $\boldsymbol{Y}$, son respectivamente

$$\begin{cases} U_1 = 0.536X_1 + 0.844X_2 \\ V_1 = 0.679Y_1 + 0.734Y_2. \end{cases}$$

Una vez que se han identificado U_1 y V_1, es posible identificar otro conjunto de ejes, U_2 y V_2, tales que:

1. La correlación entre los nuevos ejes, U_2 y V_2, sea máxima.
2. El segundo conjunto de nuevos ejes U_2 y V_2 esté incorrelacionado con los nuevos ejes iniciales U_1 y V_1, respectivamente; es decir, que:

$$Cov(U_1, U_2) = 0,\ Cov(V_1, V_2) = 0,$$
$$Cov(U_1, V_2) = 0 \text{ y } Cov(U_2, V_1) = 0.$$

Se demuestra que las variables U_2 y V_2, obtenidas bajo las condiciones anteriores, forman ángulos con $X1$ y Y_1 iguales a 138.33^o y 135.30^o, respectivamente. Los valores para las 24 observaciones, en este par de variables, se calculan mediante las ecuaciones:

$$\begin{cases} U_2 = \cos 138.33^o X_1 + \operatorname{sen} 138.33^o X_2 = -0.747X_1 + 0.665X_2 \\ V_2 = \cos 135.30^o Y_1 + \operatorname{sen} 135.30^o Y_2 = -0.711Y_1 + 0.703Y_2. \end{cases}$$

De esta manera, las combinaciones lineales de las variables $X's$ que están más altamente correlacionadas con las combinaciones lineales de las variables $Y's$, son respectivamente

$$\begin{cases} U_1 = 0.536X_1 + 0.844X_2 \\ U_2 = -0.747X_1 + 0.665X_2 \end{cases} \quad \text{y} \quad \begin{cases} V_1 = 0.679Y_1 + 0.734Y_2 \\ V_2 = -0.711Y_1 + 0.703Y_2. \end{cases}$$

Este procedimiento se debe continuar hasta tanto no se identifiquen nuevas variables. En este caso, no es posible identificar más variables, pues la

dimensión de los espacios es dos. En un caso más general, donde el número de variables del espacio $\mathbb{X}$ es m y el de $\mathbb{Y}$ es p, el número de nuevas variables es el valor mínimo entre m y p.

El nuevo sistema de variables, para cada conjunto de variables $\boldsymbol{X}$ y $\boldsymbol{Y}$, respectivamente, que satisface las condiciones anteriores, se muestra en las figuras 9.2a y 9.2b.

En la terminología del ACC, a las ecuaciones de proyección anteriores se les denomina *ecuaciones canónicas*. A las variables U y V, expresadas en las ecuaciones canónicas, se les llama *variables canónicas*. Así, U_1 y V_1, es el primer conjunto de variables canónicas y, U_2 y V_2, es el segundo conjunto de variables canónicas. La correlación entre cada par de variables canónicas se llama la *correlación canónica*.

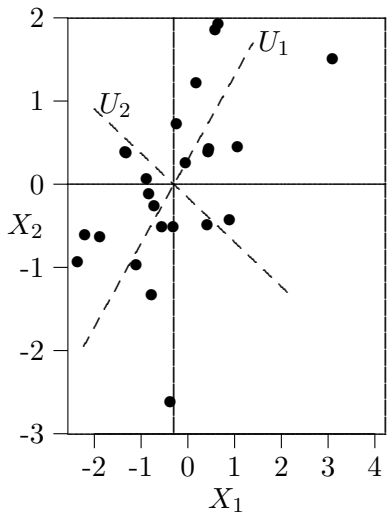


Figura 9.2a *Variables canónicas en* $\boldsymbol{X}$.

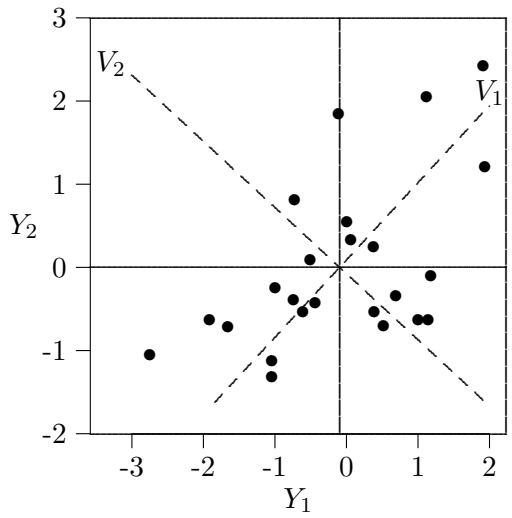


Figura 9.2b *Variables canónicas en* $\boldsymbol{Y}$.

En resumen, el objetivo de la correlación canónica es identificar nuevos ejes, U_i y V_i, donde U_i es una combinación lineal de las $X's$ y V_i es una combinación lineal de las $Y's$, tales que: (1) la correlación entre U_i y V_i sea máxima, y (2) en cada uno de estos conjuntos, las variables sean incorrelacionadas.

Geométricamente, para el caso de los vectores X_1, X_2, Y_1 y Y_2 considerados, se ubicarán en un espacio de dimensión cuatro, "incrustado" en un espacio de dimensión 24. Nuevamente, como no es posible graficar un espacio de dimensión cuatro, se dibujan, X_1 y X_2, en un plano, y a Y_1 y Y_2, en otro plano.

El propósito del ACC es identificar U_1, el cual "cae" en el mismo espacio bidimensional de $\boldsymbol{X}$, y V_1, el cual "cae" en el mismo espacio bidimensional de $\boldsymbol{Y}$, tal que el ángulo entre U_1 y V_1 sea *mínimo*. Es decir, se busca que el coseno del ángulo determinado entre U_1 y V_1, el cual equivale a la correlación entre este par de combinaciones de variables (correlación canónica), sea máximo. El siguiente par de ejes, U_2 y V_2, se determina de forma tal que el ángulo entre ellos sea mínimo.

El procedimiento anterior se puede ilustrar mediante la siguiente comparación: supóngase que se tiene un libro abierto, donde cada cara (plano) corresponde a cada uno de los espacios de las variables X y Y. Se trata entonces de buscar el ángulo mínimo posible determinado por las caras del libro en esta posición. La figura 9.3 ilustra este procedimiento, allí se ha trazado uno, entre todos los posibles ángulos que se pueden construir, que corresponde al mínimo; es decir, al que tiene el coseno más grande y en consecuencia la mayor correlación.

Se puede notar que el objetivo del ACC tiene bastante similitud con el de componentes principales sobre un conjunto de variables. La diferencia es el criterio usado para identificar los nuevos ejes. En el análisis por componentes principales, el primer eje nuevo, resulta en una nueva variable que recoge la mayor cantidad de variabilidad de los datos. En el análisis de correlación canónica, se identifica un nuevo eje para cada conjunto de variables, tal que la correlación entre los nuevos ejes sea máxima.

Es posible que con unas pocas variables canónicas sea suficiente para representar adecuadamente los dos conjuntos de variables. En este sentido se puede considerar al ACC como una técnica de reducción de datos, pues "reduce" simultáneamente los espacios representados por los dos conjuntos de variables.

Una de las primeras inquietudes que debe plantearse quien desee aplicar el ACC es acerca de la adecuación de los datos para desarrollar esta técnica; es decir: ¿qué grado de asociación tienen los dos conjuntos de datos? Esta pregunta equivale a plantear la hipótesis nula de no asociación lineal, o independencia bajo normalidad de los datos, entre las variables X y las variables Y.

En la sección (4.3.4) se muestran las estadísticas (4.19a o 4.19b) con las cuales se puede desarrollar la prueba de independencia entre los dos conjuntos de variables. En caso de no rechazar la hipótesis de independencia, el ACC no es pertinente; en caso contrario, surge el interrogante sobre cuál es el número de variables canónicas necesario para describir la relación lineal entre los dos conjuntos de variables X y Y. Rencher (1998, pág. 324)

desarrolla una prueba basada en la estadística (4.19b), con la cual se puede asegurar si la relación entre los dos conjuntos de variables se debe a las primeras r variables canónicas.

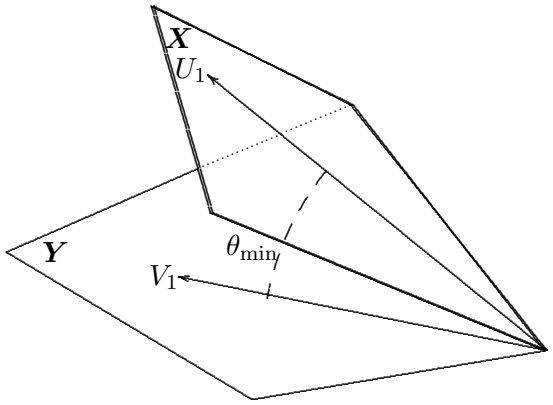


Figura 9.3 *Esquema geométrico del análisis de correlación canónica.*

9.3 Procedimiento para el análisis canónico

9.3.1 Modelo poblacional

Sean $X_1, X_2, \ldots, X_m$ e $Y_1, Y_2, \ldots, Y_p$, dos conjuntos de variables, uno de variables explicativas (independientes) y el otro de variables dependientes (respuesta), respectivamente. Por conveniencia $m \geq p$. Si $\boldsymbol{Z}$ es el vector de variables de tamaño $(1 \times (m+p))$, se puede considerar que éste ha sido particionado verticalmente en la forma $\boldsymbol{Z} = (\boldsymbol{X} \| \boldsymbol{Y})$, donde $\boldsymbol{X}$ contiene m-variables y $\boldsymbol{Y}$ las p restantes. Sin pérdida de generalidad, se asume que $\mathcal{E}(\boldsymbol{Z}) = 0$. La matriz de covarianzas del vector $\boldsymbol{Z}$ se particiona en forma análoga a la hecha anteriormente; es decir,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \vdots & \boldsymbol{\Sigma}_{XY} \\ \ldots\ldots & & \ldots\ldots \\ \boldsymbol{\Sigma}_{YX} & \vdots & \boldsymbol{\Sigma}_{YY} \end{pmatrix}. \tag{9.1}$$

Las matrices $\boldsymbol{\Sigma}_{XX}$ y $\boldsymbol{\Sigma}_{YY}$ son las matrices de las covarianzas "dentro" de cada conjunto de variables y las matrices $\boldsymbol{\Sigma}_{XY}$ y $\boldsymbol{\Sigma}_{YX}$ son las matrices

de las covarianzas "entre" los conjuntos. El esquema siguiente muestra la partición

$$\begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_m \\ \\ Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{array} \begin{array}{c} \begin{array}{cc} X_1\ X_2 \cdots X_m & \quad Y_1\ Y_2 \cdots Y_p \end{array} \\ \begin{pmatrix} & \vdots & \\ (\boldsymbol{\Sigma}_{XX})_{m\times m} & \vdots & (\boldsymbol{\Sigma}_{XY})_{m\times p} \\ & \vdots & \\ \cdots\cdots\cdots & & \cdots\cdots\cdots \\ & \vdots & \\ (\boldsymbol{\Sigma}_{YX})_{p\times m} & \vdots & (\boldsymbol{\Sigma}_{YY})_{p\times p} \\ & \vdots & \end{pmatrix} \end{array}. \tag{9.2}$$

El objetivo del análisis canónico es encontrar una combinación lineal de las m variables predictoras $\boldsymbol{X}$ (independientes) que maximice la correlación con una combinación lineal de las p variables respuesta $\boldsymbol{Y}$ (dependientes). Explícitamente, se trata de encontrar entre las siguientes combinaciones lineales

$$\begin{cases} U_1 = \boldsymbol{\alpha}_1' X = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1m}X_m \\ U_2 = \boldsymbol{\alpha}_2' X = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2m}X_m \\ \vdots \\ U_r = \boldsymbol{\alpha}_r' X = \alpha_{r1}X_1 + \alpha_{r2}X_2 + \cdots + \alpha_{rm}X_m, \end{cases} \tag{9.3a}$$

aquélla que tenga la correlación más alta con alguna de las siguientes combinaciones lineales

$$\begin{cases} V_1 = \boldsymbol{\gamma}_1' Y = \gamma_{11}Y_1 + \gamma_{12}Y_2 + \cdots + \gamma_{1p}Y_p \\ V_2 = \boldsymbol{\gamma}_2' Y = \gamma_{21}Y_1 + \gamma_{22}Y_2 + \cdots + \gamma_{2p}Y_p \\ \vdots \\ V_r = \boldsymbol{\gamma}_r' Y = \gamma_{r1}Y_1 + \gamma_{r2}Y_2 + \cdots + \gamma_{rp}Y_p, \end{cases} \tag{9.4b}$$

con $r = \min\{m, p\}$. Las combinaciones lineales se escogen de tal forma que: la correlación entre U_1 y V_1 sea máxima; la correlación entre U_2 y V_2 sea máxima con la restricción que estas variables estén no correlacionadas con U_1 y V_1; la correlación entre U_3 y V_3 sea máxima sujeta a la no correlación con U_1, V_1, U_2 y V_2, y así sucesivamente. Cada par de variables $(U_1, V_1), (U_2, V_2), \ldots, (U_r, V_r)$ representa, independientemente, la relación entre los conjuntos de variables $\boldsymbol{X}$ e $\boldsymbol{Y}$. El primer par (U_1, V_1) tiene la correlación más alta y es el más importante; el segundo par (U_2, V_2) tiene

la segunda correlación más alta, y así sucesivamente, el r-ésimo par (U_r, V_r) tiene la r-ésima correlación más alta (en orden descendente).

El procedimiento para maximizar la correlación es un problema de cálculo, el cual se esquematiza a continuación.

En forma condensada se escriben las correlaciones arriba señaladas como: $U = \boldsymbol{\alpha}'X$ y $V = \boldsymbol{\gamma}'Y$, respectivamente. La correlación entre U y V está dada por

$$\boldsymbol{\rho}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{XY}\boldsymbol{\gamma}}{\left\{\left(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{XX}\boldsymbol{\alpha}\right)\left(\boldsymbol{\gamma}'\boldsymbol{\Sigma}_{YY}\boldsymbol{\gamma}\right)\right\}^{\frac{1}{2}}}. \tag{9.4}$$

Como $\boldsymbol{\rho}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ es invariante por transformaciones de escala sobre $\boldsymbol{\alpha}$ y $\boldsymbol{\gamma}$; es decir, equivale a trabajar con $\boldsymbol{\alpha}$ y $\boldsymbol{\gamma}$ normalizados, entonces, se requiere que $\boldsymbol{\alpha}$ y $\boldsymbol{\gamma}$ sean tales que U y V tengan varianza uno; es decir, que

$$\begin{cases} var(U) = var(\boldsymbol{\alpha}'X) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{XX}\boldsymbol{\alpha} = 1, & \text{y} \\ var(V) = var(\boldsymbol{\gamma}'Y) = \boldsymbol{\gamma}'\boldsymbol{\Sigma}_{YY}\boldsymbol{\gamma} = 1, \end{cases}$$

con $\mathcal{E}(U) = \mathcal{E}(V) = 0$. Entonces, maximizar (9.5) es equivalente a maximizar,

$$\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{XY}\boldsymbol{\gamma}, \tag{9.5}$$

con las restricciones

$$\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{XX}\boldsymbol{\alpha} = 1 \quad \text{y} \quad \boldsymbol{\gamma}'\boldsymbol{\Sigma}_{YY}\boldsymbol{\gamma} = 1; \tag{9.6}$$

por multiplicadores de Lagrange el problema se transforma en maximizar

$$\boldsymbol{\varphi} = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{XY}\boldsymbol{\gamma} - \frac{1}{2}\phi(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{XX}\boldsymbol{\alpha} - 1) - \frac{1}{2}\mu(\boldsymbol{\gamma}'\boldsymbol{\Sigma}_{YY}\boldsymbol{\gamma} - 1), \tag{9.7}$$

con ϕ y μ los respectivos multiplicadores de Lagrange.

Diferenciando con respecto a $\boldsymbol{\alpha}$ y $\boldsymbol{\gamma}$ resultan las siguientes ecuaciones

$$\begin{cases} \boldsymbol{\Sigma}_{XY}\boldsymbol{\gamma} - \phi\boldsymbol{\Sigma}_{XX}\boldsymbol{\alpha} = 0, \\ \boldsymbol{\Sigma}'_{XY}\boldsymbol{\alpha} - \mu\boldsymbol{\Sigma}_{YY}\boldsymbol{\gamma} = 0. \end{cases} \tag{9.8}$$

Se demuestra que estas ecuaciones son equivalentes a

$$(\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX} - \rho^2\boldsymbol{I})\boldsymbol{\alpha} = 0, \tag{9.9a}$$

o también con

$$(\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} - \rho^2\boldsymbol{I})\boldsymbol{\gamma} = 0, \tag{9.10b}$$

donde $\rho = \mu = \phi$.

Las matrices

$$\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX} \quad \text{o} \quad \boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}. \tag{9.10}$$

tienen los mismos valores propios.

Se nota por λ_1 al valor propio más grande encontrado en esta etapa, el cual equivale al cuadrado de la correlación más grande entre U y V (donde $\lambda_1 = \rho_1^2$). Al sustituir este primer valor propio en las ecuaciones (9.10a) y (9.10b) se obtienen los vectores propios $\boldsymbol{\alpha}$ y $\boldsymbol{\gamma}$.

Una segunda combinación lineal de las $X's$ y otra de las $Y's$, no correlacionadas éstas con las primeras U_1 y V_1, se determina a través del segundo valor propio más grande λ_2, por un procedimiento análogo al anterior. De manera iterativa, en la r-ésima etapa se consiguen los pares de combinaciones (variables canónicas) $U_1 = \boldsymbol{\alpha}_1 X$ y $V_1 = \boldsymbol{\gamma}_1 Y$, $\ldots$, $U_r = \boldsymbol{\alpha}_r X$ y $V_r = \boldsymbol{\gamma}_r Y$, con los respectivos valores propios $\lambda_1, \lambda_2, \ldots, \lambda_r$.

Para resumir, sea $\boldsymbol{Z} = (\boldsymbol{X} \| \boldsymbol{Y})$ un vector con matriz de covarianzas $\boldsymbol{\Sigma}$. El cuadrado de la r-ésima *correlación canónica*, entre X e Y, es el r-ésimo valor propio más grande de alguna de las dos matrices contenidas en (9.11). Los coeficientes de $\boldsymbol{\alpha}_r X$ y $\boldsymbol{\gamma}_r Y$ definen el r-ésimo par de variables canónicas que satisfacen (9.10a) y (9.10b), respectivamnete, con $\lambda = \lambda_r$. Los valores propios $\lambda_1 > \lambda_2 > \cdots > \lambda_r$ son los cuadrados de las correlaciones entre las variables canónicas, o equivalentemente, las correlaciones canónicas son iguales a la raíz cuadrada de los respectivos valores propios; así:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \boldsymbol{\rho}(\boldsymbol{\alpha}_r X, \boldsymbol{\gamma}_r Y) = \sqrt{\lambda_r}.$$

Del desarrollo anterior se establece la siguiente relación entre los coeficientes de las combinaciones lineales en cada uno de los conjuntos de variables

$$\boldsymbol{\alpha} = \frac{\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\gamma}}{\sqrt{\lambda}} \tag{9.11a}$$

y

$$\boldsymbol{\gamma} = \frac{\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\alpha}}{\sqrt{\lambda}}. \tag{9.11b}$$

No es necesario encontrar las soluciones para los dos sistemas de ecuaciones dados en (9.9a) y (9.9b), ya que al encontrar un conjunto de coeficientes, en forma simétrica mediante (9.11a) y (9.11b) se encuentran los otros.

9.3.2 Análisis canónico para una muestra

La presentación anterior es más teórica que práctica, pues rara vez se conocen las matrices $\boldsymbol{\Sigma}_{XX}$, $\boldsymbol{\Sigma}_{YY}$, $\boldsymbol{\Sigma}_{XY}$ y $\boldsymbol{\Sigma}_{YX}$. Lo corriente es disponer de un

conjunto de datos que corresponden a $(m+p)$ respuestas o medidas de n-individuos; de tal forma que la matriz de datos $\mathbb{Z} = (\mathbb{X} \| \mathbb{Y})$ es dada por:

$$\mathbb{Z} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1m} & \vdots & Y_{11} & Y_{12} & \cdots & Y_{1p} \\ X_{21} & X_{22} & \cdots & X_{2m} & \vdots & Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ X_{n1} & X_{n2} & \cdots & X_{nm} & \vdots & Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{pmatrix} \tag{9.12}$$

La matriz de covarianzas $\boldsymbol{\Sigma}$ se estima por $\boldsymbol{S}$ y se particiona como se indica a continuación

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{XX} & \vdots & \boldsymbol{S}_{XY} \\ \cdots\cdots & & \cdots\cdots \\ \boldsymbol{S}_{YX} & \vdots & \boldsymbol{S}_{YY} \end{pmatrix}. \tag{9.13}$$

Las variables canónicas asociadas a los datos muestrales se escriben en la forma

$$U_i = \mathbf{a}^{(i)'} X \text{ y } \; V_i = \mathbf{b}^{(i)'} Y, \quad \text{para } i = 1, 2 \ldots, r. \tag{9.14}$$

De manera análoga, el cuadrado de la r-ésima correlación entre las $X's$ y las $Y's$, es el r-ésimo valor propio más grande de

$$\boldsymbol{S}_{XX}^{-1}\boldsymbol{S}_{XY}\boldsymbol{S}_{YY}^{-1}\boldsymbol{S}_{YX} \;\text{ o }\; \boldsymbol{S}_{YY}^{-1}\boldsymbol{S}_{YX}\boldsymbol{S}_{XX}^{-1}\boldsymbol{S}_{XY}. \tag{9.15}$$

Cada par de variables canónicas es determinado por los vectores $\mathbf{a}^{(i)'}$ y $\mathbf{b}^{(i)'}$.

Como ocurre en la generación de componentes principales, las variables canónicas se obtienen a partir de las *matrices de correlación*; paralelamente con el desarrollo anterior, las matrices para la generación de los valores propios son ahora

$$\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY}\boldsymbol{R}_{YY}^{-1}\boldsymbol{R}_{YX} \;\text{ o }\; \boldsymbol{R}_{YY}^{-1}\boldsymbol{R}_{YX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY}. \tag{9.16}$$

donde la matriz de correlación de la matriz de datos $\mathbb{Z}$, se ha particionado en la forma siguiente

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{XX} & \vdots & \boldsymbol{R}_{XY} \\ \cdots\cdots & & \cdots\cdots \\ \boldsymbol{R}_{YX} & \vdots & \boldsymbol{R}_{YY} \end{pmatrix}. \tag{9.17}$$

La determinación de las variables canónicas a partir de las matrices de correlación se sugiere cuando las escalas de medición registradas para las variables hacen difícil la interpretación (no conmensurabilidad).

9.3.3 Análisis canónico y análisis de regresión

Al iniciar este capítulo se comentó acerca de la relación entre el análisis de correlación canónica y el análisis de regresión. Para el caso de regresión lineal simple ($m = p = 1$), de las expresiones dadas en (9.17), y como $\boldsymbol{R}_{XX} = \boldsymbol{R}_{YY} = 1$, se tiene que

$$\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY}\boldsymbol{R}_{YY}^{-1}\boldsymbol{R}_{YX} = \boldsymbol{R}_{YY}^{-1}\boldsymbol{R}_{YX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY} = \boldsymbol{R}_{YX}\boldsymbol{R}_{XY} = r^2, \quad (9.18)$$

donde r^2, como se esperaba, es el cuadrado del coeficiente de correlación entre la variable X y la variable Y.

Ahora, en regresión lineal múltiple se tienen m variables explicativas frente a una variable respuesta (p=1). Por un razonamiento similar se concluye que

$$\boldsymbol{R}_{YY}^{-1}\boldsymbol{R}_{YX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY} = \boldsymbol{R}_{YX}^{-1}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY}. \quad (9.19)$$

Para un modelo de regresión lineal múltiple se obtiene que la expresión (9.19) es equivalente a

$$\boldsymbol{R}_{YX}\hat{\boldsymbol{\beta}}, \quad (9.20)$$

donde $\hat{\boldsymbol{\beta}}$ es un vector de tamaño $(m \times 1)$ que contiene la estimación de los m parámetros del modelo de regresión. La expresión (9.20) corresponde al coeficiente de determinación.

De la relación entre el ACC y el análisis de regresión, se puede medir la "importancia" o el aporte de cada variable respecto a su variable canónica. Considerada cada variable canónica como un modelo de regresión múltiple, se mide el peso que tiene cada variable dentro de su respectivo conjunto con relación a la respectiva variable canónica a través del coeficiente de correlación producto-momento. Cada coeficiente de correlación refleja el grado con el que cada variable canónica representa una variable.

Para el i-ésimo par de variables canónicas (U_i, V_i) los pesos que expresan el grado de asociación entre las variables y sus variables canónicas se obtienen, respectivamente, mediante las siguientes expresiones

$$\begin{aligned} \mathbf{r}_X^i &= \boldsymbol{R}_{XX}\mathbf{a}^{(i)} \\ \mathbf{r}_Y^i &= \boldsymbol{R}_{YY}\mathbf{b}^{(i)} \quad \text{con } i = 1, \ldots, r, \end{aligned} \quad (9.21)$$

donde $\mathbf{a}^{(i)}$ y $\mathbf{b}^{(i)}$ son los vectores de coeficientes de la i-ésima variable canónica para las variables X e Y, respectivamente.

9.3.4 Interpretación geométrica del ACC

Sean $\mathbf{a}^{(i)}$ y $\mathbf{b}^{(i)}$ los vectores de coeficientes que determinan el i-ésimo par de variables canónicas U_i y V_i, para $i = 1, \ldots, r$. Los n valores de $\mathbf{a}^{(i)}$ (o de $\mathbf{b}^{(i)}$) para todas las observaciones (individuos), son las componentes de $\mathbf{a}^{(i)'}X$ (o de $\mathbf{b}^{(i)'}Y$). Los vectores $\mathbf{a}'X$ y $\mathbf{b}'Y$ representan dos puntos de $\mathbb{R}^n$ (o individuos), pertenecientes a los subespacios $\mathbb{R}^m$ y $\mathbb{R}^p$ generados por las columnas de las X y las Y respectivamente.

Encontrar el par de variables canónicas, significa buscar el *ángulo mínimo* entre los subespacios $\mathbb{R}^m$ y $\mathbb{R}^p$. Más formalmente, se trata de buscar los coeficientes de $\mathbf{a}$ y $\mathbf{b}$ tales que el coseno del ángulo formado por $\mathbf{a}'X$ y $\mathbf{b}'Y$ sea máximo (ecuación 9.5).

Supóngase que las variables X y Y han sido estandarizadas y nótense por X^* e Y^* su respectiva estandarización, entonces las matrices de correlación quedan definidas por

$$\boldsymbol{R}_{XX} = X^*X^{*'}, \quad \boldsymbol{R}_{YY} = Y^*Y^{*'}, \quad \boldsymbol{R}_{XY} = X^*Y^{*'}, \quad \boldsymbol{R}_{YX} = Y^*X^{*'}, \tag{9.22}$$

en las ecuaciones equivalentes a (9.12) para el caso muestral, se puede premultiplicar por $X^{*'}$ y por $Y^{*'}$, respectivamente, y se obtiene

$$\begin{aligned} X^{*'}\mathbf{a} &= \frac{X^{*'}(X^*X^{*'})^{-1}X^{*'}Y^{*'}\mathbf{b}}{\sqrt{\lambda}} \\ Y^{*'}\mathbf{b} &= \frac{Y^{*'}(Y^*Y^{*'})^{-1}Y^*X^{*'}\mathbf{a}}{\sqrt{\lambda}}. \end{aligned} \tag{9.23}$$

Nótese que las matrices $X^{*'}(X^*X^{*'})^{-1}X^*$ e $Y^{*'}(Y^*Y^{*'})^{-1}Y^*$ son simétricas e idempotentes, de donde (sección (A.2)) las combinaciones lineales $X^{*'}\mathbf{a}$ y $Y^{*'}\mathbf{b}$ resultan ser una proyección ortogonal de puntos de $\mathbb{R}^n$ (individuos) sobre los subespacios generados por las variables X y Y; es decir, sobre $\mathbb{R}^m$ y $\mathbb{R}^p$. Las relaciones dadas por (9.24) establecen que $X^{*'}\mathbf{a}$ e $Y^{*'}\mathbf{b}$ son la una proyección de la otra.

Observación:

> $X^{*'}\mathbf{a} = \mathbf{a}'X^*$, ya que éstos son vectores del mismo espacio. Lo mismo se puede afirmar para las $Y^*\mathbf{b}$.

Ejemplo 9.1 A continuación se desarrolla el análisis de correlación canónica mediante los datos de la tabla 9.1. De acuerdo con los datos, las

matrices de covarianzas son:

$$\boldsymbol{S}_{XX} = \begin{pmatrix} 1.106807971 & 0.568582246 \\ 0.568582246 & 1.066773732 \end{pmatrix}, \quad \boldsymbol{S}_{YY} = \begin{pmatrix} 1.037247645 & 0.567465091 \\ 0.567465091 & 1.022143520 \end{pmatrix}$$
$$\boldsymbol{S}_{XY} = \begin{pmatrix} 0.760840942 & 0.702467355 \\ 0.794296558 & 0.845212373 \end{pmatrix}.$$

La matriz asociada con la primera expresión de (9.16) es

$$\boldsymbol{S}_{XX}^{-1}\boldsymbol{S}_{XY}\boldsymbol{S}_{YY}^{-1}\boldsymbol{S}_{YX} = \begin{pmatrix} 0.3416894 & 0.3698822 \\ 0.5188631 & 0.5951638 \end{pmatrix}.$$

Los valores propios de la matriz anterior corresponden a la solución de la siguiente ecuación:

$$\begin{vmatrix} 0.3416894 - \lambda & 0.3698822 \\ 0.5188631 & 0.5951638 - \lambda \end{vmatrix} = 0$$
$$(0.3416894 - \lambda)(0.5951638 - \lambda) - (0.3698822)(0.5188631) = 0$$
$$\lambda^2 - 0.9368532\lambda + 0.011442937 = 0.$$

Las soluciones de la ecuación anterior son $\lambda_1 = 0.9244754$ y $\lambda_2 = 0.0123778$.

Nótese que 0.9244754 es el cuadrado de la correlación entre las variables canónicas U_1 y V_1; es decir $\boldsymbol{\rho}(U_1, V_1) = \sqrt{0.9244754} = 0.9614964$, como se indica en la tabla 9.2 ($\theta_{U_1V_1} \approx 15.95^o$). De manera análoga, 0.0123 es el cuadrado de la correlación entre las variables canónicas U_2 y V_2.

Sustituyendo el valor λ_1 en la ecuación anterior se obtiene

$$\begin{pmatrix} 0.3416894 - 0.9244754 & 0.3698822 \\ 0.5188631 & 0.5951638 - 0.9244754 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0,$$

sistema equivalente con

$$\begin{cases} -0.5827860a_1 + 0.3698822a_2 = 0, \\ 0.5188631a_1 - 0.3293116a_2 = 0. \end{cases}$$

Este sistema de ecuaciones se reduce a:

$$a_1 = 0.634679282a_2,$$
$$a_1 = 0.634679167a_2.$$

La solución puede obtenerse al asumir que el vector $\mathbf{a}^{(1)'} = (a_1, a_2)$ es unitario; es decir, que $a_1^2 + a_2^2 = 1$, con lo cual la solución es:

$$\mathbf{a}^{(1)'} = (0.5358629, 0.8443050).$$

La varianza de la combinación lineal $\mathbf{a}^{(1)'}X$ es:

$$var(\mathbf{a}^{(1)'}X) = \mathbf{a}^{(1)'}\boldsymbol{S}_{XX}\mathbf{a}^{(1)} = 1.5927588.$$

De esta manera, el vector $\mathbf{a}^{(1)}$ se puede estandarizar dividiendo por la desviación estándar de $\mathbf{a}^{(1)'}X$, es decir, por $\sqrt{1.5927588} = 1.2620455$. Así, la primera combinación lineal es la siguiente:

$$U_1 = \mathbf{a}^{(1)'}X = 0.4245987011X_1 + 0.6689973036X_2.$$

De acuerdo con la expresión (9.12), una vez que se ha calculado $\mathbf{a}^{(1)'}$, se puede obtener $\mathbf{b}^{(1)'}$ a partir de la siguiente expresión:

$$\mathbf{b}^{(1)} = \frac{\boldsymbol{S}_{YY}^{-1}\boldsymbol{S}_{YX}\mathbf{a}}{\sqrt{\lambda}} = \begin{pmatrix} 0.6814145 \\ 0.7308324 \end{pmatrix}.$$

El vector $\mathbf{b}^{(1)}$ también se reescala dividiendo por la desviación estándar de la combinación lineal $\mathbf{b}^{(1)}Y$; es decir, por:

$$\sqrt{var(\mathbf{b}^{(1)'}Y)} = \sqrt{\mathbf{b}^{(1)'}\boldsymbol{S}_{YY}\mathbf{b}^{(1)}} = 1.284526.$$

En consecuencia, la combinación lineal de las $Y's$, que se correlaciona más altamente con la combinación de las $X's$, es la siguiente:

$$V_1 = \mathbf{b}^{(1)'}Y = 0.5399285948Y_1 + 0.5790856145Y_2.$$

Los coeficientes para las combinaciones lineales U_2 y V_2 se obtienen de manera semejante mediante el valor propio $\lambda_2 = 0.0123778$. El análisis puede también hacerse desde las matrices de correlación:

$$\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY}\boldsymbol{R}_{YY}^{-1}\boldsymbol{R}_{YX} \text{ o } \boldsymbol{R}_{YY}^{-1}\boldsymbol{R}_{YX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY}.\text{☑}$$

Ejemplo 9.2 Los datos (tomados de Manly (1986, pág. 165)) que contiene la tabla 9.3 corresponden a mediciones hechas sobre 16 colonias de mariposas *Euphydryaus editha*.

El primer conjunto de variables está conformado por las variables registradas en el habitat de estos insectos (ambientales):

Y_1: *altitud*,

Y_2: *precipitación anual*,

Y_3: *temperatura máxima anual*,

Y_4: *temperatura mínima anual*.

El segundo conjunto de variables está constituido por seis variables genéticas, dadas como porcentajes de seis frecuencias genéticas de la *fosfoglucosa isomerasa* (fgi), determinadas por electroforesis (de X_1 a X_6), así:

X_1: es el porcentaje de genes con movilidad 0.40,

X_2: es el porcentaje de genes con movilidad 0.60,

X_3: es el porcentaje de genes con movilidad 0.80,

X_4: es el porcentaje de genes con movilidad 1.00,

X_5: es el porcentaje de genes con movilidad 1.16,

X_6: es el porcentaje de genes con movilidad 1.30.

Tabla 9.3 Mediciones sobre mariposas

Colonia	Y_1	Y_2	Y_3	Y_4	X_1	X_2	X_3	X_4	X_5	X_6
SS	500	43	98	17	0	3	22	57	17	1
SB	800	20	92	32	0	16	20	38	13	13
WSB	570	28	98	26	0	6	28	46	17	3
JRC	550	28	98	26	0	4	19	47	27	3
JRH	550	28	98	26	0	1	8	50	35	6
SJ	380	15	99	28	0	2	19	44	32	3
CR	930	21	99	28	0	0	15	50	27	8
UO	650	10	101	27	10	21	40	25	4	0
LO	600	10	101	27	14	26	32	28	0	0
DP	1500	19	99	23	0	1	6	80	12	1
PZ	1750	22	101	27	1	4	34	33	22	6
MC	2000	58	100	18	0	7	14	66	13	0
IF	2500	34	102	16	0	9	15	47	21	8
AF	2000	21	105	20	3	7	17	32	27	14
GH	7850	42	84	5	0	5	7	84	4	0
GL	10500	50	81	-12	0	3	1	92	4	0

Se trata de explorar la posible asociación entre estos factores ambientales y las características genéticas presentes en estos insectos; es decir, establecer la existencia de una posible adaptación de una especie a las condiciones

que su habitat le ofrece (en un sentido darwinista). Para este propósito, se quiere establecer la "mejor" relación entre una combinación lineal de las variables ambientales (las $Y's$) y una combinación lineal de las variables genéticas (las $X's$). Aunque las segundas son las variables respuesta y las primeras las explicativas, admítase, por ahora, un cambio en la notación tradicional; en este caso $p = 6$ y $m = 4$. Obsérvese que en cada colonia la suma de las frecuencias genéticas suma 100%, se puede suprimir entonces cualquiera de las X; pues esta será igual a 100 menos la suma de los demás porcentajes, se suprime para desarrollar este ejercicio la variable X_6. La matriz de correlación entre las variables se presenta en la forma (9.17) a continuación

$$
\left(\begin{array}{ccccc:cccc}
\multicolumn{5}{c:}{\boldsymbol{R}_{XX}} & \multicolumn{4}{c}{\boldsymbol{R}_{XY}} \\
1.000 & 0.855 & 0.618 & -0.532 & -0.506 & -0.203 & -0.530 & 0.295 & 0.221 \\
0.855 & 1.000 & 0.615 & -0.548 & -0.597 & -0.190 & -0.410 & 0.173 & 0.246 \\
0.618 & 0.615 & 1.000 & -0.824 & -0.127 & -0.573 & -0.550 & -0.536 & 0.593 \\
-0.532 & -0.548 & -0.824 & 1.000 & -0.264 & 0.727 & 0.699 & -0.717 & -0.759 \\
-0.506 & -0.597 & -0.127 & -0.264 & 1.000 & -0.458 & -0.138 & 0.438 & 0.412 \\
\hdots & \hdots & \hdots & \hdots & \hdots & \hdots & \hdots & \hdots & \hdots \\
\multicolumn{5}{c:}{\boldsymbol{R}_{YX}} & \multicolumn{4}{c}{\boldsymbol{R}_{YY}} \\
-0.203 & -0.190 & -0.573 & 0.727 & -0.458 & 1.000 & 0.568 & -0.828 & -0.936 \\
-0.530 & -0.410 & -0.550 & 0.699 & -0.138 & 0.568 & 1.000 & -0.479 & -0.705 \\
0.295 & 0.173 & 0.536 & -0.717 & 0.438 & -0.828 & -0.479 & 1.000 & 0.719 \\
0.221 & 0.246 & 0.593 & -0.759 & 0.412 & -0.936 & -0.705 & 0.719 & 1.000
\end{array}\right).
$$

Los valores propios se obtienen para la matriz $\boldsymbol{R}_{YY}^{-1}\boldsymbol{R}_{YX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY}$ (de la expresión 9.17).

Después del desarrollo algebraico respectivo (Apéndice A.2.2), los valores propios de la matriz anterior son: $\lambda_1 = 0.7731$, $\lambda_2 = 0.5570$, $\lambda_3 = 0.1694$ y $\lambda_4 = 0.0472$. Nótese que $r = 4 = \min\{5, 4\}$. La raíz cuadrada de los valores propios equivale a la correlación entre las variables canónicas respectivas, así $\sqrt{\lambda_1} = \sqrt{0.7731} = 0.879$ es la correlación entre U_1 y V_1; en forma similar se establece la correlación entre los demás pares de variables canónicas; las cuales se presentan en la siguiente matriz

Correlación entre las U y las V

$$\begin{array}{c} \\ U_1 \\ U_2 \\ U_3 \\ U_4 \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & V_3 & V_4 \end{array} \\ \begin{pmatrix} 0.879 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.746 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.411 & 0.0000 \\ 0.000 & 0.000 & 0.000 & 0.217 \end{pmatrix} \end{array}.$$

De las ecuaciones presentadas en (9.12) se obtienen los coeficientes de las combinaciones lineales, y así, las variables canónicas asociadas con los dos conjuntos de variables son:

$$\begin{cases} U_1 = \mathbf{a}^{(1)'}X = -0.675X_1 + 0.909X_2 + 0.376X_3 + 1.442X_4 + 0.269X_5, \\ V_1 = \mathbf{b}^{(1)'}Y = -0.114Y_1 + 0.619Y_2 - 0.693Y_3 + 0.048Y_4 \end{cases}$$

$$\begin{cases} U_2 = \mathbf{a}^{(2)'}X = -1.087X_1 + 3.034X_2 + 2.216X_3 + 3.439X_4 + 2.928X_5, \\ V_2 = \mathbf{b}^{(2)'}Y = -0.777Y_1 + 0.980Y_2 - 0.562Y_3 + 0.928Y_4 \end{cases}$$

$$\begin{cases} U_3 = \mathbf{a}^{(3)'}X = 1.530X_1 + 2.049X_2 + 2.231X_3 + 4.916X_4 + 3.611X_5, \\ V_3 = \mathbf{b}^{(3)'}Y = -3.654Y_1 - 0.601Y_2 - 0.565Y_3 - 3.623Y_4 \end{cases}$$

$$\begin{cases} U_4 = \mathbf{a}^{(4)'}X = 0.284X_1 - 2.331X_2 - 0.867X_3 - 1.907X_4 - 1.133X_5, \\ V_4 = \mathbf{b}^{(4)'}Y = 1.594Y_1 + 0.860Y_2 + 1.599Y_3 + 0.742Y_4. \end{cases}$$

Aunque en general la interpretación de las variables canónicas no es sencilla, pues requiere de un amplio conocimiento de las variables que intervienen en el problema, se intentará darle un sentido a los resultados aquí obtenidos.

El siguiente es un significado de los pares de variables canónicas, para este caso. En el par de variables (U_1, V_1), U_1 muestra un contraste entre la variable X_1 y el resto de variables genéticas. Representa la escasez de genes con movilidad de 0.40. En la combinación lineal V_1 el coeficiente de Y_2 (precipitación) es positivo y alto, mientras que el coeficiente de Y_3 (temperatura máxima) es negativo y grande en valor absoluto. Se puede afirmar entonces que la escasez de genes con movilidad de 0.40 está altamente asociada con la ocurrencia de altas precipitaciones y caídas en la temperatura máxima.

Las correlaciones entre las variables canónicas y las respectivas variables, se obtienen mediante la ecuación (9.22). Las siguientes matrices contienen las correlaciones entre cada una de las variables canónicas y las respectivas variables genéticas y ambientales

Correlación entre las U y las X

$$\begin{array}{c} \\ U_1 \\ U_2 \\ U_3 \\ U_4 \end{array} \begin{array}{c} \begin{array}{ccccc} X_1 & X_2 & X_3 & X_4 & X_5 \end{array} \\ \begin{pmatrix} -0.568 & -0.387 & -0.703 & 0.922 & -0.361 \\ -0.433 & -0.164 & 0.209 & -0.243 & 0.478 \\ -0.221 & 0.121 & 0.069 & -0.191 & -0.035 \\ 0.657 & 0.899 & 0.411 & -0.231 & -0.728 \end{pmatrix} \end{array},$$

Correlación entre las V y las Y

$$\begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \\ V_4 \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & Y_3 & Y_4 \end{array} \\ \begin{pmatrix} -0.763 & 0.853 & -0.861 & -0.780 \\ -0.625 & 0.155 & 0.280 & 0.561 \\ 0.137 & -0.148 & -0.142 & 0.185 \\ -0.065 & -0.476 & -0.401 & 0.207 \end{pmatrix} \end{array}.$$

En la matriz de correlación entre U y X se aprecian, entre otras, las siguientes correlaciones: U_1 y X_1: -0.57; U_1 y X_2: -0.39; U_1 y X_3: -0.70; U_1 y X_4: 0.92 y U_1 y X_5: -0.36. Así, U_1 está altamente correlacionado, de manera directa, con X_4 e inversamente correlacionado con las demás variables. Se puede interpretar entonces a U_1 como un indicador de genes de movilidad 1.00.

Las correlaciones entre la variable canónica V_1 y las $Y's$ son: con Y_1: -0.76; con Y_2: 0.85; con Y_3: -0.86 y con Y_4: -0.78. Con esto se puede considerar a V_1 asociado indirectamente con la altitud y temperaturas máximas y mínimas como también, directamente con la precipitación anual.

Para la interpretación del par U_1 y V_1, de acuerdo con las correlaciones, se puede afirmar que el porcentaje de genes con movilidad 1.00, es alto para las colonias de mariposas que viven en regiones de grandes precipitaciones y altitudes pero con bajas temperaturas. Por un procedimiento similar se hace el análisis de las demás variables canónicas, con el auxilio de un genetista o de un entomólogo. ☑

9.4 Rutina SAS para el análisis de correlación canónica

Se desarrolla el análisis de correlación canónica entre los dos conjuntos de variables presentadas en la tabla 9.1, mediante el PROC CANCORR del paquete SAS. Este procedimiento construye las variables canónicas a partir de las matrices de covarianzas (contenidas en (9.16)) y a partir de las matrices de correlación (contenidas en (9.17)), respectivamente. Al frente (o debajo) de cada instrucción se explica su propósito dentro de los símbolos /∗ y ∗/.

```
TITLE1 'Análisis de correlación canónica';
TITLE2 'de los datos de la tabla 9.1';
DATA Tabla9_1; INPUT X1 X2 Y1 Y2 ; /*variables X1, X2, Y1 y Y2 */
CARDS; /*datos*/
       1.051  -0.435  0.083    0.538
      -0.419  -1.335  -1.347  -0.723
         ⋮       ⋮       ⋮       ⋮
0.421 1.215 -0.017 1.838;
PROC CANCORR ALL VNAME='X variables' WNAME='Y variables';
/*ALL imprime estadísticas simples, correlaciones entre las variables y la redundancia del ACC*/
/*VDEP o WDEP especifican, los rótulos para la variable X y las Y, respectivamente*/
VAR X1 X2; /*variables del primer conjunto */
WITH Y1 Y2; /*variables del segundo conjunto*/
RUN;
```

9.5 Procesamiento de datos con R

Se realiza el ejemplo 9.2.

```
# lectura de datos
datosY<-scan()
500  43  98  17   800 20  92   32   570 28  98   26   550 28  98
26   550 28  98   26  380 15   99   28  930 21   99   28  650 10
101  27  600 10   101 27  1500 19   99  23  1750 22   101 27
2000 58  100 18  2500 34  102 16   2000 21  105 20   7850 42  84
5    10500   50  81   -12

datosY<-matrix(datos,ncol=4,byrow=TRUE)
colnames(datosY)<-paste("Y",1:4,sep="")
datosX<-scan()
0   3   22  57  17  1   0   16  20  38  13  13  0   6   28  46
17  3   0   4   19  47  27  3   0   1   8   50  35  6   0   2  19
44  32  3   0   0   15  50  27  8   10  21  40  25  4   0   14
26  32  28  0   0   0   1   6   80  12  1   1   4   34  33  22
6   0   7   14  66  13  0   0   9   15  47  21  8   3   7   17
32  27  14  0   5   7   84  4   0   0   3   1   94  4   0
```

```
datosX<-matrix(datosX,ncol=6,byrow=TRUE)
colnames(datosX)<-paste("X",1:6,sep="")
ejemp9_2<-data.frame(cbind(datosX,datosY))
# matriz de correlaciones de los datos
round(cor(ejemp9_2),3)
# análisis de correlación canónica
corcan<-cancor(datosX,datosY)
# correlaciones
corcan$cor
# coeficientes estimados para las variables X
corcan$xcoef
# coeficientes estimados para las variables Y
corcan$ycoef
# valores usados para ajustar X
corcan$xcenter
# valores usados para ajustar Y
corcan$ycenter
```

Capítulo 10

Escalamiento multidimensional

10.1 Introducción

El *escalamiento multidimensional* (EM) es una técnica que, partiendo de las distancias o similitudes establecidas entre un conjunto de n objetos, intenta la construcción de una representación de estos en un espacio, generalmente un plano. La comparación tradicional de la forma como trabaja el escalamiento multidimensional es la construcción de un mapa (representación en un plano) en el cual se ubican unas ciudades de una región, teniendo las distancias entre ellas.

Si se tienen dos objetos, éstos quedan ubicados sobre una línea recta. Tres objetos pueden situarse en una línea recta o en un plano. Cuatro o más objetos pueden "dibujarse" en el espacio tridimensional o en un espacio de dimensión superior. La siguiente figura muestra un mapa[1] para los objetos A, B y C de acuerdo con la matriz distancias dada por

$$\boldsymbol{D} = \begin{pmatrix} & A & B & C \\ A & 0 & 4 & 3 \\ B & 4 & 0 & 3 \\ C & 3 & 3 & 0 \end{pmatrix}.$$

La figura 10.1 ilustra tres de las infinitas formas como pueden situarse en un mapa los objetos A, B y C de acuerdo con la matriz de distancias $\boldsymbol{D}$; los segmentos que unen los tres puntos se dibujan para facilitar la apreciación del "mapa".

[1] En mercadeo se acostumbra llamar "mapa perceptual"

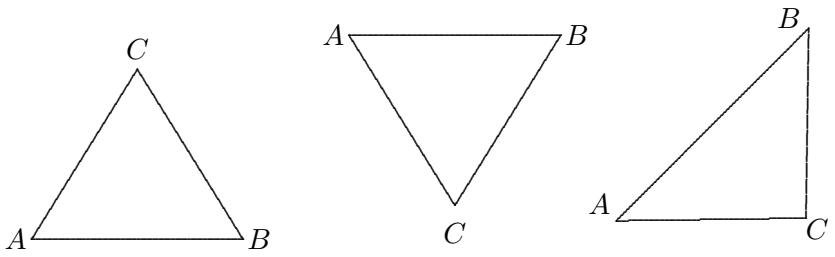


Figura 10.1 *Mapa de la similitud entre tres objetos.*

La técnica del *escalamiento multidimensional* (EM) emplea la *proximidad* entre objetos. Una proximidad es un número que indica la *similitud* o diferencia de dos objetos; es decir, el grado de similitud con el que son percibidos los objetos. El resultado principal de la metodología es una representación de los objetos en un espacio que, generalmente, tiene una dimensión menor al número de objetos y variables. En resumen, se trata de que a partir de una matriz de distancias o de similitudes entre objetos en un sistema de ejes referenciados (por ejemplo, factores), encontrar las coordenadas que "mejor" ubiquen tales objetos en un plano.

Entre algunas aplicaciones del escalamiento multidimensional (EM) se pueden señalar las siguientes:

- Identificar una tipología de productos (bienes o servicios), de acuerdo con algunos atributos percibidos por los consumidores.
- En antropología, permite estudiar las diferencias culturales de varios grupos, de acuerdo con sus creencias, lenguaje e información atribuible.
- En una contienda electoral se quiere encontrar las similitudes entre los candidatos, registrando la percepción de los potenciales electores.

Los principales propósitos del escalamiento multidimensional son los siguientes

i) Es un método que representa las (di)similitudes de los datos como distancias en un espacio (coordenadas) para hacer que éstos sean accesibles a la inspección visual y la exploración.

ii) Es una técnica que permite verificar si las diferencias, que distinguen a unos objetos de otros, se reflejen en la representación conseguida.

iii) Es una aproximación analítica a los datos que permite descubrir las dimesiones relevantes presentes en las (di) similitudes.

iv) Es un modelo que explica los criteros de las (di)similitudes en términos de una regla que "emula" un tipo de distancia particular.

Para reproducir las similitudes (o disimilaridades $\delta_{ii'}$) entre los individuos u objetos i e i' para $i, i' = 1, \ldots, n$, se pueden necesitar hasta $(n-1)$ dimensiones; el propósito del escalamiento multidimensional es encontrar una configuración en una dimensionalidad lo más baja posible, que reproduzca las similitudes (o las disimilaridades) dadas. Naturalmente la disposición en dos dimensiones tiene la gran ventaja de que los datos pueden ser fácilmente ubicados en el plano para mostrar algún patrón de asociación, en particular, la conformación de grupos o conglomerados.

Los principales procedimientos de escalamiento multidimensional son los siguientes:

1. *Clásico*, en el cual se asume que las distancias son de tipo euclidiano, y por lo tanto se corresponden con las disimilaridades; es decir, $d_{ii'} = \delta_{ii'}$. Se hace uso de la descomposición espectral de la matriz de disimilaridades doblemente centrada (sección (10.2)), para determinar el sistema de ejes referencial. Otra forma para desarrollar escalamiento clásico es mediante *mínimos cuadrados*, los cuales transforman las disimilaridades (o proximidades) en distancias $d_{ii'}$, mediante una función $f(\delta_{ii'})$, donde f es una función continua que debe preservar el orden de la disimilaridad; es decir, debe ser continua y monótona (creciente o decreciente). Por ejemplo, si se quiere que $d_{ii'} \approx f(\delta_{ii'}) = \alpha + \beta\delta_{ii'}$, se debe minimizar la distancia entre las $d_{ii'}$ y los respectivos puntos de una línea recta (función f); es decir, se deben encontrar los valores de α y β que minimicen la expresión

$$\frac{\sum_{i=1}^{n}\sum_{i'=1}^{n}(d_{ii'} - (\alpha + \beta\delta_{ii'}))^2}{\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{ii'}^2}.$$

Nótese que el símbolo $\approx$ se emplea para indicar un "ajuste", a manera de regresión, de los $\delta_{ii'}$ sobre los $d_{ii'}$.

2. *Ordinal o no métrico*, se emplea cuando la transformación de las disimilaridades no conservan la magnitud de las variables pero mantienen las propiedades de orden o monotonía; esto es, si se tiene que

$$\delta_{ii'} < \delta_{jj'}, \text{ entonces } f(\delta_{ii'}) < f(\delta_{jj'})$$

para todos los objetos $1 \leq i, i'j, j' \leq n$. Es decir, se ha preservado el orden de las disimilaridades; de aquí el calificativo de transformación *no métrica* (sección (10.3)).

3. *Análisis por acoplamiento o Procusto*, cuando el EM se ha desarrollado sobre algún conjunto de datos de disimilaridades, a través de dos configuraciones, los dos gráficos resultantes (capas) representan el mismo conjunto de objetos. Es el caso de objetos cuyas ubicaciones obedecen a dos tiempos o épocas distintas, como también la ubicación de productos de acuerdo con la percepción de dos grupos de personas diferentes.

El *análisis por acoplamiento* dilata, translada, refleja y rota una de las capas o configuraciones de puntos para mezclarla, tanto como sea posible, con el otro arreglo. El propósito es la comparación de las dos configuraciones. En otras palabras, se trata de comparar dos mapas que representan los mismos objetos (sección (10.4)).

Las medidas de *similitud* y *disimilaridad* requeridas se condensan en matrices de tamaño $(n \times n)$; la matriz de disimilaridad se nota por $\mathbf{\Delta} = (\delta_{ij})$. Las medidas de similitud son, frecuentemente, *coeficientes de similitud* y toman valores generalmente en el intervalo $[0, 1]$. Las medidas de similitud (δ') o de disimilaridad (δ) están estrechamente relacionadas en forma inversa; es decir, $\delta' = k - \delta$, donde k es una constante (generalmente igual a 1). Formalmente, δ' y δ son funciones del conjunto de pares de individuos en un espacio euclidiano. No siempre las medidas δ' y δ surgen por apreciaciones o percepciones, como ocurre en las encuestas de opinión o en la calificación de atributos sobre objetos, sino que pueden obtenerse a partir de una matriz de datos de tamaño $(n \times p)$.

En el capítulo 7 se presentan medidas de similitud tales como la distancia euclidiana, de Mahalanobis y de Minkowski, para variables continuas y los coeficientes de similitud para medidas en escala ordinal o nominal. En las tablas 10.1 y 10.2 se resumen algunas de las medidas de disimilaridad y similitud de uso más frecuente.

Para el caso de variables dicotómicas, la construcción de los coeficientes de similitud se hace conforme a los presentados en la sección (7.2.3). Por comodidad se transcribe nuevamente el concepto de similitud medido por tales coeficientes.

Al comparar dos objetos de acuerdo con un conjunto de p atributos dicotómicos $(p = a + b + c + d)$, se tienen cuatro posibilidades:

Tabla 10.1 Medidas de disimilaridad para datos cuantitativos

Nombre	Distancia o disimilaridad ($\delta_{ii'}$)
Distancia euclidiana	$\delta_{ii'} = \left\{\sum_j (X_{ij} - X_{i'j})^2\right\}^{\frac{1}{2}}$
Distancia euclidiana ponderada	$\delta_{ii'} = \left\{\sum_i w_j (X_{ij} - X_{i'j})^2\right\}^{\frac{1}{2}}$
Distancia de Mahalanobis	$\delta_{ii'} = \left\{(X_i - X_{i'})' \mathbf{\Sigma}^{-1} (X_i - X_{i'})\right\}^{\frac{1}{2}}$
Métrica de la ciudad	$\delta_{ii'} = \sum_j \lvert X_{ij} - X_{i'j}\rvert$
Métrica de Minkowski	$\delta_{ii'} = \left\{\sum_i \lvert X_{ij} - X_{i'j}\rvert^{\lambda}\right\}^{\frac{1}{\lambda}}$ con $\lambda \geq 1$
Métrica de Canberra	$\delta_{ii'} = \sum_j \lvert X_{ij} - X_{i'j}\rvert / (X_{ij} + X_{i'j})$
Métrica de Bray-Curtis	$\delta_{ii'} = \frac{1}{p} \dfrac{\sum_j \lvert X_{ij} - X_{i'j}\rvert}{\sum_j (X_{ij} - X_{i'j})}$
Distancia de Bhattacharyya	$\delta_{ii'} = \left\{\sum_j (X_{ij}^{\frac{1}{2}} - X_{i'j}^{\frac{1}{2}})\right\}^{\frac{1}{2}}$
Separación angular	$\delta_{ii'} = \dfrac{\sum_j (X_{ij} X_{i'j})}{\left\{(\sum_j X_{ij}^2)(\sum_j X_{i'j}^2)\right\}^{\frac{1}{2}}}$
Correlación	$\delta_{ii'} = 1 - \dfrac{\sum_j (X_{ij} - \bar{X}_i)(X_{i'j} - \bar{X}_{i'})}{\left\{\sum_j (X_{ij} - \bar{X}_i)^2 \sum_j (X_{i'j} - \bar{X}_{i'})^2\right\}^{\frac{1}{2}}}$

1. Que ambos tengan presente el carácter comparado $(1, 1)$.
2. Que ambos tengan ausente el carácter comparado $(0, 0)$.
3. Que el primero tenga el carácter presente y el segundo ausente $(1, 0)$.
4. Que el primero tenga el carácter ausente y el segundo presente $(0, 1)$.

La frecuencia con la cual se presentan estas cuatro características se resume en una tabla 2x2 como la siguiente.

	Objeto i′	
Objeto i	1	0
1	(a)	(b)
0	(c)	(d)

Con estas frecuencias se construyen coeficientes de similitud tales como los que se muestran en la tabla 10.2.

Tabla 10.2 Coeficientes de similitud para datos binarios

Nombre	Similitud ($\delta'_{ii'}$)
Czekanowski, Sϕrensen, Dice	$\delta'_{ii'} = \frac{2a}{2a+b+c}$
Hamman	$\delta'_{ii'} = \frac{(a+d)-(b+c)}{a+b+c+d}$
Coeficiente de Jaccard	$\delta'_{ii'} = \frac{a}{a+b+c}$
Kulezynski	$\delta'_{ii'} = \frac{a}{a+b}$
Mountford	$\delta'_{ii'} = \frac{2a}{a(b+c)+2bc}$
Mozley, Margalef	$\delta'_{ii'} = \frac{a(a+b+c+d)}{(a+b)(a+c)}$
Ochiai	$\delta'_{ii'} = \frac{a}{[(a+b)(a+c)]^{\frac{1}{2}}}$
Phi	$\delta'_{ii'} = \frac{ad-bc}{[(a+b)(a+c)(b+d)(c+d)]^{\frac{1}{2}}}$
Rogers, Tanimoto	$\delta'_{ii'} = \frac{a+d}{a+2b+2c+d}$
Coeficiente asociación simple	$\delta'_{ii'} = \frac{a+d}{a+b+c+d}$
Rusell, Rao	$\delta'_{ii'} = \frac{a}{a+b+c+d}$
Yule	$\delta'_{ii'} = \frac{ad-bc}{ad+bc}$

Ejemplo 10.1 Sobre las especies león, jirafa, vaca, oveja y humanos se observaron, en forma de presencia/ausencia, los seis atributos siguientes: (A) la especie tiene cola, (B) la especie es un animal silvestre, (C) la especie tiene cuello largo, (D) la especie es un animal de granja (E) la especie come otros animales y (F) la especie camina sobre cuatro patas. Los datos se muestran en la tabla siguiente.

Especie	Atributo					
	A	B	C	D	E	F
León	1	1	0	0	1	1
Jirafa	1	1	1	0	0	1
Vaca	1	0	0	1	0	1
Oveja	1	0	0	1	0	1
Humano	0	0	0	0	1	0

Para el león y la jirafa las frecuencias de la similitudes son $a = 3$, $b = 1$, $c = 1$ y $d = 1$. El coeficiente de asociación simple para estos dos animales es $\delta'_{ii'} = \frac{a+d}{a+b+c+d} = \frac{3+1}{3+1+1+1} = \frac{4}{6}$. Esta medida de similitud se transforma en una medida de distancia (disimilaridad) al hacer $d_{ii'} = 1 - \delta'_{ii'}$. Así, la distancia entre un león y una jirafa es de 1/3 (los más cercanos). La tabla siguiente contiene tal distancia entre los seis animales. Para una

	León	Jirafa	Vaca	Oveja	Humano
León	-	1/3	1/2	1/2	1/2
Jirafa	1/3	-	1/2	1/2	5/6
Vaca	1/2	1/2	-	0	1/3
Oveja	1/2	1/2	0	-	2/3
Humano	1/2	5/6	2/3	2/3	—

variable j de tipo *nominal* ($j = 1, \ldots, p$), si los objetos i e i' comparten alguna de sus modalidades, entonces, se considera $\delta'_{ii'j} = 1$, ó $\delta'_{ii'j} = 0$ si no "caen" en la misma categoría. Una medida de similitud entre dos objetos, con respecto a p-variables es el promedio de ellas; es decir, $\sum_{j=1}^{p} \delta'_{ii'}/p$. Además, si se tiene información adicional sobre las relaciones entre varias categorías, entonces es necesario que $\delta'_{ii'j}$ exprese tal relación mediante un valor apropiado.

La última situación se puede ilustrar a través del siguiente ejemplo (Cox y Cox, 1994 pág. 12). Supóngase que se considera la variable "forma de una botella" con las categorías "estándar" (st), "corta y cilíndrica" (cc), "alta y cilíndrica" (ac) y "sección cuadrada" (sc). Los siguientes puntajes de similitud pueden ser apropiados para relacionar dos botellas i e i'. Nótese la similitud entre las botellas cilíndricas (1.0, 0.5 y 0.3), a comparación de la similitud entre las botellas cilíndricas y una de sección cuadrada (0.0).

		$Botella_{i'}$			
		st	cc	ac	sc
	st	1.0	0.5	0.5	0.0
$Botella_i$	cc	0.5	1.0	0.3	0.0
	ac	0.5	0.3	1.0	0.0
	sc	0.0	0.0	0.0	1.0

Para el caso de *variables ordinales* con k categorías se pueden construir $(k-1)$ variables indicadoras para representar estas categorías. De esta forma las variables indicadoras pueden emplearse para obtener coeficientes de similitud como el $\delta'_{ii'j}$ anterior. Por ejemplo, nuevamente con relación a una botella, si la variable es "altura de la botella" con las modalidades: pequeña, estándar, alta, larga y delgada, entonces la variable *altura de la botella* puede categorizarse como se muestra enseguida:

	V. indicadora		
$Modalidad$	I_1	I_2	I_3
pequeña :	0	0	0
estándar :	1	0	0
alta :	1	1	0
larga y delgada :	1	1	1

Sobre estas variables indicadoras se puede aplicar alguno de los coeficientes de similitud anteriores, por ejemplo, al comparar mediante el *coeficiente de asociación simple* las botellas i e i' respecto a las modalidades *estándar*, y *larga y delgada*, el valor del coeficiente es: $\delta'_{ii'} = \frac{a+d}{a+b+c+d} = \frac{1+0}{1+0+2+0} = 0.33$.

Como se expresó arriba, una medida de similitud puede *transformarse* en una medida de disimilaridad. Las transformaciones más comunes son:

$$\delta_{ii'} = 1 - \delta'_{ii'}, \quad \delta_{ii'} = k - \delta'_{ii'} \quad \text{y} \quad \delta_{ii'} = \{2(1-\delta'_{ii'})\}^{\frac{1}{2}}.$$

Las cuales se escogen de acuerdo con el problema a tratar.

10.2 Escalamiento clásico

El escalamiento clásico es un método algebraico, para encontrar y reconstruir la configuración de objetos a partir de sus disimilaridades, este método es apropiado cuando las disimilaridades son distancias euclidianas o aproximadamente éstas (Chatfield y Collins, 1986, pág. 198).

Supóngase n objetos con disimilaridades $\{\delta_{ii'}\}$. Con el EM se intenta encontrar un conjunto de puntos en un espacio donde cada punto represente uno de los objetos y las distancias entre los puntos $(d_{ii'})$ sean tales que

$$d_{ii'} \approx f(\delta_{ii'}),$$

donde f es una función continua y monótona (creciente o decreciente) de la disimilaridad; esta función puede ser la identidad.

Como se insinúa en la figura 10.1, dado un conjunto de distancias euclidianas la representación de los puntos concordantes con éstas no es única, pues no hay una *localización* y *orientación* única de la configuración. La localización generalmente se obvia tansladando el origen del arreglo al centroide o "centro de gravedad" de los datos. Para el problema de la orientación la configuración se obtiene mediante cualquier transformación ortogonal, por ejemplo una rotación rígida del tipo ACP, la cual deja invariante las distancias y los ángulos entre los puntos.

10.2.1 Cálculo de las coordenadas a partir de las distancias euclidianas

Conocidas las coordenadas de n-puntos en un espacio euclidiano de p dimensiones, se pueden calcular las distancias euclidianas entre estos puntos. Esto se hace a través de la matriz de datos $\mathbb{X}$ o mediante la matriz $\boldsymbol{B} = \mathbb{X}\mathbb{X}'$, la cual contiene las sumas de cuadrados y productos cruzados (entre individuos), explícitamente

$$
\begin{aligned}
\boldsymbol{B} &= \mathbb{X}\mathbb{X}' \\
&= \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1p} \\
X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{ip} \\
\vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{nj} & \cdots & X_{np}
\end{pmatrix}
\begin{pmatrix}
X_{11} & X_{21} & \cdots & X_{i'1} & \cdots & X_{n1} \\
X_{12} & X_{22} & \cdots & X_{i'2} & \cdots & X_{n2} \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
X_{1j} & X_{2j} & \cdots & X_{i'j} & \cdots & X_{nj} \\
\vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\
X_{1p} & X_{2p} & \cdots & X_{i'p} & \cdots & X_{np}
\end{pmatrix} \\
&= \begin{pmatrix}
\sum_{j=1}^{p} X_{1j}^2 & \sum_{j=1}^{p} X_{1j}X_{2j} & \cdots & \sum_{j=1}^{p} X_{1j}X_{i'j} & \cdots & \sum_{j=1}^{p} X_{1j}X_{nj} \\
\sum_{j=1}^{p} X_{2j}X_{1j} & \sum_{j=1}^{p} X_{2j}^2 & \cdots & \sum_{j=1}^{p} X_{2j}X_{i'j} & \cdots & \sum_{j=1}^{p} X_{2j}X_{nj} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\sum_{j=1}^{p} X_{ij}X_{1j} & \sum_{j=1}^{p} X_{ij}X_{2j} & \cdots & \boxed{\sum_{j=1}^{p} X_{ij}X_{i'j}} & \cdots & \sum_{j=1}^{p} X_{ij}X_{nj} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
\sum_{j=1}^{p} X_{nj}X_{1j} & \sum_{j=1}^{p} X_{nj}X_{2j} & \cdots & \sum_{j=1}^{p} X_{nj}X_{i'j} \cdots & & \sum_{j=1}^{p} X_{nj}^2
\end{pmatrix}.
\end{aligned}
$$

El término $(b_{ii'})$ de la matriz $\boldsymbol{B}$ está dado por

$$b_{ii'} = \sum_{j}^{p} X_{ij}X_{i'j}, \text{ para } i, i' = 1, \ldots, n. \tag{10.1}$$

La matriz que contiene estas distancias es la matriz $\boldsymbol{D}$ de tamaño $(n \times n)$ o la matriz de dismilaridades $\boldsymbol{\Delta}$, cuyo elemento $d_{ii'}$. Por el término $d^2_{ii'}$ se entiende el "*cuadrado de la distancia euclidiana entre los objetos" i e i'*; es decir,

$$\begin{aligned} d^2_{ii'} &= \sum_{j=1}^{p}(X_{ij} - X_{i'j})^2 \\ &= \sum_{j=1}^{p} X^2_{ij} + \sum_{j=1}^{p} X^2_{i'j} - 2\sum_{j=1}^{p} X_{ij}X_{i'j} \\ &= b_{ii} + b_{i'i'} - 2b_{ii'}. \end{aligned} \tag{10.2}$$

De esta forma, mediante la matriz $\boldsymbol{B}$ es posible encontrar la matriz de distancias a través de la ecuación (10.2). Pero como se ha insistido el problema del escalamiento multidimensional es precisamente el recíproco; es decir, conocidas las distancias entre los objetos, se deben buscar sus coordenadas con respecto a un espacio donde queden "mejor" representados. Supóngase que se conoce la matriz de distancias y por tanto el cuadrado de éstas, para encontrar las coordenadas de los objetos, primero se encuentra la matriz $\boldsymbol{B}$ y luego se factoriza de la forma $\boldsymbol{B} = \mathbb{X}\mathbb{X}'$, y a partir de esta representación se debe hallar la matriz $\mathbb{X}$.

Los elementos $b_{ii'}$ están asociados con la matriz cuyo elemento genérico está dado por (10.2), bajo la restricción de que los datos deben estar centrados alrededor de cero; es decir, que $\bar{\mathbb{X}} = \mathbf{0}$, entonces $\sum_{j=1}^{p} X_{ij} = 0$ para todo $j = 1, \ldots, p$. Al sumar en la ecuación (10.2) respecto a i, i' y los dos i e i', respectivamente, se obtiene

$$\sum_{i=1}^{n} d^2_{ii'} = \sum_{i=1}^{n} b_{ii'} + nb_{i'i'} \tag{10.3a}$$

$$\sum_{i'=1}^{n} d^2_{ii'} = nb_{ii} + \sum_{i=1}^{n} b_{ii'} \tag{10.3b}$$

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2_{ii'} = 2n\sum_{i=1}^{n} b_{ii'}. \tag{10.3c}$$

La solución del sistema de ecuaciones compuesto por (10.2) y (10.3) es

$$b_{ii'} = -\frac{1}{2}\Big[d^2_{ii'} - d^2_{i.} - d^2_{.i'} + d^2_{..}\Big], \tag{10.4}$$

donde $d^2_{i.}$, $d^2_{.i'}$ y $d^2_{..}$ son el promedio por fila, columna y global de la matriz de distancias al cuadrado, respectivamente. Así, cada una de las entradas de la matriz $\boldsymbol{B}$ se obtiene a través de la ecuación (10.4).

Ahora, para obtener la matriz de coordenadas $\mathbb{X}$ a partir de la matriz $\boldsymbol{B}$, se procede como a continuación se describe. Dado que la matriz $\boldsymbol{B}$ es de tamaño $(n \times n)$, semidefinida positiva, simétrica y de rango p (con $p \leq n$), entonces, $\boldsymbol{B}$ tiene p-valores propios no nulos y $(n-p)$ valores propios nulos. En consecuencia, la matriz $\boldsymbol{B}$ se puede escribir, de acuerdo con la descomposición espectral (A2.27), de la forma siguiente

$$\boldsymbol{B} = \mathbb{X}\mathbb{X}' = \boldsymbol{L}\boldsymbol{\Lambda}\boldsymbol{L}', \tag{10.5}$$

donde $\boldsymbol{\Lambda} = \text{Diag}(\lambda_1, \lambda_2, \cdots \lambda_n)$, es la matriz diagonal de los valores propios $\{\lambda_i\}$ de $\boldsymbol{B}$, y $\boldsymbol{L} = [l_1, l_2, \ldots, l_n]$ la matriz de vectores propios normalizados. Por conveniencia los valores propios se han ordenado en forma descendente; es decir, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$.

Como hay $(n-p)$ valores propios nulos, entonces la matriz $\boldsymbol{B}$ puede escribirse

$$\boldsymbol{B} = \boldsymbol{L}_1\boldsymbol{\Lambda}_1\boldsymbol{L}_1', \tag{10.5a}$$

donde $\boldsymbol{\Lambda}_1 = \text{Diag}(\lambda_1, \lambda_2, \cdots \lambda_p)$ y $\boldsymbol{L}_1 = [l_1, l_2, \ldots, l_p]$.

Por tanto, las coordenadas de la matriz $\mathbb{X}$ están dadas por

$$\mathbb{X} = \boldsymbol{L}_1\boldsymbol{\Lambda}_1^{\frac{1}{2}} = \left[\sqrt{\lambda_1}l_1, \sqrt{\lambda_2}l_2, \ldots, \sqrt{\lambda_p}l_p\right] = [\mathfrak{f}_1, \mathfrak{f}_2, \ldots, \mathfrak{f}_p]$$

donde $\boldsymbol{\Lambda}^{\frac{1}{2}} = \text{Diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_p})$.

Nótese que puede darse $r(\boldsymbol{B}) = r(\mathbb{X}\mathbb{X}') = k \leq p$, en tal caso, la configuración obtenida es una matriz $\mathbb{X}$ de tamaño $(n \times k)$.

De esta forma las coordenadas de los puntos han sido "recuperadas" con base en la matriz de distancias entre los puntos.

El procedimiento de escalamiento clásico se puede resumir en los siguientes pasos.

1. Computar la matriz $\boldsymbol{D}$ y $\boldsymbol{D}^2$ (o también $\boldsymbol{\Delta}$ y $\boldsymbol{\Delta}^2$).
2. Aplicar el doble centrado sobre esta matriz para obtener la matriz $\boldsymbol{B}$, cuyo elemento genérico es dado por la igualdad (10.4), es decir,

$$\boldsymbol{B} = -\frac{1}{2}\boldsymbol{J}\boldsymbol{D}^2\boldsymbol{J}, \text{ donde } \boldsymbol{J} = \boldsymbol{I} - n^{-1}(\boldsymbol{1}\boldsymbol{1}').$$

3. Obtener los valores propios de $\boldsymbol{B}$, de acuerdo con la descomposición espectral dada en (10.5); es decir,

$$\boldsymbol{B} = \mathbb{X}\mathbb{X}' = \boldsymbol{L}\boldsymbol{\Lambda}\boldsymbol{L}'.$$

4. Sea r la dimensionalidad de la solución. La matriz $\mathbf{\Lambda}_1$ denota la matriz diagonal con los primeros r valores propios positivos y $\boldsymbol{L}_1$ las primeras r columnas de $\boldsymbol{L}$. Entonces la matriz de coordenadas, desde el escalamiento clásico, está dada por

$$\mathbb{X} = \boldsymbol{L}_1 \mathbf{\Lambda}_1^{\frac{1}{2}}$$

10.2.2 Relación entre escalamiento clásico (EM) y análisis de componentes principales (ACP)

El EM clásico está orientado al análisis de una matriz de disimilaridades de tamaño $(n \times n)$, la cual se puede aproximar a una matriz de distancias euclidianas; es decir, $\delta_{ii'} \approx d_{ii'}$. Para investigar la conexión entre el ACP y el EM se asume, sin pérdida de generalidad, que el primero es desarrollado a partir de una matrix $\mathbb{X}$ corregida por la media (Chatfield y Collins 1986, pág. 200), y para el EM clásico se construye una matriz de distancias euclidianas de tamaño $(n \times n)$ y se desarrolla el análisis arriba descrito. Si la matriz $\mathbb{X}$ es de rango $k < min\{n, p\}$, se puede obtener una nueva configuración de los datos, con una matriz $\mathbb{X}^*$ de tamaño $(n \times k)$, la cual no siempre es igual a la matriz de datos originales. El análisis, como en ACP (capítulo 5), consiste en encontrar los valores propios de la matriz $\mathbb{X}\mathbb{X}'$.

Para mostrar la conexión entre las dos técnicas se retoma el procedimiento seguido en la sección (5.2). Los valores propios de la matriz de covarianzas son proporcionales a $\mathbb{X}'\mathbb{X}$. Sean $\{\lambda_i\}$ y $\{l_i\}$ los valores y vectores propios de la matriz $\mathbb{X}'\mathbb{X}$, entonces

$$(\mathbb{X}'\mathbb{X})l_i = \lambda_i l_i$$

premultiplicando por $\mathbb{X}$, se obtiene

$$(\mathbb{X}\mathbb{X}')\mathbb{X}l_i = \lambda_i \mathbb{X}l_i$$

De esta ecuación se observa que los valores propios de la matriz $\mathbb{X}\mathbb{X}'$ son los mismos que los de la matriz $\mathbb{X}'\mathbb{X}$, mientras que los vectores propios l_i^* están relacionados con los de $\mathbb{X}'\mathbb{X}$ por una simple transformación lineal de la forma Xl_i, así los li deben ser proporcionales a $\mathbb{X}l_i$. Nótese que l_i^* es de tamaño $(n \times 1)$, mientras que l_i es de tamaño $(p \times 1)$. El vector $\mathbb{X}l_i$ suministra las componentes de las coordenadas de los individuos respecto al i-ésimo eje principal, y la suma de los cuadrados de sus componentes $(\mathbb{X}l_i)$ es igual a λ_i.

En conclusión, la siguiente igualdad se tiene

$$\sqrt{\lambda_i} l_i^* = \mathbb{X} l_i,$$

excepto, posiblemente, por el signo.

Así las coordenadas se pueden obtener directamente desde los valores propios de la matriz $\mathbb{X}\mathbb{X}'$, a través de la relación

$$\mathfrak{f}_i = \sqrt{\lambda_i} l_i^*.$$

De esta forma se concluye que los resultados del ACP son equivalentes al EM clásico si las distancias obtenidas desde la matriz de datos son euclidianas.

A través de los siguientes ejemplos se pretende mostrar la interpretación de algunos resultados del EM clásico.

Ejemplo 10.2 Una de las aplicaciones más sencillas del EM es la reconstrucción de un mapa, desde las distancias entre sus puntos. Se intenta reconstruir el mapa de 13 ciudades de Colombia desde la matriz de distancias entre éstas, en kilómetros carreteables[2]. Las ciudades y su respetiva sigla son: Barranquilla (Bl), Bogotá (Bt), Bucaramanga (Bg), Cali (Ca), Cartagena (Cg), Cúcuta (Cu), Manizález (Mz), Medellin (Ml), Pasto (Pt), Pereira (Pr), Quibdó (Qd), Riohacha (Rh) y Santa Marta (Sm).

La matriz de distancias entre las ciudades es la siguiente:

	Bl	Bt	Bg	Ca	Cg	Cu	Mz	Ml	Pt	Pr	Qd	Rh	Sm
Bl	0												
Bt	1302	0											
Bg	793	439	0										
Ca	1212	484	923	0									
Cg	124	1178	917	1088	0								
Cu	926	649	210	1133	1050	0							
Mz	1003	299	738	275	879	984	0						
Ml	750	552	1543	462	626	1201	253	0					
Pt	1612	884	1323	400	1488	1533	675	862	0				
Pr	1054	330	769	224	930	979	51	304	624	0			
Qd	998	800	1791	710	874	1449	501	248	1110	552	0		
Rh	284	1147	708	1403	408	1024	1194	941	1803	1245	1189	0	
Sm	83	1139	700	1305	217	833	1096	843	1705	1147	1091	191	0

La disposición en dos dimensiones facilita la interpretación. Mediante el procedimiento *MDS* del paquete estadístico *SAS* se obtiene la configuración de las 13 ciudades, partiendo de la matriz de distancias por carretera entre éstas. Se puede apreciar una alta aproximación con la representación

[2]Datos suministrados por el Instituto Geográfico Agustin Codazzi, 1998.

geográfica, como se muestra en la figura 10.2. La gráfica se ha construido tomando *"Dimensión 1"* en el eje vertical y *"Dimensión 2"* en el eje horizontal. Es preciso advertir que el EM no tendrá mucha importancia en problemas como el anterior, en donde por cartografía se sabe previamente la ubicación de los objetos o individuos; la intención no es más que ilustrativa.

A veces la interpretación con relación a ejes dispuestos en forma canónica no es sencilla, resulta entonces ventajoso realizar una rotación arbitraria de los ejes, que permita extraer más información de tal representación.

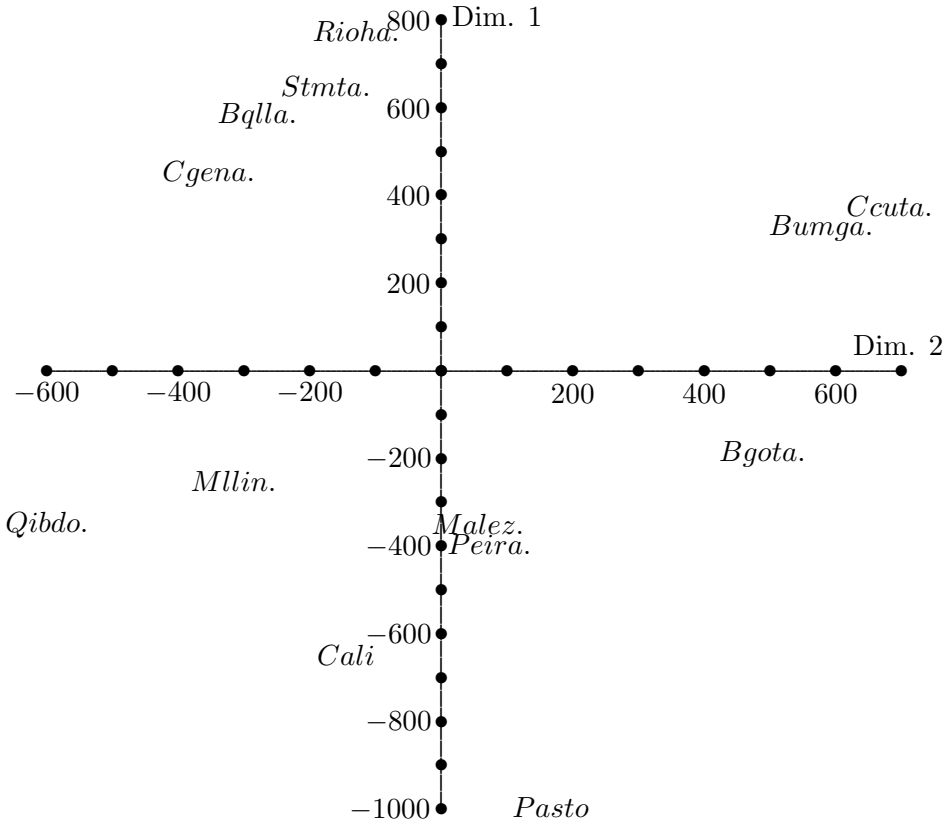


Figura 10.2 *Mapa de Colombia (Región Andina) construido por EM.*

Las coordenadas del arreglo anterior, suministradas por el procedimiento *MDS* del *SAS*, son

$Ciudad:$	Bl	Bt	Bg	Ca	Cg	Cu	Mz
$Dim.1:$	582.97	−188.4	327.0	−645.8	450.6	376.4	−348.5
$Dim.2:$	−279.8	488.2	577.4	−144.9	−353.7	681.7	32.2

$Ciudad:$	Ml	Pt	Pr	Qd	Rh	Sm
$Dim.1:$	−248.7	−994.5	−393.3	−350.0	782.4	650.2
$Dim.2:$	−315.8	166.3	53.1	−599.8	−129.3	−175.8 ☑

La búsqueda del significado para la configuración obtenida es uno de los principales propósitos del EM. Kruskal y Wish (1978, pág. 36) sugieren emplear la regresión lineal entre las variables asociadas a las coordendas de la configuración (variables regresoras) y alguna variable ligada con éstas (variable dependiente). Otra interpretación se obtiene sobre la conformación de grupos o conglomerados de puntos.

Para configuraciones en tres dimensiones puede tenerse más problemas en la interpretación. Una estrategia útil es la configuración en dos dimensiones simultáneas con los pares de ejes; por ejemplo *eje 1 vs. eje 2*, *eje 1 vs. eje 3* o *eje 2 vs. eje 3*.

Ejemplo 10.3 En un grupo de 30 estudiantes se preguntó acerca de la tasa de disimilaridad, en una escala de 0 a 9, entre los rostros de una mujer que actúa en cuatro escenas diferentes representadas en cuatro láminas. La disimilaridad fue definida como "una diferencia en expresión emocional o felicidad". Las escenas son las siguientes (Borg y Groenen 1997, págs. 209-210)

1. Tristeza por la muerte de la madre (✠).
2. Saboreando un refresco (⌣).
3. Una sorpresa agradable (♣).
4. Amor maternal hacia un bebé de brazos (♡).

La matriz de disimilaridad y su cuadrado, para emplearla de acuerdo con la igualdad (10.4), son respectivamente

$$\boldsymbol{\Delta} = \begin{pmatrix} 0.00 & 4.05 & 8.25 & 5.57 \\ 4.05 & 0.00 & 2.54 & 2.69 \\ 8.25 & 2.54 & 0.00 & 2.11 \\ 5.57 & 2.69 & 2.11 & 0.00 \end{pmatrix}, \text{ tal que } \boldsymbol{\Delta}^2 = \begin{pmatrix} 0.00 & 16.40 & 68.06 & 31.02 \\ 16.40 & 0.00 & 6.45 & 7.24 \\ 68.06 & 6.45 & 0.00 & 4.45 \\ 31.02 & 7.24 & 4.45 & 0.00 \end{pmatrix}.$$

El segundo paso es centrar doblemente; así:

$$
\begin{aligned}
\boldsymbol{B} &= -\frac{1}{2}\boldsymbol{J}\boldsymbol{\Delta}^2\boldsymbol{J} \\
&= -\frac{1}{2}\begin{pmatrix} 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{pmatrix}\begin{pmatrix} 0.00 & 16.40 & 68.06 & 31.02 \\ 16.40 & 0.00 & 6.45 & 7.24 \\ 68.06 & 6.45 & 0.00 & 4.45 \\ 31.02 & 7.24 & 4.45 & 0.00 \end{pmatrix} \\
&\quad \cdot \begin{pmatrix} 3/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \end{pmatrix} \\
&= \begin{pmatrix} 20.52 & 1.64 & -18.08 & -4.09 \\ 1.64 & -0.83 & 2.05 & -2.87 \\ -18.08 & 2.05 & 11.39 & 4.63 \\ -4.09 & -2.87 & 4.63 & 2.33 \end{pmatrix}.
\end{aligned}
$$

El tercer paso es calcular los valores propios de la descomposición de la matriz $\boldsymbol{B}$ y con éstos se obtienen las matrices $\boldsymbol{L}$ y $\boldsymbol{\Lambda}$. Éstas son:

$$
\boldsymbol{L} = \begin{pmatrix} 0.77 & 0.04 & 0.50 & -0.39 \\ 0.01 & -0.61 & 0.50 & 0.61 \\ -0.61 & -0.19 & 0.50 & -0.59 \\ -0.18 & 0.76 & 0.50 & 0.37 \end{pmatrix}
$$

y

$$
\boldsymbol{\Lambda} = \begin{pmatrix} 35.71 & 0.00 & 0.00 & 0.00 \\ 0.00 & 3.27 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & -5.57 \end{pmatrix}.
$$

Hay dos valores propios positivos, uno cero por el doble centrado y uno negativo. Para este caso, se pueden reconstruir a lo más dos dimensiones.

El último paso señala que la configuración (coordenadas) de la matriz $\mathbb{X}$ se encuentra mediante la expresión

$$
\mathbb{X} = \boldsymbol{L}_1\boldsymbol{\Lambda}_1^{\frac{1}{2}} = \begin{pmatrix} 0.77 & 0.04 \\ 0.01 & -0.61 \\ -0.61 & -0.19 \\ -0.18 & 0.76 \end{pmatrix}\begin{pmatrix} 5.98 & 0.00 \\ 0.00 & 1.81 \end{pmatrix} = \begin{pmatrix} 4.62 & 0.07 \\ 0.09 & -1.11 \\ -3.63 & -0.34 \\ -1.08 & 1.38 \end{pmatrix}.
$$

En la figura 10.3 se ubica la matriz $\mathbb{X}$ que contiene las coordenadas de las cuatro expresiones faciales.

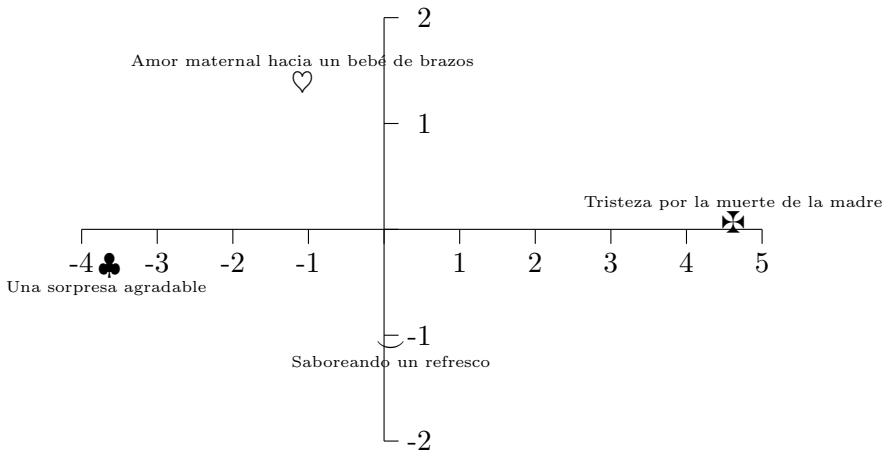


Figura 10.3 *Posicionamiento de las cuatro expresiones faciales.*

No obstante, que se trata de un ejemplo simplificado se pueden aventurar algunas conclusiones sobre la ubicación de estos "estímulos" y los ejes de referencia. De una parte, nótese que sobre el eje horizontal está lo relacionado con afectos mientras que en la parte inferior se ubica lo placentero pero un poco más tangible (material); el eje horizontal determina el estado de ánimo, triste del lado derecho y agrado del lado izquierdo. Además, la ubicación de los estados de ánimo, tristeza por muerte y alegría por el hijo, dan cuenta de la reciprocidad del afecto hijo-madre y madre-hijo, pero, en estados opuestos. ☑

10.3 Escalamiento ordinal o no métrico

En la sección (10.2) se consideró el problema de encontrar k-vectores

$[\mathfrak{f}_1, \mathfrak{f}_2, \ldots, \mathfrak{f}_k]$ en un espacio cuyas distancias $d_{ii'}$ fueran lo más cercanas posible al conjunto de disimilaridades $\{\delta_{ii'}\}$. De otra manera, el objetivo es encontrar n-puntos cuyas distancias estén en concordancia con las disimilaridades dadas para los n-objetos o individuos. El valor de k no se tiene de antemano, generalmente se prueba con valores de 2 o de 3 dimensiones para la configuración o "mapa" de los n puntos.

La diferencia básica del escalamiento *no métrico* con el EM clásico, es que se emplea tan sólo el *rango o puesto* que ocupa cada disimilaridad con respecto a las demás, muy útil para casos en los cuales las disimilaridades están en una escala nominal u ordinal.

La asignación de rangos a las observaciones se presenta en muchas aplicaciones, para las cuales el valor exactamente numérico de la medición no tiene mucho significado o importancia. Es el caso, por ejemplo, de los conceptos emitidos por una persona para calificar la calidad de un objeto o para establecer el grado de disimilaridad entre dos objetos o individuos.

Algún grado de distorsión (ruido) puede admitirse en la representación, siempre que el orden de acuerdo con el rango del $d_{ii'}$, sea el mismo del respectivo $\delta_{ii'}$. Se puede ampliar entonces la búsqueda de la configuración consiguiendo una transformación f tal que

$$d_{ii'} \approx f(\delta_{ii'}),$$

con f una función monótona creciente, es decir $\delta_{ii'} < \delta_{jj'}$ si y solo si $f(\delta_{ii'}) < f(\delta_{jj'})$.

La ubicación de un conjunto con n-objetos, como el caso clásico, no se puede establecer de manera única (gráfica 10.1). Para superar el problema de localización de la configuración, se translada el centroide de ésta al origen.

El procedimiento general para el escalamiento *ordinal o no métrico* es el siguiente. En primer lugar se "adivina" (o mejor se intenta) una configuración en el espacio k dimensional, se calculan las distancias euclidianas entre cada par de puntos i e i', notadas por $\hat{d}_{ii'}$, en este espacio se comparan las distancias con las disimilaridades observadas inicialmente. Si el puesto o rango de las $\hat{d}_{ii'}$ es el mismo que el de las $\delta_{ii'}$, entonces se ha conseguido una buena configuración de los objetos.

Por ejemplo, supóngase que se tienen 4 objetos para los cuales las disimilaridades están en el siguiente orden

$$\delta_{14} < \delta_{24} < \delta_{13} < \delta_{34} < \delta_{12} < \delta_{23},$$

y además, que las distancias ajustadas en el espacio bidimensional son tales que

$$\hat{d}_{14} \leq \hat{d}_{24} \leq \hat{d}_{13} \leq \hat{d}_{34} \leq \hat{d}_{12} \leq \hat{d}_{23}.$$

Encontrar una configuración como la anterior, en la cual se mantenga el orden entre las $\hat{d}_{ii'}$ y las $\delta_{ii'}$ es una situación casi imposible o supremamente rara en la práctica. A cambio, hay que conformarse con una configuración en un espacio de dimensión determinada, donde el orden de las $\hat{d}_{ii'}$ ajustadas, esté tan cerca como sea posible al orden de las $\delta_{ii'}$. La palabra "cerca" se mide por la concordancia entre el orden de las distancias ajustadas en el espacio de las $\hat{d}'s$ y el orden de los $\delta's$ para los objetos i e i'.

La *bondad del ajuste* de cualquier configuración propuesta se encuentra mediante la construcción de una regresión mínimo cuadrática que relacione las $\hat{d}'s$ y las $\delta's$. Una estrategia gráfica para medir este ajuste consiste en disponer sobre el eje horizontal, las disimilaridades y sobre el respectivo eje vertical las distancias; éste se conoce con el nombre de *diagrama de Shepard*. Los $\binom{n}{2} = n(n-1)/2$ puntos dispuestos en el diagrama deben, en el mejor de los casos, nuevamente configurarse en una línea poligonal creciente. La figura 10.4 muestra el diagrama de Shepard para los cuatro objetos anteriores. En el caso (a) se tiene un ajuste "perfecto", mientras que en el caso (b) las distancias y las disimilaridades no concuerdan tan perfectamente, pues la relación de orden es alterada entre $\hat{d}_{24}$ y $\hat{d}_{13}$, y entre $\hat{d}_{12}$ y $\hat{d}_{23}$, la cual no está en correspondencia con el orden creciente de las disimilaridades. El orden de las disimilaridades y las distancias es, respectivamente,

$$\delta_{23} > \delta_{12} > \delta_{34} > \delta_{13} > \delta_{24} > \delta_{14} \quad \text{y}$$
$$\hat{d}_{12} > \hat{d}_{23} > \hat{d}_{34} > \hat{d}_{24} > \hat{d}_{13} > \hat{d}_{14}.$$

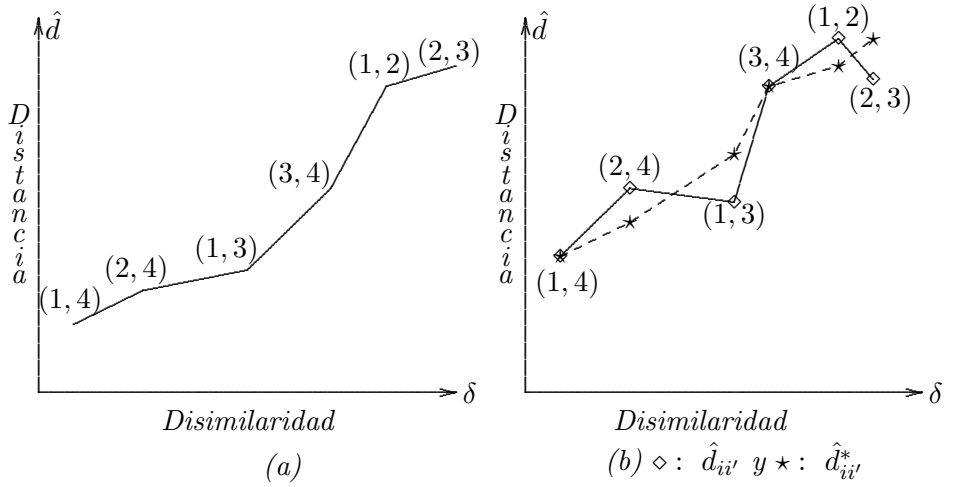


Figura 10.4 *Diagramas de Shepard.*

Se puede ajustar un nuevo conjunto de distancias $\hat{d}^*_{ii'}$ tal que cumpla el requerimiento de *concordancia monotónica* (línea discontinua); es decir,

$$\hat{d}^*_{23} \geq \hat{d}^*_{12} \geq \hat{d}^*_{34} \geq \hat{d}^*_{13} \geq \hat{d}^*_{24} \geq \hat{d}^*_{14}.$$

Para asegurar qué tan buena es la configuración que se obtiene, se calcula una estadística llamada *"stress"*, la cual se define por

$$\mathcal{S} = \frac{\sum_{i \neq i'} (d_{ii'} - \hat{d}_{ii'})^2}{\sum_{i \neq i'} d^2_{ii'}}. \tag{10.6}$$

El stress $\mathcal{S}$ es simplemente la suma de cuadrados residuales normalizados, de tal forma que su rango es el intervalo $[0, 1]$. El denominador de (10.6) es un factor de escala, inferido desde los $\delta' s$, el cual, para algunos investigadores, puede ser $\sum_{i \neq i'} (d_{ii'} - \bar{d})^2$ en lugar de $\sum_{i \neq i'} d^2_{ii'}$.

El stress se expresa frecuentemente en porcentaje; es decir, multiplicando por 100 el valor de $\mathcal{S}$. Una relación perfecta entre los $d' s$ y los $\delta' s$ produce un stress igual a 0. Se acepta como una configuración *"buena"* aquella cuyos valor de stress es 0.05 (5%) o menos, mientras que valores superiores a 0.1 (10%) se consideran como configuraciones o ajustes *"pobres"*.

Una vez que se ha hecho una primera configuración, los puntos son removidos para tratar de reducir el stress. La diferenciabilidad de $\mathcal{S}$ permite desarrollar un proceso iterativo, semejante al de "mayor descenso", para tratar de encontrar el ajuste que produzca el mínimo valor de $\mathcal{S}$. El problema de estos procedimientos es que no trabajan bien cuando en inmediaciones o vecindades del mínimo la función a minimizar no es cuadrática. Modernamente se tienen procedimientos de cálculo numérico, a los cuales se les ha desarrollado la programación computacional pertinente e incorporado a los diferentes paquetes estadísticos.

El ajuste anterior se puede obtener mediante un procedimiento conocido como *escalamiento óptimo*. Una de las transformaciones más empleadas es la transformación *mínimo cuadrática de Kruskal*, esta transformación produce disimilaridades en concordancia monótona con las distancias, en el sentido mínimo cuadrático. El siguiente ejemplo muestra cómo se trabaja el procedimiento de Kruskal. Se considera que este procedimiento es más de tipo recursivo que analítico (Chatfield y Collins, 1986, pág. 205).

- Parte I: Se muestra una matriz de disimilaridades entre seis objetos junto con las disimilaridades ordenadas en forma ascendente. Dentro de un cuadro se señala el orden con el respectivo subíndice de

las disimilaridades que ocupan el puesto uno, dos, hasta el quince, respectivamente.

	A	B	C	D	E	F
A	0					
B	$\boxed{3_{(2)}}$	0				
C	5	$\boxed{2_{(1)}}$	0			
D	10	6	22	0		
E	8	9	14	16	0	
F	$\boxed{24_{(15)}}$	20	13	21	19	0

2 3 5 6 8 9 10 13 14 16 19 20 21 22 24

- Parte II: De acuerdo con las disimilaridades se "adivina" una configuración inicial y se calculan las distancias euclidianas. Para este ejemplo, se dispusieron en un plano los seis objetos de acuerdo con las disimilaridades y se ajustó a "ojo" una línea de regresión de los $d_{ii'}$ sobre los $\delta_{ii'}$. La matriz de distancias es:

	A	B	C	D	E	F
A	0.0					
B	$\boxed{0.9_{(2)}}$	0				
C	1.2	$\boxed{0.7(1)}$	0			
D	2.0	1.5	3.5	0		
E	1.3	1.9	2.5	2.9	0	
F	$\boxed{4.0_{(15)}}$	3.1	2.3	3.2	2.7	0

- Parte III: Se disponen las distancias en los puestos señalados en la Parte I. Nótese que la sucesión de números no es estrictamente creciente, por ejemplo las distancias 1.5 y 1.3 no conservan el orden de las anteriores; por lo tanto se reemplazan por la distancia promedio. Algo semejante ocurre con las distancias 2.9 y 2.7. Puede ocurrir que la media de tres o más distancias no corresponda con el mismo orden que la distancia siguiente; se calcula entonces el promedio de estas tres (o más) distancias. El procedimiento termina cuando se haya obtenido una sucesión de números *no decreciente*.

$$0.7\ \ 0.9\ \ 1.2\ \ \underbrace{1.5\ \ 1.3}_{\text{prom.}=1.4}\ \ 1.9\ \ 2.0\ \ 2.3\ \ 2.5\ \ \underbrace{2.9\ \ 2.7}_{\text{prom.}=2.8}\ \ 3.1\ \ 3.2\ \ 3.5\ \ 4.0$$

- Parte IV: Una vez que se ha logrado un orden no decreciente (relación mayor o igual) de todos los números asociados a las distancias, se conforman estos con la estructura matricial inicial, así:

$$0.7\ \ 0.9\ \ 1.2\ \ \underbrace{1.4\ \ 1.4}\ \ 1.9\ \ 2.0\ \ 2.3\ \ 2.5\ \ \underbrace{2.8\ \ 2.8}\ \ 3.1\ \ 3.2\ \ 3.5\ \ 4.0$$

	A	B	C	D	E	F
A	0.0					
B	0.9	0				
C	1.2	0.7	0			
D	2.0	1.4	3.5	0		
E	1.4	1.9	2.5	2.8	0	
F	4.0	3.1	2.3	3.2	2.8	0

En resumen, el proceso de EM ordinal empieza con una configuración (escogida tal vez en forma arbitraria o aleatoria) de los puntos en una dimensión particular. La configuración es removida iterativa y paralelamente con una disminución de la medida del stress, bajo la restricción de una relación monótona entre las disimilaridades y las distancias ajustadas. El proceso termina cuando el valor del stress esté dentro de un límite propuesto (converja).

10.4 Determinación de la dimensionalidad

Con fines descriptivos es bastante cómodo desarrollar el escalamiento sobre un espacio de dimensión dos. Para configuraciones en espacios de dimensión tres, se dispone del procedimiento MDS del paquete SAS o del paquete ESTADISTICA.

La decisión sobre la dimensión apropiada para la configuración puede ser simplemente aquella que produzca el menor valor para el stress. Un procedimiento alternativo es el propuesto por Kruskal (Cox y Cox 1994, pág. 69), el cual sugiere ensayar con varios valores de la dimensionalidad p y graficarlo frente a su respectivo valor del stress. El stress decrece en tanto p aumenta, Kruskal sugiere que el valor adecuado para p es aquel donde "la estadística $\mathcal{S}$" se muestre en forma de *"codo"*. La figura 10.5 muestra que una dimensión apropiada para una situación particular puede ser $p = 3$.

Se deben tener algunas precauciones en la elección de la dimensionalidad, Kruskal y Wish (1978, págs. 48-60) hacen, entre otras, las siguientes recomendaciones:

1. Los paquetes estadísticos tienen procedimientos que minimizan sistemáticamente el stress (tal como SAS, MD-SCAL, SSA, o el ALSCAL). Aunque un valor numérico arrojado por un paquete puede indicar un buen ajuste para una dimensión determinada, este mismo valor puede ser malo para otro paquete.

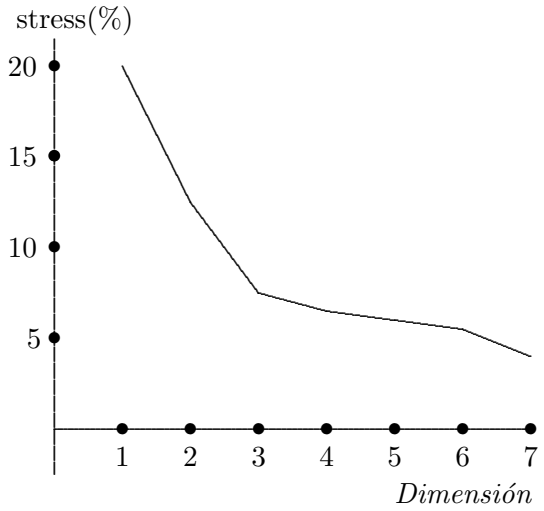


Figura 10.5 *Selección de la dimensionalidad.*

2. Cada valor del stress resulta de un procedimiento computacional iterativo; es decir, de un procedimiento en el cual la configuración es modificada, paso a paso, para estar bastante cerca a los datos. Una terminación prematura del proceso iterativo puede suministrar un stress en un mínimo local, el cual resulte mayor que el verdadero valor mínimo del stress.

3. Aunque el stress sea 0 o significativamente cercano a 0 (por ejemplo 0.01 o menos), la posibilidad de una solución parcial o completamente degenerada debe ser investigada.

4. Es importante examinar el gráfico del stress frente a la dimensionalidad p, para ver si tiene una apariencia normal; es decir, el stress debe decrecer conforme al aumento de la dimensionalidad. Los puntos usualmente forman un polígono convexo; es decir, el segmento que une cualquier par de puntos está por encima de los puntos intermedios. Una alteración a esta forma puede sugerir la existencia de un mínimo local o una convergencia incompleta.

5. El valor del stress es sensible al número de objetos y a la dimensionalidad p. Como se puede apreciar existe analogía con la elección del número de componentes principales del ACP presentado en el capítulo 5, o con el número de factores a retener del capítulo 6.

La *interpretación* es otro criterio que se debe tener en cuenta. Puede presentarse que una configuración en dos dimensiones no sugiera interpretación alguna, mientras que en tres dimensiones se tenga una interpretación más completa. Así, la manera como la interpretación cambia de una dimensión a la siguiente puede ser compleja, sin embargo, la interpretación juega un papel central en la elección de la dimensionalidad, dentro de un rango razonable de ajuste. No obstante, el hecho de que un investigador particular no pueda interpretar una dimensión, no necesariamente significa que la dimensión carezca de interpretación.

Cuando una configuración bidimensional no reconstruya una proyección perpendicular de un espacio tridimensional, puede ocurrir que en ninguna dirección sobre una configuración bidimensional sea interpretable, mientras que en una, dos, tres, o más direcciones sobre un espacio tridimensional se encuentren mejores interpretaciones. Existen ayudas computacionales tal como el paquete NTSYS-PC, con las cuales se puede apreciar la disposición tridimensional de los objetos y algunas opciones de proyección bidimensional.

En el caso de tres dimensiones, cuando la inspección visual no es suficiente para descubrir la dirección de la "mejor" proyección, se puede emplear la regresión lineal, la correlación canónica y los métodos factoriales para buscarla. A continuación se resume el proceso de regresión lineal mútiple:

1. Obtener el promedio para cada objeto respecto a la característica de interés.

2. Hallar la regresión del atributo en cuestión, para cada uno de los objetos sobre las coordenadas del espacio de configuración. Las coordenadas corresponden a las variables regresoras o explicativas. Las regresiones vienen dadas por

$$a_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_k X_{ik}, \text{ para } i = 1, 2, \ldots, n$$

   donde i es el i-ésimo objeto y k la dimensión del espacio.

3. Calcular el *coeficiente de correlación múltiple*, el cual suministra la correlación entre la proyección de los objetos y los atributos. Valores bajos de este coeficiente sugieren que su representación en esta dimensión no es adecuada.

La *correlación canónica*, tal como se aborda en el capítulo 9, busca determinar el grado de asociación lineal entre dos conjuntos de variables. El objetivo es determinar dos combinaciones lineales, una por cada conjunto,

tal que la correlación, producto momento entre las dos combinaciones lineales, sea lo más grande posible. En este escenario, un conjunto de variables corresponde a los atributos iniciales de los individuos y el otro a las coordenadas de los individuos respecto a la dimensionalidad escogida. Como en el caso anterior la magnitud de la correlación justifica la dimensionalidad.

10.5 Análisis de acoplamiento ("Procusto")

El nombre original de esta técnica es *Procusto*, corresponde al salteador de grandes caminos, de acuerdo con la mitología griega, quien interceptaba a los viajeros que se encontraban en su camino y los llevaba a su casa. Allí los obligaba a acostarse: los pequeños en una cama grande y los grandes en una cama pequeña. Luego estiraba a los primeros, para que se adaptaran a las dimensiones del lecho y cortaba las extremidades de los segundos con el mismo objetivo.

En forma semejante, pero menos tortuosa, el *análisis por acoplamiento* dilata, translada, refleja y rota una de las capas o configuraciones de puntos para "mezclarla", tanto como sea posible, con otra configuración.

La técnica consiste en comparar una configuración con otra, sobre un mismo espacio euclidiano, y producir una medida de comparación. Supóngase que n-puntos (objetos, individuos, o estímulos) en un espacio euclidiano q-dimensional, están representados por una matriz de datos $\mathbb{X}$ de tamaño $(n \times q)$, la cual debe compararse con otra configuración de n-puntos ubicados en el espacio p-dimensional $(p \geq q)$ con matriz de coordenadas $\mathbb{Y}$ de tamaño $(n \times p)$. Se asume que el r-ésimo punto de la primera configuración está en relación uno a uno con el r-ésimo punto de la segunda configuración. En primer lugar, como la matriz $\mathbb{X}$ tiene menos columnas que la matriz $\mathbb{Y}$, se colocan $(p-q)$ columnas de ceros en la matriz $\mathbb{X}$; de tal forma que las dos configuraciones queden ubicadas sobre un mismo espacio de dimensión p. La suma de las distancias entre los puntos del conjunto $\mathbb{Y}$ y los correspondientes de $\mathbb{X}$ está dada por

$$\boldsymbol{R}^2 = \sum_{r=1}^{n} (Y_r - X_r)'(Y_r - X_r), \tag{10.7}$$

donde $\mathbb{X} = [X_1, X_2, \ldots, X_n]'$, $\mathbb{Y} = [Y_1, Y_2, \ldots, Y_n]'$, con X_r y Y_r los vectores de coordenadas del r-ésimo punto en los dos espacios de dimensión p.

Los puntos de $\mathbb{X}$ son transladados, dilatados y rotados sobre nuevas coordenadas $\mathbb{X}'$, donde el r-ésimo punto, resultado de las transformaciones anteriores, es

$$X'_r = \rho \boldsymbol{A}' X_r + b. \tag{10.8}$$

La matriz $\boldsymbol{A}$, es una matriz ortogonal que produce una rotación rígida, b es el vector de translación y el vector ρ es la dilatación. Los movimientos o transformaciones anteriores se desarrollan de tal forma que minimizan la suma de distancias entre los puntos de $\mathbb{Y}$ y los "nuevos" de $\mathbb{X}'$; es decir,

$$\begin{aligned} \boldsymbol{R}^2 &= \sum_{r=1}^{n} (Y_r - X'_r)'(Y_r - X'_r) \\ &= \sum_{r=1}^{n} (Y_r - \rho \boldsymbol{A}' X_r - b)'(Y_r - \rho \boldsymbol{A}' X_r - b). \end{aligned} \tag{10.9}$$

La translación, dilatación y rotación óptimas del conjunto representado por $\mathbb{X}$ sobre el conjunto representado por $\mathbb{Y}$, se obtiene después de algunas consideraciones de cálculo, a través de los siguientes pasos (Cox y Cox 1994, págs. 93-96):

1. Sustraer el vector de medias para cada una de las configuraciones, con el fin de *transladar* los datos a los centroides.

2. Encontrar la matriz de *rotación* $\boldsymbol{A} = (\mathbb{X}'\mathbb{Y}\mathbb{Y}'\mathbb{X})^{\frac{1}{2}}(\mathbb{Y}\mathbb{X})^{-1}$ y rotar la configuración $\mathbb{X}$ a la configuración $\mathbb{X}\boldsymbol{A}$.

3. Escalar (dilatar o contraer) la configuración $\mathbb{X}$, a través de la multiplicación de cada una de sus coordenadas por ρ, donde $\rho = tra\{\mathbb{X}'\mathbb{Y}\mathbb{Y}'\mathbb{X}\}/tra\{\mathbb{X}'\mathbb{X}\}$.

4. Calcular el valor minimizado y escalado de

$$\boldsymbol{R}^2 = 1 - \{\mathrm{tra}(\mathbb{X}'\mathbb{Y}\mathbb{Y}'\mathbb{X})^{\frac{1}{2}}\}^2 / \{\mathrm{tra}(\mathbb{X}'\mathbb{X})\,\mathrm{tra}(\mathbb{Y}'\mathbb{Y})\}, \tag{10.10}$$

ésta es una medida de la calidad del ajuste que se le conoce con el nombre de *estadística de Procusto*.

En resumen, la técnica Procusto trata con dos configuraciones de puntos que representan el mismo conjunto de n objetos. El acoplamiento se hace tomando una de las configuraciones como fija y la otra se mueve (translación y rotación) hasta que se "acomode" lo más cerca posible a la otra. Las configuraciones iniciales, la translación a un origen común y la rotación de los ejes, se muestran en la figura 10.6 como etapas (a), (b), y (c) respectivamente.

Cox y Cox (1994, cap. 6) desarrollan una serie de ejemplos utilizando este análisis.

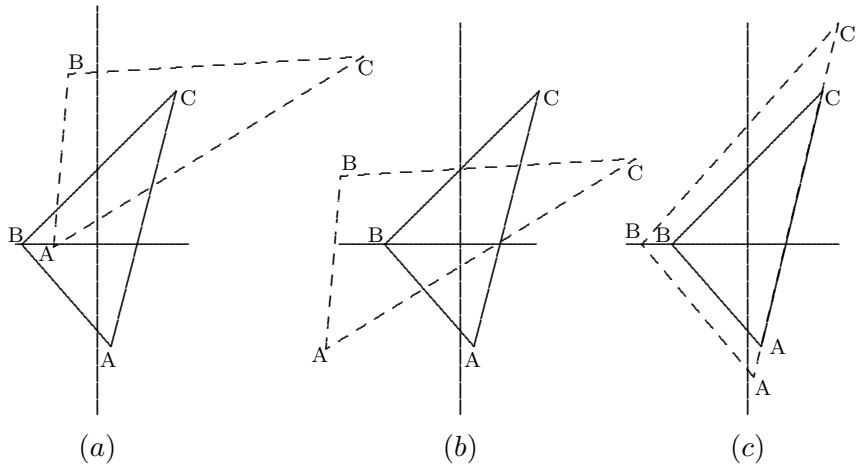


Figura 10.6 *Método de acoplamiento (Procusto).*

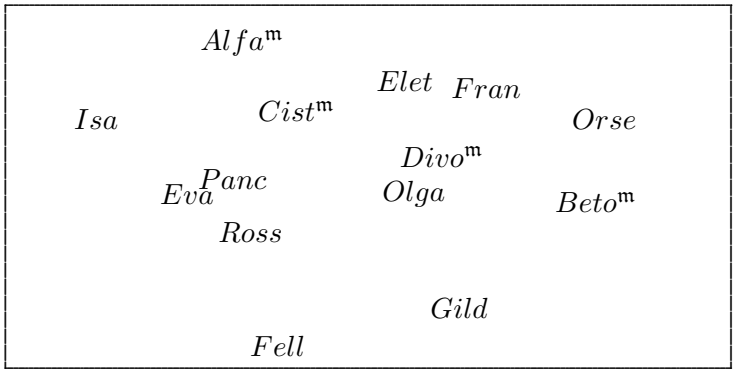
En uno de los ejemplos se estudia la estructura de proximidad en una colonia de 14 micos japoneses. Las relaciones de proximidad fueron hechas cada 60 segundos. Si dos micos estaban a una distancia máxima de 1.5 m. y se manifestaban tolerantes el uno al otro, se les calificaba como "cercanos". Las disimilaridades se calcularon para cada par de micos (91 parejas), basados sobre la cantidad de tiempo que cada par de micos estuviera cerca el uno del otro. Las proximidades se tratan dentro del EM no métrico. Las proximidades fueron medidas separadamente en época de *apareamiento* y en época de *no apareamiento*.

Los 14 micos son descritos por su nombre, edad y sexo en el siguiente cuadro

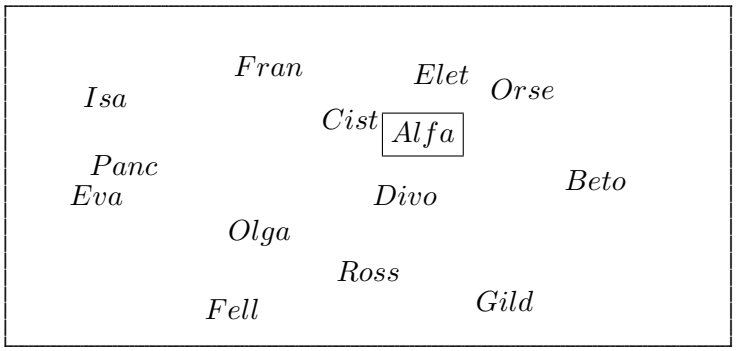
$Nombre$	$Edad/sexo$	$Nombre$	$Edad/sexo$
$Alfa$	$Adulto/macho$	$Olga$	$Adulto/hembra$
$Fran$	$Adulto/hembra$	$Orse$	$Inf-juv/hembra$
$Fell$	$Inf-juv/macho$	$Ross$	$Adulto/hembra$
$Panc$	$Adulto/hembra$	$Divo$	$Joven/macho$
Isa	$Adulto/hembra$	$Cist$	$Joven/macho$
$Gild$	$Adolsc/hembra$	$Elet$	$Adulto/hembra$
$Beto$	$Joven/macho$	Eva	$inf-juv/hembra$

La figura 10.7 (a) muestra la configuración de los micos en época de no apareamiento, mientras que la figura 10.7 (b) corresponde a la época de

apareamiento. Las dos configuraciones han sido alineadas usando el método de acoplamiento de Procusto. Aunque el stress fue alto (28% y 29%), se pueden señalar algunas interpretaciones de la configuración. Una es que los tres infantes juveniles (Fell, Orse y Eva) están en la periferia en ambos mapas. Los machos, excepto Fell, se disponen sobre una misma línea que deja a cada lado las hembras; esto se indica en la figura 10.7 (a) con la letra $\mathtt{m}$. Nótese que Alfa, el único macho adulto, está en la periferia en la primera configuración, mientras que en la segunda se ubica en la parte central.



(a): Epoca de no apareamiento



(b): Epoca de apareamiento

Figura 10.7 *Configuraciones obtenidas mediante análisis de Procusto.*

10.6 Cálculo y cómputo empleado en el EM

El escalamiento, vía mínimos cuadrados, suministra una transformación monótona de las disimilaridades $f(\delta_{ii'})$ previa al encuentro de una configuración. De esta forma, se encuentra una configuración del tipo $\{X_{ii'}\}$ tal que la cantidad

$$\mathcal{S} = \frac{\sum_{i \neq i'} \omega_{ii'}[d_{ii'} - f(\delta_{ii'})]^2}{\sum_{i \neq i'} d_{ii'}^2}, \tag{10.11}$$

llamada *stress* sea mínima, donde $\{\omega_{ii'}\}$ son ponderaciones adecuadamente seleccionadas. La distancia $d_{ii'}$ no necesariamente debe ser euclidiana. La minimización de $\mathcal{S}$ se hace a través de métodos numéricos, en particular mediante el método del *gradiente*. Entre los métodos alternativos al del gradiente se encuentran el *ALSCAL* el cual se resume a continuación.

► ALSCAL

Es el método de escalamiento vía mínimos cuadrados alternantes](Alternating Least squares SCALing) desarrollado por Takane, Young y Leeuw (1977). El ALSCAL puede aplicarse en datos con las siguientes características:

1. Están en escala nominal, ordinal, de intervalo y de razón.
2. Son completos o tienen valores faltantes.
3. Son simétricos o asimétricos.
4. Están condicionados o incondicionados.
5. Tienen replicaciones o no son replicados.
6. Son continuos o discretos.

El problema del escalamiento se puede establecer como la búsqueda de una función ϕ, que aplica sobre las disimilaridades $\{\delta_{ii'j}\}$ un conjunto de distancias $\{\hat{d}_{ii'j}\}$, de modo que

$$\phi(\delta_{ii'j}^2) = \hat{d}_{ii'j}^2 \tag{10.12}$$

donde los $\{\hat{d}_{ii'j}^2\}$ son los estimadores mínimo cuadráticos de $\{d_{ii'j}^2\}$; se obtienen por la minimización de la función de pérdida llamada *SSTRESS* denotada por $\mathcal{SS}$ y definida por

$$\mathcal{SS} = \sum_i \sum_{i'} \sum_j (d_{ii'j}^2 - \hat{d}_{ii'j}^2)^2. \tag{10.13}$$

Nótese la diferencia de SSTRESS y STRESS, en el primero se emplean las distancias al cuadrado, mientras que en el segundo no.

La minimización del SSTRESS dada en (10.13) se hace a través de los *mínimos cuadrados alternantes*. Cada iteración del algoritmo tiene dos etapas: una de escalamiento óptimo y otra de estimación del modelo. El SSTRESS se puede escribir, de acuerdo con (10.11), como una función de las coordenadas $\mathbb{X}$, las ponderaciones ω, y las distancias ajustadas $\hat{d}$. En forma matricial, el SSTRESS es una función de la forma $\boldsymbol{SS}(\mathbb{X}, \boldsymbol{W}, \hat{\boldsymbol{D}})$. Así, en la etapa del escalamiento óptimo se encuentran las distancias mínimo cuadráticas $\hat{\boldsymbol{D}}$ manteniendo fijas las matrices $\mathbb{X}$ y $\boldsymbol{W}$ y en la siguiente etapa, estimación del modelo, se calculan las nuevas coordenadas $\mathbb{X}$ y ponderaciones $\boldsymbol{W}$ para una matriz fija $\hat{\boldsymbol{D}}$.

A continuación se resume el algoritmo ALSCAL:

1. Encontrar la configuración inicial de $\mathbb{X}$ y las ponderaciones $\boldsymbol{W}$.
2. Etapa de *escalamiento óptimo*: se calcula la matriz de disimilaridades $\boldsymbol{D}$ y la matriz de disimilaridades al cuadrado $\boldsymbol{D}^{\star} = (\delta^2_{ii',j})$ se normaliza.
3. Determinar si el SSTRESS es convergente.
4. Etapa de estimación del modelo: minimizar $\mathcal{SS} = (\boldsymbol{W}|\mathbb{X}, \boldsymbol{D}^{\star})$ sobre $\boldsymbol{W}$; y luego minimizar $\mathcal{SS} = (\mathbb{X}|\boldsymbol{W}, \boldsymbol{D}^{\star})$ sobre $\mathbb{X}$.
5. Volver a 2.

Además del procedimiento anterior, existen otros tales como el MINISSA, POLYCON, KYST, INDSCAL/SINDSCAL y MULTISCALE.

En resumen, el ALSCAL es un procedimiento que puede desarrollarse para escalamiento métrico, escalamiento no métrico y en la técnica del *desdoblamiento* multidimensional (Cox y Cox 1994, capítulo 7). Este algoritmo se encuentra disponible en los paquetes estadísticos *SAS* y *SPSS(X)*. El paquete SAS emplea los procedimientos *MDS*, *ALSCAL* y *MLSCALE* para el desarrollo del escalamiento multidimensional.

10.7 Rutina SAS para el escalamiento multidimensional

El procedimiento MDS (MultiDimensional Scaling) es una rutina computacional útil para estimar, entre otras, las coordenadas de un conjunto de

objetos en un espacio de dimensión determinada (menor que la del conjunto inicial), mediante una matriz de distancias entre los pares de objetos o estímulos; se indica que "una" matriz simétrica de distancias o matrcices asimétricas de (di)similitudes. El procedimiento MDS tiene una opción con la cual se escoge una determinada distancia.

```
OPTIONS NODATE NONUMBER;
/*Para que en la salida no aparezca fecha ni paginación*/
TITLE 'Mapa de Colombia';
/*Escribe el título: "Mapa de Colombia"*/
DATA EJEM10\_1;
/*Datos de distancia entre ciudades. Ejemplo 10.2. */


INPUT (Bl  Bt  Bg  Ca  Cg  Cu  Mz  Ml  Pt  Pr  Qd  Rh  Sm ) (5.)
@70 CIUDAD $2.;
/*Ciudades, cuyos rótulos de longitud 2 se deben escribir desde la columna 70 */
CARDS; /*Para ingresar la matriz de distancias entre las 13 ciudades */
```

0													Bl
1302	0												Bt
793	439	0											Bg
1212	484	923	0										Ca
124	1178	917	1088	0									Cg
926	649	210	1133	1050	0								Cu
1003	299	738	275	879	984	0							Mz
750	552	1543	462	626	1201	253	0						Ml
1612	884	1323	400	1488	1533	675	862	0					Pt
1054	330	769	224	930	979	51	304	624	0				Pr
998	800	1791	710	874	1449	501	248	1110	552	0			Qd
284	1147	708	1403	408	1024	1194	941	1803	1245	1189	0		Rh
83	1139	700	1305	217	833	1096	843	1705	1147	1091	191	0	Sm

```
/*Se debe escribir de esta manera la matriz de distancias, */
/*de cuerdo con el formato dado en el INPUT */


;
PROC MDS DATA=EJEM10_1 LEVEL=ABSOLUTE OUT=EJEM_RES;
/*Hace el análisis en un nivel de medida absoluta */
/*El archivo EJEM_RES contiene las coordenadas sobre el mapa para dibujarlas con
PLOT */
ID CIUDAD;


/*Copia los nombres de las ciudades en el archivo EJEM_RES */
PROC PLOT DATA=EJEM_RES VTOH=1.7;
/*PLOT ubica en un plano las ciudades. VTOH razón entre líneas a la distancia entre
los caracteres */
PLOT DIM2*DIM1 $ CIUDAD / HAXIS=BY 500 VAXIS=BY 500;
/*HAXIS VAXIS, ejes horizontal y vertical con las mismas unidades */
```

```
  WHERE  _TYPE_='CONFIG';
PROC PRINT DATA=EJEM_RES;'
/* imprime las coordenadas de las ciudades respecto a los dos nuevos ejes */
RUN;
```

10.8 Procesamiento de datos con R

Escalamiento multidimensional

La función para el escalamiento clásico es `cmdscale`.

La sintaxis de la función es:

```
cmdscale(d, k = 2, eig = FALSE, add =FALSE,x.ret = FALSE)
```

con argumentos

d es una matriz de distancias o una matriz que contiene e disimilitudes

k es la dimensión del espacio en el cual se representan los datos; debe ser $\{1, 2, \ldots, n-1\}$.

eig indica si los valores propios son requeridos.

add indicador logico para incluir una constante aditiva c, que adicionada a las disimilitudes no diagonales, hace que todos n-1 valores propios son no negativos.

x.ret indica si el centrado doble de la matriz de distancias se requiere.

Se considera el ejemplo 10.3 de la matriz de distancias entre las 13 ciudades colombianas. La matriz se debe escribir en forma completa, es decir a la matriz que aparece en el ejemplo 10.3 se le deben agregar los términos que están por encima de la diagonal. Esto se puede en R mediante la función `cmdscale`. Primero se deben copiar las distancias desde la matriz completa anterior (`clipboard`). Luego se deben emplear los siguientes commandos:
`source("http://personality-project.`
`org/r/useful.r")`

```
#para obtener algunas funciones extra, incluyendo la función
read.clipboard()

ciudades<- read.clipboard(header="TRUE")
# toma los datos desde el clipboard ciudades
```

```
# muestra los datos
city.location <- cmdscale(cities, k=2)

# solicita una representación bidimensional
round(city.location,0)

# imprime la localización
plot(city.location,type="n", xlab="Dimension 1",
ylab="Dimension 2",main ="cmdscale(cities)")

# coloca una ventana de gráficos
text(city.location,labels=names(cities))

# coloca las ciudades en un mapa
# ¡intente hacerlo con los datos del ejemplo 10.3!
```

Capítulo 11

Análisis de correspondencias

11.1 Introducción

Es común encontrar casos cuyas matrices de datos tienen filas y columnas asociadas con modalidades de variables categóricas. Las entradas de esta matriz contienen la frecuencia absoluta o relativa de los individuos que toman tales valores en cada una de las respectivas modalidades. A estas matrices se les conoce también con el nombre de *tablas de contingencia*[1]. Un análisis de la información contenida en las filas o en las columnas se hace a través del *análisis de correspondencias*, el cual en adelante se notará como AC . Esta técnica puede ser vista como el procedimiento que encuentra la "mejor" representación para dos conjuntos de datos, los dispuestos en filas, o en las columnas de la respectiva matriz de datos (Lebart, Morineau y Warwick, 1984, pág. 30). De otra manera, el análisis de correspondencias, tal como el ACP, busca obtener una tipología de las filas o una tipología de las columnas y relacionarlas entre sí. Lo anterior justifica el uso del término *correspondencia*, pues la técnica busca las filas (o columnas) que *se correspondan* en información; es decir, que algunas filas (o columnas) pueden estar suministrando información equivalente respecto a un conjunto de individuos. Una de las tareas es encontrar tales filas (o columnas) e interpretar la información allí consignada.

En resumen, en lugar de comparar filas/columnas utilizando probabilidades condicionales, el *análisis de correspondencias* procede a obtener un pequeño número de dimensiones (factores), de tal forma que la primera dimensión

[1]Término introducido por Pearson en 1904, como una medida de "la desviación total de la clasificación respecto a la independencia probabilística".

explique la mayor parte de la asociación total entre filas y columnas (medidas mediante un coeficiente ji-cuadrado), la segunda dimensión explique la mayor parte del residuo de la asociación no explicada por la primera, y así sucesivamente con el resto de las dimensiones. El número máximo de dimensiones es igual al menor número de categorías de cualquiera de las dos variables (fila o columna), menos uno, pero por lo común dos o tres dimensiones son suficientes para representar con rigor la asociación entre las dos variables. En este sentido las dismensiones son conceptualmente similares a las componentes principales.

El *análisis de correspondencias* se desarrolla mediante el trabajo sobre dos tablas de datos: una primera tabla contiene las frecuencias respecto a las modalidades de dos variables; usualmente se denomina *análisis de correspondencias binarias*; el segundo tipo de tabla contiene la información sobre varias variables; el análisis se conoce como de *correspondencias múltiples*. En la primera parte se dedicará al desarrollo del AC binario o simple; el análisis de correspondencias múltiple se presenta en la segunda parte de este capítulo.

A manea de ejemplo, considérese la matriz de frecuencias (n_{ij}) contenida en la tabla 11.1, tomada de Thompson (1995)[2]. En esta tabla las filas ($i = 1, 2, 3, 4$) son el color de los ojos y las columnas ($j = 1, 2, 3, 4, 5$) el color del cabello, cuyas modalidades varian de claro a oscuro. Para encontrar la representación más adecuada de estos datos, es necesario comparar las filas y las columnas de la tabla. Tal comparación implica hacer uso de una medida de distancia apropiada. El análisis de correspondencias permite describir las proximidades existentes entre los perfiles, color del cabello (perfil fila) y color de los ojos (perfil columna), de acuerdo con la partición que se haga de los individuos, sea por filas o por columnas.

La matriz de densidades o frecuencias relativas (f_{ij}) y las densidades marginales de filas ($f_{i.}$) y columnas ($f_{.j}$) es mostrada en la Tabla 11.2. Los números son dados como porcentaje y representan el $f_{ij}100\%$. Los números a la derecha de cada fila, presentan las densidades marginales, como el porcentaje $f_{i.}100\%$, y la última fila representa las densidades marginales por columna $f_{.j}100\%$. En resumen, la mayoría de las personas tienen el color de los ojos medio (32.93%) y el color de cabello más común es también medio (39.66%).

[2]Ronald A. Fisher en 1940 estudió estos datos como tablas de contingencia.

Tabla 11.1 Frecuencias absolutas

	Color de cabello					
Color de ojos	Rubio (ru)	Rojo (r)	Medio (m)	Oscuro (o)	Negro (n)	*Total* ($n_{i.}$)
Claros (C)	688	116	584	188	4	1580
Azules (A)	326	38	241	110	3	718
Medio (M)	343	84	909	412	26	1774
Oscuros (O)	98	48	403	681	85	1315
Total ($n_{.j}$)	1455	286	2137	1391	118	5387

El origen del análisis de correspondencias se puede remontar a los trabajos Hirschfeld (1935) y de Fisher (1940) sobre tablas de contingencia, pero el verdadero responsable de esta técnica estadística es Benzecri (1964, 1973 y 1976); tal como se cita en Lebart, Morineau y Fénelon (1985, pág. 276). Cox y Cox (1995, pág. 126) presentan el AC como un método de escalamiento multidimensional sobre las filas y las columnas de una tabla de contingencia o matriz de datos cuyas entradas deben ser no negativas. En reconocimiento a la escuela francesa se mantienen en este texto algunos de sus términos, los cuales tienen sus respectivas nominaciones en la escuela anglosajona.

Tabla 11.2 Frecuencias relativas

	Color de cabello					
Color de ojos	Rubio (ru)	Rojo (r)	Medio (m)	Oscuro (o)	Negro (n)	*Total* ($f_{i.}$)
Claros (C)	12.77	2.15	10.84	3.49	0.07	29.32
Azules (A)	6.05	0.71	4.47	2.04	0.06	13.33
Medio (M)	6.37	1.56	16.87	7.65	0.48	32.93
Oscuros (O)	1.82	0.89	7.48	12.65	1.58	24.42
Total ($f_{.j}$)	27.01	5.31	39.66	25.83	2.19	100.00

Se presenta en este capítulo, en forma esquemática, la técnica del análisis de correspondencias. Por ser una técnica estadística relativamente nueva en nuestro medio, la escritura de esta parte sigue el estilo de la literatura citada para cada caso.

11.2 Representación geométrica de una tabla de contingencia

En una tabla de contingencia (matriz de datos) pueden considerarse dos espacios, el espacio fila ($\mathbb{R}^p$) o el espacio columna ($\mathbb{R}^n$). Para el ejemplo anterior, el espacio *color de los ojos* ($\mathbb{R}^4$) y el espacio *color del cabello* ($\mathbb{R}^5$), respectivamente.

La matriz de datos $\mathbb{X}$, tiene n-filas y p-columnas, n_{ij} representa el número de individuos de la fila i y la columna j. En el ejemplo, n_{ij} es el número de individuos con el color de los ojos i y color del cabello j.

El número total de individuos por fila se nota por

$$n_{i.} = \sum_{j=1}^{p} n_{ij}, \quad \text{para } i = 1, \cdots, n. \tag{11.1}$$

El número total de individuos por columna se nota por

$$n_{.j} = \sum_{i=1}^{n} n_{ij}, \quad \text{para } j = 1, \cdots, p. \tag{11.2}$$

El número total de individuos de la tabla está dado por

$$N = \sum_{i=1}^{n}\sum_{j=1}^{p} n_{ij} = \sum_{i=1}^{n} n_{i.} = \sum_{j=1}^{p} n_{.j}. \tag{11.3}$$

Las frecuencias relativas absolutas y marginales se notan como sigue

$$f_{ij} = \frac{n_{ij}}{N}; \quad f_{i.} = \sum_{j=1}^{p} f_{ij} = \frac{n_{i.}}{N}; \text{ y } f_{.j} = \sum_{i=1}^{n} f_{ij} = \frac{n_{.j}}{N}. \tag{11.4}$$

Con lo anterior se puede apreciar que la matriz $\mathbb{X}$ de elementos n_{ij} se ha transformado en la matriz de elementos f_{ij}; esta última se nota por $\boldsymbol{F} = (f_{ij})$.

Las frecuencias relativas condicionales, de columna respecto a filas (perfiles) y fila respecto a columnas, se escriben, respectivamente, como sigue:

$$f_{i|j} = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}} \text{ y } f_{j|i} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}, \text{ para } i = 1, \ldots, n \; j = 1, \ldots, p. \tag{11.5}$$

En el espacio fila ($\mathbb{R}^p$) o *nube de puntos fila*, el i-ésimo vector (perfil fila) tiene coordenadas

$$\left(\frac{n_{i1}}{n_{i.}}, \ldots, \frac{n_{ip}}{n_{i.}}\right) = \left(\frac{f_{i1}}{f_{i.}}, \ldots, \frac{f_{ip}}{f_{i.}}\right) = \left(f_{1|i}, \ldots, f_{p|i}\right); \; i = 1, \ldots, n. \tag{11.6}$$

La nube de puntos fila (perfil fila) queda determinada por la matriz $\boldsymbol{D}_n^{-1}\boldsymbol{F}$, donde la matriz $\boldsymbol{D}_n = \text{Diag}(f_{i.})$, matriz diagonal que contiene las frecuencias marginales por fila o "pesos" $f_{i.}$. Se observa que cada punto o perfil fila está afectado por su peso $f_{i.}$.

El *centroide* o *baricentro (centro de gravedad)* de la nube de puntos fila se representa por $\mathcal{G}_f$, sus coordenadas son las frecuencias marginales; es decir, $\mathcal{G}_f = (f_{.1}, \ldots, f_{.p})$.

De manera similar, en el espacio columna ($\mathbb{R}^n$) o *nube de puntos columna* , el j-ésimo vector (perfil columna) tiene coordenadas

$$\left(\frac{n_{1j}}{n_{.j}}, \cdots, \frac{n_{nj}}{n_{.j}}\right) = \left(\frac{f_{1j}}{f_{.j}}, \ldots, \frac{f_{nj}}{f_{.j}}\right) = \Big(f_{1|j}, \ldots, f_{n|j}\Big);\ j = 1, \ldots, p. \tag{11.7}$$

De esta manera, la nube de puntos columna queda representada por la matriz $\boldsymbol{F}\boldsymbol{D}_p^{-1}$, donde $\boldsymbol{D}_p = \text{Diag}(f_{.j})$, es una matriz diagonal que contiene las frecuencias marginales por columna o "pesos" $f_{.j}$. Se nota también, que cada uno de estos puntos está afectado por los respectivos pesos $f_{.j}$.

También, el *centroide* o *baricentro* de la nube de puntos columna se representa por $\mathcal{G}_c$, sus coordenadas son las frecuencias marginales; es decir, $\mathcal{G}_c = (f_{1.}, \ldots, f_{n.})$.

En forma gráfica se puede representar lo anterior mediante el esquema de la figura 11.1.

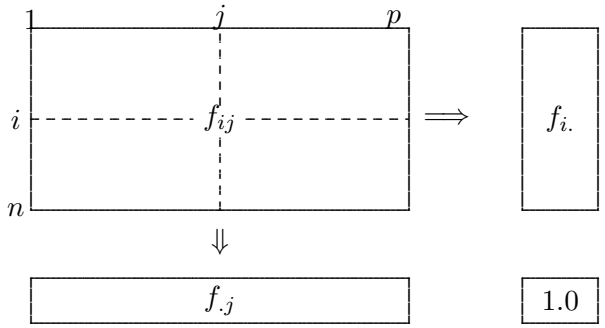


Figura 11.1 *Tabla de frecuencias y sus marginales.*

11.2.1 Perfiles fila y columna

Las ecuaciones 11.4 y 11.5 equivalen a las densidades marginales y condicionales, respectivamente. De la tabla que contiene la frecuencia por celdas n_{ij} para cada fila i, el vector de densidades condicionales de tamaño $(p \times 1)$ es determinado a través de $n_{ij}/n_{i.}$, con $j = 1, \ldots, p$ y se nota por $f_{j|i}$. Estas densidades condicionales por fila son llamadas *perfiles fila*. Paralelamente, el vector columna de densidades condicionales $n_{ij}/n_{.j}$, con $i = 1, \ldots, n$ y es notado por $f_{i|j}$. Las tablas 11.3 y 11.4 contienen los perfiles fila y columna, respectivamente. Así, la tabla 11.3 muestra la distribución del color del cabello por cada uno de los colores de los ojos; recíprocamente la tabla 11.4 suministra la distribución del color de ojos manteniendo constante el color del cabello.

Tabla 11.3 Perfil fila

			Color de cabello			
Color de ojos	Rubio (ru)	Rojo (r)	Medio (m)	Oscuro (o)	Negro (n)	*Total*
Claros (C)	0.4354	0.0734	0.3697	0.1190	0.0025	1.0000
Azules (A)	0.4540	0.0529	0.3357	0.1532	0.0042	1.0000
Medio (M)	0.1933	0.0474	0.5124	0.2322	0.0147	1.0000
Oscuros (O)	0.0745	0.0365	0.3065	0.5179	0.0646	1.0000
Centroide columna	0.2701	0.0531	0.3966	0.2583	0.0219	1.0000

La distribución de frecuencias condicionadas, del color de cabello de acuerdo con el color de los ojos de las personas estudiadas, se representa en el vector $(n_{ij}/n_{i.} = f_{j|i})$, éste se ilustra en la figura 11.2. Alternamente, se ilustra la distribución condicional de frecuencias del color de los ojos respecto al color del cabello $(n_{ij}/n_{.j} = f_{i|j})$ en la figura 11.3.

Los perfiles fila y columna pueden ser comparados con las distribuciones columna y fila con el respectivo peso, para juzgar su "apartamiento" de la indepedencia. La gráfica del perfil *color de ojos respecto al color del cabello* muestra una alta similitud entre los perfiles ojos claros y ojos azules, lo mismo, aunque un poco más baja, la similitud o proximidad entre los perfiles ojos medios y oscuros (figura 11.2).

Para el perfil color del cabello, se encuentra una alta semejanza entre los perfiles cabello rubio y rojo y entre los cabellos oscuro y negro; el perfil cabello medio es bastante diferente de los demás, como se muestra en la figura 11.3.

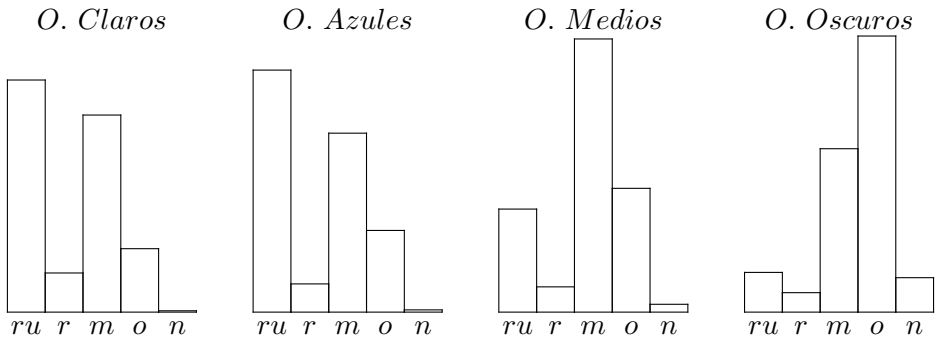


Figura 11.2 *Perfiles fila.*

Tabla 11.4 Perfil Columna

	Color de cabello					
Color de ojos	Rubio (ru)	Rojo (r)	Medio (m)	Oscuro (o)	Negro (n)	*Total*
Claros (C)	0.4729	0.4056	0.2733	0.1352	0.0339	0.2932
Azules (A)	0.2241	0.1329	0.1128	0.0791	0.0255	0.1333
Medio (M)	0.2356	0.2937	0.4254	0.2961	0.2203	0.3293
Oscuros(O)	0.0674	0.1678	0.1885	0.4896	0.7203	0.2442
Centroide columna	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

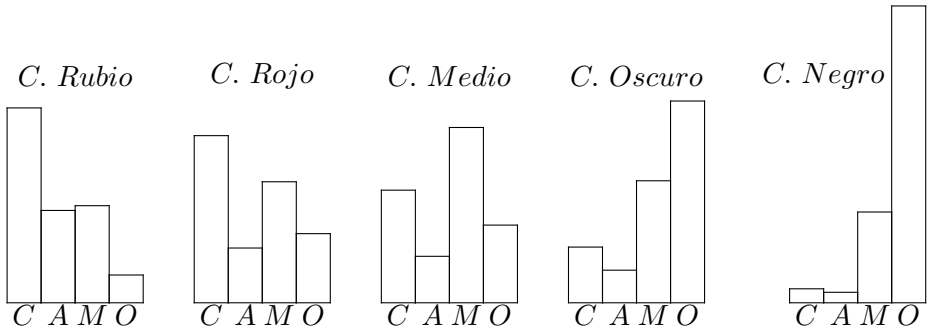


Figura 11.3 *Perfiles columna.*

11.3 Semejanza entre perfiles: distancia ji-cuadrado

Una vez que se han definido las dos nubes de puntos, espacio fila ($\mathbb{R}^p$) y espacio columna ($\mathbb{R}^n$), se debe decidir como medir la distancia entre ellos. En el análisis de correspondencias, la semejanza entre dos líneas (o entre dos columnas) está dada por la distancia entre sus perfiles (Escofier y Pagés, 1990). Esta distancia es conocida con el nombre de distancia *ji-cuadrado*, se nota χ^2. Se define en forma análoga la distancia entre perfiles fila y columna, respectivamente.

La distancia entre dos perfiles fila i e i' está dada por

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \tag{11.8}$$

Similarmente, la distancia entre dos perfiles columna j y j' es

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 \tag{11.9}$$

Nótese que (11.8) y (11.9) miden la distancia entre dos distribuciones multinomiales; es decir, permite comparar los histogramas (distribuciones empíricas) por cada par de filas o columnas.

Las distancias dadas en las igualdades (11.8) y (11.9) difieren de la distancia euclidiana en que cada cuadrado es ponderado por el inverso de la frecuencia para cada modalidad; es decir, se ponderan las distintas coordenadas, de manera que se le da más "importancia" a las categorías o modalidades con menor frecuencia y menos "importancia" a las que tengan alta frecuencia.

Las distancias anteriores se traducen en que el AC da prioridad a las modalidades raras, por cuanto éstas, por su escasez, son más diferenciadoras que las otras.

La distancia ji-cuadrado es equivalente a la distancia euclidiana usual; es decir, tan sólo es necesario transformar adecuadamente las coordenadas de los vectores de perfiles para obtener el cuadrado de la distancia euclidiana entre tales puntos. Así, para dos perfiles fila i e i', su distancia está dada por:

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 = \sum_{j=1}^{p} \left(\sqrt{\frac{1}{f_{.j}}} \left(\frac{f_{ij}}{f_{i.}} \right) - \sqrt{\frac{1}{f_{.j}}} \left(\frac{f_{i'j}}{f_{i'.}} \right) \right)^2 . \tag{11.10a}$$

Un resultado semejante se tiene para la distancia entre dos perfiles columna j y j', éste es:

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 = \sum_{i=1}^{n} \left(\sqrt{\frac{1}{f_{i.}}} \left(\frac{f_{ij}}{f_{.j}} \right) - \sqrt{\frac{1}{f_{i.}}} \left(\frac{f_{ij'}}{f_{.j'}} \right) \right)^2 . \tag{11.10b}$$

11.3.1 Equivalencia distribucional

Esta propiedad permite juntar o agregar dos modalidades, con perfiles idénticos o proporcionales (linealmente dependientes) de una misma variable, en una nueva modalidad cuya ponderación es la suma de los pesos asociados a cada modalidad; sin que se alteren las distancias entre las modalidades de esta variable, ni las distancias entre las modalidades de la otra variable. Así por ejemplo, considérese que los perfiles fila i_1 e i_2 con pesos $f_{i_1.}$ y $f_{i_2.}$ son idénticos en $\mathbb{R}^p$, éstos se unen en un nuevo perfil fila cuyo peso es $f_{i_1.} + f_{i_2.}$. De otra manera, dos (o más) perfiles homogéneos pueden confundirse en uno solo, sin que se modifique la estructura de la nube de puntos.

Lo mismo ocurre al juntar modalidades o perfiles columna. Esta propiedad garantiza cierta invarianza de los resultados del AC con relación a la selección de modalidades para una variable; siempre que las modalidades agrupadas tengan perfiles semejantes. En resumen, no hay pérdida de información al unir o dividir modalidades homogéneas de una misma variable. La demostración de esta propiedad se puede consultar en Lebart, Morineau y Piron (1995, págs. 81-82).

11.4 Ajuste de las dos nubes de puntos

11.4.1 Ajuste de la nube de puntos fila en $\mathbb{R}^p$

El problema consiste en encontrar un subespacio ($\mathbb{R}^q$) de dimensión menor que el espacio fila ($\mathbb{R}^p$), es decir, $q < p$, que conserve el máximo de la información de la nube de puntos original; una medida de la cantidad de información es la cantidad de varianza o inercia[3] retenida por el subespacio ($\mathbb{R}^q$). De la misma forma que el ACP, el AC procede a buscar una sucesión de ejes ortogonales sobre los cuales la nube de puntos es proyectada.

[3]En física la inercia de un punto X_i de masa p_i, respecto a su centro de gravedad $g = \bar{X}$, es $I_g = \sum_i p_i \|X_i - g\|^2$; equivale a la varianza.

El interés sobre las modalidades de la primera variable consiste en la yuxtaposición de los perfiles fila. Cada perfil fila es un arreglo de p valores numéricos, el cual se representa por un punto del espacio $\mathbb{R}^p$, cada una de las p dimensiones está asociada a una de las modalidades de la segunda variable. La distancia χ^2 define la cercanía entre los perfiles fila, o como se ha advertido, la distancia entre dos histogramas (distribuciones).

Las distancias entre los puntos en el subespacio imagen, deben ser lo más semejantes a las distancias entre los puntos de la nube inicial. Este objetivo es similar al ajuste de la nube de individuos para el ACP; es decir, que la nube analizada debe centrarse, de tal forma que su *baricentro* o *centroide* $\mathcal{G}_f$, sea escogido como el origen del sistema de coordenadas.

Respecto al centroide de la nube, la clase definida por la modalidad i se representa por un punto cuya coordenada sobre el j-ésimo eje es igual a: $f_{ij}/f_{i.} - f_{.j} = f_{j|i} - f_{.j}$. La posición de este punto representa la diferencia entre la distribución de la clase i y el total en las modalidades de la segunda variable. De esta manera, la búsqueda de las direcciones de máxima varianza o *inercia* de la nube centrada, pone en evidencia las clases que más se apartan en el conjunto de perfiles de la población.

Cada perfil está previsto de un peso igual a su frecuencia marginal $f_{i.}$. Los pesos o ponderaciones intervienen, en primer lugar, en el cálculo del baricentro de la nube y en segundo lugar, en el criterio de ajuste de los ejes.

Por un procedimiento similar al que se desarrolló para componentes principales (sección (5.2)), se bosqueja el cálculo para la determinación de los ejes principales y las "nuevas" coordenadas de los puntos proyectados que conforman la nube. Los detalles se pueden consultar en Escofier y Pagés (1990), Jobson (1992) y Saporta (1990).

Sea $\mathbb{X}$ la matriz de datos de tamaño $(n \times p)$. Sin pérdida de generalidad, considérese primero la nube de puntos fila en $\mathbb{R}^q$. El problema consiste en buscar un subespacio $\mathbb{R}^q$ de menor dimensión ($\mathbb{R}^q \subseteq \mathbb{R}^p$), que conserve la máxima información de la nube original.

Esto se logra buscando un subespacio, $\boldsymbol{H}$, en el que la inercia de los puntos proyectados sea máxima, lo que equivale a maximizar la expresión:

$$\sum_i f_{i.} d^2_{\boldsymbol{H}}(i, \mathcal{G}_f), \tag{11.11}$$

donde $d^2_{\boldsymbol{H}}(i, \mathcal{G}_f)$ es la distancia al cuadrado entre el perfil fila i y su respectivo centroide $\mathcal{G}_f$, el cual está contenido en $\boldsymbol{H}$.

Mediante el AC se busca primero la recta que esté en la dirección de un vector unitario u_1, sobre la cual se recoja *la máxima inercia proyectada*.

Una vez se ha encontrado esta recta, se busca otra, ortogonal a la primera y en la dirección de un segundo vector unitario u_2, que recoja *la máxima inercia restante proyectada*. Hecho lo anterior se busca una tercera recta ortogonal a las dos primeras, y en la dirección de un vector unitario u_3, que reuna *la máxima inercia restante proyectada* y así sucesivamente. Una vez se termina este procedimento constructivo; es decir, en el p-ésimo paso, se obtiene una descomposición de la inercia total de la nube de puntos fila original, en direcciones ortogonales. El subespacio $\boldsymbol{H}$ se genera por los vectores unitarios u_i.

Se demuestra que los vectores $u_1, u_2, \ldots, u_p$, que determinan la posición y dirección de los *ejes principales*, son generados por los respectivos valores propios de la matriz

$$\boldsymbol{S} = \boldsymbol{F}'\boldsymbol{D}_n^{-1}\boldsymbol{F}\boldsymbol{D}_p^{-1}, \tag{11.12}$$

en el orden $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, los cuales son soluciones del sistema

$$\boldsymbol{S}u = \lambda u. \tag{11.13}$$

El término general $s_{jj'}$ de la matriz $\boldsymbol{S}$, se escribe en la forma

$$s_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i.} f_{.j'}}.$$

La inercia recogida en cada eje, igual que en el ACP, corresponde al valor propio asociado al eje; es decir,

$$I_T = \lambda_1 + \lambda_2 + \cdots + \lambda_p. \tag{11.14}$$

Nótese que la matriz $\boldsymbol{S}$ no es una matriz simétrica. Este problema se puede resolver como se muestra a continuación.

La matriz $\boldsymbol{S}$ se define, de acuerdo con (11.12), como $\boldsymbol{S} = \boldsymbol{F}'\boldsymbol{D}_n^{-1}\boldsymbol{F}\boldsymbol{D}_p^{-1}$.

Sea $\tilde{\boldsymbol{A}} = \boldsymbol{F}'\boldsymbol{D}_n^{-1}\boldsymbol{F}$, la cual es simétrica. Como la matriz $\boldsymbol{D}_p^{-1}$ es diagonal se puede expresar en la forma

$$\boldsymbol{D}_p^{-1} = \boldsymbol{D}_p^{-1/2}\boldsymbol{D}_p^{-1/2}.$$

Por tanto

$$\boldsymbol{S} = \tilde{\boldsymbol{A}}\boldsymbol{D}_p^{-1/2}\boldsymbol{D}_p^{-1/2}.$$

La ecuación (11.13) es equivalente a

$$\tilde{\boldsymbol{A}}\boldsymbol{D}_p^{-1/2}\boldsymbol{D}_p^{-1/2}u = \lambda u,$$

multiplicando a la izquierda de cada miembro de la igualdad anterior por $\boldsymbol{D}_p^{-1/2}$ y llamando $\boldsymbol{D}_p^{-1/2}u = w$, se obtiene

$$\boldsymbol{D}_p^{-1/2}\tilde{\boldsymbol{A}}\boldsymbol{D}_p^{-1/2}w = \lambda w.$$

De manera que la matriz

$$\boldsymbol{S}^* = \boldsymbol{D}_p^{-1/2}\tilde{\boldsymbol{A}}\boldsymbol{D}_p^{-1/2} \tag{11.15}$$

es simétrica y tiene los mismos valores propios que la matriz $\boldsymbol{S}$. Con esta última matriz resulta más sencillo obtener los valores y vectores propios, los cuales sugieren la cantidad de inercia y la dirección de los ejes principales.

Una última observación es que la línea que une el origen con el centro de gravedad G (fila o columna) es un vector propio de la matriz $\boldsymbol{S}$ con relación al valor propio $\lambda = 1$, el cual tiene la forma $g = (f_{.1}, \ldots, f_{.j}, \ldots, f_{.p})$ en el espacio fila. Mediante la forma general del elemento $s_{jj'}$ de $\boldsymbol{S}$ (sección (11.4.1)) se muestra que $\boldsymbol{S}g = g$; es decir, que 1 es un valor propio de $\boldsymbol{S}$. Por tanto, es suficiente diagonalizar la matriz $\boldsymbol{S}^*$ y dejar de lado el valor propio igual a 1 y su correspondiente eje tanto en $\mathbb{R}^p$ como en $\mathbb{R}^n$.

11.4.2 Relación con el ajuste de la nube de puntos columna en $\mathbb{R}^n$

Un papel análogo juegan los datos dispuestos en columna; es decir, aquellos que están en *correspondencia* con los datos fila, de aquí que el análisis en $\mathbb{R}^n$ puede deducirse del desarrollado para $\mathbb{R}^p$ mediante el intercambio de los subíndices i y j.

Las coordenadas de un punto columna j (o vector de $\mathbb{R}^n$) tienen la forma $f_{ij}/f_{.j}\sqrt{f_{i.}}$, para $i = 1, \ldots, n$.

A partir de la matriz de datos $\mathbb{X}$, de tamaño $(n \times p)$, se trata de buscar un subespacio de dimensión menor que n, tal que recoja la máxima cantidad de información de la nube original. Esto se logra, nuevamente, buscando un subespacio, $\boldsymbol{H}^*$, en el que la inercia de los puntos proyectados sobre éste sea máxima; es decir, maximizar la expresión:

$$\sum_j f_{.j} d^2_{\boldsymbol{H}^*}(j, \mathcal{G}_c), \tag{11.16}$$

donde $d^2_{\boldsymbol{H}^*}(j, \mathcal{G}_c)$ es la distancia al cuadrado entre el perfil columna j y el respectivo centroide de las columnas $\mathcal{G}_c$.

Los vectores $v_1, v_2, \ldots, v_n$, que determinan la posición y dirección de los *ejes principales* y generan el subespacio H^*, se obtienen de los respectivos valores propios de la matriz

$$\boldsymbol{S}^* = \boldsymbol{F}\boldsymbol{D}_p^{-1}\boldsymbol{F}'\boldsymbol{D}_n^{-1}. \tag{11.17}$$

Retomando la ecuación (11.13)

$$\begin{aligned} \boldsymbol{S}u &= \lambda u, \\ \boldsymbol{F}'\boldsymbol{D}_n^{-1}\boldsymbol{F}\boldsymbol{D}_p^{-1}u &= \lambda u. \end{aligned}$$

Premultiplicando en ambos lados por $\boldsymbol{F}\boldsymbol{D}_p^{-1}$:

$$\boldsymbol{F}\boldsymbol{D}_p^{-1}\boldsymbol{F}'\boldsymbol{D}_n^{-1}(\boldsymbol{F}\boldsymbol{D}_p^{-1}u) = \lambda(\boldsymbol{F}\boldsymbol{D}_p^{-1}u).$$

Así, se observa que el vector v es proporcional a $\boldsymbol{F}\boldsymbol{D}_p^{-1}u$. Como la norma de $\boldsymbol{F}\boldsymbol{D}_p^{-1}u$ respecto a $\boldsymbol{D}_n^{-1}$ es igual a λ, y además, $v'\boldsymbol{D}_n^{-1}v = 1$, se tiene entonces la siguiente relación entre los vectores propios que generan los subespacios H^* y H, respectivamente

$$\begin{cases} v = \frac{1}{\sqrt{\lambda}}\boldsymbol{F}\boldsymbol{D}_p^{-1}u, \\ u = \frac{1}{\sqrt{\lambda}}\boldsymbol{F}'\boldsymbol{D}_n^{-1}v. \end{cases}$$

Las dos relaciones anteriores muestran que las coordenadas de los puntos sobre un determinado eje principal en un espacio, son proporcionales a las componentes del factor del otro espacio correspondientes al *mismo* valor propio. En general, denominando $\psi_{i\alpha}$ la proyección de la i-ésima fila sobre el eje α, y $\varphi_{j\alpha}$ la proyección de la columna j-ésima sobre el eje α, se tienen las siguientes relaciones

$$\begin{cases} \psi_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha}}\boldsymbol{D}_n^{-1}\boldsymbol{F}\varphi_{j\alpha}, \\ \\ \varphi_{j\alpha} = \frac{1}{\sqrt{\lambda_\alpha}}\boldsymbol{D}_p^{-1}\boldsymbol{F}'\psi_{i\alpha}. \end{cases} \tag{11.18}$$

Las ecuaciones (11.18), son llamadas *ecuaciones de transición*, y pueden reescribirse en términos de las coordenadas de proyección de la siguiente forma:

$$\begin{cases} \hat{\psi}_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{j=1}^{p}\frac{f_{ij}}{f_{i.}}\hat{\varphi}_{j\alpha}, \\ \\ \hat{\varphi}_{i\alpha} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{i=1}^{n}\frac{f_{ij}}{f_{.j}}\hat{\psi}_{i\alpha}. \end{cases} \tag{11.19}$$

Estas últimas ecuaciones ponen en relación las dos representaciones gráficas obtenidas. Así, existe una relación llamada *pseudo-baricéntrica*, la cual especifica que las coordenadas de un punto fila pueden encontrarse como

el baricentro de todas las coordenadas de los puntos columna, tomando como ponderaciones los elementos del perfil de la fila en cuestión y multiplicándolas por un factor de expansión.

Otra interpretación, de acuerdo con las dos últimas ecuaciones, es la siguiente: un punto fila, aparece próximo de aquellas columnas en las cuales su perfil (frecuencia condicional) presenta máximos y aparece alejado de aquellas en las que el perfil tiene los mínimos. En forma simétrica, un punto columna aparece cercano de aquellas filas en las que su perfil presenta valores más altos y está alejado de las filas en las que su perfil tiene valores más bajos. También, cuanto más extremos aparezcan los puntos más seguridad habrá sobre la composición de su perfil.

Las relaciones cuasi-baricéntricas (11.19) permiten la representación simultánea de filas y columnas. Aunque no tiene sentido la distancia entre un punto fila y un punto columna, pues éstos pertenecen a espacios diferentes, el AC permite ubicar e interpretar un punto de un espacio (fila o columna) con respecto a los puntos del otro espacio. Como ilustración, admítase que se tienen dos hojas de acetato y en cada una de ellas se han dibujado las proyecciones de los espacios fila y columna, por la propiedad mencionada es posible superponer las dos láminas para ayudarse en la interpretación y búsqueda de resultados.

11.4.3 Reconstrucción de la tabla de frecuencias

En forma semejante al desarrollo hecho en el ACP, se reconstruye la matriz de frecuencias (ecuación (5.4)). Esta matriz $\boldsymbol{F}^* = (f_{ij}/\sqrt{f_{i.}f_{.j}})$ se puede obtener aproximadamente mediante

$$\boldsymbol{F}^* \approx \mathbb{X}^* = \sum_{\alpha=1}^{q} \sqrt{\lambda_\alpha} v_\alpha u'_\alpha. \tag{11.20}$$

De las anteriores relaciones (11.18) y sustituyendo u_α y v_α por sus respectivas proyecciones, después de algunas simplificaciones se obtiene la fórmula de reconstrucción de la matriz $\boldsymbol{F} = (f_{ij})$, con

$$f_{ij} = f_{i.}f_{.j}\left\{1 + \sum_{\alpha>1} \sqrt{\lambda_\alpha}\psi_{i\alpha}\varphi_{j\alpha}\right\}. \tag{11.21}$$

11.4.4 Ubicación de elementos suplementarios

A veces, como una estrategia para la interpretación, se pueden adicionar a la matriz de datos filas (individuos) o columnas (variables), de los cuales se

conocen sus características. El objetivo es proyectarlos en las respectivas nubes (individuos o variables); la posición de éstos (individuos o variables suplementarios) es útil para interpretar los "nuevos" ejes y los grupos que conforman tanto los individuos como las variables iniciales (*activos*). Éstos se pueden considerar como "marcadores", en el sentido de que la ubicación de los demás respecto a tales elementos ayuda a esclarecer los diferentes perfiles de grupos (de variables u objetos) que se conforman; aquí se aplica el aforismo que reza: "dime con quien andas y te diré quien eres". Se obtiene así, una tabla ampliada por un cierto número de columnas (o filas) *suplementarias*. Se trata entonces de posicionar los perfiles de estos nuevos puntos-columna respecto a los p puntos ya situados en $\mathbb{R}^n$, como se ilustra en la figura 11.4.

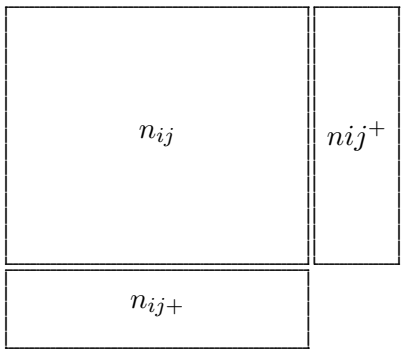


Figura 11.4 *Elementos suplementarios.*

Para las columnas suplementarias, sea n_{ij}^+ la i-ésima coordenada de la j-ésima columna suplementaria; su perfil está dado por:

$$\{n_{ij}^+/n_{.j}^+\},\ \text{con}\ n_{.j}^+ = \sum_{i=1}^{n} n_{ij}^+.$$

La proyección del punto j sobre el eje α, de acuerdo con (11.19), es:

$$\hat{\varphi}_{i\alpha}^+ = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{n} \frac{n_{ij}^+}{n_{.j}^+} \hat{\psi}_{i\alpha}.$$

Análogamente, para una línea suplementaria i, se tiene:

$$\hat{\psi}_{i\alpha}^+ = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{p} \frac{n_{ij}^+}{n_{i.}^+} \hat{\varphi}_{j\alpha}.$$

El interés de proyectar variables suplementarias está en enriquecer la interpretación de los gráficos factoriales obtenidos. El mismo procedimiento se sigue para individuos suplementarios.

11.4.5 Interpretación de los ejes factoriales

El problema central, una vez se ha reducido la dimensionalidad del conjunto de datos, es la asignación de un nombre a los primeros ejes factoriales, para interpretar las proyecciones sobre los planos factoriales, junto con la superposición, de acuerdo con las relaciones de transición.

La asignación de un nombre está en relación con la *contribución absoluta* de cada eje a la variabilidad total, la cual expresa la proporción de la varianza (inercia) con que una modalidad de la variable contribuye a la varianza "retenida" por el eje. En la asignación del nombre, también se consideran las *contribuciones relativas (cosenos cuadrados)* o correlaciones entre elemento-factor, que expresan las contribuciones de un factor en la "explicación" de la dispersión de un elemento.

Mediante las contribuciones absolutas se puede saber qué variables son las responsables de la construcción de un factor, las contribuciones relativas muestran cuales son las características exclusivas de este factor.

Los ejes no aparecen por azar, sino que identifican las direcciones de mayor dispersión (mayor inercia) con respecto a la nube de puntos, siendo la inercia proyectada sobre cada eje igual a su valor propio (λ_α); es decir,

$$\lambda_\alpha = f_{1.}\psi_{1\alpha}^2 + \cdots + f_{n.}\psi_{n\alpha}^2.$$

La contribución de cada punto i (fila) en la inercia de cada eje α está dada por:

$$\mathcal{CA}_{i\alpha} = \frac{f_{i.}\psi_{i\alpha}^2}{\lambda_\alpha}, \text{ para } i = 1, \ldots, n \tag{11.22a}$$

este cociente muestra la contribución del elemento i (fila) al eje α, permite establecer en cuánta proporción un punto i contribuye a la inercia λ_α de la nube de puntos proyectada sobre el eje α.

Así, para interpretar un eje se deben identificar los puntos de mayor contribución, sin perder de vista que la contribución media de un punto i es $1/n$, separando los puntos de acuerdo con el signo de su coordenada respecto al eje.

La interpretación puede hacerse a partir de los puntos fila, como se ha insistido, o también por parte de los puntos columna. De esta misma forma, se define la contribución del elemento j (columna) al eje α mediante:

$$\mathcal{CA}_{j\alpha} = \frac{f_{.j}\psi_{j\alpha}^2}{\lambda_\alpha}, \text{ para } j = 1, \ldots, p \tag{11.22b}$$

Ahora la inquietud es, ¿Qué tan bien queda representado cada punto en los ejes factoriales obtenidos?. Como se tienen los puntos en la base representada por los ejes factoriales, se puede medir la calidad de representación de un punto sobre un eje (contribución relativa) mediante el cociente

$$\mathcal{CR}_\alpha(i) = \frac{\psi_{i\alpha}^2}{d_{(i,G)}} = \cos_\alpha^2(\omega_i), \tag{11.23a}$$

que es el coseno al cuadrado del ángulo (ω_i) formado por el punto i con el eje α. De otra manera, se trata de la relación entre una variable multinomial (p-modalidades) y un eje factorial. Ésta es la contribución relativa o *coseno cuadrado*. Un coseno cuadrado próximo a 1 identifica un ángulo cercano a 0^0 o a 180^0.

Los cosenos cuadrados son aditivos respecto a los ejes factoriales (pues $\sum_\alpha cos_\alpha^2(\omega) = 1$), luego permiten medir la calidad de la representación de los puntos en el espacio definido por los primeros ejes factoriales y la detección de puntos mal representados en los ejes seleccionados. Valores de estos cosenos al cuadrado próximos a 1 dan cuenta de puntos que influyen o están asociados altamente con el respectivo eje.

De manera similar se mide la contribución relativa del eje factorial α a la posición del punto j (columna), es decir, mediante el coseno al cuadrado del ángulo (ω_j) formado entre el eje α y el vector j; esta expresión es:

$$\mathcal{CR}_\alpha(j) = \frac{\psi_{j\alpha}^2}{d_{(j,G)}} = \cos_\alpha^2(\omega_j), \tag{11.23b}$$

similarmente, valores bajos de $\mathcal{CR}_\alpha(j)$ indican una contribución "pobre" del eje α en la posición del punto j.

Ejemplo 11.1 Retomando la tabla de contingencia para el color de ojos y cabello en una muestra de 5387 personas (ahora tabla 11.5).

La nube de puntos fila queda representada por

$$\boldsymbol{D}_4^{-1}\boldsymbol{F} = \begin{pmatrix} 1/1580 & 0 & 0 & 0 \\ 0 & 1/718 & 0 & 0 \\ 0 & 0 & 1/1774 & 0 \\ 0 & 0 & 0 & 1/1315 \end{pmatrix} \begin{pmatrix} 688 & 116 & 584 & 188 & 4 \\ 326 & 38 & 241 & 110 & 3 \\ 343 & 84 & 909 & 412 & 26 \\ 98 & 48 & 403 & 681 & 85 \end{pmatrix}$$

$$= \begin{pmatrix} 0.435 & 0.073 & 0.369 & 0.118 & 0.002 \\ 0.454 & 0.053 & 0.336 & 0.153 & 0.004 \\ 0.193 & 0.047 & 0.512 & 0.232 & 0.015 \\ 0.075 & 0.037 & 0.306 & 0.518 & 0.065 \end{pmatrix},$$

con $\boldsymbol{D}_4 = \text{Diag}(f_{i.})$, matriz diagonal que contiene las frecuencias marginales por fila $f_{i.}$.

Tabla 11.5 Color de ojos vs. color del cabello

	Color de cabello					
Color de ojos	Rubio (ru)	Rojo (r)	Medio (m)	Oscuro (o)	Negro (n)	*Total*
Claros (C)	688	116	584	188	4	1580
Azules (A)	326	38	241	110	3	718
Medio (M)	343	84	909	412	26	1774
Oscuros (O)	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

El *centroide* o *baricentro* de la nube de puntos fila se representa por $\mathcal{G}_f$, y sus coordenadas son iguales a las frecuencias marginales; es decir,

$$\mathcal{G}_f = (f_{.1}, \ldots, f_{.5}) = (0.2700, 0.0530, 0.3967, 0.2582, 0.2190).$$

La matriz a diagonalizar es dada por la ecuación (11.15)

$$\boldsymbol{S}^* = \boldsymbol{D}_5^{-1/2}\hat{\boldsymbol{A}}\boldsymbol{D}_5^{-1/2}$$

$$= \begin{pmatrix} 0.358182 & 0.135761 & 0.322935 & 0.184305 & 0.034908 \\ 0.135761 & 0.056843 & 0.145023 & 0.101453 & 0.026053 \\ 0.322935 & 0.145023 & 0.414569 & 0.305195 & 0.083349 \\ 0.184305 & 0.101453 & 0.305195 & 0.350518 & 0.125863 \\ 0.034908 & 0.026053 & 0.083349 & 0.1258637 & 0.049989 \end{pmatrix}.$$

Los valores propios de $\boldsymbol{S}^*$ son, en forma decreciente, 1.0000, 0.1992, 0.0301, 0.0009 y 0.0000. Como se explicó anteriormente el valor propio igual a

1.0000 es descartado. En el siguiente cuadro se resumen los valores propios junto con la inercia individual y acumulada retenida por cada valor propio.

$ValorPropio$	$Porcentaje$	$Porc.\ Acum.$	
0.1992	86.56	86.56	$*********$
0.0301	13.07	99.63	$***$
0.0009	0.37	100.00	$*$
0.0000	0.00	100.00	$*$

La tabla anterior indica que con la primera dimensión se reune el 86.6% de la varianza y que con la segunda dimensión se reune casi toda su variabilidad; es decir, 99.6%.

Las coordenadas para la "reconstrucción" de la matriz $\mathbb{X}^*$ se obtienen de acuerdo con la ecuación (11.18), los resultados para la descomposición por filas (color de ojos) o columnas (color del cabello) se resumen en la tabla 11.6

La figura 11.5 representa la proyección de los puntos fila y columna (tabla 11.6) en el primer plano factorial. La primera dimensión está relacionada con el color del cabello, variando, de izquierda a derecha, desde el color oscuro al claro, respectivamente. Se puede apreciar que los datos referentes a los ojos siguen un "patrón" similar al del cabello, con colores oscuros a la izquierda y claros a la derecha. Los puntos para *azul* y *rubio* están razonablemente próximos; aunque algunas veces es difícil determinar si las personas tienen ojos claros o azules por problemas de pigmentación.

Tabla 11.6 Coordenadas, color de ojos y del cabello

Coordenadas fila Color de ojos			Coordenadas columna Color del cabello		
	Dim. 1	Dim. 2		Dim. 1	Dim. 2
Claros	0.44	0.09	Rubio	0.54	0.17
Azules	0.40	0.17	Rojo	0.23	0.05
Medios	-0.30	-0.25	Medio	0.04	-0.21
Oscuros	-0.70	0.13	Oscuro	-0.59	0.10
			Negro	-1.09	0.29

En resumen, la dirección del color es de izquierda a derecha, y va de *claro* a *oscuro*; tanto para el cabello como para los ojos.

El procedimiento para el análisis de correspondencias simple o binaria se puede resumir en las siguientes etapas, las cuales se ilustran en la figura 11.6.

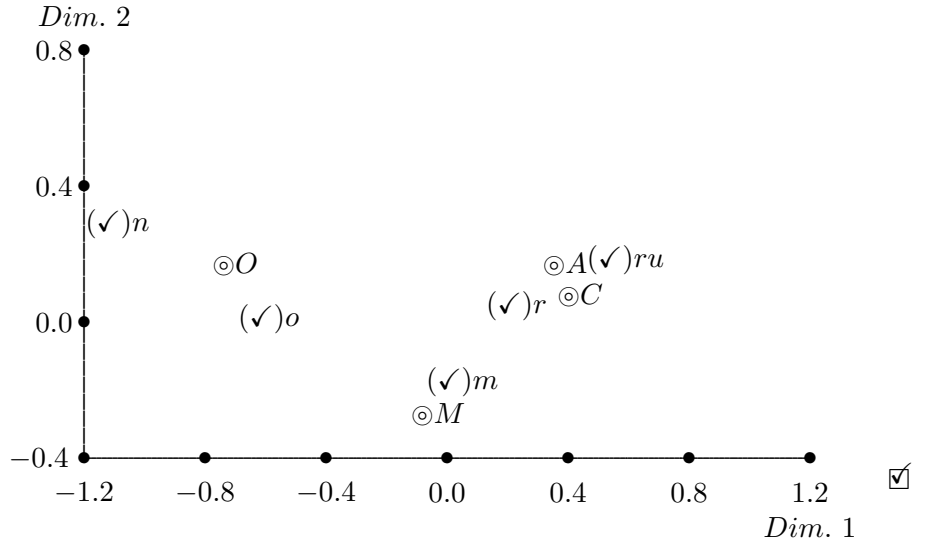


Figura 11.5 *Representación de los datos color de ojos (◎) y del cabello (✓).*

1. Se parte de los datos originales, las filas y columnas juegan papeles simétricos; éstas son las modalidades de las dos variables, respectivamente. La suma de todos los términos de la tabla es n, el cual es el número total de individuos o efectivos.

2. Se construye una tabla de las frecuencias relativas las cuales conforman las probabilidades. Las frecuencias marginales, fila o columna, dadas por los vectores $(f_{i.} : \ i = 1, \ldots, n)$ y $(f_{.j} : \ j = 1, \ldots, p)$, son las probabilidades marginales o perfiles fila y/o columna, respectivamente.

3.-4. Para estudiar las líneas de la tabla, se les transforma en perfiles fila. De manera semejante se procede con las columnas. Se dispone entonces de dos tablas, una para los perfiles fila y otra para los perfiles columna. Un perfil se interpreta como una probabilidad condicional. El perfil medio es la distribución asociada con la que se presenta en el numeral 2.

5. Un perfil-fila es un arreglo de p-números y está representado por un punto de $\mathbb{R}^p$. La nube de puntos $\mathcal{H}_c$, de los perfiles fila, está en un

hiperplano $\mathcal{H}_f$ de vectores tales que la suma de sus componentes es igual a 1. Cada perfil fila i es afectado por los puntos $f_{i.}$; de manera que la nube $\mathcal{H}_f$ está "equilibrada" en los perfiles medios o baricentro $\mathcal{G}_i$. En la nube $\mathcal{H}_f$ se busca la semejanza entre los perfiles, medida a través de una distancia χ^2.

6. La representación de los perfiles columna de $\mathbb{R}^n$ se hace de forma análoga a la representación de los perfiles fila en $\mathbb{R}^p$.

7. El análisis factorial de la nube consiste en poner en evidencia una sucesión de direcciones ortogonales, tales que la inercia, con relación al origen O de la proyección de la nube de puntos sobre tales direcciones sea máxima.

8. Simétricamente, se desarrolla un procedimiento análogo para las columnas.

9.-10 Los planos factoriales, determinados por dos factores sobre las filas o sobre las columnas, proporcionan imágenes aproximadas de las nubes $\mathcal{H}_f$ y $\mathcal{H}_c$, sobre este plano, la distancia entre dos puntos se interpreta como la semejanza entre los perfiles de esos puntos. El origen de los ejes se considera como el perfil promedio.

11. Las relaciones de transición expresan los resultados de un análisis factorial, por ejemplo los del espacio fila en función del espacio columna y recíprocamente, los del espacio columna en función del espacio fila.

12. Una vez que se han realizado las transiciones, las interpretaciones de los planos factoriales que representan a $\mathcal{H}_f$ y $\mathcal{H}_c$ deben hacerse conjuntamente. Ésta es la comodidad de las superposiciones, la interpretación de esta representación simultánea se facilita por la propiedad del doble baricento.

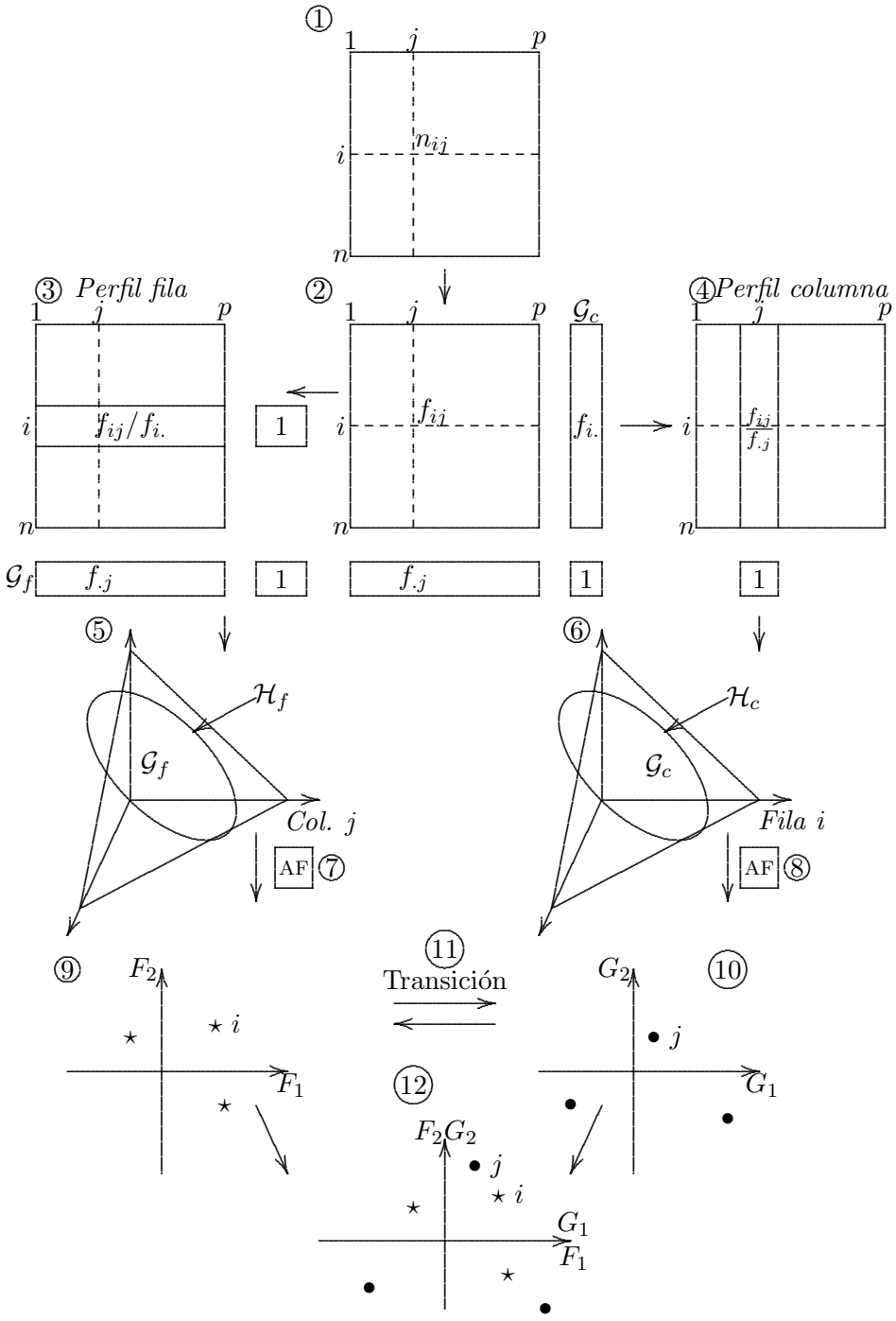


Figura 11.6 *Esquema del análisis de correspondencias*[2].

[2] Tomado de Escofier y Pagés (1990, pág. 42)

11.5 Análisis de correspondencias múltiples-ACM

El AC se ha ocupado, principalmente, de tablas de contingencia bidimensionales. El análisis de correspondencias puede extenderse a tablas de tres o más entradas, caso en el cual se aprecia más su afinidad con el método de componentes principales. Las filas de estas tablas se consideran como los objetos o individuos y las columnas como las modalidades de las variables categóricas en estudio. Es el caso de las encuestas, donde las filas son individuos, grupos humanos o instituciones y las columnas modalidades de respuesta a las preguntas formuladas en el cuestionario o instrumento. El *análisis de correspondencias múltiple* es un análisis de correspondencias simple aplicado no solo a una tabla de contingencia sino a una tabla disyuntiva completa, en el sentido de que una variable categórica asigna a cada individuo de una población una modalidad, y, en consecuencia, particiona (de manera disyuntiva y exhaustiva) a los individuos de la población.

A pesar de sus semejanzas con el análisis de correspondencias simple, el ACM tiene algunas particularidades, debido a la naturaleza misma de la tabla disyuntiva completa ($\mathbb{X}$). En esta sección se enuncian los principios del ACM, cuando éste se desarrolla sobre la tabla disyuntiva completa y después se muestra la equivalencia con el análisis de la tabla de Burt ($\mathbb{B}$).

11.5.1 Tablas de datos

A manera de ilustración, considérese un conjunto de n individuos a los cuales se les registra:

El grupo de edad
: Modalidades: joven (1), adulto (2), anciano (3)

Género
: Modalidades: masculino (1), femenino (2)

Nivel de estudios o escolaridad
: Modalidades: primaria (1), secundaria (2), universitaria (3), otra (4)

Categoría socioeconómica
: Modalidades: bajo (1), medio, (2), alto (3)

Posesión de vivienda
: Modalidades: propietario (1), no propietario (2).

Se tiene entonces una matriz de datos $\boldsymbol{R}$ con 10 filas (individuos) y cinco columnas. Las entradas de esta matriz son los códigos asociados a cada

modalidad de respuesta por pregunta. La siguiente es una de las matrices que surge de las posibles modalidades asumidas por los n individuos

$$\boldsymbol{R} = \begin{pmatrix} 2 & 1 & 2 & 2 & 1 \\ 1 & 2 & 3 & 3 & 2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 3 & 1 & 4 & 3 & 1 \end{pmatrix}$$

Así, la primera fila de la matriz $\boldsymbol{R}$ señala a un hombre adulto, con estudios de secundaria, de estrato socioeconómico medio quien tiene vivienda propia.

Esta matriz o tabla de datos no es tratable vía análisis de correspondencias múltiples; pues la suma de estos números en filas o en columnas no tienen ningún sentido. Una salida para el análisis de esta tabla es una recodificación. Esta recodificación se logra cruzando los individuos con las combinaciones de modalidades para cada una de las preguntas; para el caso se tienen 5 preguntas con 3, 2, 4, 3 y 2 modalidades respectivamente; es decir, $3 \cdot 2 \cdot 4 \cdot 3 \cdot 2 = 144$ posibles respuestas de los individuos.

Mediante el uso de variables indicadoras se convierte una tabla múltiple en una tabla de doble entrada. Supóngase, en general, que a una tabla con k-variables (o preguntas) donde cada una tiene p_i *modalidades* o *categorías* (para $i = 1, \ldots, k$), se asocia, de manera adecuada, una variable indicadora a cada una de las modalidades asociadas con cada una de las variables columna de la tabla. La codificación dada por p_i, hace corresponder tantas variables binarias como modalidades tenga la variable categórica. El total de modalidades es igual a $\sum_{i=1}^{k} p_i = p$.

Un individuo particular se codifica con uno (1) si el individuo posee el atributo de la respectiva modalidad y con cero (0) en las demás modalidades de la misma variable, pues se asume que las modalidades son excluyentes. Resulta entonces una matriz $\mathbb{X}$ de tamaño $(n \times p)$ formada por bloques columna, cada uno de los cuales hace referencia a una variable registrada sobre los n individuos.

Para la matriz $\boldsymbol{R}$ anterior la codificación es como la que se muestra en la figura 11.7, donde las modalidades de cada variable se consideran ahora como variables de tipo dicotómico; cada individuo toma sólo el valor de 1 en una única modalidad y de 0 en las demás modalidades de la misma variable.

La suma en cada una de las filas es constante, en este caso $p = 5$, mientras que la suma en las columnas n_j $(j = 1, \ldots, 14)$ suministra el número de individuos que participan en cada una de las 14 modalidades. La tabla

o matriz $\mathbb{X}$ con n-filas y p-columnas describe las k-respuestas de los n-individuos a través de un código binario (0 o 1) y se le llama *tabla disyuntiva completa*. Esta tabla es la unión de k tablas (una por pregunta). Así, para el ejemplo anterior $\mathbb{X} = [\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3, \mathbb{X}_4, \mathbb{X}_5]$. En general,

$$\mathbb{X} = [\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_k]. \tag{11.24}$$

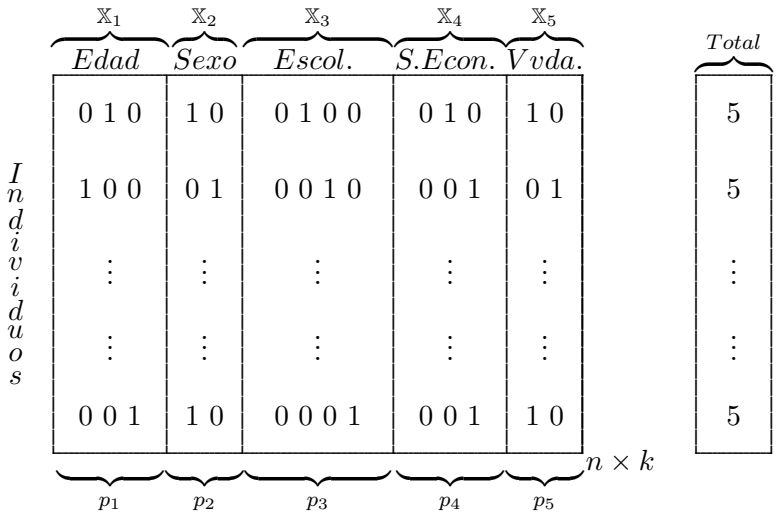


Figura 11.7 *Tabla múltiple.*

Cada una de las tablas $\mathbb{X}_j$, $j = 1, \ldots, k$, describe la partición de los n individuos de acuerdo con sus respuestas a la pregunta j. De otra manera $\mathbb{X}_j = (x_{im})$, donde

$$x_{im} = \begin{cases} 1, & \text{si el } i\text{-ésimo individuo tiene la modalidad } m \text{ de la pregunta } j, \\ 0, & \text{si el } i\text{-ésimo individuo no tiene la modalidad } m \text{ de la pregunta } j. \end{cases}$$

▶ Tabla de Burt

Para cada pregunta o variable, sus p_j respuestas o modalidades permiten particionar la muestra en máximo p_j clases. Para dos variables, con modalidades p_i y p_j, la partición del conjunto de individuos viene determinada por las celdas o casillas de la tabla de contingencia que éstas conforman; esta partición tiene $p_i \times p_j$ clases. Esto puede generalizarse al caso de más de dos variables.

Recuérdese que una tabla disyuntiva completa $\mathbb{X}$ es aquella cuya codificación para las entradas señala la pertenencia de cada individuo a una y solo una de las modalidades de cada variable, de manera que aparece 1 únicamente en la modalidad que asume cada individuo en la respectiva variable. A partir de la tabla disyuntiva completa $\mathbb{X}$ se construye una tabla simétrica $\mathbb{B}$ de tamaño $(p \times p)$ que contiene las frecuencias para los cruces entre todas las k variables. Esta tabla es

$$\mathbb{B} = \mathbb{X}'\mathbb{X}, \tag{11.25}$$

la cual se le conoce como *tabla de contingencia Burt* asociada a la tabla disyuntiva completa $\mathbb{X}$. Un esquema de la tabla de Burt se presenta en la figura 11.8. El término general de $\mathbb{B}$ se escribe

$$b_{jj'} = \sum_{i=1}^{n} x_{ij}x_{ij'}.$$

Las marginales son

$$b_j = \sum_{j'=1}^{p} b_{jj'} = kx_{.j}, \text{ para todo } j \leq p.$$

La frecuencia total es igual a

$$b = k^2 x_{.j}$$

La tabla $\mathbb{B}$ está conformada por k^2 bloques, donde:

- El bloque $\boldsymbol{X}'_j\boldsymbol{X}_{j'}$ de tamaño $(p_j \times p_{j'})$ corresponde a la tabla de contingencia que cruza las respuestas a las preguntas (variables) j y j'.
- El j-ésimo bloque cuadrado $\boldsymbol{X}'_j\boldsymbol{X}_j$ se obtiene mediante el cruce de cada variable consigo misma. Ésta es una matriz diagonal de tamaño $(p_j \times p_j)$; la matriz es diagonal dado que dos o más modalidades de una misma pregunta no pueden ser seleccionados simultáneamente. Los términos sobre la diagonal son las frecuencias de las modalidades de la pregunta j.

Sobre la diagonal de la tabla de Burt $\mathbb{B}$, de la figura 11.8, se han insinuado matrices diagonales. Éstas se notan por $\boldsymbol{D}_j = P\mathbb{X}'_jP\mathbb{X}_j$; $j = 1, \ldots, k$ y son matrices de tamaño $(p_j \times p_j)$. Dichas matrices deben ser diagonales puesto que un individuo no puede estar ubicado de manera simultánea en dos o más

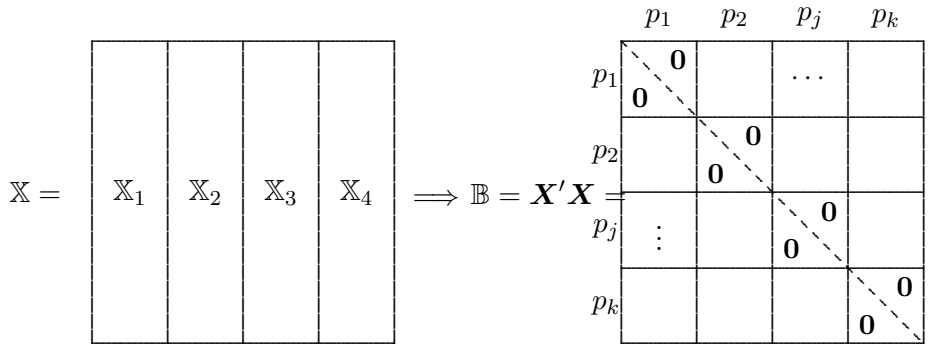


Figura 11.8 *Construcción de la tabla de Burt.*

modalidades para una misma pregunta o variable. Los elementos o términos de la diagonal son las frecuencias de las modalidades de la pregunta j; es decir, es el número de individuos por modalidad en la pregunta j.Nótese que la suma de estas frecuencias (traza) es la misma para todas las matrices de la diagonal y es igual al número de individuos u objetos; a menos que haya información faltante en algunas de las modalidades.

Las matrices que están fuera de la diagonal principal de $\mathbb{B}$ son las mismas tablas de contingencia entre las respectivas variables fila y columna de la tabla de Burt.

Se nota por $\boldsymbol{D}$ a la matriz diagonal de tamaño $(p \times p)$; es decir, sobre la diagonal están las frecuencias correspondientes a cada una de las modalidades

$$d_{jj} = b_{jj} = x_{.j},$$
$$d_{jj'} = 0 \text{ para todo } j \neq j'.$$

La matriz $\boldsymbol{D}$ se puede considerar que está conformada por k^2 bloques. Las únicas matrices no nulas son las matrices diagonales $\boldsymbol{D}_j = \boldsymbol{X}_j'\boldsymbol{X}_j$; $j = 1, \ldots, k$ las cuales están dispuestas sobre la diagonal principal de $\boldsymbol{D}$.

En resumen, una tabla de Burt yuxtapone todas las tablas de contingencia de las variables cruzadas por pares. La tabla de Burt es simétrica por bloques, las tablas de la diagonal son a su vez diagonales y contienen las frecuencias marginales de cada una de las variables, las tablas fuera de la diagonal son las tablas de contingencia de las variables que las definen.

Ejemplo 11.2 En un grupo de 20 individuos se hizo una encuesta acerca de las cinco variables socioeconómicas descritas anteriormente. A continua-

ción se muestra la matriz de datos con su código condensado $\boldsymbol{R}$, la tabla de datos disyunta completa $\mathbb{X}$, la tabla de Burt $\mathbb{B}$ y la tabla diagonal $\boldsymbol{D}$.

$$\boldsymbol{R} = \left[\begin{array}{ccccc}
2 & 1 & 2 & 2 & 1 \\
3 & 2 & 1 & 2 & 1 \\
3 & 1 & 4 & 2 & 1 \\
3 & 2 & 2 & 2 & 1 \\
2 & 1 & 1 & 2 & 1 \\
1 & 1 & 1 & 1 & 2 \\
2 & 1 & 2 & 2 & 2 \\
2 & 2 & 2 & 2 & 1 \\
3 & 1 & 3 & 3 & 1 \\
2 & 2 & 4 & 2 & 1 \\
1 & 2 & 3 & 3 & 1 \\
1 & 1 & 2 & 3 & 2 \\
2 & 2 & 2 & 2 & 1 \\
3 & 2 & 2 & 2 & 1 \\
3 & 1 & 4 & 3 & 1 \\
1 & 2 & 2 & 2 & 1 \\
3 & 2 & 3 & 2 & 1 \\
3 & 1 & 1 & 1 & 2 \\
2 & 2 & 2 & 1 & 2 \\
1 & 1 & 3 & 2 & 1
\end{array}\right], \quad X = \left[\begin{array}{ccc|cc|cccc|cccc|cc}
0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0
\end{array}\right]$$

La tabla de Burt $\mathbb{B}$ y la matriz diagonal $\boldsymbol{D}$ son, respectivamente,

$$\mathbb{B} = \left[\begin{array}{ccc|cc|cccc|cccc|cc}
5 & 0 & 0 & 3 & 2 & 1 & 2 & 2 & 0 & 1 & 2 & 2 & 3 & 2 \\
0 & 7 & 0 & 3 & 4 & 1 & 5 & 0 & 1 & 1 & 6 & 0 & 5 & 2 \\
0 & 0 & 8 & 4 & 4 & 2 & 2 & 2 & 2 & 1 & 5 & 2 & 7 & 1 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
3 & 3 & 4 & 10 & 0 & 3 & 3 & 2 & 2 & 2 & 5 & 3 & 6 & 4 \\
2 & 4 & 4 & 0 & 10 & 1 & 6 & 2 & 1 & 1 & 8 & 1 & 9 & 1 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
1 & 1 & 2 & 3 & 1 & 4 & 0 & 0 & 0 & 2 & 2 & 0 & 2 & 2 \\
2 & 5 & 2 & 3 & 6 & 0 & 9 & 0 & 0 & 1 & 7 & 1 & 6 & 3 \\
2 & 0 & 2 & 2 & 2 & 0 & 0 & 4 & 0 & 0 & 2 & 2 & 4 & 0 \\
0 & 1 & 2 & 2 & 1 & 0 & 0 & 0 & 3 & 0 & 2 & 1 & 3 & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
1 & 1 & 1 & 2 & 1 & 2 & 1 & 0 & 0 & 3 & 0 & 0 & 0 & 3 \\
2 & 6 & 5 & 5 & 8 & 2 & 7 & 2 & 2 & 0 & 13 & 0 & 12 & 1 \\
2 & 0 & 2 & 3 & 1 & 0 & 1 & 2 & 1 & 0 & 0 & 4 & 3 & 1 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
3 & 5 & 7 & 6 & 9 & 2 & 6 & 4 & 3 & 0 & 12 & 3 & 15 & 0 \\
2 & 2 & 1 & 4 & 1 & 2 & 3 & 0 & 0 & 3 & 1 & 1 & 0 & 5
\end{array}\right]$$

$\Downarrow$

$$\boldsymbol{D} = \left[\begin{array}{ccc|cc|cccc|ccc|cc}
5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 13 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 15 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5
\end{array}\right]$$

11.5.2 Bases del análisis de correspondencias múltiples

El análisis de correspondencias múltiples compara individuos a través de las modalidades de las variables que los identifican en el estudio. Un grupo de individuos es similar si éstos asumen aproximadamente las mismas variables. La asociación entre variables se presenta porque son casi que los mismos individuos quienes asumen las mismas modalidades de diferentes variables.

El análisis de correspondencias múltiples encuentra asociaciones entre variables de tipo categórico a través de las respectivas modalidades de éstas.

Como en el análisis de correspondencias binarias, el análisis para el caso múltiple considera la nube de puntos fila (n puntos) y la nube de puntos columna (p puntos). En el primer caso los ejes son las variables y en el segundo los individuos. En consecuencia los principios de ACM son los mismos que los del AC simples, éstos son (Lebart, Morineau y Piron 1995, págs. 113-127):

- transformar la tabla de datos en perfiles fila y perfiles columna,
- ajustar los datos o puntos ponderados por sus perfiles marginales (fila o columna),
- estar dada la distancia entre perfiles por la ji-cuadrado.

▶ Criterio de ajuste y distancia ji-cuadrado

Los individuos están afectados por la misma ponderación $1/n$. Cada una de las modalidades j está ponderada por su frecuencia; es decir, $n_j = x_{.j}/nk$, con $x_{.j} = \sum_{i=1}^{n} x_{ij}$.

Las respectivas distancias ji-cuadrado, entre modalidades y entre individuos, aplicada a una tabla disyuntiva completa están dadas por

$$\begin{cases} d^2(j,j') = \sum_{i=1}^{n} n\left(\frac{x_{ij}}{x_{.j}} - \frac{x_{ij'}}{x_{.j'}}\right)^2, & \text{modalidades } j \text{ y } j' \text{ (en } \mathbb{R}^n), \\ d^2(i,i') = \frac{1}{k}\sum_{j=1}^{p} \frac{n}{x_{.j}}\left(x_{ij} - x_{i'j}\right)^2, & \text{individuos } i \text{ e } i' \text{ (en } \mathbb{R}^p). \end{cases}$$

Así, dos modalidades que son seleccionadas por los mismos individuos coinciden (pues $x_{ij} = x_{ij'}$). Además, las modalidades de frecuencia baja (las "raras") están alejadas de las otras modalidades. En forma semejante, dos individuos están próximos si ellos han seleccionado las mismas modalidades. Los individuos están alejados si no han respondido de la misma manera.

▶ Ejes factoriales y factores

Como el procedimiento seguido en la sección (11.4), para el análisis de correspondencias simples, la notación es

$\boldsymbol{F}\dfrac{1}{nk}\mathbb{X}$, cuyo término general es $f_{ij} = \dfrac{x_{ij}}{nk}$.

$\boldsymbol{D}_p = \dfrac{1}{nk}\boldsymbol{D}$, cuyo término general es $f_{.j} = \delta_{ij}\dfrac{x_{.j}}{nk}$, $\delta_{ij} = 1$ si $i = j$ y 0 si $i \neq j$.

$\boldsymbol{D}_n = \dfrac{1}{n}\boldsymbol{I}_n$, el término general es $f_{i.} = \dfrac{\delta_{ij}}{n}$.

Los ejes factoriales se encuentran a través de los valores y vectores propios de la matriz (similar a (11.12)):

$$\boldsymbol{S} = \boldsymbol{F}'\boldsymbol{D}_n^{-1}\boldsymbol{F}\boldsymbol{D}_p^{-1} = \frac{1}{k}\mathbb{X}'\mathbb{X}\boldsymbol{D},$$

cuyo término general es

$$s_{jj'} = \frac{1}{kx_{.j'}}\sum_{i=1}^{n} x_{ij}x_{ij'}.$$

En el espacio fila o de los individuos (en $\mathbb{R}^p$), la ecuación del α-ésimo eje factorial $\boldsymbol{u}_\alpha$ es:

$$\frac{1}{k}\mathbb{X}'\mathbb{X}\boldsymbol{D}^{-1}\boldsymbol{u}_\alpha = \lambda_\alpha\boldsymbol{u}_\alpha. \tag{11.26}$$

Para el espacio columna o de las modalidades (en $\mathbb{R}^n$), el α-ésimo eje factorial ψ_α se esribe

$$\frac{1}{k}\mathbb{X}\boldsymbol{D}^{-1}\mathbb{X}'\psi_\alpha = \lambda_\alpha\psi_\alpha. \tag{11.27}$$

donde los factores φ_α y ψ_α (de norma λ_α) representan las coordenadas de los puntos fila y de los puntos columna sobre el eje factorial α.

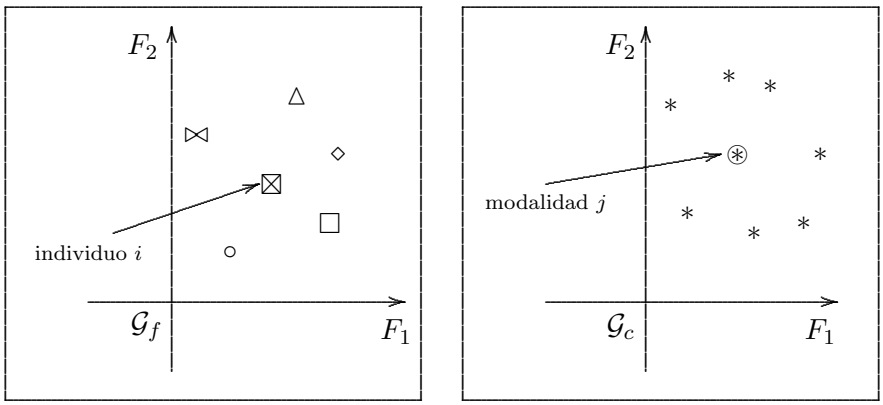
Las relaciones de transición (como en (11.18)) entre los factores φ_α y ψ_α son:

$$\begin{cases} \varphi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\boldsymbol{D}^{-1}\mathbb{X}'\psi_\alpha, \\ \\ \psi_\alpha = \frac{1}{k\sqrt{\lambda_\alpha}}\mathbb{X}\varphi_{j\alpha}. \end{cases} \tag{11.28}$$

Las coordenadas factoriales de un individuo i sobre el eje α están dadas por:

$$\psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{j=1}^{p}\frac{x_{ij}}{x_{i.}}\varphi_{\alpha j} = \frac{1}{k\sqrt{\lambda_\alpha}}\sum_{j\in\mathcal{P}(i)}\varphi_{\alpha j},$$

donde $\mathcal{P}(i)$ es el conjunto de las modalidades seleccionadas por el individuo i. Con excepción del coeficiente $1/\sqrt{\lambda_\alpha}$, el individuo i se encuentra en el punto medio de la nube de modalidades que él ha seleccionado (figura 11.9a).



(a) *(b)*

Figura 11.9 *Proyección de individuos y modalidades*

De forma análoga, la coordenada de la modalidad j sobre el eje α está dada por

$$\begin{aligned}\varphi_{\alpha i} &= \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{x_{ij}}{x_{.j}} \varphi_{\alpha i} \\ &= \frac{1}{x_{.j}\sqrt{\lambda_\alpha}} \sum_{i \in \mathcal{I}(j)} \varphi_{\alpha i}, \end{aligned} \tag{11.29}$$

donde $\mathcal{I}(j)$ es el conjunto de los individuos que selecionaron la modalidad j. Antes de la "dilatación" sobre el eje α, la modalidad j se encuentra en el punto medio de la nube de los individuos que le han seleccionado como respuesta (figura 11.9b).

La nube de modalidades en $\mathbb{R}^n$ se puede descomponer en *subnubes*; así, la j-ésima nube corresponde al conjunto de las p_j modalidades de la variable j. Estas subnubes tienen su centro de gravedad en $\mathcal{G}_f$, el mismo de la nube global.

En resumen, el análisis de correspondencia múltiples se dirige a buscar aquellas variables o factores "cercanas" (altamente correlacionadas) con todos los grupos de modalidades. El factor F_1 representa el primer factor común al conjunto de variables categóricas iniciales. Los demás factores se obtienen con la condición de ortogonalidad sobre los anteriores.

Los factores $F_1,\ F_2, \ldots, F_k$ ubicados en el espacio de las modalidades, son los ejes en el espacio de los individuos; de tal forma que su proyección sobre estos "nuevos" ejes retienen la máxima variabilidad. Se puede observar la similitud, por lo menos conceptual, con el análisis de componentes principales, con una importante diferencia, y es que aquí cada variable está constituida por un subgrupo de variables binarias.

Observación:

- La tabla de Burt $\mathbb{B}$ es un caso particular de tablas de contingencia, las cuales se pueden asociar con las caras de un hipercubo de contingencia.
- El análisis de correspondencias aplicado a una tabla disyuntiva completa $\mathbb{X}$ es equivalente a la tabla de Burt $\mathbb{B}$ y produce los mismos factores.

En seguida se destacan algunas propiedades de este análisis.

▶ Propiedades del análisis de correspondencias múltiples

1. Es una representación gráfica de la asociación entre variables categóricas dos a dos; en consecuencia el análisis de correspondencias simple es un caso especial para un par de variables en particular.

2. A diferencia del análisis de componentes principales, los primeros ejes, aún en forma creciente, explican una pequeña parte de la variabilidad total.

3. La distancia de una modalidad al origen en el ACM es inversamente proporcional a su participación n_j. Es decir, modalidades con participación baja (n_j pequeño) aparecen más alejadas del origen que las modalidades de mayor frecuencia.

4. Las modalidades o categorías de una variable están centradas; es decir, el centro de las modalidades de una misma variable es el origen del "nuevo" sistema de coordenadas. Así, las modalidades de una variable dicotómica se ubicarán en forma opuesta al origen.

5. El ACM es una descomposición de la nube de puntos de la varianza o inercia total del espacio de individuos (filas) o del espacio de las modalidades (columnas), en ciertas direcciones ortogonales, de tal forma que en cada dirección se maximice la inercia explicada.

6. Así como en el ACP la influencia de cada variable está dada por su varianza, las modalidades situadas a mayor distancia tienen la mayor inercia, luego son las más influyentes y de acuerdo con la propiedad (3.), son las que tienen menor número de individuos.

7. Tal como en el AC simple, existe una relación de transición entre la "nueva" variable del espacio de los individuos y la de las modalidades. Ésta se expresa a través de la *propiedad baricéntrica* dada en (11.31).

8. La proyección de un individuo es el centro de gravedad de las modalidades que éste ha escogido (a una distancia $\frac{1}{\sqrt{\lambda_\alpha}}$ del origen). Simétricamente, la proyección de una modalidad es el centro de gravedad de los individuos que la han escogido (a una distancia $\frac{1}{\sqrt{\lambda_\alpha}}$ del origen).

▶ Reglas para la interpretación

Decir que existen afinidades entre respuestas, equivale a decir que hay individuos que han seleccionado simultáneamente todas o casi todas, las mismas respuestas.

El análisis de correspondencias múltiples pone en evidencia a los individuos con perfiles semejantes respecto a los atributos seleccionados para su descripción. De acuerdo con las distancias entre elementos de la tabla disyuntiva completa y las relaciones baricéntricas, se expresa:

- *La cercanía entre individuos en términos de semejanzas*; es decir, dos individuos son semejantes si han seleccionado globalmente las mismas modalidades.
- *La proximidad entre modalidades de variables diferentes en términos de asociación*; es decir, estas modalidades corresponden a puntos medios de los individuos que las han seleccionado, y son próximas porque están ligadas a los mismos individuos o individuos parecidos.
- *La proximidad entre dos modalidades de una misma variable en términos de semejanza*; por construcción, las modalidades de una misma variable son excluyentes. Si ellas están cerca, su proximidad se interpreta en términos de semejanza entre los grupos de individuos que las han seleccionado (con respecto a las otras variables activas del análisis).

Las reglas de interpretación de los resultados, tales como coordenadas, contribuciones, cosenos cuadrados, son casi las mismas que las dispuestas para el análisis de correspondencias simples.

La conceptualización de cada variable debe ser tenida en cuenta al momento de la interpretación, ésta se debe hacer a través de las modalidades que la conforman. No debe olvidarse que los análisis están orientados por una teoría o marco conceptual, desde donde se "ponen en escena" los datos.

La contribución de una variable a un factor α se calcula sumando las contribuciones de las respectivas modalidades sobre ese factor; es decir,

$$\mathcal{CR}_\alpha = \sum_{h \in j} \mathcal{CR}_\alpha(h).$$

Así, se debe prestar atención a las variables que participan en la definición del factor, de acuerdo con las modalidades más "responsables" de los ejes factoriales.

► Individuos y variables suplementarios

La utilización de elementos suplementarios, sean individuos y/o variables, en el ACM permiten considerar información adicional que facilita la búsqueda de una tipología de los elementos activos; toda vez que se conozcan las

características de los individuos (o variables) suplementarios. Los elementos suplementarios se hacen intervenir en una tabla disyuntiva completa para:

- Enriquecer la interpretación de los ejes mediante variables que no han participado en su conformación. Se proyectarán entonces en el espacio de las variables los centros de grupos de individuos definidos por las modalidades de variables suplementarias.
- Adoptar una óptica de pronóstico, proyectando las variables suplementarias en el espacio de los individuos; las variables activas hacen el papel de variables explicativas. Se pueden proyectar a los individuos suplementarios en el espacio de las variables, para ubicarlos con respecto a los individuos activos o con respecto a grupos de individuos activos a manera de discriminación o separación de grupos.

Ejemplo 11.3 Se consideran los datos del ejemplo 11.2 sobre los 20 individuos a quienes se les registró las variables: grupo etáreo, género, escolaridad, estrato socioeconómico y posesión de vivienda. El análisis se hace a través del procedimiento CORRESP del paquete SAS. Se construyen algunas tablas de contingencia y se determinan los factores, que junto con algunos indicadores sirven para interpretar y juzgar la calidad de los ejes factoriales. A pesar de insistir en la idealización o simulación de los datos, se aventuran algunas conclusiones derivadas del análisis de correspondencias múltiple para estos datos.

Valor propio	*Porcen.*	*Porcen. Acumul.*	5 10 15 20 25 ●●●●○●●●●○●●●●○●●●●○●●●●○●●●
0.64761	23.97%	23.97%	**********
0.62382	22.24%	46.21%	**********
0.57554	18.93%	65.14%	**********
0.47050	12.65%	77.79%	**********
0.39452	8.89%	86.68%	*********
0.37784	8.16%	94.84%	********
0.30070	5.16%	100.00%	*****

En la tabla anterior se observa que se tienen siete valores propios no nulos, pues el número de variables activas es $k = 4$ y el número de modalidades es $p = 3 + 2 + 4 + 2 = 11$, de donde $p - k = 7$. Aunque no se consignaron aquí, la inercia ligada a cada valor propio varía entre 0.41940 para el valor propio más grande y 0.09042 para el más pequeño. Esto no debe sorprender ya que los códigos binarios asignados a las modalidades de una misma

variable resultan, así sea artificialmente, ortogonales. Ya se advirtió sobre el cuidado de emplear los valores propios y las tasas de inercia como indicadores del número de ejes apropiados; sin embargo, a pesar de los casi siempre resultados pesimistas encontrados con éstos, pues obsérvese que con los dos primeros ejes reunen el 46.21% de la inercia total. Para efectos de interpretación de los datos, se puede y se debe hacer el análisis sobre el primer plano factorial y sobre otros planos tales como el $factor$ 1 vs el $factor$ 3, por ejemplo.

La tabla 11.7 contiene las variables, las modalidades con sus respectivas etiquetas, las coordenadas de las modalidades sobre los dos primeros factores y los cuadrados de los cosenos de las modalidades sobre los dos primeros ejes factoriales.

Tabla 11.7 Coordenadas y contribuciones de las modalidades

Variable	Modalidad	Factor 1	Factor 2	Cosenos cuadrados	
Edad					
	∘ joven	0.57924	0.49117	0.111839	0.080417
	⊙ adulto	0.19298	-1.01751	0.020054	0.557487
	⊛ viejo	-0.53088	0.58334	0.187891	0.226857
Género					
	♠ hombre	0.51883	0.53084	0.269182	0.281789
	♡ mujer	-0.51883	-0.53084	0.269182	0.281789
Escolaridad					
	⊡ prima.	1.01657	0.72719	0.258352	0.132200
	⊟ secun.	0.16571	-0.91692	0.022466	0.687880
	⊞ univer.	-0.70487	1.00251	0.124211	0.251254
	⊠ otro	-0.91271	0.44450	0.147006	0.034868
Vivienda					
	$\widehat{\text{II}}$ propie.	-0.50180	0.02075	0.755407	0.001291
	II noprop.	1.50540	-0.06224	0.755407	0.001291
		Variable suplementaria			
Estrato SE.					
	⊖ bajo	1.48530	0.20099	0.389312	0.007129
	⊘ medio	-0.29588	-0.28078	0.162585	0.146414
	⊕ alto	-0.15236	0.76180	0.005803	0.145085

Con relación al primer factor se nota que está definido por la posesión de vivienda. Situación que se corrobora con los cosenos cuadrados; recuérdese que un valor de éstos cercano a 1.0 indica un ángulo de la modalidad con el respectivo eje próximo a 0.0; es decir, una alta asociación entre la modalidad y el eje. También se destaca la diferenciación mostrada entre el grupo etáreo "viejo" y los demás; con una proximidad a la posesión de vivienda, lo que sugiere una relación directa entre la tenencia de vivienda y la edad. Una conclusión similar se puede establecer para la edad y el nivel de escolaridad,

los datos exhiben que el nivel de escolaridad superior (universitaria y otro) están asociadas con edades avanzadas.

El segundo factor, se observa que es determinado por la escolaridad superior y secundaria. La variable suplemetaria, nivel socioeconómico, refuerza la asociación de este eje con tales aspectos. Respecto al género se puede afirmar, a partir de estos datos, que no definen los ejes (se ubican el la bisectriz principal). Para la variable edad la modalidad "joven" es indiferente en la definición de alguno de los dos ejes (se ubica en la bisectriz principal), en cambio las modalidades adulto y viejo son opuestas y están altamente ligadas con el segundo eje.

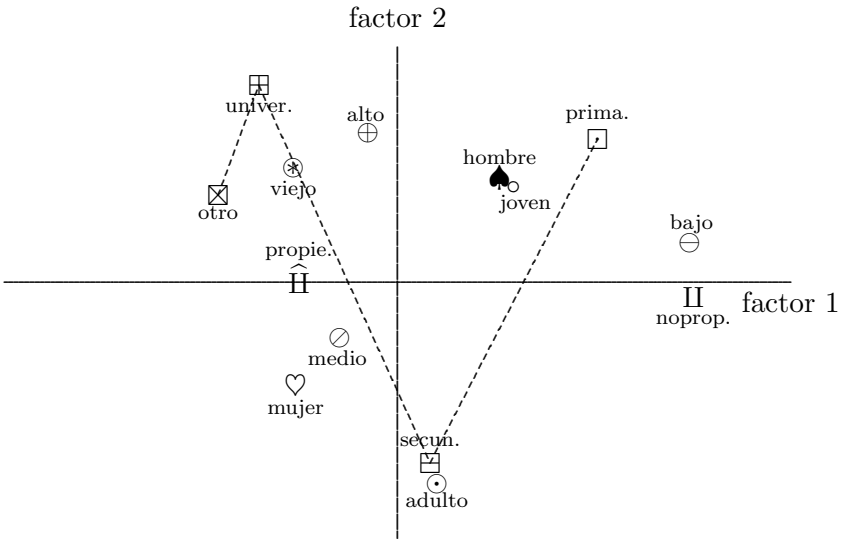


Figura 11.10 *Variables en el primer plano factorial.*

La figura 11.10 muestra la disposición de las modalidades en el primer plano factorial. Se observa que el primer eje factorial (factor 1) está altamente determinado por la variable posesión de casa propia. Así, este eje determina dos tipologías de individuos, del *lado izquierdo* se puede afirmar que están quienes poseen un nivel de escolaridad universitario o más, con vivienda propia y por lo tanto en estratos socioeconómicos medios y altos, mientras que del *lado derecho* se encuentran quienes tienen un nivel de escolaridad a lo más de secundaria y que no tienen vivienda propia. El segundo eje factorial (factor 2) está definido por las modalidades asociadas a la escolaridad, discriminada ésta por la modalidad secundaria frente a la modalidad universitario u otro.

Se han unido, en forma ordenada, las modalidades mediante una línea poligonal. Con estas líneas se puede leer y descubrir relaciones entre las modalidades. Nótese, por ejemplo, que la línea que une las modalidades de escolaridad, tiene un orden decreciente de izquierda a derecha (por magnitud de escolaridad) y pasa cerca o paralela a las modalidades con las que se asocia directamente. Un ejercicio similar puede hacerse con las demás modalidades. ☑

11.6 Rutina SAS para análisis de correspondencias

El procedimiento PROC CORRESP es una rutina computacional del paquete SAS para desarrollar análisis de correspondencias simples o múltiples. El análisis puede hacerse con base en una tabla de contingencia, una tabla de Burt o a partir de los datos categóricos originales.

```
DATA EJEM11_2;/*Ejemplo 11.2*/
INPUT NOMBRE $ EDAD $ GENERO $ ESCOL $ SOCIEC $ VVDA $ @@;
/*Variables categóricas */
CARDS;
2 1 2 2 1  3 2 1 2 1  3 1 4 2 1  3 2 2 2 1  2 1 1 2 1  1 1 1 1 2  2 1 2 2 2
2 2 2 2 1  3 1 3 3 1  2 2 4 2 1  1 2 3 3 1  1 1 2 3 2  2 2 2 2 1  3 2 2 2 1
3 1 4 3 1  1 2 2 2 1  3 2 3 2 1  3 1 1 1 2  2 2 2 1 2  1 1 3 2 1
;
DATA EJE\_MODI;
SET EJEM11_2; /* Nombre de cada categoría por variable*/
IF EDAD='1' THEN EDAD='JOVEN'; IF EDAD='2' THEN EDAD='ADULTO';
IF EDAD='3' THEN EDAD='VIEJO';
IF GENERO='1' THEN GENERO='HOMBRE'; ELSE GENERO='MUJER';
IF ESCOL='1' THEN ESCOL='PRIMA';
IF ESCOL='2' THEN ESCOL='SECUN';
IF ESCOL='3' THEN ESCOL='UNIVER';
 IF  ESCOL='4' THEN ESCOL='OTRO';
IF SOCIEC='1' THEN SOCIEC='BAJO';
 IF SOCIEC='2' THEN SOCIEC='MEDIO';
IF SOCIEC='3' THEN SOCIEC='ALTO';
IF VVDA='1' THEN VVDA='PROPIE';  ELSE VVDA='NOPRO';
PROC CORRESP DATA=ACM1 OUTC=EJES OBSERVED MCA;
/*Procedimiento para el análisis de correspondencias múltiples,*/
/*EJES contiene las coordenadas de las modalidades de variables activas
                    y suplementarias*/
```

```
/*OBSERVED imprime tabla de contingencia*/
/*MCA indica análisis de correspondencias múltiples*/
TABLES EDAD GENERO ESCOL SOCIEC VVDA;
/*TABLES crea una tabla de contingencia o de Burt desde la variables dadas en el
INPUT*/
SUPPLEMENTARY SOCIEC; /*indica la(s) variables suplementaria(s) */
DATA EJES1;
SET EJES;
Y=DIM2;
   X=DIM1;
   XSYS = '2';
   YSYS = '2';
   TEXT = _NAME_ ;
   SIZE =2;
  LABEL Y='FACTOR 2'
        X='FACTOR 2';
  KEEP X Y TEXT XSYS YSYS SIZE;
PROC GPLOT DATA=EJES1;
  SYMBOL V=NONE;
 AXIS1 LENGTH=8 IN ORDER=-2 TO 2 BY 0.5;
   PLOT Y*X=1/ANNOTATE=EJES1 FRAME HAXIS=AXIS1 VAXIS=AXIS1
              HREF=0 VREF=0;
/*Rutina para ubicar las modalidades en el primer plano factorial */
RUN;
```

11.7 Procesamiento de datos con R

```
# análisis de correspondencias simples
# Ejemplo cápítulo 11
### introducir tabla 11.1
t11.1<-matrix(c(688,326,343, 98,
                116, 38, 84, 48,
                584,241,909,403,
                188,110,412,681,
                  4,  3, 26, 85 ),nrow=4)
dimnames(t11.1)<-list(col.ojos=c("Claros","Azules","Medios",
                    "Oscuros"),col.cabello=c("Rubio","Rojo",
                    "Medio","Oscuro","Negro"))
# se agregan las marginales
# tabla 11.1 addmargins(t11.1)
options("digits"=2)
```

```
# sugerencia sobre el numero de cifras significativas
# tabla 11.2 addmargins( 100*prop.table(t11.1) ) options("digits"=4)
# sugerencia sobre el numero de cifras significativas
# tabla 11.3 addmargins( prop.table(t11.1,1),2 )
# tabla 11.4 addmargins( prop.table(t11.1,2),1 )
# en lo que sigue se realizan los graficos de los perfiles
# fila y columna require(reshape) require(lattice)
# se organizan los datos para realizar el gráfico.
datosf<-melt( prop.table(t11.1,1)  )

# figura 11.3 barchart(value~col.cabello|col.ojos,
data=datosf,layout=c(4,1),xlab="Color del cabello",
main="Perfiles Fila")
datosc<-melt( prop.table(t11.1,2)  )

# figura 11.4
barchart(value~col.ojos|col.cabello ,data=datosc,layout=c(5,1),
xlab="Color del cabello",main="Perfiles Columna")

# El análisis de correspondencia simple se encuentra dentro
# de la librería "ca".
# El análisis de correspondencias propiamente dicho library(ca).
require(ca)
acs<-ca(t11.1) summary(acs)

# En Rows:  de la salida anterior, marcadas con k=1 y k=2 estás
# las cordenadas fila de la tabla 11.6 pero multiplicadas por
# mil y con el signo contrario. si se quiere recuperar esa
# información, tal como aparece en dicha tabla, se hace lo
# siguiente para obtener, explicitamente la  tabla 11.6
res<-summary(acs)
cord.filas<--cbind(res$rows[,5],res$rows[,8])/1000

# corrdenasdas para las filas, tabla 11.6
cord.filas

# En Columns: de la salida de summary(acs), marcadas con k=1
# y k=2 están las cordenada fila de la tabla 11.6 pero
# multiplicadas por mil y con el signo contrario.  Si se
# quiere recuperar esa  información, tal como aparece en
# dicha tabla,  se hace lo siguiente
cord.col<--cbind(res$columns[,5],res$columns[,8])/1000

# corrdenasdas para las columnas, tabla 11.6
cord.col
# biplot
plot(acs)
res$columns[,5]
names(res)
res$row
names(acs)
```

Apéndice A

Álgebra de matrices

A.1 Introducción

La derivación, desarrollo y comprensión de los diferentes temas tratados en el texto se posibilita, en gran parte, mediante el empleo del álgebra lineal. Por esta razón se hace una presentación condensada de los elementos esenciales de esta área. Los temas considerados en este aparte deliberadamente tienen el enfoque hacia la estadística, es decir que son un caso particular de una teoría más general como el álgebra lineal. Se enfatiza en los conceptos y los resultados más no en su demostración, para un tratamiento formal se pueden consultar Graybill (1969), Searle (1982), Magnus (1990) y Harville (1997); textos de álgebra lineal con un tratamiento exclusivo para la estadística.

A.1.1 Vectores

Un vector es un arreglo de números dispuestos en filas o en columnas. Si el arreglo tiene n números se dice que el vector tiene *tamaño* $(n \times 1)$ o $(1 \times n)$, según se trate de un *vector columna* o un *vector fila*; en cualquiera de los dos casos se dice que es un elemento de $\mathbb{R}^n$. Ellos se escriben respectivamente en la forma

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{y} \quad Y = \begin{pmatrix} y_1, & y_2, \cdots, & y_n \end{pmatrix}. \tag{A1.1}$$

Un vector columna se obtiene por la transposición de un vector fila, y recíprocamente, un vector fila corresponde al transpuesto de un vector

columna. Se nota por X' (o X^T), el vector de tamaño $(1 \times n)$ que corresponde al transpuesto del vector X de tamaño $(n \times 1)$; es decir,

$$X' = \left(x_1,\ x_2, \cdots,\ x_n\right). \tag{A1.2}$$

Observación:

- En el texto los vectores se consideran como vectores columna de tamaño $(n \times 1)$, en caso contrario se toma el transpuesto.
- El vector cuyos componentes son todos cero se denomina vector *nulo* o *cero*. El vector nulo se nota por $\mathbf{0}$. Análogamente, el vector de *unos* está conformado por unos; se nota $\mathbf{1} = \boldsymbol{j}$. Explícitamente

$$\mathbf{0'} = \left(0,\ 0, \cdots,\ 0\right), \quad \mathbf{1'} = \left(1,\ 1, \cdots,\ 1\right).$$

La *suma (o resta)* de dos vectores del mismo tamaño es el vector cuyas componentes son la suma (o resta) de las respectivas componentes. Esto es

$$X \pm Y = \begin{pmatrix} x_1 \pm y_1 \\ x_2 \pm y_2 \\ \vdots \\ x_n \pm y_n \end{pmatrix}. \tag{A1.3}$$

La suma entre vectores de $\mathbb{R}^n$ tiene las siguientes propiedades:

(i) Clausurativa: Si X y Y son vectores de $\mathbb{R}^n$ entonces $X + Y$ también está en $\mathbb{R}^n$.

(ii) Conmutativa: $X + Y = Y + X$, para todo par de vectores X y Y de $\mathbb{R}^n$.

(iii) Asociativa: $(X + Y) + Z = X + (Y + Z)$, para cualquier X, Y y Z vectores de $\mathbb{R}^n$.

(iv̈) Identidad: Existe el vector *nulo* $\mathbf{0}$ en $\mathbb{R}^n$, tal que $X + \mathbf{0} = \mathbf{0} + X = X$, para todo vector X de $\mathbb{R}^n$.

(v̈) Opuesto: para todo vector X de $\mathbb{R}^n$ existe el vector opuesto $-X$ en $\mathbb{R}^n$ tal que $X + (-X) = (-X) + X = \mathbf{0}$

La multiplicación de un número (escalar) por un vector, es el vector cuyas componentes se conforman por el producto entre cada componente del vector y el número real; se conoce como *la multiplicación por un escalar*. Sea

λ un número (escalar) y X un vector, entonces

$$\lambda X = \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{pmatrix}. \tag{A1.4}$$

La multiplicación por un número (escalar) tiene las siguientes propiedades:

1. Si X es un elemento de $\mathbb{R}^n$ y λ es un número, entonces λX está en $\mathbb{R}^n$.

2. $\lambda(\mu X) = (\lambda\mu)X$, para λ y μ números cualesquiera.

3. $\lambda(X + Y) = \lambda X + \lambda Y$ y $(\lambda + \mu)X = \lambda X + \mu X$.

Al conjunto $\mathbb{R}^n$ por satisfacer las propieades (i) a (v̈) y (1) a (3) se le llama un *espacio vectorial* y a sus elementos vectores. Este concepto se extiende a cualquier conjunto $\mathcal{V}$ y cualquier conjunto de números $\mathbb{K}$ (escalares). El conjunto de los números reales $\mathbb{R}$ es un espacio vectorial; el cual coincide con los escalares.

Observación:

- De las propiedades anteriores se sigue que:

  $$1 \cdot X = X \; ; 0 \cdot X = \mathbf{0}; \; (-1)X = -X; -(X + Y) = -X - Y; \\ (\lambda - \mu)X = \lambda X - \mu X,$$

  entre otras.

- Un subconjunto $\mathbb{A}$ de $\mathbb{R}^n$ es un *subespacio vectorial* si a su vez es un espacio vectorial; de otra forma, si para cualquier X y Y de $\mathbb{A}$ y λ un escalar, se tiene que $(X + Y)$ y (λX) están en $\mathbb{A}$.

  De esta forma, por ejemplo, la proyección de puntos de $\mathbb{R}^n$ sobre el plano $X_1 \times X_2$, que corresponde a los puntos cuyas coordenadas son de la forma $(x_1, x_2, 0, \cdots, 0)$, es un subespacio de $\mathbb{R}^n$; a veces se confunde con $\mathbb{R}^2$ pero esto no es del todo correcto, otra cosa es que tengan una estructura vectorial isomorfa.

La expresión (A1.4) se puede generalizar a un número finito de escalares y vectores, de esta forma: sean $\lambda_1,\ \lambda_2, \cdots,\ \lambda_k$ escalares y $X_1,\ X_2, \cdots,\ X_k$ vectores, entonces el vector

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots \lambda_k X_k, \tag{A1.5}$$

es una *combinación lineal* de los vectores $X_1,\ X_2, \cdots,\ X_k$.

Un conjunto de vectores es *linealmente independiente* (LI) si la combinación lineal (A1.5) de vectores no nulos, es igual al vector nulo únicamente cuando todos los escalares son cero; es decir,

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots \lambda_k X_k = \mathbf{0},\ \text{si y solo si},\ \lambda_1 = \lambda_2 = \cdots = \lambda_k = 0. \tag{A1.6}$$

En caso contrario, los vectores son *linealmente dependientes* (LD). De otra forma, un conjunto de vectores es LD si alguno de estos vectores se puede expresar como una combinación lineal de los demás.

Dados dos vectores X y Y de $\mathbb{R}^n$, se define el *producto escalar*, *producto interior* o *producto punto*, notado por $\langle X, Y\rangle = X \cdot Y$, mediante

$$\langle X, Y\rangle = X \cdot Y = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{i=1}^{n} x_i y_i. \tag{A1.6}$$

La longitud de un vector X también se llama *norma*, y se nota por $\|X\|$. Para un vector X de $\mathbb{R}^n$ su norma es

$$\|X\| = \sqrt{\langle X, X\rangle} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}. \tag{A1.7}$$

Un vector cuya norma sea igual a 1, se denomina *unitario*. Cualquier vector X no nulo se puede transformar en un vector unitario al multiplicarlo por el inverso de su norma; así

$$\frac{X}{\|X\|},\ \text{ es un vector unitario.}$$

Se puede emplear el concepto de norma para obtener la distancia entre dos vectores. La distancia entre los vectores X y Y corresponde a la norma del vector diferencia, y se nota $d(X, Y)$. Ésta es

$$d(X, Y) = \|X - Y\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}. \tag{A1.8}$$

Nótese la siguiente relación entre los conceptos de norma, distancia y producto interior

$$d(X, Y) = \|X - Y\| = \sqrt{\langle X - Y, X - Y\rangle}. \tag{A1.9}$$

El ángulo θ, determinado por dos vectores no nulos X y Y de $\mathbb{R}^n$, se obtiene de la siguiente expresión

$$\cos\theta = \frac{\langle X, Y\rangle}{\|X\|\|Y\|} = \frac{X \cdot Y}{\|X\|\|Y\|}. \tag{A1.10}$$

Una manera alterna para definir el producto interior entre dos vectores es

$$\langle X, Y\rangle = \|X\|\|Y\|\cos\theta, \tag{A1.11}$$

de donde se puede afirmar fácilmente que dos vectores son *ortogonales* si y sólo si el producto interior entre ellos es cero. En (A1.11), si $X \cdot Y = 0$ entonces $\theta = \pi/2$, recíprocamente, si en (A1.10) $\theta = \pi/2$ se concluye que $X \cdot Y = 0$. Si los vectores son unitarios y ortogonales se llaman *ortonormales*.

Observación:

Una propiedad importante es la *desigualdad de Cauchy-Schwarz*, la cual se presenta en tres versiones equivalentes, ella establece que, para $X,\ Y \in \mathbb{R}^n$,

$$\begin{aligned} &(i)\ \ (\langle X, Y\rangle)^2 \le (\langle X, X\rangle)(\langle Y, Y\rangle), \\ &(ii)\ \ (X'Y)^2 \le (X'X)(Y'Y),\ \text{o} \\ &(iii)\ \ |\langle X.Y\rangle| \le \|X\|\|Y\|. \end{aligned} \tag{A1.12}$$

La *proyección ortogonal* de un vector X sobre un vector Y es el vector X_p, donde:

$$X_p = \frac{\langle X, Y\rangle}{\|Y\|^2} \cdot Y = k \cdot Y, \tag{A1.13}$$

con k un escalar igual a $\langle X, Y\rangle/\|Y\|^2$. La figura A.1, muestra la proyección de un vector $X = (x_1, x_2)$ sobre un vector $Y = (y_1, y_2)$ en $\mathbb{R}^2$.

Ejemplo A.1 La mayoría de aplicaciones de la estadística multivariada contempla vectores de componentes reales; es decir, de $\mathbb{R}^n$. Para facilitar la comprensión de los conceptos anteriores se desarrollan varios casos en $\mathbb{R}^2$; esto permite mostrar algunos conceptos geométricos contenidos en los vectores.

Sean los vectores

$$X = \begin{pmatrix} 1 \\ 4 \end{pmatrix}, \quad Y = \begin{pmatrix} 5 \\ 1 \end{pmatrix} \quad y \quad Z = \begin{pmatrix} 4 \\ -2 \end{pmatrix}$$

la suma entre X y Y es igual a

$$X + Y = \begin{pmatrix} 1 \\ 4 \end{pmatrix} + \begin{pmatrix} 5 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 5 \end{pmatrix}.$$

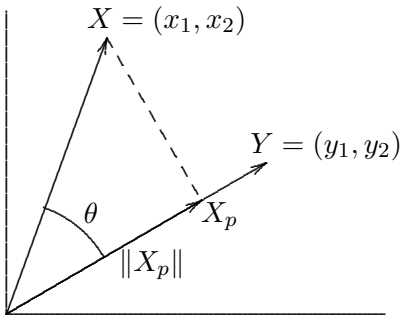


Figura A.1 *Proyección ortogonal.*

La suma de los tres vectores es

$$X + Y + Z = \begin{pmatrix} 1 \\ 4 \end{pmatrix} + \begin{pmatrix} 5 \\ 1 \end{pmatrix} + \begin{pmatrix} 4 \\ -2 \end{pmatrix} = \begin{pmatrix} 10 \\ 3 \end{pmatrix}.$$

La combinación lineal

$$X + Y - 2Z = \begin{pmatrix} 1 \\ 4 \end{pmatrix} + \begin{pmatrix} 5 \\ 1 \end{pmatrix} - 2\begin{pmatrix} 4 \\ -2 \end{pmatrix} = \begin{pmatrix} -2 \\ 9 \end{pmatrix}.$$

Los vectores X y Y son linealmente independientes, pues la ecuación

$$\lambda_1 X + \lambda_2 Y = 0,$$

o equivalentemente

$$\begin{aligned} \lambda_1 + 5\lambda_2 &= 0 \\ 4\lambda_1 + \lambda_2 &= 0, \end{aligned}$$

tiene solución única $\lambda_1 = \lambda_2 = 0$. Los tres vectores no son linealmente independientes[1], el sistema $\lambda_1 X + \lambda_2 Y + \lambda_3 Z = 0$, que corresponde a

$$\begin{aligned} \lambda_1 + 5\lambda_2 + 4\lambda_3 &= 0 \\ 4\lambda_1 + \lambda_2 - 2\lambda_3 &= 0, \end{aligned}$$

tiene soluciones diferentes de $(0, 0, 0)$.

La longitud o norma de los vectores X, Y y Z es, respectivamente

$$\begin{aligned} &\|X\| = \sqrt{1^2 + 4^2} = \sqrt{17},\ \|Y\| = \sqrt{5^2 + 1^2} = \sqrt{26} \text{ y} \\ &\|Z\| = \sqrt{4^2 + (-2)^2} = 2\sqrt{5}. \end{aligned}$$

[1]En espacios de dimensión k, conjuntos con más de k vectores son LD.

La distancia entre los vectores es, respectivamente

$$d(X,Y) = \|X - Y\| = \sqrt{(1-5)^2 + (4-1)^2} = 5.$$
$$d(X,Z) = \|X - Z\| = \sqrt{(1-4)^2 + (4+2)^2} = 3\sqrt{5}.$$
$$d(Y,Z) = \|Y - Z\| = \sqrt{(5-4)^2 + (1+2)^2} = \sqrt{10}.$$

Los ángulos conformados entre los vectores se obtienen, para cada par, de esta manera:

- $\cos\theta_{XY} = \dfrac{X \cdot Y}{\|X\|\|Y\|} = \dfrac{(1\times 5) + (4\times 1)}{(\sqrt{17})(\sqrt{26})} = 0.42808$, así $\theta_{XY} \approx 65^0$.
- $\cos\theta_{XZ} = \dfrac{X \cdot Z}{\|X\|\|Z\|} = \dfrac{(1\times 4) + (4\times (-2))}{(\sqrt{17})(2\sqrt{5})} = -0.21693$, así $\theta_{XZ} \approx 103^0$.
- $\cos\theta_{YZ} = \dfrac{Y \cdot Z}{\|Y\|\|Z\|} = \dfrac{(5\times 4) + (1\times (-2))}{(\sqrt{26})(2\sqrt{5})} = 0.78935$, así $\theta_{YZ} \approx 38^0$.

En la figura A.2 se ilustran algunos de los procedimientos anteriores. Se puede apreciar que la suma $(X + Y)$ corresponde al vector dispuesto sobre la diagonal principal del paralelogramo determinado por los vectores X y Y; para más de dos vectores la suma se hace similarmente, aplicando la propiedad asociativa, así , $X + Y + Z$ se ubica en la diagonal principal del paralelogramo determinado por los vectores $(X+Y)$ y Z, es decir, se aplica la propiedad asociativa $X + Y + Z = (X + Y) + Z$. La diferencia $(X - Y)$ es el vector trazado sobre la diagonal secundaria del mismo paralelogramo. La multiplicación por un escalar "alarga" o "contrae" el vector de acuerdo con la magnitud del escalar y en la dirección determinada por su signo. ☑

A.2 Matrices

▶ Definiciones

Una $\boldsymbol{A}$ matriz de *tamaño* $(n \times p)$ es un arreglo rectangular de números[2] dispuestos en n-filas y en p-columnas; se escribe de la siguiente forma

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}. \tag{A2.1}$$

Es usual la notación de una matriz $\boldsymbol{A}$ en términos de su elemento genérico a_{ij}; es decir, $\boldsymbol{A} = (a_{ij})$, $i = 1, \cdots, n$ y $j = 1, \cdots, p$.

[2]Los números pueden ser reales $\mathbb{R}$ o complejos $\mathbb{C}$.

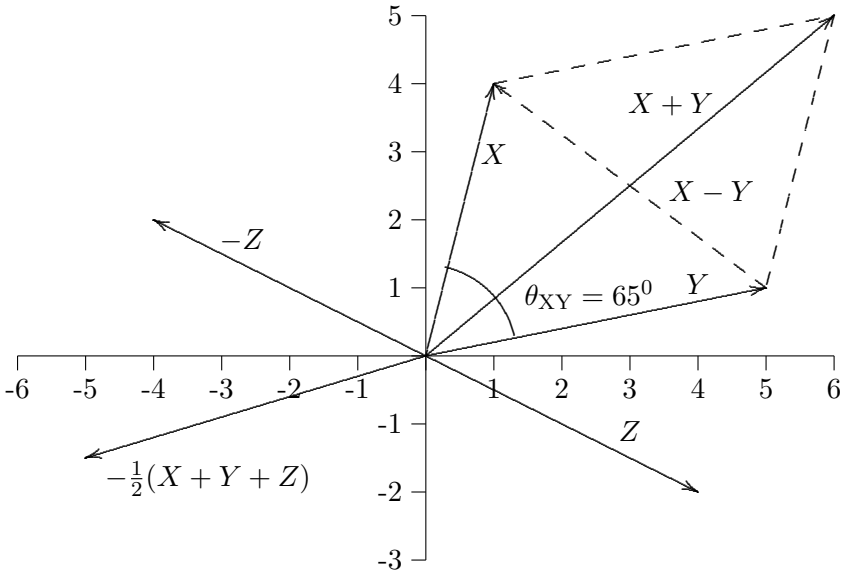


Figura A.2 *Operaciones entre vectores.*

Una matriz es también un arreglo de n vectores fila de tamaño $(1 \times p)$, o de p vectores columna de tamaño $(n \times 1)$; cuando se aborde una matriz en esta forma, se hará referencia al espacio fila o al espacio columna, respectivamente.

Se nota la i-ésima fila de la matriz $\boldsymbol{A}$ por $a_{(i)}$ y su j-ésima columna por $a^{(j)}$. Con esta notación la matriz $\boldsymbol{A}$ de tamaño $(n \times p)$ se puede escribir como

$$\boldsymbol{A} = \begin{pmatrix} a_{(1)} \\ a_{(2)} \\ \vdots \\ a_{(n)} \end{pmatrix} = \begin{pmatrix} a^{(1)} & a^{(2)} & \cdots & a^{(p)} \end{pmatrix}. \tag{A2.1a}$$

Recíprocamente, se pueden considerar los vectores como un caso especial de las matrices, donde n o p son iguales a uno.

Dos matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ son *iguales* si tienen el mismo tamaño y los elementos de posiciones correspondientes son iguales. De tal forma, las matrices $\boldsymbol{A} = (a_{ij})$ y $\boldsymbol{B} = (b_{ij})$ son iguales ($\boldsymbol{A} = \boldsymbol{B}$), si y sólo si, $a_{ij} = b_{ij}$ para todo i y j.

Matrices para las cuales el número de filas es igual al número de columnas se llaman *matrices cuadradas*. Si $\boldsymbol{A}$ es una matriz cuadrada de tamaño

$(p \times p)$, entonces se dice que es de tamaño p . Los elementos a_{ii} de una matriz cuadrada conforman la *diagonal principal*.

La *transpuesta* de una matriz, es una matriz cuyas filas son las columnas de la original, y en consecuencia, sus columnas son las filas de la original. De otra manera, la *transpuesta* de una matriz $\boldsymbol{A} = (a_{ij})$ de tamaño $(n \times p)$ es una matriz $\boldsymbol{A'} = (a_{ji})$ de tamaño $(p \times n)$ (se nota también por $\boldsymbol{A}^T$) de tamaño $(p \times n)$.

Existen algunos vectores y matrices especiales, que aparecen frecuentemente en el trabajo estadístico multivariado.

• La *matriz nula o cero* tiene todos sus elementos iguales cero;

$$\mathbf{0} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}. \tag{A2.2}$$

• Un vector o una matriz constituidos por unos se denotan, respectivamente, por

$$\mathbf{1} = \boldsymbol{j} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{y} \quad \boldsymbol{J} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}.$$

• Una matriz cuadrada cuyos elementos fuera de la diagonal son todos cero se denomina *matriz diagonal*; es decir, una matriz $\boldsymbol{D} = (d_{ij})$ es diagonal si $d_{ij} = 0$ para $i \neq j$. Se escribe $\text{Diag}(\boldsymbol{D}) = \text{Diag}(d_{11}, \cdots, d_{pp}) = (d_{ii})$. Explícitamente:

$$\text{Diag}(\boldsymbol{D}) = (d_{ii}) = \begin{pmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{pp} \end{pmatrix}. \tag{A2.3}$$

• La transformación *diagonal* asigna a una matriz cuadrada A la matriz diagonal con elementos a_{ii} sobre la diagonal principal; se nota $\text{diag}(A) = (a_{ii})$. Si D es una matriz diagonal, entonces $\text{diag}(D) = D$.

• Una matriz *simétrica*, es una matriz cuadrada tal que su transpuesta es igual a la matriz original $(\boldsymbol{A'} = \boldsymbol{A})$; es decir, $\boldsymbol{A}$ es simétrica si $a_{ij} = a_{ji}$ para $i, j = 1, \cdots, p$.

• Una matriz cuadrada es *triangular superior* si todos los elementos por debajo de la diagonal son cero; es decir, si $a_{ij} = 0$ para $i > j$. Así:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1p} \\ 0 & a_{22} & a_{23} & \cdots & a_{2p} \\ 0 & 0 & a_{33} & \cdots & a_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{pp} \end{pmatrix}. \tag{A2.4}$$

• Recíprocamente, una matriz cuadrada es *triangular inferior* si todos los elementos por encima de la diagonal son cero; es decir, si $a_{ij} = 0$ para $i < j$. Explícitamente

$$\begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & a_{p3} & \cdots & a_{pp} \end{pmatrix}. \tag{A2.5}$$

• La matriz *identidad* es una matriz diagonal con todos los elementos de la diagonal principal iguales a uno. Se nota y escribe así

$$\boldsymbol{I}_p = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}. \tag{A2.6}$$

Esta matriz, a demás de ser una matriz diagonal, es tanto triangular superior, como triangular inferior.

► Operaciones con matrices

⊞ *Suma*

Sean $\boldsymbol{A}$ y $\boldsymbol{B}$ matrices de tamaño [3]. $(n \times p)$. Se define la *suma* (o la resta) entre $\boldsymbol{A}$ y $\boldsymbol{B}$ por

$$\boldsymbol{A} \pm \boldsymbol{B} = (a_{ij}) \pm (b_{ij}) = (a_{ij} \pm b_{ij}) = \begin{pmatrix} a_{11} \pm b_{11} & a_{12} \pm b_{12} & \cdots & a_{1p} \pm b_{1p} \\ a_{21} \pm b_{21} & a_{22} \pm b_{22} & \cdots & a_{2p} \pm b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} \pm b_{n1} & a_{n2} \pm b_{n2} & \cdots & a_{np} \pm b_{np} \end{pmatrix}. \tag{A2.7}$$

Como en los vectores, las matrices satisfacen las siguientes propiedades respecto a la suma:

[3] Las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ son conformables para la suma (o la resta) solo si las matrices tienen el mismo tamaño

i) *Conmutativa*: $\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$, para todo par de matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ conformables para la suma.

ii) *Asociativa*: $(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C})$, para cualquier $\boldsymbol{A}$, $\boldsymbol{B}$ y $\boldsymbol{C}$ matrices conformables para la suma.

iii) *Identidad*: existe la matriz *nula* $\mathbf{0}$, tal que $\boldsymbol{A} + \mathbf{0} = \mathbf{0} + \boldsymbol{A}$, para toda matriz $\boldsymbol{A}$.

iv) *Opuesta*: para toda matriz $\boldsymbol{A}$ existe la matriz opuesta aditiva, notada por $-\boldsymbol{A}$, tal que $\boldsymbol{A} + (-\boldsymbol{A}) = (-\boldsymbol{A}) + \boldsymbol{A} = \mathbf{0}$.

La demostración de cada una de estas propiedades se hace teniendo en cuenta que la suma entre matrices se define en términos de la suma entre sus respectivas entradas, las cuales son números reales, y que éstos cumplen con las propiedades enunciadas para el caso matricial.

⊞ *Multiplicación por un escalar*

La multiplicación de una matriz $\boldsymbol{A}$ por un escalar λ es igual a la matriz que resulta de multiplicar cada elemento de $\boldsymbol{A}$ por λ. En general se tiene que:

$$\lambda \boldsymbol{A} = (\lambda a_{ij}) = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \cdots & \lambda a_{1p} \\ \lambda_{21} & \lambda a_{22} & \cdots & \lambda a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{n1} & \lambda a_{n2} & \cdots & \lambda a_{np} \end{pmatrix}. \tag{A2.8}$$

A continuación se describen las propiedades básicas por un escalar. Sean $\boldsymbol{A}$ y $\boldsymbol{B}$ matrices de tamaño $n \times p$, y, λ_1 y λ_2 un escalares.

i) $\lambda \boldsymbol{A}$ es una matriz $n \times p$.

ii) $(\lambda_1 + \lambda_2)\boldsymbol{A} = \lambda_1 \boldsymbol{A} + \lambda_2 \boldsymbol{A}$.

iii) $\lambda_1(\boldsymbol{A} + \boldsymbol{B}) = \lambda_1 \boldsymbol{A} + \lambda_1 \boldsymbol{B}$.

iv) $\lambda_1(\lambda_2 \boldsymbol{A}) = (\lambda_1 \lambda_2)\boldsymbol{A}$.

v) $1\boldsymbol{A} = \boldsymbol{A}$.

Observación:

De acuerdo con las propiedades anteriores para la suma entre matrices y la multiplicación por un escalar, se tiene que el conjunto $\mathbb{M}$ de las

matrices de tamaño $(n \times p)$ es un espacio vectorial sobre el cojunto de escalares $\mathbb{R}$.

⊞ *Producto*

Si la matriz $\boldsymbol{A}$ es de tamaño $(n \times k)$ y la matriz $\boldsymbol{B}$ es de tamaño $(k \times p)$; es decir, la matriz $\boldsymbol{A}$ tiene un número de columnas igual al número de filas de la matriz $\boldsymbol{B}$, entonces se dice que son conformables respecto el *producto* entre matrices. El elemento genérico c_{ij}, correspondiente al producto entre la matriz $\boldsymbol{A}$ y la matriz $\boldsymbol{B}$ se esquematiza enseguida

$$\begin{aligned}
\boldsymbol{A}B &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ldots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nk} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1j} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2j} & \cdots & b_{2p} \\ \vdots & \vdots & \ldots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{kj} & \cdots & b_{kp} \end{pmatrix} \\
&= \begin{pmatrix} \vdots \\ \boxed{a_{i1} \quad a_{i2} \cdots \quad a_{ik}} \\ \vdots \end{pmatrix}_{nk} \begin{pmatrix} \cdots & \boxed{\begin{matrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{kj} \end{matrix}} & \cdots \end{pmatrix}_{kp} \\
&= \begin{pmatrix} \vdots \\ \cdots \quad \boxed{c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots a_{ik}b_{kj}} \quad \cdots \\ \vdots \end{pmatrix}_{np} . \qquad \text{(A2.9)}
\end{aligned}$$

En la última parte de (A2.9) se observa el producto interior entre el i-ésimo vector fila de $\boldsymbol{A}$ y el transpuesto del j-ésimo vector columna de $\boldsymbol{B}$. El producto entre estas dos matrices se presenta en una forma más condensada en la siguiente expresión

$$\boldsymbol{AB} = (a_{(i)})(b^{(j)}) = (c_{ij}) = \left(\sum_{h=1}^{k} a_{ih} b_{hj} \right), \quad i = 1, \cdots, n, \quad j = 1, \cdots, p. \qquad \text{(A2.10)}$$

Un caso especial del producto entre matrices cuadradas es la multiplicación de una matriz por si misma, este producto se nota $\boldsymbol{AA} = \boldsymbol{A}^2$. De manera más general

$$\underbrace{\boldsymbol{AA} \cdots \boldsymbol{A}}_{k-\text{veces}} = \boldsymbol{A}^k, \text{ con } k \text{ número entero no negativo.}$$

Para $k = 0$, $\boldsymbol{A}^0 = \boldsymbol{I}$. La potenciación se extenderá a todos los enteros, más adelante cuando se defina la matriz inversa ($\boldsymbol{A}^{-1}$).

Una matriz $\boldsymbol{A}$ es *idempotente* si $\boldsymbol{A}^2 = \boldsymbol{A}$. Se demuestra que si una matriz $\boldsymbol{A}$ es idempotente, entonces la matriz $(\boldsymbol{I} - \boldsymbol{A})$ también es idempotente.

El producto entre matrices cumple las propiedades que a continuación se describen, se asume que los productos y sumas de matrices son conformables,

$$\begin{aligned} \textit{Asociativa}: (\boldsymbol{A}B)\boldsymbol{C} &= \boldsymbol{A}(\boldsymbol{B}\boldsymbol{C}). \\ \textit{Distributiva a derecha}: \boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) &= \boldsymbol{A}B + \boldsymbol{A}\boldsymbol{C}. \\ \textit{Distributiva a izquierda}: (\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C} &= \boldsymbol{A}\boldsymbol{C} + \boldsymbol{B}\boldsymbol{C}. \\ \textit{Identidad}: \boldsymbol{I}\boldsymbol{A} &= \boldsymbol{A}\boldsymbol{I} = \boldsymbol{A}. \end{aligned} \tag{A2.11}$$

La transposición de una matriz tiene, entre otras, las siguientes propiedades

$$\begin{aligned} (\boldsymbol{A}')' &= \boldsymbol{A} \\ (\lambda\boldsymbol{A} + \mu\boldsymbol{B})' &= \lambda\boldsymbol{A}' + \mu\boldsymbol{B}', \quad \lambda \text{ y } \mu \text{ escalares} \\ (\boldsymbol{A}B)' &= \boldsymbol{B}'\boldsymbol{A}'. \end{aligned} \tag{A2.12}$$

A continuación se muestran algunos productos especiales.

$$\begin{aligned} \boldsymbol{a}\boldsymbol{b}' &= \begin{pmatrix} a_1 \\ a_2 \\ \vdots a_n \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \cdots & b_p \end{pmatrix} \\ &= \begin{pmatrix} a_1 b1 & a_1 b_2 & \cdots & a_1 b_p \\ a_2 b1 & a_2 b_2 & \cdots & a_2 b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_n b1 & a_n b_2 & \cdots & a_n b_p \end{pmatrix}. \end{aligned}$$

Un caso especial del producto anterior es:

$$\boldsymbol{j}\boldsymbol{j}' = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \boldsymbol{J},$$

donde $\boldsymbol{j}$ y $\boldsymbol{J}$ se definen como se hizo arriba. Otros productos que involucran a $\boldsymbol{j}$ son:

$$\boldsymbol{a}'\boldsymbol{j} = \sum_{i=1}^{n} a_i$$

$$\boldsymbol{A}\boldsymbol{j} = \begin{pmatrix} \sum_{j=1}^{p} a_{1j} \\ \sum_{j=1}^{p} a_{2j} \\ \vdots \\ \sum_{j=1}^{p} a_{nj} \end{pmatrix}.$$

Ejemplo A.2 Sean las matrices

$$\boldsymbol{A} = \begin{pmatrix} 2 & -3 & 1 & 0 \\ 1 & 5 & 4 & 5 \\ 3 & 0 & -1 & 6 \end{pmatrix} \text{ y } \boldsymbol{B} = \begin{pmatrix} 4 & 7 & -1 & 2 \\ 0 & 6 & 3 & 1 \\ -2 & 1 & 1 & 4 \end{pmatrix}.$$

Nótese que las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ son conformables para la suma y para los siguientes productos $\boldsymbol{A}'\boldsymbol{B}$ y $\boldsymbol{A}\boldsymbol{B}'$, entre otros ($\boldsymbol{A}\boldsymbol{B}$ no es conformable).

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} 2+4 & -3+7 & 1+(-1) & 0+2 \\ 1+0 & 5+6 & 4+3 & 5+1 \\ 3+(-2) & 0+1 & -1+1 & 6+4 \end{pmatrix} = \begin{pmatrix} 6 & 4 & 0 & 2 \\ 1 & 11 & 7 & 6 \\ 1 & 1 & 0 & 10 \end{pmatrix}.$$

El producto entre $\boldsymbol{A}'$ y $\boldsymbol{B}$ es

$$\boldsymbol{A}'\boldsymbol{B} = \begin{pmatrix} 2 & 1 & 3 \\ -3 & 5 & 0 \\ 1 & 4 & -1 \\ 0 & 5 & 6 \end{pmatrix} \begin{pmatrix} 4 & 7 & -1 & 2 \\ 0 & 6 & 3 & 1 \\ -2 & 1 & 1 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 23 & 4 & 17 \\ -12 & 9 & 18 & -1 \\ 6 & 30 & 10 & 2 \\ -12 & 36 & 21 & 29 \end{pmatrix}.$$

El producto

$$\boldsymbol{A}\boldsymbol{A}' = \begin{pmatrix} 2 & -3 & 1 & 0 \\ 1 & 5 & 4 & 5 \\ 3 & 0 & -1 & 6 \end{pmatrix} \begin{pmatrix} 2 & 1 & 3 \\ -3 & 5 & 0 \\ 1 & 4 & -1 \\ 0 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 14 & -9 & 5 \\ -9 & 67 & 29 \\ 5 & 29 & 46 \end{pmatrix}. \text{ ☑}$$

Nótese que la matriz $(\boldsymbol{A}\boldsymbol{A}')$ es simétrica, en general, las matrices $(\boldsymbol{A}\boldsymbol{A}')$ y $(\boldsymbol{A}'\boldsymbol{A})$ son simétricas, puesto que $(\boldsymbol{A}\boldsymbol{A}')' = (\boldsymbol{A}')'\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{A}'$, similarmente se muestra que $(\boldsymbol{A}'\boldsymbol{A})$ y $\frac{1}{2}(\boldsymbol{A}' + \boldsymbol{A})$ son matrices simétricas.

⊞ *Traza*

La *traza* de una matriz cuadrada $\boldsymbol{A}$ de tamaño $(p \times p)$ es la suma de los elementos de su diagonal principal. Así,

$$\text{tra}(\boldsymbol{A}) = a_{11} + a_{22} + \cdots + a_{pp} = \sum_{i=1}^{p} a_{ii}. \qquad \text{(A2.13)}$$

Algunas propiedades de la función traza son las siguientes:

i) $\text{tra}(\boldsymbol{A}') = \text{tra}(\boldsymbol{A})$, $\boldsymbol{A}$ matriz cuadrada.

ii) $\text{tra}(\lambda\boldsymbol{A}) = \lambda\,\text{tra}(\boldsymbol{A})$, para λ un escalar y $\boldsymbol{A}$ una matriz cuadrada.

iii) $\text{tra}(\boldsymbol{A}+\boldsymbol{B}) = \text{tra}(\boldsymbol{A}) + \text{tra}(\boldsymbol{B})$, $\boldsymbol{A}$ y $\boldsymbol{B}$ matrices cuadradas y conformables para la suma.

iv̈) $\text{tra}(\boldsymbol{AB}) = \text{tra}(\boldsymbol{BA})$.

⊞ *Determinante*

El *determinante* de una matriz cuadrada , es un número importante para el análisis y aplicación de algunas técnicas multivariadas. Aunque existe actualmente un buen número de procedimientos de cómputo, es inevitable presentar en un plano intuitivo su definición formal.

Dada una matriz cuadrada $\boldsymbol{A}$ de tamaño p, el determinante de $\boldsymbol{A}$, notado por $|\boldsymbol{A}|$ o $\det(\boldsymbol{A})$, está definido por

$$|\boldsymbol{A}| = \sum(-1)^{f(j_1,j_2,\cdots,j_p)} \prod_{i=1}^{p} a_{ij_i}, \tag{A2.14}$$

la suma es sobre todas las permutaciones $(j_1,,\cdots,j_p)$ de los enteros de 1 a p y $f(j_1,\cdots,j_p)$ es el número de transposiciones requeridas para ir de $(1,\cdots,p)$ a $(j_1,\cdots,j_p)$.

Una transposición consiste en el intercambio de dos números. Nótese que al escribir j_i en (A2.14), se señala que el producto toma un único elemento por fila y columna.

Se demuestra que el intercambio es siempre un número par o un número impar. De manera que $(-1)^{f(j_1,j_2,\cdots,j_p)}$ es 1 o -1, respectivamente.

Para una matriz de tamaño (2×2), $\boldsymbol{A} = \left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right)$, las posibles permutaciones de los enteros 1 y 2 son $(1,2)$ y $(2,1)$, los posibles productos, así definidos, en la matriz $\boldsymbol{A}$ son: $a_{11}a_{22}$, $a_{12}a_{21}$. En la primera se deben hacer 0 permutaciones del arreglo $(1,2)$ para llegar al arreglo $(1,2)$, mientras que en la segunda se debe hacer una permutación para transformar $(1,2)$ en $(2,1)$; entonces los signos de los productos son $+$ y $-$, respectivamente. En consecuencia, el determinante de la matriz $\boldsymbol{A}$ de acuerdo con la expresión (A2.14) es $|\boldsymbol{A}| = a_{11}a_{22} - a_{12}a_{21}$.

Para una matriz de tamaño (3×3), $\boldsymbol{A} = \left(\begin{smallmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{smallmatrix}\right)$, los productos de las entradas de $\boldsymbol{A}$, en la forma considerada en (A2.14), con la permutación

de los segundos subíndices (j_i), el signo y el producto con signo se ilustra enseguida

Producto	Permutación	Signo	Producto con signo
$a_{11}a_{22}a_{33}$	$(1,2,3)$	$+$	$a_{11}a_{22}a_{33}$
$a_{11}a_{23}a_{32}$	$(1,3,2)$	$-$	$-a_{11}a_{23}a_{32}$
$a_{12}a_{21}a_{33}$	$(2,1,3)$	$-$	$-a_{12}a_{21}a_{33}$
$a_{12}a_{23}a_{31}$	$(2,3,1)$	$+$	$a_{12}a_{23}a_{31}$
$a_{13}a_{21}a_{32}$	$(3,1,2)$	$+$	$a_{13}a_{21}a_{32}$
$a_{13}a_{22}a_{31}$	$(3,2,1)$	$-$	$-a_{13}a_{22}a_{31}$

El determinante de esta matriz es la suma de estos productos (A2.14).

A continuación se ilustra el cálculo del determinante para matrices de tamaño 2 y 3. El determinante se calcula como la suma de los productos de los elementos de las diagonales principales menos los productos de los elementos de las diagonales secundarias. Para matrices de tamaño (3×3), se repiten las dos primeras filas o las dos primeras columnas, y se calculan sobre la matriz así conformada los productos de los elementos en cada diagonal.

$$\det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \overbrace{a_{11}\cdot a_{22}}^{\text{Dg. ppal.}} - \underbrace{a_{12}\cdot a_{21}}_{\text{Dg. sec.}}$$

$$\det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \overbrace{a_{11}\cdot a_{22}\cdot a_{33} + a_{12}\cdot a_{23}\cdot a_{31} + a_{13}\cdot a_{21}\cdot a_{32}}^{\text{Dg. ppal.}}$$

$$\underbrace{-a_{11}\cdot a_{23}\cdot a_{32} - a_{12}\cdot a_{21}\cdot a_{33} - a_{13}\cdot a_{22}\cdot a_{31}}_{\text{Dg. sec.}}$$

Las siguientes son algunas definiciones conducentes al cálculo del determinante en matrices de tamaño superior a 3.

El *menor* del elemento a_{ij} de una matriz $\boldsymbol{A}$ de tamaño $(p \times p)$, está definido como el determinante de la matriz que se obtiene al suprimir la fila i y la columna j de la matriz $\boldsymbol{A}$. Esta cantidad se nota por A_{ij}.

El *cofactor* del elemento a_{ij} de una matriz $\boldsymbol{A}$ de tamaño $(p \times p)$, es $(-1)^{i+j}A_{ij}$. Se nota por c_{ij}

Para la matriz anterior de tamaño (3×3) , el menor y el cofactor de a_{12} son respectivamente,

$$A_{12} = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} \quad \text{y} \quad c_{12} = (-1)^{(1+2)}A_{12} = -(a_{21}a_{33} - a_{23}a_{31}).$$

Finalmente, el determinante para una matriz $\boldsymbol{A}$ de tamaño $(p \times p)$ es el siguiente

$$|\boldsymbol{A}| = \sum_{j=1}^{p} (-1)^{i+j} a_{ij} A_{ij} = \sum_{j=1}^{p} (-1)^{i+j} a_{jk} A_{jk}. \tag{A2.15}$$

De acuerdo con la definición (A2.15) la expansión en cofactores se puede hacer sobre cualquier columna o fila de la matriz $\boldsymbol{A}$.

Algunas propiedades del determinante se consignan enseguida:

i) Si λ es un escalar, entonces $|\lambda \boldsymbol{A}| = \lambda^p |\boldsymbol{A}|$.

ii) $|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$.

iii) $|\boldsymbol{A}'| = |\boldsymbol{A}|$.

iv̈) Si los elementos de una fila (o columna) de una matriz $\boldsymbol{A}$ son todos cero, entonces $|\boldsymbol{A}| = 0$.

v̈) Si una fila (o columna) de una $\boldsymbol{A}$ es múltiplo de otra (LD), entonces $|\boldsymbol{A}| = 0$.

⊞ *Inversa*

La inversión de una matriz es análoga al proceso aritmético de división. Es decir, el proceso con el cual, dado un escalar $\lambda \neq 0$ se busca otro, notado λ^{-1}, tal que $\lambda \times \lambda^{-1} = 1$. Similarmente dada una matriz cuadrada $\boldsymbol{A} \neq \boldsymbol{0}$, entonces su inversa notada $\boldsymbol{A}^{-1}$, es tal que $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$; con $\boldsymbol{I}$ la matriz identidad.

La inversa de matrices está definida solo para matrices cuadradas, aunque hay matrices cuadradas que no tienen inversa[4]. Cuando la inversa de $\boldsymbol{A}$ existe, es tanto a la izquierda como a la derecha, así $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$. Cuando una matriz tiene inversa se dice que es *invertible*.

La inversa de una matriz, con determinante no nulo, se calcula a través de la siguiente expresión

$$\boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|} \operatorname{Adj}(\boldsymbol{A}), \tag{A2.16}$$

donde $\operatorname{Adj}(\boldsymbol{A})$ es la *adjunta* de $\boldsymbol{A}$ y corresponde a la transpuesta de la matriz de cofactores; es decir, la transpuesta de la matriz que se obtiene al reemplazar las entradas a_{ij} de $\boldsymbol{A}$ por los respectivos cofactores A_{ij}.

[4]Una extensión es la inversa generalizada, Searle (1982).

La inversa de una matriz de tamaño (2×2) se calcula mediante:

$$\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$

Se destacan las siguientes propiedades respecto a la matriz inversa.

i) Si existe la inversa de una matriz $\boldsymbol{A}$, ésta es única.

ii) Para que una matriz $\boldsymbol{A}$ tenga inversa, es condición necesaria y suficiente que su determinante sea diferente de cero. Una matriz invertible se denomina *no singular* y en caso contrario *singular*.

iii) Para cualquier escalar $\lambda \neq 0$, $(\lambda \boldsymbol{A})^{-1} = \lambda^{-1}\boldsymbol{A}^{-1}$.

iv) $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

v) $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.

vi) $\boldsymbol{A}^{-n} = (\boldsymbol{A}^{-1})^n$, para $n \geq 0$.

⊞ *Rango*

El rango de una matriz $\boldsymbol{A}$ de tamaño $(n \times p)$ es el número máximo de filas (o columnas) linealmente independientes. Si el rango de $\boldsymbol{A}$ es r se nota $r(\boldsymbol{A}) = r$.

Las siguientes propiedades son útiles para la sustentación de algunas metodologías multivariadas.

i) El rango fila de una matriz $\boldsymbol{A}$ es igual a su rango columna.

ii) $0 \leq r(\boldsymbol{A}) \leq \min\{n, p\}$.

iii) $r(\boldsymbol{A}') = r(\boldsymbol{A})$.

iv) $r(\boldsymbol{A} + \boldsymbol{B}) \leq r(\boldsymbol{A}) + r(\boldsymbol{B})$.

Las siguientes proposiciones son equivalentes.

a) $\boldsymbol{A}$ es invertible.

b) $|\boldsymbol{A}| \neq 0$.

c) El sistema $\boldsymbol{A}X = \boldsymbol{0}$ tiene únicamente la solución trivial $X = \boldsymbol{0}$.

d) El sistema $\boldsymbol{A}X = \boldsymbol{b}$ es consistente para cualquier vector $\boldsymbol{b}$ de tamaño $(p \times 1)$.

e) Los vectores fila (o columna) de $\boldsymbol{A}$ son linealmente independientes.

Cuando una matriz $\boldsymbol{A}$ satisface alguna de las cinco propiedades anteriores, se dice que $\boldsymbol{A}$ es una matriz de *rango completo*.

La proposición anterior (d) se refiere a la solucion del sistema de ecuaciones

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p = b_2 \\ \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p = b_p. \end{cases} \tag{A2.17}$$

La matriz de coeficientes $\boldsymbol{A}$, junto con el vector columna $\boldsymbol{b}$ conforman la *matriz aumentada*; ésta es

$$(\boldsymbol{A} : \boldsymbol{b}) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2p} & b_2 \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots & \cdots \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} & b_p \end{pmatrix}. \tag{A2.18}$$

Las ecuaciones lineales anteriores son dependientes o independientes según sean dependientes o independientes las filas de la matriz aumentada $(\boldsymbol{A} : \boldsymbol{b})$. El sistema de ecuaciones (A2.18) tiene una solución, si y sólo si, la matriz de los coeficientes $\boldsymbol{A}$ y la matriz aumentada $(\boldsymbol{A} : \boldsymbol{b})$ tienen el mismo rango. La solución del sistema viene dada por

$$X = \boldsymbol{A}^{-1}\boldsymbol{b}, \tag{A2.19}$$

éste es el significado de *consistencia* de un sistema de ecuaciones.

Al sistema de ecuaciones descrito en la proposición (c), caso especial de la (d), se le conoce como sistema *homogéneo* de ecuaciones.

Un sistema homogéneo de ecuaciones $\boldsymbol{A}X = \boldsymbol{0}$ tiene soluciones no triviales $(X \neq \boldsymbol{0})$ si y sólo si

$$r(\boldsymbol{A}) < p \quad \text{o equivalentemente, si y sólo si,} \quad |\boldsymbol{A}| = 0, \tag{A2.20}$$

caso en cual la matriz $\boldsymbol{A}$ es singular.

⊞ *Matrices ortogonales*

En la sección (A1) se explica el concepto de ortogonalidad entre vectores. Tratando las matrices como un arreglo de vectores fila (o columna) se aplica este concepto sobre tales vectores para obtener las matrices *ortogonales*.

Una matriz $\boldsymbol{A}$ de tamaño $(p \times p)$ es una *matriz ortogonal* si sus columnas son vectores ortogonales y unitarios[5].

Formalmente la matriz $\boldsymbol{A}$ es ortogonal si y sólo si $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$; es decir, si $\boldsymbol{A}' = \boldsymbol{A}^{-1}$.

Las matrices ortogonales tienen, entre otras, las siguientes propiedades:

i) $|\boldsymbol{A}| = \pm 1$

ii) El producto de un número finito de matrices ortogonales es ortogonal.

iii) La inversa y en consecuencia la transpuesta de una matriz ortogonal es ortogonal.

iv) Dada la matriz $\boldsymbol{A}$ y la matriz *ortogonal* $\boldsymbol{P}$, entonces $|\boldsymbol{A}| = |\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}|$.

⊞ *Transformaciones lineales*

A continuación se presenta la noción de *transformación lineal* desde una visión matricial.

Sea $\boldsymbol{A}$ una matriz de tamaño $(n \times p)$ y sea X un vector de $\mathbb{R}^p$, la ecuación

$$Y = \boldsymbol{A}X. \tag{A2.21}$$

define una *transformación lineal* de $\mathbb{R}^p$ en $\mathbb{R}^n$; es decir, el vector X se transforma mediante la matriz $\boldsymbol{A}$ en el vector Y. En forma práctica, la transformación lineal "envía" un vector X del espacio $\mathbb{R}^p$ al vector Y del espacio $\mathbb{R}^n$.

El siguiente diagrama ilustra el concepto de transformación lineal

$$\begin{aligned} \mathbb{R}^p &\longrightarrow \mathbb{R}^n \\ X &\longrightarrow Y = \boldsymbol{A}X. \end{aligned} \tag{A2.22}$$

Este tipo de transformaciones también se llaman *lineales homogéneas*, pues transforman el vector nulo de $\mathbb{R}^n$ en el vector nulo de $\mathbb{R}^p$ y lineal por que preservan la operaciones de multiplicación por un escalar y suma de vectores en los respectivos espacios vectoriales. Es decir, $\mathbf{0} \in \mathbb{R}^p$, vector $(p \times 1)$, es

[5] Debería hablarse de ortonormalidad.

transformado por $\boldsymbol{A}$ en $\mathbf{0} \in \mathbb{R}^n$, vector $(n \times 1)$, y, para λ_1 y λ_2 escalares, X_1 y X_2 vectores de $\mathbb{R}^p$,

$$\boldsymbol{A}(\lambda_1 X_1 + \lambda_2 X_2) = \lambda_1 \boldsymbol{A} X_1 + \lambda_2 \boldsymbol{A} X_2, \tag{A2.23}$$

el cual es un vector de $\mathbb{R}^n$.

La transformación esquematizada en (A2.22) muestra la estrecha relación entre las tranformaciones de $\mathbb{R}^p$ en $\mathbb{R}^n$ y las matrices; es decir que a toda tranformación de $\mathbb{R}^p$ en $\mathbb{R}^n$ se le puede asociar una matriz $\boldsymbol{A}$ de tamaño $(n \times p)$; y, recíprocamente, toda matriz $\boldsymbol{A}$ de tamaño $(n \times p)$ induce una transformación de $\mathbb{R}^p$ en $\mathbb{R}^n$. Así, las transformaciones lineales en espacios finitos se pueden considerar a través de las respectivas matrices.

Una transformación lineal de un espacio en si mismo se llama un *operador lineal* en tal espacio.

Ejemplo A.3 La transformación $Y : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$, definida por $Y = \boldsymbol{A}X$, donde la matriz $\boldsymbol{A}$ está dada por

$$\boldsymbol{A} = \begin{pmatrix} \cos\theta & -\operatorname{sen}\theta \\ \operatorname{sen}\theta & \cos\theta \end{pmatrix}$$

es una transformación lineal. La transformación Y sobre un vector X corresponde a la *rotación* de $X = (x_1, x_2)$ un ángulo θ. Esta transformación lineal es también un *operador lineal* en $\mathbb{R}^2$. La figura A.3 muestra la tranformación , por la rotación Y, de un vector (x_1, x_2) en el vector (x_1', x_2').

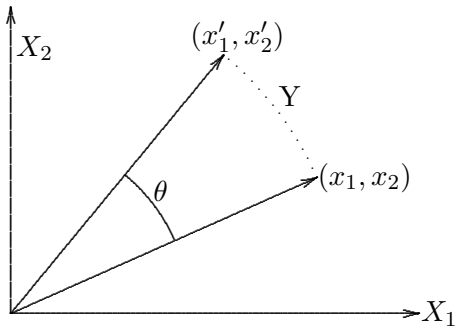


Figura A.3 *Transformación lineal por rotación.*

Observación:

> Una transformación lineal importante es aquella que asigna a cada vector X de $\mathbb{R}^n$ su *proyección ortogonal* en un subespacio V de $\mathbb{R}^n$. Una condición necesaria y suficiente para que una proyección sea ortogonal es que la matriz $\boldsymbol{A}$, asociada con la transformación lineal, sea simétrica e idempotente; es decir, que $Y = \boldsymbol{A}X$ es una proyección ortogonal si y sólo si $\boldsymbol{A'} = \boldsymbol{A}$ y $\boldsymbol{A}^2 = \boldsymbol{A}$.

⊞ *Valores y vectores propios*

Una de las transformaciones lineales de mayor interés en la estadística multivariada es aquella que "contrae" o "dilata" a un vector X, naturalmente X debe ser diferente del vector nulo.

La transformación corresponde a la multiplicación de X por un escalar λ. Si $|\lambda| < 1$ el resultado es una contracción del vector X, de lo contrario es una dilatación de X. El problema se plantea así:

Dada la transformación definida por la matriz cuadrada $\boldsymbol{A}$ de tamaño $(p \times p)$, encontrar los vectores X de $\mathbb{R}^p$, tal que

$$\boldsymbol{A}X = \lambda X, \text{ para } \lambda \neq 0. \tag{A2.24}$$

Al escalar λ de (A2.24) se le llama *valor propio* o *valor característico* de $\boldsymbol{A}$ y a X el respectivo *vector propio* o *vector característico*.

En un lenguaje geométrico-estadístico, se trata de buscar aquellos vectores, que al ser transformados por $\boldsymbol{A}$ no cambian su sentido (permanecen en la misma recta); esto es importante en estadística, pues con estos vectores resulta posible identificar la dirección en que se conserva la información más importante contenida en los datos. Encontrar este vector significa hallar la dirección en la que se encuentra un buena parte de la información contenida en los datos. La figura A.4 muestra una interpretación geométrica de la transformación expresada en A2.24.

Resolver la ecuación (A2.24) es equivalente a encontrar la solución de

$$(\boldsymbol{A} - \lambda \boldsymbol{I})X = \boldsymbol{0}, \tag{A2.25}$$

respecto a λ, con $X \neq \boldsymbol{0}$.

El sistema anterior tiene soluciones diferentes a la solución nula, si y sólo si, el determinante de la matriz $(\boldsymbol{A} - \lambda \boldsymbol{I})$ es igual a cero; es decir,

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0. \tag{A2.26}$$

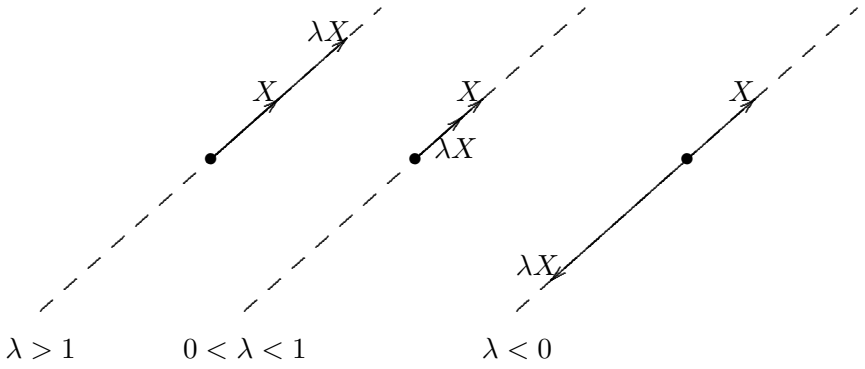


Figura A.4 *Representación de* $\boldsymbol{A}X = \lambda X$*, valor propio* (λ) *y vector propio* (X)*.*

La ecuación (A2.26) recibe el nombre de *ecuación característica* y las raices de esta ecuación son los *valores propios* de la matriz $\boldsymbol{A}$. Un vector X asociado al valor propio λ es llamado el *vector propio* de $\boldsymbol{A}$.

Cuando la matriz $\boldsymbol{A}$ es simétrica, todos sus valores propios son números reales; en caso contrario pueden ser números complejos.

A continuación se resumen las propiedades sobre los valores propios, de uso más frecuente en estadística multivariada.

i) Una matriz $\boldsymbol{A}$ tiene al menos un valor propio igual a cero si y sólo si $\boldsymbol{A}$ es singular, esto equivale a decir que $|\boldsymbol{A}| = 0$.

ii) Si $\boldsymbol{A}$ es una matriz simétrica con valores en los números reales, los vectores propios correspondientes a valores propios diferentes son *ortogonales*.

iii) Cualquier matriz simétrica $\boldsymbol{A}$ puede ser escrita como

$$\boldsymbol{A} = \boldsymbol{P}\Lambda\boldsymbol{P}', \tag{A2.27}$$

donde Λ es una matriz diagonal formada por los valores propios de $\boldsymbol{A}$ y $\boldsymbol{P}$ es una matriz ortogonal cuyas columnas son los vectores propios unitarios asociados con los elementos de la diagonal de Λ. Esta propiedad se conoce con el nombre de *teorema de la descomposición espectral*.

iv) Si $\boldsymbol{A}$ es una matriz simétrica, entonces $r(\boldsymbol{A})$ es igual al número de sus valores propios no nulos.

v̈) Si $\lambda_1, \lambda_2, \cdots, \lambda_p$ son los valores propios de la matriz $\boldsymbol{A}$, entonces

$$\text{tra}(\boldsymbol{A}) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \lambda_i, \tag{A.1a}$$

$$|\boldsymbol{A}| = \lambda_1 \cdot \lambda_2 \cdot \cdots \cdot \lambda_p = \prod_{i=1}^{p} \lambda_i. \tag{A.1b}$$

v̈i) Si λ es un valor propio de la matriz $\boldsymbol{A}$, entonces λ^k es un valor propio de la matriz $\boldsymbol{A}^k$. Los valores propios del polinomio matricial $a_o\boldsymbol{I} + a_1\boldsymbol{A} + a_2\boldsymbol{A}^2 + \cdots + a_k\boldsymbol{A}^k$ corresponden al polinomio de la forma $a_o + a_1\lambda + a_2\lambda^2 + \cdots + a_k\lambda^k$.

v̈ii) Si $\boldsymbol{A}$ es una matriz de tamaño $(n \times p)$ y de rango r, entonces $\boldsymbol{A}$ puede escribirse en la forma,

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}', \tag{A2.29}$$

donde $\boldsymbol{\Delta} = \text{Diag}(\delta_1, \ldots, \delta_r)$, con $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_r \geq 0$, $\boldsymbol{U}$ una matriz ortogonal de tamaño $(n \times r)$, y $\boldsymbol{V}$ una matriz ortonormal de tamaño $(p \times r)$; es decir, $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_r$.

Los valores $\{\delta_i\}$ se llaman los *valores singulares* de $\boldsymbol{A}$. Si $\boldsymbol{U}$ y $\boldsymbol{V}$ se escriben en términos de sus vectores columna, $\boldsymbol{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r\}$, $\boldsymbol{V} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r\}$, entonces $\{\boldsymbol{u}_i\}$ son los vectores singulares a izquierda de $\boldsymbol{A}$ y $\{\boldsymbol{v}_i\}$ son los vectores singulares a derecha de $\boldsymbol{A}$. La matriz $\boldsymbol{A}$ puede escribirse en la forma,

$$\boldsymbol{A} = \sum_{i=1}^{r} \delta_i \boldsymbol{u}_i \boldsymbol{v}_i'. \tag{A2.30}$$

A (A2.29) o (A2.30) se les conoce con el nombre de *descomposición en valor singular* de la matriz $\boldsymbol{A}$.

Además, se demuestra que $\{\delta_i^2\}$ son los valores propios no nulos de la matriz $\boldsymbol{A}\boldsymbol{A}'$ y también de la matriz $\boldsymbol{A}'\boldsymbol{A}$; es decir, $\delta_i = \sqrt{\lambda_i}$, con λ_i valor propio no nulo de $\boldsymbol{A}\boldsymbol{A}'$. Los vectores $\{\boldsymbol{u}_i\}$ son los correspondientes vectores propios normalizados de $\boldsymbol{A}\boldsymbol{A}'$, y los $\{\boldsymbol{v}_i\}$ los correspondientes vectores propios normalizados de $\boldsymbol{A}'\boldsymbol{A}$.

v̈ii) Las matrices $\boldsymbol{A}$ y $\boldsymbol{A}'$ tienen el mismo conjunto de valores propios pero un vector propio de $\boldsymbol{A}$ no necesariamente es un vector propio de $\boldsymbol{A}'$.

Ejemplo A.4 Dada la matriz $\boldsymbol{A} = \begin{pmatrix} 3 & -2 & 0 \\ -2 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$, calcular (i) El determinante, (ii) Su inversa (iii) La traza (iv̈) Los valores y vectores propios (v̈) Diagonalizar, si es posible la matriz $\boldsymbol{A}$.

i) El determinante de $\boldsymbol{A}$ es fácil encontrarlo haciendo la expansión por los cofactores de los elementos de la tercera fila, pues es la que contiene más ceros.

$$|\boldsymbol{A}| = 0 \times (-1)^{3+1} \begin{vmatrix} -2 & 0 \\ 3 & 0 \end{vmatrix} + 0 \times (-1)^{3+2} \begin{vmatrix} 3 & 0 \\ -2 & 0 \end{vmatrix} + 5 \times (-1)^{3+3} \begin{vmatrix} 3 & -2 \\ -2 & 3 \end{vmatrix}$$
$$= 5 \times (9 - 4) = 25.$$

ii) Por ser el determinante diferente de cero, la matriz $\boldsymbol{A}$ es no singular; es decir, tiene inversa. La inversa se calcula mediante (A2.16)

$$\boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|} \operatorname{Adj}(\boldsymbol{A}) = \frac{1}{25} \begin{pmatrix} 15 & 10 & 0 \\ 10 & 15 & 0 \\ 0 & 0 & 5 \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & \frac{2}{5} & 0 \\ \frac{2}{5} & \frac{3}{5} & 0 \\ 0 & 0 & \frac{1}{5} \end{pmatrix} = \frac{1}{5} \begin{pmatrix} 3 & 2 & 0 \\ 2 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

iii) La traza de $\boldsymbol{A}$ es

$$\operatorname{tra}(\boldsymbol{A}) = tra \begin{pmatrix} 3 & -2 & 0 \\ -2 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix} = 3 + 3 + 5 = 11.$$

iv̈) Los valores propios de $\boldsymbol{A}$ se obtienen al resolver la ecuación característica

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$$
$$\begin{vmatrix} 3-\lambda & -2 & 0 \\ -2 & 3-\lambda & 0 \\ 0 & 0 & 5-\lambda \end{vmatrix} = 0$$
$$(1 - \lambda)(5 - \lambda)^2 = 0.$$

Por consiguiente los valores propios de $\boldsymbol{A}$ son $\lambda = 1$ y $\lambda = 5$. El valor propio $\lambda = 5$ ocurre dos veces, se dice entonces que tiene *multiplicidad* igual a 2; en general la *multiplicidad* de un valor propio es el número de veces que éste es solución de la ecuación característica.

Por definición X es un vector propio de $\boldsymbol{A}$ al cual le corresponde el valor propio λ, ahora $\boldsymbol{A}X = \lambda X$, si y sólo si, X es una solución no nula de $(\boldsymbol{A} - \lambda \boldsymbol{I})X = 0$; es decir, solución no trivial de

$$\begin{pmatrix} 3-\lambda & -2 & 0 \\ -2 & 3-\lambda & 0 \\ 0 & 0 & 5-\lambda \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Si $\lambda = 1$, la ecuación anterior es

$$\begin{pmatrix} 2 & -2 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

al resolver el sistema se obtiene $X_1 = r$, $X_2 = r$ y $X_3 = 0$, con r un escalar distinto de cero. De esta forma los vectores propios asociados con el valor propio $\lambda = 1$ tienen la forma

$$X = \begin{pmatrix} r \\ r \\ 0 \end{pmatrix} = r \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

Para $\lambda = 5$ el sistema de ecuaciones se transforma en

$$\begin{pmatrix} -2 & -2 & 0 \\ -2 & -2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

cuya solución es $X_1 = -s$, $X_2 = s$ y $X_3 = t$ con s y t escalares ambos no nulos. Los vectores característicos no nulos ligados a $\lambda = 5$ toman la forma

$$X = \begin{pmatrix} -s \\ s \\ t \end{pmatrix} = \begin{pmatrix} -s \\ s \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ t \end{pmatrix} = s \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + t \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

v̈) Como la matriz $\boldsymbol{A}$ es simétrica, entonces, de acuerdo con (A2.27), es diagonizable ortogonalmente. Los vectores propios V_1, V_2 y V_3 son linealmente independientes

$$V_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \quad V_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \text{y} \quad V_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

Éstos se obtienen al hacer $r = s = t = 1$; pero pueden tomar cualquier otro valor, diferentes de cero. La matriz $\boldsymbol{P}$ se conforma al transformar los vectores anteriores en unitarios (esto se consigue dividiendo a cada uno por su norma), ésta es entonces

$$\boldsymbol{P} = \begin{pmatrix} \frac{-1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{pmatrix}.$$

La matriz $\boldsymbol{P}$ es ortogonal, pues se verifica que $\boldsymbol{PP} = \boldsymbol{P'P} = \boldsymbol{I}$. La diagonalización se obtiene al aplicar (A2.27), así

$$\begin{aligned}\boldsymbol{P'AP} &= \Lambda \\ &= \begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{pmatrix} \begin{pmatrix} 3 & -2 & 0 \\ -2 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} \frac{-1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix}.\end{aligned}$$

Se observa que la última matriz es diagonal con los valores propios sobre la diagonal principal, también se verifica que el determinante $\boldsymbol{A}$ es igual al producto de sus valores propios, es decir $|\boldsymbol{A}| = 25$ y que la traza de $\boldsymbol{A}$ es igual a la suma de sus valores propios, $\text{tra}(\boldsymbol{A}) = 11$. ☑

⊞ *Formas cuadráticas*

Sea $\boldsymbol{A}$ una matriz simétrica de tamaño $(p \times p)$ y X un vector de tamaño $(p \times 1)$, la función

$$Q(X) = X'\boldsymbol{A}X, \tag{A2.31}$$

se llama una *forma cuadrática* de X. $Q(X)$ es un escalar y puede ser expresado alternativamente por la ecuación

$$Q(X) = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}x_i x_j, \tag{A2.32}$$

con a_{ij} elemento de la matriz $\boldsymbol{A}$, x_i y x_j elementos del vector X.

Para $p = 2$ la ecuación (A2.29) toma la forma

$$Q(X) = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = ax_1^2 + 2bx_1x_2 + cx_2^2$$

Esta forma cuadrática está ligada a la ecuación general de segundo grado

$$ax_1^2 + 2bx_1x_2 + cx_2^2 + dx_1 + ex_2 + f = 0, \tag{A2.33}$$

que representa las llamadas *secciones cónicas* (elipse, parábola e hipérbole), de acuerdo con que al menos uno de los números a, b o c sea diferente de cero. Los números d, e y f determinan el centro y radio (tamaño) de la gráfica de (A2.33) en $\mathbb{R}^2$. Si $d = e = 0$, la gráfica tiene su centro en el punto $(0, 0)$.

Si $Q(X) > 0$ para todo $X \neq 0$, se dice que $\boldsymbol{A}$ es *definida positiva*. Si $Q(X) \geq 0$ para todo $X \neq 0$, $\boldsymbol{A}$ se llama *semidefinida positiva*. Si $\boldsymbol{A}$ es

definida positiva se nota $\boldsymbol{A} > 0$ y si $\boldsymbol{A}$ es semidefinida positiva, se nota $\boldsymbol{A} \geq 0$.

Se resaltan las siguientes propiedades para las formas cuadráticas.

i) Si $\boldsymbol{A} > 0$, entonces todos sus valores propios $\lambda_1, \lambda_2, \cdots, \lambda_p$ son positivos. Si $\boldsymbol{A} \geq 0$, entonces $\lambda_i \geq 0$ para $i = 1, 2 \cdots, p$ y $\lambda_i = 0$ para algún i.

ii) Si $\boldsymbol{A} > 0$, entonces $\boldsymbol{A}$ es no singular y en consecuencia $|\boldsymbol{A}| > 0$.

iii) Si $\boldsymbol{A} > 0$ entonces $\boldsymbol{A}^{-1} > 0$.

iv) Si $\boldsymbol{A} > 0$ y $\boldsymbol{C}$ es una matriz no singular $(p \times p)$, entonces $\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C} > 0$.

En semejanza con los números reales, para las matrices definidas no negativas existe una única matriz que corresponde a su *raíz cuadrada*; es decir, que para la matriz $A \geq 0$ existe una única matriz $B \geq 0$ tal que

$$\boldsymbol{B}^2 = \boldsymbol{A}. \tag{A2.34}$$

Se nota $\boldsymbol{A}^{1/2} = \boldsymbol{B}$. Ahora, si $\boldsymbol{A} > 0$ entonces $\boldsymbol{A}^{-1} > 0$ y $\boldsymbol{A}^{1/2} > 0$. Además, $(\boldsymbol{A}^{-1})^{1/2} = (\boldsymbol{A}^{1/2})^{-1} = \boldsymbol{A}^{-1/2}$. Contrario a lo que se espera $\boldsymbol{A}^{1/2}$ no es la matriz cuyos elementos son la raíz cuadrada de los respectivos de $\boldsymbol{A}$; esto sólo se tiene en matrices diagonales $\boldsymbol{D} = \text{Diag}(d_{ii}) \geq 0$.

Ejemplo A.5 Se ilustran analítica y gráficamente los conceptos de valor propio, vector propio, ortogonalidad y forma cuadrática, en $\mathbb{R}^2$, mediante el siguiente caso particular.

Considérese la ecuación

$$5x_1^2 - 4x_1x_2 + 8x_2^2 + \frac{20}{\sqrt{5}}x_1 - \frac{80}{\sqrt{5}}x_2 + 4 = 0,$$

la cual representa una elipse. La forma matricial de esta ecuación es

$$\begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 5 & -2 \\ -2 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \frac{20}{\sqrt{5}} & \frac{-80}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + 4 = 0$$

$$X'\boldsymbol{A}X + KX + 4 = 0.$$

Los valores propios de $\boldsymbol{A}$ se obtienen al resolver la ecuación $|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$, así,

$$\begin{vmatrix} 5-\lambda & -2 \\ -2 & 8-\lambda \end{vmatrix} = (4-\lambda)(9-\lambda) = 0.$$

Éstos son $\lambda_1 = 4$ y $\lambda_2 = 9$; por las propiedades señaladas arriba se puede afirmar que la matriz $\boldsymbol{A}$ es definida positiva (pues $\lambda_1, \lambda_2 > 0$). Los vectores característicos correspondientes a $\lambda_1 = 4$, resultan de la solución no trivial de

$$\begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

la solución es

$$V_1 = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2t \\ t \end{pmatrix} = t \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

Similarmente, para $\lambda_2 = 9$, los vectores propios son de la forma

$$V_2 = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -t \\ 2t \end{pmatrix} = t \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

Los vectores V_1 y V_2 normalizados se transforman, respectivamente, en:

$$P_1 = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}, \text{ y } P_2 = \begin{pmatrix} \frac{-1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}.$$

La matriz $\boldsymbol{P}$, cuyas columnas son los vectores propios ortonormales, P_1 y P_2, es

$$\boldsymbol{P} = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix} = \begin{pmatrix} \cos\theta & -\operatorname{sen}\theta \\ \operatorname{sen}\theta & \cos\theta \end{pmatrix}, \text{ con } \theta = 26.56^o;$$

la cual diagonaliza a la matriz $\boldsymbol{A}$.

La matriz $\boldsymbol{P}$ corresponde a una transformación por *rotación "rígida"* de los ejes X_1 y X_2 (un ángulo $\theta = 26.56^o$), la cual se define por $\tilde{X} = \boldsymbol{P}'X$ o $X = \boldsymbol{P}\tilde{X}$; al sustituir en la ecuación de la elipse se obtiene

$$(\boldsymbol{P}\tilde{X})'\boldsymbol{A}(\boldsymbol{P}\tilde{X}) + K(\boldsymbol{P}\tilde{X}) + 4 = 0$$
$$(\tilde{X})'(\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P}\tilde{X} + (K\boldsymbol{P})\tilde{X} + 4 = 0.$$

Dado que

$$\boldsymbol{P}'\boldsymbol{A}\boldsymbol{P} = \begin{pmatrix} 4 & 0 \\ 0 & 9 \end{pmatrix} \text{ y } K\boldsymbol{P} = \begin{pmatrix} \frac{20}{\sqrt{5}} & \frac{-80}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \end{pmatrix} = \begin{pmatrix} -8 & -36 \end{pmatrix},$$

la ecuación de la elipse se puede escribir como

$$4\tilde{x}_1^2 + 9\tilde{x}_2^2 - 8\tilde{x}_1 - 36\tilde{x}_2 + 4 = 0$$

se puede apreciar que el efecto de la rotación o la diagonalización es la eliminación del término x_1x_2, el cual indica una asociación entre dichas variables.

Para que esta cónica tenga su origen en el punto $(0, 0)$ es necesario trasladar el sistema de coordenadas $\tilde{X}_1 \times \tilde{X}_2$. La traslación se sugiere después de llevar la ecuación anterior a la forma canónica[6]., algebraicamente se hace mediante la completación a trinomios cuadrados perfectos; así, la ecuación de la elipse anterior es

$$4(\tilde{x}_1^2 - 2\tilde{x}_1) + 9(\tilde{x}_2^2 - 4\tilde{x}_2) = -4,$$

al completar a trinomio cuadrado perfecto dentro de cada paréntesis se obtiene

$$\begin{aligned} 4(\tilde{x}_1^2 - 2\tilde{x}_1 + 1) + 9(\tilde{x}_2^2 - 4\tilde{x}_2 + 4) &= -4 + 4 + 36 \\ 4(\tilde{x}_1 - 1)^2 + 9(\tilde{x}_2 - 2)^2 &= 36. \end{aligned}$$

La translación de los ejes $\tilde{X}_1$ y $\tilde{X}_2$ a los "nuevos" ejes X_1^* y X_2^* viene dada por

$$X_1^* = \tilde{X}_1 - 1 \quad X_2^* = \tilde{X}_2 - 2.$$

La ecuación resultante, finalmente es

$$4x_1^{*2} + 9x_2^{*2} = 36 \text{ o } \frac{x_1^{*2}}{9} + \frac{x_2^{*2}}{4} = 1,$$

la cual corresponde a la ecuación de una elipse en posición normal (canónica) respecto al sistema de coordenadas $X_1^* \times X_2^*$. La figura A.5 ilustra el proceso anterior junto con el resulado final ☑.

⊞ *Descomposición de Cholesky*

Una matriz $\boldsymbol{A}$ definida positiva se puede factorizar como

$$\boldsymbol{A} = \boldsymbol{T}'\boldsymbol{T}, \tag{A2.35}$$

donde $\boldsymbol{T}$ es una matriz no singular triangular superior. Una forma de obtener la matriz $\boldsymbol{T}$ es mediante la *descomposición de Cholesky*, cuyo procedimento se explica a continuación.

Sean $\boldsymbol{A} = (a_{ij})$ y $\boldsymbol{T} = (t_{ij})$ matrices de tamaño $p \times p$. Entonces los elementos de la matriz $\boldsymbol{T}$ se encuentran como sigue:

- $t_{11} = \sqrt{a_{11}}$, $t_{1j} = \frac{a_{1j}}{t_{11}}$, para $2 \leq j \leq n$;

[6]La ecuación canónica de la elipse con centro $(0, 0)$ es: $\frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} = 1$

- $t_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} t_{ki}^2}$, para $2 \leq i \leq n$;
- $t_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} t_{ki} t_{kj}}{tii}$, para $2 \leq i < j \leq n$;
- $t_{ij} = 0$, para $1 \leq j < i \leq n$.

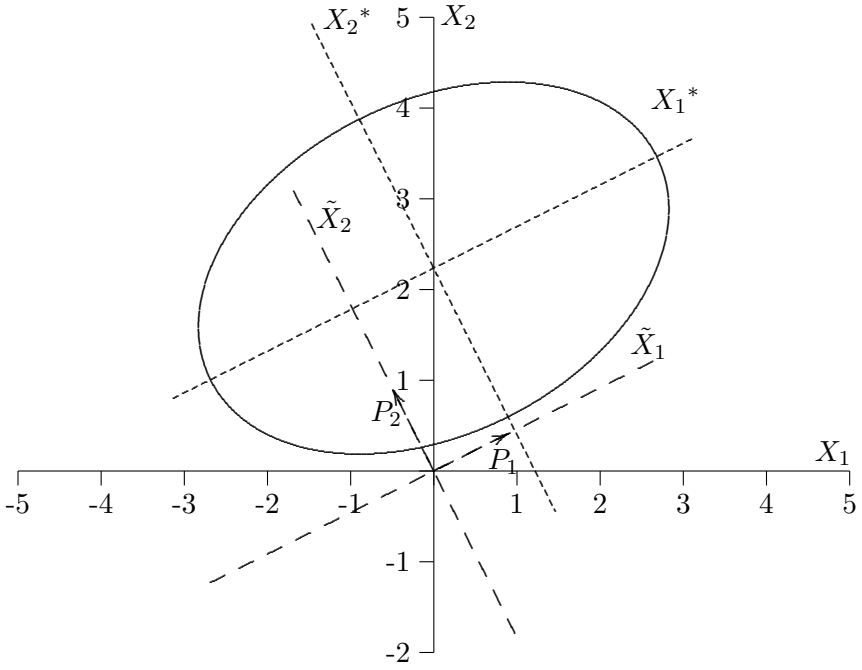


Figura A.5 *Translación y rotación.*

Ejemplo A.6 Sea $\boldsymbol{A}$ la siguiente matriz

$$\boldsymbol{A} = \begin{pmatrix} 3 & 0 & -3 \\ 0 & 6 & 3 \\ -3 & 3 & 6 \end{pmatrix}.$$

Por el método de Cholesky, se obtiene

- $t_{11} = \sqrt{3}$, $t_{12} = \frac{0}{\sqrt{3}} = 0$, $t_{13} = \frac{-3}{\sqrt{3}} = -\sqrt{3}$;
- $t_{22} = \sqrt{6 - (0^2)} = \sqrt{6}$;
- $t_{23} = \frac{3-(0)(-\sqrt{3})}{\sqrt{6}} = \sqrt{1.5}$;

- $t_{33} = \sqrt{6 - \left[(-\sqrt{3})^2 + (\sqrt{1.5})^2\right]} = \sqrt{1.5}$;
- $t_{21} = t_{31} = t_{32} = 0$.

De donde la matriz $\boldsymbol{T}$ es

$$\boldsymbol{T} = \begin{pmatrix} \sqrt{3} & 0 & -\sqrt{3} \\ 0 & \sqrt{6} & \sqrt{1.5} \\ 0 & 0 & \sqrt{1.5} \end{pmatrix}.$$

Se satisface que

$$\boldsymbol{T}'\boldsymbol{T} = \begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{6} & 0 \\ -\sqrt{3} & \sqrt{1.5} & \sqrt{1.5} \end{pmatrix} \begin{pmatrix} \sqrt{3} & 0 & -\sqrt{3} \\ 0 & \sqrt{6} & \sqrt{1.5} \\ 0 & 0 & \sqrt{1.5} \end{pmatrix} = \boldsymbol{A}.$$

⊞ *Partición de una matriz*

A veces resulta más cómodo expresar una matriz en forma de "submatrices", es decir, tal que sus elementos conformen matrices de tamaño más pequeño (sea por filas, columnas o ambos) que la original. En general, sea $\boldsymbol{A}$ una matriz de tamaño $(n \times p)$, la matriz $\boldsymbol{A}$ se puede escribir así:

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1j} & \cdots & \boldsymbol{A}_{1p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \boldsymbol{A}_{i1} & \boldsymbol{A}_{i2} & \cdots & \boldsymbol{A}_{ij} & \cdots & \boldsymbol{A}_{ip} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \boldsymbol{A}_{n1} & \boldsymbol{A}_{n2} & \cdots & \boldsymbol{A}_{nj} & \cdots & \boldsymbol{A}_{np} \end{pmatrix}, \tag{A2.36}$$

donde la "submatriz" $\boldsymbol{A}_{ij}$ es de tamaño $(n_i \times p_j)$, con $\sum_{i=1}^{n} n_i = n$ y $\sum_{j=1}^{p} p_j = p$.

La suma y producto entre este tipo de matrices se conforma de manera semejante a como se describió en (A2.7) y (A2.9). De esta forma, si las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ se particionan similarmente entonces

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} \boldsymbol{A}_{11} + \boldsymbol{B}_{11} & \cdots & \boldsymbol{A}_{1j} + \boldsymbol{B}_{1j} & \cdots & \boldsymbol{A}_{1p} + \boldsymbol{B}_{1p} \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ \boldsymbol{A}_{n1} + \boldsymbol{B}_{n1} & \cdots & \boldsymbol{A}_{nj} + \boldsymbol{B}_{nj} & \cdots & \boldsymbol{A}_{np} + \boldsymbol{B}_{np} \end{pmatrix}. \tag{A2.37}$$

Si las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$, son de tamaño $(m \times n)$ y $(n \times p)$, respectivamente, y se particionan adecuadamente para el producto, éste es

$$\boldsymbol{AB} = \begin{pmatrix} \sum_{j=1}^{n} \boldsymbol{A}_{1j}\boldsymbol{B}_{j1} & \cdots & \sum_{j=1}^{n} \boldsymbol{A}_{1j}\boldsymbol{B}_{jp} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{n} \boldsymbol{A}_{mj}\boldsymbol{B}_{j1} & \cdots & \sum_{j=1}^{n} \boldsymbol{A}_{mj}\boldsymbol{B}_{jp} \end{pmatrix}. \tag{A2.38}$$

Para una matriz $\boldsymbol{A}$ particionada en la siguiente forma

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix}, \tag{A2.39}$$

donde $\boldsymbol{A}_{11}$ y $\boldsymbol{A}_{22}$ son matrices *no singulares*, la inversa de $\boldsymbol{A}$ se calcula mediante

$$\boldsymbol{A}^{-1} = \begin{bmatrix} (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} & -(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1} & \boldsymbol{A}_{22}^{-1} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}(\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \end{bmatrix} \tag{A2.40}$$

El determinante de la matriz $\boldsymbol{A}$ se puede calcular a partir de la partición (A2.39), para los casos en que las submatrices $\boldsymbol{A}_{ii}$, $i = 1, 2$ sean no singulares. Es decir,

$$\begin{aligned} |\boldsymbol{A}| &= |\boldsymbol{A}_{11}| \cdot |\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}|, \text{o} \\ |\boldsymbol{A}| &= |\boldsymbol{A}_{22}| \cdot |\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}|. \end{aligned} \tag{A2.41}$$

Un caso especial del resultado anterior es el siguiente:

$$\begin{vmatrix} 1 & -y' \\ y & C \end{vmatrix} = \begin{vmatrix} C & y \\ -y' & 1 \end{vmatrix} \tag{A2.41a}$$

lo cual es equivalente, por (A2.41), con la expresión:

$$|C + yy'| = |C|(1 + y'C^{-1}y). \tag{A2.41b}$$

⊞ *Sumas y productos directos*

Sean $\boldsymbol{A}$ y $\boldsymbol{B}$ matrices de tamaño $(n_1 \times p_1)$ y $(n_2 \times p_2)$, respectivamente. La *suma directa* entre las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ es definida por

$$\boldsymbol{A} \oplus \boldsymbol{B} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{pmatrix}.$$

la cual es una matriz de tamaño $(n_1 + n_2) \times (p_1 + p_2)$. Las matrices nulas son de tamaño $(n_1 \times p_2)$ y $(n_2 \times p_1)$, respectivamente. En forma general,

$$\bigoplus_{i=1}^{k} \boldsymbol{A}_i = \begin{pmatrix} \boldsymbol{A}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{A}_k \end{pmatrix} = \text{Diag}\{\boldsymbol{A}_i\}, \text{ para } i = 1, \cdots, k. \tag{A2.42}$$

Se muestran algunas propiedades de la suma directa entre matrices, tomando como referencia la suma usual.

i) $(\boldsymbol{A} \oplus \boldsymbol{B})' = \boldsymbol{A}' \oplus \boldsymbol{B}'$.

ii) $\boldsymbol{A} \oplus (-\boldsymbol{A}) \neq 0$ a menos que $\boldsymbol{A} = 0$.

iii) $(\boldsymbol{A} \oplus \boldsymbol{B}) + (\boldsymbol{C} \oplus \boldsymbol{D}) = (\boldsymbol{A} + \boldsymbol{B}) \oplus (\boldsymbol{C} + \boldsymbol{D})$, siempre que las matrices sean conformables para la suma.

iv̈) $(\boldsymbol{A} \oplus \boldsymbol{B})(\boldsymbol{C} \oplus \boldsymbol{D}) = (\boldsymbol{AC}) \oplus (\boldsymbol{BD})$, asegurando la conformabilidad respecto al producto.

v̈) $(\boldsymbol{A} \oplus \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \oplus \boldsymbol{B}^{-1}$

v̈i) La suma directa $(\boldsymbol{A} \oplus \boldsymbol{B})$ es cuadrada y de tamaño $((n+p) \times (p+n))$, solo si $\boldsymbol{A}$ es de tamaño $(n \times p)$ y $\boldsymbol{B}$ es de tamaño $(p \times n)$.

v̈ii) El determinante de $(\boldsymbol{A} \oplus \boldsymbol{B})$ es igual a $|\boldsymbol{A}||\boldsymbol{B}|$ si $\boldsymbol{A}$ y $\boldsymbol{B}$ son matrices cuadradas, de otra forma es cero o no existe.

El *producto directo*[7] entre la matriz $\boldsymbol{A}$ $\boldsymbol{B}$, de tamaño $(n_1 \times p_1)$ y $(n_2 \times p_2)$ respectivamente, se define como

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1p_1}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2p_1}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n_1 1}\boldsymbol{B} & a_{n_1 2}\boldsymbol{B} & \cdots & a_{n_1 p_1}\boldsymbol{B} \end{pmatrix}. \tag{A2.43}$$

El producto directo entre estas matrices da como resultado una matriz de tamaño
$(n_1 \cdot n_2 \times p_1 \cdot p_2)$, que consta de todos los posibles productos de un elemento de la matriz $\boldsymbol{A}$ por un elemento de la matriz $\boldsymbol{B}$.

[7] Llamado también producto Kronecker.

Entre las propiedades del producto directo se destacan las siguientes.

i) $(\boldsymbol{A} \otimes \boldsymbol{B})' = \boldsymbol{A}' \otimes \boldsymbol{B}'$.

ii) Para X y Y vectores: $X' \otimes Y = YX' = Y \otimes X'$.

iii) Para λ un escalar: $\lambda \otimes \boldsymbol{A} = \lambda \boldsymbol{A} = \boldsymbol{A} \otimes \lambda = \boldsymbol{A}\lambda$.

iv) $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$

v) $|\boldsymbol{A} \otimes \boldsymbol{B}| = |\boldsymbol{A}|^{p_2} \cdot |\boldsymbol{B}|^{p_1}$ donde las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ son matrices cuadradas de tamaño p_1 y p_2, respectivamente.

vi) Los valores propios de $\boldsymbol{A} \otimes \boldsymbol{B}$ son los productos de los valores propios de $\boldsymbol{A}$ con los valores propios de $\boldsymbol{B}$.

Ejemplo A.7 Se ilustra la suma directa y el producto directo entre matrices con los dos casos siguientes.

$$\begin{pmatrix} 2 & 4 \\ 5 & 3 \end{pmatrix} \oplus \begin{pmatrix} 1 & 0 & 3 \\ 4 & 6 & 10 \\ 5 & 8 & 9 \end{pmatrix} = \left(\begin{array}{cc:ccc} 2 & 4 & 0 & 0 & 0 \\ 5 & 3 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 1 & 0 & 3 \\ 0 & 0 & 4 & 6 & 10 \\ 0 & 0 & 5 & 8 & 9 \end{array}\right).$$

$$\begin{pmatrix} 2 & 4 \\ 5 & 3 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 & 3 \\ 4 & 6 & 10 \\ 5 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 2\begin{pmatrix} 1 & 0 & 3 \\ 4 & 6 & 10 \\ 5 & 8 & 9 \end{pmatrix} & 4\begin{pmatrix} 1 & 0 & 3 \\ 4 & 6 & 10 \\ 5 & 8 & 9 \end{pmatrix} \\ 5\begin{pmatrix} 1 & 0 & 3 \\ 4 & 6 & 10 \\ 5 & 8 & 9 \end{pmatrix} & 3\begin{pmatrix} 1 & 0 & 3 \\ 4 & 6 & 10 \\ 5 & 8 & 9 \end{pmatrix} \end{pmatrix}$$

$$= \left(\begin{array}{ccc:ccc} 2 & 0 & 6 & 4 & 0 & 12 \\ 8 & 12 & 20 & 16 & 24 & 40 \\ 10 & 16 & 18 & 20 & 32 & 36 \\ \hdashline 5 & 0 & 15 & 3 & 0 & 9 \\ 20 & 30 & 50 & 12 & 18 & 30 \\ 25 & 40 & 45 & 15 & 24 & 27 \end{array}\right).$$

⊞ *Diferenciación con vectores y matrices*

Se presenta la derivada de un escalar (campo escalar), la derivada de un vector (campo vectorial) y la derivada asociada a una forma cuadrática. Otros resultados del cálculo, tales como de derivadas para determinantes, inversas y trazas, se desarrollan en forma condensada.

Sea f una función que asigna a un vector $X \in \mathbb{R}^p$ un número real, esquemáticamente

$$f: \mathbb{R}^p : \xrightarrow{\text{f}} \mathbb{R}$$

$$X = (x_1, \cdots, x_p) \longrightarrow f(X).$$

Se define la derivada de $f(X)$ con respecto al vector X de tamaño $p \times 1$ como la matriz

$$\frac{\partial f(X)}{\partial X} = \left(\frac{\partial f(X)}{\partial x_{ij}} \right). \tag{A2.44}$$

- Para $f(X) = \boldsymbol{a}'X = a_1x_1 + \cdots + a_px_p$ donde $\boldsymbol{a}$ y X son vectores de $\mathbb{R}^p$, la derivada de la función f respecto al vector X, de acuerdo con (A2.44), está dada por

$$\frac{\partial f(X)}{\partial X} = \frac{\partial}{\partial X}(\boldsymbol{a}'X) = \frac{\partial}{\partial X}(X'\boldsymbol{a}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \boldsymbol{a}. \tag{A2.45}$$

- La derivada de $Y' = X'\boldsymbol{A}$ (campo vectorial), con X vector de $\mathbb{R}^p$ y $\boldsymbol{A}$ matriz de tamaño $(p \times p)$, se obtiene aplicando la derivada (A2.44) sobre cada uno de los elementos del vector Y'. Explícitamente, el vector Y' se puede escribir como

$$Y' = \begin{pmatrix} y_1 & y_2 & \cdots & y_p \end{pmatrix} = \begin{pmatrix} X'a^1 & X'a^2 & \cdots & X'a^p \end{pmatrix}, \tag{A2.46}$$

donde el i-ésimo elemento de Y es $Y_i = X'a^i$; a^i es la i-ésima columna de $\boldsymbol{A}$, $i = 1, \cdots, p$. Aplicando (A2.44) a cada elemento de (A2.45) resulta

$$\begin{aligned} \frac{\partial Y'}{\partial X} &= \begin{pmatrix} \frac{\partial Y_1}{\partial X} & \frac{\partial Y_2}{\partial X} & \cdots & \frac{\partial Y_p}{\partial X} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial X'a^1}{\partial X} & \frac{\partial X'a^2}{\partial X} & \cdots & \frac{\partial X'a^p}{\partial X} \end{pmatrix} = \begin{pmatrix} a^1 & a^2 & \cdots & a^p \end{pmatrix} \\ &= \boldsymbol{A}. \end{aligned}$$

- La derivada de la *forma cuadrática* $Q(X) = X'\boldsymbol{A}X$ es

$$\frac{\partial Q}{\partial X} = \frac{(\partial X'\boldsymbol{A}X)}{\partial X} = 2\boldsymbol{A}X. \tag{A2.47}$$

- La *derivada de la inversa* de una matriz no singular $\boldsymbol{X}$ de tamaño $(p \times p)$ respecto a su elemento x_{ij}, se deduce de la siguiente forma:
  - Si $\boldsymbol{X}^{-1}$ es la matriz inversa de $\boldsymbol{X}$, entonces:

$$\boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{I}.$$

  - Por la propiedad para la derivada de un producto, aplicada en la expresión anterior conduce a:

$$\frac{\partial \boldsymbol{X}}{\partial x_{ij}} \cdot \boldsymbol{X}^{-1} + \boldsymbol{X} \cdot \frac{\partial \boldsymbol{X}^{-1}}{\partial x_{ij}} = \boldsymbol{0}.$$

  - Despejando el término de interés $\frac{\partial \boldsymbol{X}^{-1}}{\partial x_{ij}}$ se obtiene:

$$\frac{\partial \boldsymbol{X}^{-1}}{\partial x_{ij}} = -\boldsymbol{X}^{-1} \cdot \frac{\partial \boldsymbol{X}}{\partial x_{ij}} \cdot \boldsymbol{X}^{-1}. \tag{A2.48}$$

    En la expresión anterior $\frac{\partial \boldsymbol{X}}{\partial x_{ij}}$ es una matriz tal que en el lugar donde se ubica la variable x_{ij} tiene un 1 y en los demás ceros; esta matriz se nota por $\boldsymbol{\Delta}_{ij}$. Aquí deben considerarse tanto el caso en que la matriz $\boldsymbol{X}$ tiene todas sus entradas diferentes como el caso en que la matriz $\boldsymbol{X}$ es simétrica. A continuación se consideran estos dos casos.
  - Si todos los elementos de $\boldsymbol{X}$ son distintos, entonces:

$$\frac{\partial \boldsymbol{X}^{-1}}{\partial x_{ij}} = -\boldsymbol{X}^{-1}\boldsymbol{\Delta}_{ij}\boldsymbol{X}^{-1}. \tag{A2.48a}$$

  - Si la matriz $\boldsymbol{X}$ es simétrica, entonces:

$$\frac{\partial \boldsymbol{X}^{-1}}{\partial x_{ij}} = \begin{cases} -\boldsymbol{X}^{-1}\boldsymbol{\Delta}_{ii}\boldsymbol{X}^{-1}; & i = j, \\ -\boldsymbol{X}^{-1}(\boldsymbol{\Delta}_{ij} + \boldsymbol{\Delta}_{ji})\boldsymbol{X}^{-1}; & i \neq j. \end{cases} \tag{A2.48b}$$

- Para una matriz $\boldsymbol{X}$ no singular de tamaño $(p \times p)$, la *derivada de su determinante* respecto al elemento x_{ij} es

$$\frac{\partial |\boldsymbol{X}|}{\partial x_{ij}} = \boldsymbol{X}_{ij}, \tag{A2.49}$$

  donde $\boldsymbol{X}_{ij}$ es el cofactor de x_{ij}. Así, la matriz de derivadas es:

$$\frac{\partial |\boldsymbol{X}|}{\partial \boldsymbol{X}} = (\boldsymbol{X}_{ij}). \tag{A2.49a}$$

- Para matrices simétricas, la matriz de derivadas es

$$\begin{aligned}\frac{\partial|\boldsymbol{X}|}{\partial \boldsymbol{X}} &= 2\,\mathrm{Adj}(\boldsymbol{X}) - \mathbf{diag}[\mathrm{Adj}(\boldsymbol{X})] \\ &= |\boldsymbol{X}|[2\boldsymbol{X}^{-1} - \mathrm{diag}(\boldsymbol{X}^{-1})] \end{aligned} \tag{A2.50}$$

  donde diag [Adj($\boldsymbol{X}$)] es la matriz diagonal de la matriz adjunta de $\boldsymbol{X}$.

  El resultado siguiente es bastante útil, por ejemplo, para la obtención de estimadores máximo verosímiles p-variantes,

$$\begin{aligned}\frac{\partial(\ln|\boldsymbol{X}|)}{\partial \boldsymbol{X}} &= \left(\frac{1}{|\boldsymbol{X}|}\right)\frac{\partial|\boldsymbol{X}|}{\partial \boldsymbol{X}} \\ &= 2\boldsymbol{X}^{-1} - \mathbf{diag}(\boldsymbol{X}^{-1}). \end{aligned} \tag{A2.51}$$

- La *derivada de la traza* de una matriz $\boldsymbol{X}$ de tamaño $(p \times p)$ es

$$\frac{\partial[\mathrm{tra}(\boldsymbol{X})]}{\partial \boldsymbol{X}} = \boldsymbol{I}, \tag{A2.52}$$

  de donde

$$\frac{\partial[\mathrm{tra}(\boldsymbol{X}\boldsymbol{A})]}{\partial \boldsymbol{X}} = \boldsymbol{A}'. \tag{A2.53}$$

- Si la matriz $\boldsymbol{X}$ es simétrica la anterior derivada es igual a:

$$\frac{\partial[\mathrm{tra}(\boldsymbol{X}\boldsymbol{A})]}{\partial \boldsymbol{X}} = \boldsymbol{A} + \boldsymbol{A}' - \mathrm{diag}(\boldsymbol{A}). \tag{A2.53a}$$

A.3 Rutina SAS para vectores y matrices

El procedimiento IML (Interactive Matrix Language) del paquete SAS contiene una serie de rutinas computacionales, con las cuales se puede hacer una buena parte del trabajo con matrices. Se presentan en esta sección los comandos y sintaxis de uso más frecuente en la estadística multvariada, tales como la creación de vectores y matrices, las operaciones entre vectores y matrices, la transformación de un archivo en una matriz, la solución de sistemas lineales de ecuaciones, entre otros.

`PROC IML;` /∗Invoca el procedimiento IML∗/

► Conformación de matrices

Las entradas de una matriz se escriben dentro de corchetes { }, separando los entradas por un espacio y las filas por una coma ",". A cada vector o matriz se le puede asignar un nombre antepuesto al signo "=".

Para matrices cuyas entradas son caracteres, éstos se pueden escribir dentro de comillas sencillas (‘ ’) o dobles (“ ”). Si se omiten las comillas, como en la matriz Clase anterior, SAS deja las entradas en mayúsculas fijas.

Las instrucciones:

```
u={2 3 -1 1}; v={0, 2,1}; A={3 1, 2 5};
B={2 4, 0 1};

C={$1 3 4, 3 2 1, 4 1 3}; Clase={Pedro, Olga, Pilar, Carlos};

PRINT u v A B C Clase;

RUN;
```

Producen los siguientes vectores y matrices:

$$u = \begin{pmatrix} 2 & 3 & -1 & 1 \end{pmatrix},\ v = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix},\ A = \begin{pmatrix} 3 & 1 \\ 2 & 5 \end{pmatrix},\ B = \begin{pmatrix} 2 & 4 \\ 0 & 1 \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 3 & 4 \\ 3 & 2 & 1 \\ 4 & 1 & 3 \end{pmatrix} \text{ y } Clase = \begin{pmatrix} PEDRO \\ OLGA \\ PILAR \\ CARLOS \end{pmatrix}$$

Con las instrucciones (a la derecha del signo =):

```
M_UNOS=J(2,3,1); M_CEROS=J(2,3,0); V_UNOS=J(1,4,1); I_2=I(2);
```

se genera una matriz de tamaño 2×3 de unos, una matriz nula, un vector de unos y una matriz identidad de tamaño 2×2. Resultan estos arreglos:

$$M\_UNOS = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},\quad M\_CEROS = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$V\_UNOS = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \text{ e } I\_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

► Traspaso de un archivo de datos SAS a una matriz

Se muestra cómo un archivo de datos SAS se puede transformar a una matriz de datos, para que así permitir el trabajo con el procedimiento IML.

El ejemplo considera el archivo llamado “EJER_1”, compuesto por las variables X_1 a X_5 y 10 observaciones. La instrucción “READ ALL INTO X” hace que éste sea considerado como la matriz X de tamaño 10×5.

```
OPTIONS NOCENTER PS=60 LS=80; /* tamaño de página*/
DATA EJER_1; /* Archivo de datos Ejer_1 */
```

```
INPUT X1 X2 X3 X4 X5 @@';/* Ingreso de las variables X1, X2, X3, X4 y X5 */
CARDS; /* Para ingresar datos */
0 1 2 3 5
1 2 5 7 8
0 1 2 8 4
0 3 5 9 7
1 5 4 2 3
1 2 4 3 9
0 1 2 8 4
1 4 4 6 3
1 2 5 7 8
0 5 2 9 2;
PROC IML;
USE EJER_1; /* invoca el archivo Ejer_1  */
READ ALL INTO X; /* Pone los datos del archivo Ejer_1 en la matriz X */
n=NROW(X); /* n es el número de observaciones */
p=NCOL(X); /* p es el número de variables*/
PRINT X n p;
RUN;
```

► Operaciones y transformaciones sobre matrices

Entre las matrices A y B, señaladas arriba, se desarrollan las operaciones y transformaciones respectivas mediante sintaxix IML, la cual se describe a continuación.

$C = A + B$; produce la suma entre A y B, $C = \begin{pmatrix} 5 & 5 \\ 2 & 6 \end{pmatrix}$

$D = A - B$; produce la resta entre A y B, $D = \begin{pmatrix} 1 & -3 \\ 2 & 4 \end{pmatrix}$

$E = A * B$; produce el producto entre A y B, $E = \begin{pmatrix} 6 & 13 \\ 4 & 13 \end{pmatrix}$

$F\_1 = A\#B$; Producto entre elementos correspondientes de A y B, $F\_1 = \begin{pmatrix} 6 & 4 \\ 0 & 5 \end{pmatrix}$

$F\_2 = A\#\#B$; Cada elemento de A elevado al respectivo de B, $F\_2 = \begin{pmatrix} 9 & 1 \\ 1 & 5 \end{pmatrix}$

$G\_1 = A\#\#2$; G_1, las entradas son el cuadrado de las entradas de A , $G\_1 = \begin{pmatrix} 9 & 1 \\ 4 & 25 \end{pmatrix}$

$G\_2 = A**3$; matriz A multiplicada por si misma tres veces (A^3) , $G\_2 = \begin{pmatrix} 49 & 51 \\ 102 & 151 \end{pmatrix}$

$H = A//B$; Dispone la matriz B debajo de la matriz A, $H = \begin{pmatrix} 3 & 1 \\ 2 & 5 \\ 2 & 4 \\ 0 & 1 \end{pmatrix}$

$K = A@B$; Producto directo (Kronecker) entre A y B, $K = \begin{pmatrix} 6 & 12 & 2 & 4 \\ 0 & 3 & 0 & 1 \\ 4 & 8 & 10 & 20 \\ 0 & 2 & 0 & 5 \end{pmatrix}$

$L = B/A$; Divide cada elemento de B por el respectivo de A, $L = \begin{pmatrix} 0.67 & 4.00 \\ 0.00 & 0.20 \end{pmatrix}$

$m = 6 : 10$; Genera el vector m con valores entre 6 y 10, $m = \begin{pmatrix} 6 & 7 & 8 & 9 & 10 \end{pmatrix}$

$n = 8 : 5$ Genera el vector n con valores entre 8 y 5, $n = \begin{pmatrix} 8 & 7 & 6 & 5 \end{pmatrix}$

$O = BLOCK(A, B)$; Matriz en bloques; A y B en la diagonal, $O = \begin{pmatrix} 3 & 1 & 0 & 0 \\ 2 & 5 & 0 & 0 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

$Det\_A = DET(A)$; Produce el determinante de A, $Det\_A = 13$

$DG\_A = DIAG(A)$; Transforma la matriz A en una matriz diagonal, $DG\_A = \begin{pmatrix} 3 & 0 \\ 0 & 5 \end{pmatrix}$

$Inv\_A = INV(A)$; Produce la inversa de A (A^{-1}), $Inv\_A = \begin{pmatrix} 0.3846 & -0.0769 \\ -0.1538 & 0.2307 \end{pmatrix}$

$T\_A = T(A)$; Produce la matriz transpuesta de A ($A^T = A$‘), $T\_A = \begin{pmatrix} 3 & 2 \\ 1 & 5 \end{pmatrix}$

$Vap\_C = EIGVAL(C)$; Obtiene los valores propios de C , $Vap\_C = 7.47,\ 1.40$ y -2.87

$Vep\_C = EIGVEC(C)$; Obtiene los vectores propios V_1, V_2 y V_3 de C (matrices simétricas),

$Vep\_C = \begin{pmatrix} V_1 & V_2 & V_3 \\ 0.61 & 0.04 & -0.79 \\ 0.45 & 0.80 & 0.39 \\ 0.65 & -0.60 & 0.47 \end{pmatrix}$

$Filas\_C = NROW(C)$; Cuenta el número de filas de C, $Filas\_C = 3$

$Colum\_C = NCOL(C)$; Cuenta el número de columnas de C, $Colum\_C = 3$

$Tra\_A = TRACE(C)$; Calcula la traza de la matriz C, $Tra\_C = 6$

$X = SOLVE(A, b)$; Resuelve el sistema $\begin{cases} 3x_1 + x_2 = 3 \\ 2x_1 + 5x_2 = 2 \end{cases}$

con $A = \begin{pmatrix} 3 & 1 \\ 2 & 5 \end{pmatrix}$ y $b = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$. Solución: $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$CALL\ SVD(U, \Delta, V, A)$; Encuentra la descomposición sigular de la matriz A (A2.29)

ésta es: $U = \begin{pmatrix} 0.42 & 0.91 \\ 0.91 & -0.42 \end{pmatrix}$ $\Delta = \begin{pmatrix} 5.83 & 0 \\ 0 & 2.23 \end{pmatrix}$ $V = \begin{pmatrix} 0.53 & 0.85 \\ 0.85 & -0.53 \end{pmatrix}$

A.4 Procesamiento de matrices con R

Conformación de matrices

Las instrucciones

```
u<-matrix(c(2,3,-1,1),nrow=1)
v<-matrix(c(0,2,1),ncol=1)
```

```
A<-matrix(c(3,2,1,5),nrow=2)
B<-matrix(c(2,0,4,1),nrow=2)
C<-matrix(c(1,3,4,3,2,1,4,1,3),nrow=3)
Clase<-matrix(c(Pedro","Olga","Pilar","Carlos"),ncol=1 )"
```

producen los siguientes vectores y matrices

$$u = \begin{pmatrix} 2 & 3 & -1 & 1 \end{pmatrix} \text{ , } v = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \text{ , } A = \begin{pmatrix} 3 & 1 \\ 2 & 5 \end{pmatrix} \text{ , } B = \begin{pmatrix} 2 & 4 \\ 0 & 1 \end{pmatrix}$$

$$C = \begin{pmatrix} 1 & 3 & 4 \\ 3 & 2 & 1 \\ 4 & 1 & 3 \end{pmatrix} \text{ y } Clase = \begin{pmatrix} Pedro \\ Olga \\ Pilar \\ Carlos \end{pmatrix}$$

Con las instrucciones

```
M_UNOS<-matrix(rep(1,6),nrow=2)
M_CEROS <-matrix(rep(0,6),nrow=2)
V_UNOS<-matrix(rep(1,4),nrow=1)
I_2<-diag(1,2)
```

se genera una matriz de tamaño 2×3 de unos, una matriz nula, un vector de unos y una matriz identidad de tamaño 2×2

$$M\_UNOS = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \text{ , } M\_CEROS = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$V\_UNOS = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \text{ e } I\_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Traspaso de un archivo de datos a una matriz

Podemos crear una matriz usando las columnas de un marco de datos (`data frame`). Para el ejemplo usaremos el marco de datos `women`, de los datos de ejemplo de `R`[8]

```
data(women)
 W <- as.matrix(women)
```

[8]Para tener un listado y una corta descripción de los marcos de datos disponibles en el paquete (librería) `datasets` de R, use el comando `data()`, si quiere un listado de los marcos de datos de ejemplo de todas las librerías instaladas use `data(package = .packages(all.available = TRUE))`

Operaciones y transformaciones sobre matrices

Entre las matrices A y B, señaladas arriba, se desarrollan las operaciones y transformaciones respectivas mediante la sintaxis de `R`, la cual se describe a continuación.

`C<-A+B;` produce la suma entre A y B, $\begin{pmatrix} 5 & 5 \\ 2 & 6 \end{pmatrix}$.

`D<-A-B;` Resta B de A, $\begin{pmatrix} 1 & -3 \\ 2 & 4 \end{pmatrix}$.

.

`E<-A%*%B;` Producto entre A y B, $\begin{pmatrix} 1 & -3 \\ 2 & 4 \end{pmatrix}$.

`F_1<- A*B;` Producto entre elementos correspondientes de A y B, $\begin{pmatrix} 6 & 4 \\ 0 & 5 \end{pmatrix}$.

`F_2<- A^B;` Cada elemento de A elevado al correspondiente de B, $\begin{pmatrix} 9 & 1 \\ 1 & 5 \end{pmatrix}$.

`G_1<-A^2;` Las entradas `G_1` son el cuadrado de las entradas de A, $\begin{pmatrix} 9 & 1 \\ 4 & 25 \end{pmatrix}$.

`G_2<-A%*%A%*%A;` Matriz A multiplicada por si misma tres veces (A^3), $\begin{pmatrix} 49 & 51 \\ 102 & 151 \end{pmatrix}$.

`H<-rbind(A,B);` Dispone la matriz B debajo de la matriz A , $H = \begin{pmatrix} 3 & 1 \\ 2 & 5 \\ 2 & 4 \\ 0 & 1 \end{pmatrix}$.

`H<-cbind(A,B);` Dispone la matriz B al lado de la matriz A , $H = \begin{pmatrix} 3 & 1 & 2 & 4 \\ 2 & 5 & 0 & 1 \end{pmatrix}$.

`K<-A%x%B;` Producto directo (Kronecker) entre A y B , $K = \begin{pmatrix} 6 & 12 & 2 & 4 \\ 0 & 3 & 0 & 1 \\ 4 & 8 & 10 & 20 \\ 0 & 2 & 0 & 5 \end{pmatrix}$.

`L<-B/A;` Divide cada elemento de B por el respectivo de A , $L = \begin{pmatrix} 0.67 & 4.00 \\ 0.00 & 0.20 \end{pmatrix}$.

`m<-6:10;` Genera el vector m con valores enteros entre 6 y 10, $m = \begin{pmatrix} 6 & 7 & 8 & 9 & 10 \end{pmatrix}$.

`n<-8:5;` Genera el vector n con valores enteros entre 8 y 5, $m = \begin{pmatrix} 8 & 7 & 6 & 5 \end{pmatrix}$.

```
library(Matrix);
```

`O<-as.matrix(bdiag(A,B));` Matriz en bloques; A y B en la diagonal

$$O = \begin{pmatrix} 3 & 1 & 0 & 0 \\ 2 & 5 & 0 & 0 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

`Det_A<-det(A);` Determinante de A , $Det\_A = 13$.

`diag(A);` Vector de tipo numérico cuyas componentes son la diagonal de A, $d = \begin{pmatrix} 3 & 5 \end{pmatrix}$.

`DG_A<-diag(diag(A));` Transforma la matriz A en una matriz diagonal,

$$DG\_A = \begin{pmatrix} 3 & 0 \\ 0 & 5 \end{pmatrix}.$$

`Inv_A<-solve(A);` Inversa de la matriz A, $Inv\_A = \begin{pmatrix} 0.3846 & -0.0769 \\ -0.1538 & 0.2308 \end{pmatrix}$. `T_A<-t(A);` Produce la transpuesta de A, $T\_A = \begin{pmatrix} 3 & 2 \\ 1 & 5 \end{pmatrix}$.

`Vap_C<-eigen(C)$values` Obtiene los valores propios de C,

$Vap\_C = \begin{pmatrix} 7.470 & 1.399 & -2.870 \end{pmatrix}$

`Vep_C<-eigen(C)$vectors` Obtiene los vectores propios de C,

$$Vep\_C = \begin{pmatrix} -0.61 & -0.04 & 0.79 \\ -0.45 & -0.80 & -0.39 \\ -0.65 & 0.60 & -0.47 \end{pmatrix}$$

`Filas_C<-nrow(C);` Cuenta el numero de filas de C , $Filas\_C = 3$.

`Colum_C<-ncol(C);` Cuenta el numero de columnas de C , $Colum\_C = 3$.

`Tra_C<-sum(diag(C));` Calcula la traza de la matriz C , $Tra\_C = 6$.

```
b<-matrix(c(3,2),ncol=1)
X<-solve(A,b)
```

resuelve el sistema $\begin{cases} 3x_1 + x_2 = 3 \\ 2x_1 + 5x_2 = 2 \end{cases}$ con $A = \begin{pmatrix} 3 & 1 \\ 2 & 5 \end{pmatrix}$ y $b = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$.

Solución $X = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

```
svd(A)$u
 svd(A)$v
 diag(svd(A)$d)
```

Encuentra la descompocición singular de la matriz A (A2.29) esta es:
$U = \begin{pmatrix} -0.416 & -0.909 \\ -0.909 & 0.416 \end{pmatrix}$, $V = \begin{pmatrix} -0.526 & -0.851 \\ -0.851 & 0.526 \end{pmatrix}$

y $\Delta = \begin{pmatrix} 5.834 & 0.000 \\ 0.000 & 2.228 \end{pmatrix}$

Apéndice B

Conceptos estadísticos básicos

B.1 Introducción

Se hace una breve revisión de los conceptos de la estadística univariada. El propósito es la explicación de algunos términos y la presentación de la notación utilizada en el texto.

Una parte está dedicada a la revisión de los modelos probabilísticos univariados básicos y la otra resume los tópicos relacionados con la inferencia estadística, desde un punto de vista clásico; aunque también se hace referencia a otras escuelas estadísticas.

B.2 Conceptos Probabilísticos

Un *espacio muestral* $\mathbb{S}$ es el conjunto de todos los posibles resultados de un experimento aleatorio ξ. Los elementos que conforman el espacio muestral se denominan *eventos o sucesos.*

Una *variable aleatoria* (*va*) es una función para la cual su dominio son los elementos del espacio muestral, y su rango el conjunto de todos los números reales. Dicho de otra manera, es una función del espacio muestral, en los números reales. Alternamente, X es una variable aleatoria si para cada número real x existe una probabilidad tal que el valor asumido por la variable aleatoria no exceda a x, notada por $P(X \leq x)$ o por $F_X(x)$, y llamada *función de distribución acumulada* de X (*fda*).

Es común notar la variable aleatoria con una letra mayúscula como X y el valor asumido por ella con su correspondiente letra minúscula x. Así

que la expresión $X = x$, significa que el valor asignado por la variable aleatoria X, a un evento s del espacio muestral, es x; en notación funcional se debe escribir $X(s) = x$, pero la manera usual en la literatura estadística es $X = x$.

Una variable aleatoria es *discreta* si su recorrido es un conjunto finito o infinito numerable. En forma práctica, una variable aleatoria es discreta si entre dos valores cualesquiera de la variable hay siempre un número finito de posibles valores.

Una variable aleatoria es *continua* si su recorrido es un intervalo de la recta numérica.

Las propiedades matemáticas de cualquier función de distribución acumulada $F_X(\cdot)$ de la varaible aleatoria X son las siguientes

i) $F_X(x_1) \leq F_X(x_2)$ para todo $x_1 \leq x_2$.

ii) $\lim_{x\to-\infty} F_X(x) = 0$ y $\lim_{x\to\infty} F_X(x) = 1$.

iii) $F_X(x)$ es continua por la derecha, es decir, si

$$\lim_{\varepsilon\to 0^+} F_X(x+\varepsilon) = F_X(x;)$$

donde $\varepsilon \to 0^+$ significa "acercarse" a 0 por el lado positivo de la recta numérica.

Una variable aleatoria es continua si su fda es continua. Ésta es otra forma de definir variable aleatoria continua. Se asume que la fda para variables aleatorias continuas es diferenciable excepto en un número finito de puntos. La derivada de $F_X(x)$, notada por $f_X(x)$, es una función no negativa llamada *función de densidad* (fdp) de X. Así cuando X es continua

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du;\ f_X(x) = \frac{d}{dx}F_X(x) = F'(x) \text{ y } \int_{-\infty}^{\infty} f_X(x)dx = 1. \tag{B.1}$$

Para variables aleatorias *discretas*, se define la *función de probabilidad* o *función de masa* por

$$f_X(x) = P(X = x) = F_X(x) - \lim_{\varepsilon\to 0} F_X(x-\varepsilon)$$

además, se debe tener que $\sum_{\forall x} f_X(x) = 1$.

Ejemplo B.1

• Supóngase que "dentro de un cuadrado de lado a se lanzan dos monedas normales". Sobre esta acción se pueden definir varios experimentos aleatorios como los siguientes:

- ξ_1: "Se observa el resultado que aparece sobre la cara superior de las monedas".

- ξ_2: Asumiendo que las monedas caen dentro del cuadrado, "se mide la distancia entre sus centros"

- El espacio muestral asociado a ξ_1 es $\mathbb{S}_{\xi_1} = \{CC, CS, SC, SS\}$; donde CS significa que aparece cara en una moneda y sello en la otra.

- El espacio muestral asociado a ξ_2 es $\mathbb{S}_{\xi_2} = \{d \in \mathbb{R} : 0 \leq d \leq a\sqrt{2}\} = [0, a\sqrt{2}]$.

- Para el primer espacio muestral $\mathbb{S}_{\xi_1}$ considérese la variable aleatoria X: "Número de caras obtenidas en un lanzamiento". En este caso $X(CC) = 2$, $X(CS) = 1$, $X(SC) = 1$ y $X(SS) = 0$. Es decir el espacio muestral $\mathbb{S}_{\xi_1}$ se ha transformado en el conjunto $\{0, 1, 2\}$ mediante la variable aleatoria discreta X.

- La tabla contiene la función de probabilidad para la variable aleatoria X.

x	0	1	2
P(X=x)	1/4	1/2	1/4

Así, $P(CS \text{ o } SC) = P(X = 1) = \frac{1}{2}$ ☑.

• Sea X la variable aleatoria "la duración" (en unidades de 100 horas) de cierto artefacto electrónico. Supóngase que X es una variable aleatoria continua y que la *fdp* f está dada por

$$f(x) = \begin{cases} 2e^{-2x}, & x > 0 \\ 0, & \text{en otro caso.} \end{cases}$$

- La probabilidad de que un artefacto de éstos dure más de una unidad de tiempo (100 horas) es

$$\begin{aligned} P(X > 1) &= 1 - P(X \leq 1) = 1 - \int_0^1 2e^{-2x} dx \\ &= 1 - \left(-e^{-2x}\right)\Big|_0^1 = 1 + e^{-2} - 1 \\ &= e^{-2} = 0.1353 \text{ ☑} \end{aligned}$$

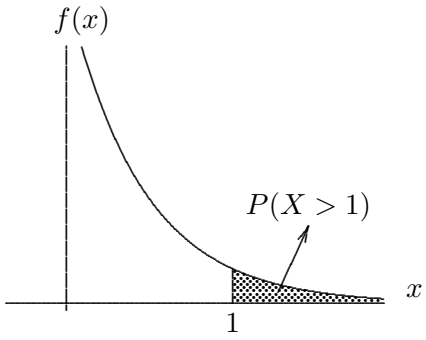


Figura B.1 *Función de densidad.*

Sobre una distribución se registran algunas caraterísticas tales como localización, dispersión, apuntamiento, simetría, entre otras. La cantidad que mide estas características sobre una distribución se le denomina *parámetro.* Enseguida se definen algunos parámetros de interés frecuente.

El *valor esperado* de la función $g(x)$ de variable aleatoria X, notado por $\mathcal{E}\{(g(x)\}$, es

$$\mathcal{E}\{g(x)\} = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) dx, & \text{si } X \text{ es continua} \\ \\ \sum_{\forall x} g(x) f_X(x), & \text{si } X \text{ es discreta,} \end{cases} \tag{B.2}$$

en particular, para $g(x) = X$ se tiene que su valor esperado, se nota μ_X.

El *k-ésimo momento* de la variable aleatoria X se define como $\mathcal{E}(X^k) = \mu'_k$, y su *k-ésimo momento, centrado en la media* μ, por

$$\mu_k = \mathcal{E}\{(X - \mu)^k\}. \tag{B.3}$$

Para $k = 2$ resulta $\mathcal{E}\{(X-\mu)^2\} = \text{var}(X) = \sigma^2$, es la varianza de la variable aleatoria X. Al desarrollar el cuadrado, la varianza de X es igual a

$$\text{var}(X) = \sigma_X^2 = \mu_2 = \mathcal{E}(X^2) - \mu^2. \tag{B.4}$$

Ejemplo B.2 Para la variable aleatoria continua X anterior, el valor esperado es

$$\mathcal{E}(X) = \int_0^{\infty} x f(x) dx = \int_0^{\infty} 2xe^{-2x} dx.$$

Integrando por partes y haciendo $u = x$ y $dv = 2e^{-2x}dx$, se obtiene que $v = -e-2x$, y $du = dx$; luego

$$\mathcal{E}(X) = \mu = \left[-xe^{-2x}\right]\Big|_0^{\infty} + \int_0^{\infty} e^{-2x} dx = \frac{1}{2}.$$

Es decir, el promedio de duración de estos artefactos es de 0.5 unidades de tiempo (50 horas)

La varianza de X se obtiene, de acuerdo con (B.4), de manera semejante. Se encuentra que $\mathcal{E}(X^2) = 1/2$, luego

$$\text{var}(X) = \mathcal{E}(X^2) - \mu^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4} \quad \boxtimes.$$

Hasta ahora se puede afirmar que si se conoce la distribución de probabilidades de una variable (discreta o continua), se pueden calcular $\mathcal{E}(X)$ y $\text{var}(X)$, si existen. El recíproco no siempre es cierto; es decir, conociendo $\mathcal{E}(X)$ y $\text{var}(X)$ no se puede reconstruir la distribución probabilidades de X. No obstante, se pueden obtener algunos valores aproximados, en probabilidad, para la concentración de una variable en torno a su media. Así, valores tales como $P(|X - \mu| \leq c)$, se calculan mediante la conocida *desigualdad de Chebyshev*, a continuación se enuncia esta desigualdad en una versión útil para los propósitos de este texto.

∘ *Desigualdad de Chebyshev.*

Sea X una variable aleatoria, con $\mathcal{E}(X) = \mu$ y $\text{var}(X) = \sigma^2$, con valores finitos, entonces, para cualquier valor k positivo

$$P[|X - \mu| \leq k\sigma] \geq 1 - \frac{1}{k^2}.$$

Por ejemplo, para $k = 2$, se puede afirmar que "la probabilidad de que la variable aleatoria X difiera de la media máximo en $k\sigma$ es al menos 0.75 (75%)", cualquiera que sea la distribución de X.

Para las dos variables aleatorias X y Y, su *covarianza* y *correlación*, son, respectivamente

$$\begin{aligned} \text{cov}(X,Y) &= \sigma_{XY} = \mathcal{E}\{(X - \mu_X)(Y - \mu_Y)\} = \mathcal{E}(XY) - \mu_X\mu_Y \\ corr(X,Y) &= \rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}. \end{aligned} \tag{B.5}$$

El coeficiente de correlación lineal ρ_{XY} es una cantidad adimensional, que toma valores entre -1 y 1; es decir, $|\rho_{XY}| \leq 1$. Valores próximos a $+1$ o a -1, sugieren la existencia de una asociación lineal entre las variables X y Y.

La *función generadora de momentos* (*fgm*) de la variable aleatoria X es

$$M_X(t) = \mathcal{E}\{\exp(tX)\}, \tag{B.6}$$

y recibe este nombre porque

$$\mu'_k = \mathcal{E}(X^k) = \frac{d^k}{dt^k} M_X(t)\Big|_{t=0} = M_X^{(k)}(0); \tag{B.7}$$

es decir, la k-ésima derivada de la fgm de X calculada en 0, es el momento de orden k centrado en 0 para la variable aleatoria X.

◦ Distribución Uniforme

Si X es una variable aleatoria continua que toma todos los valores en el intervalo $[a, b]$, a y b finitos, tiene distribución *uniforme* si su *fdp* está dada por

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{en otra parte.} \end{cases}$$

Se nota $X \sim U[a, b]$. La figura B.2 muestra esta función de densidad.

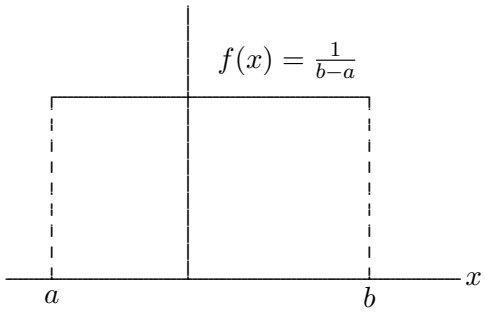


Figura B.2 *Distribución uniforme.*

◦ Distribución Normal

Si una variable aleatoria X tiene como fdp la siguiente expresión

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty, \tag{B.8}$$

entonces se dice que la variable aleatoria X tiene *distribución normal* de media μ y varianza σ^2 y se nota $X \sim n(\mu, \sigma^2)$.

Si $\mu = 0$ y $\sigma^2 = 1$, entonces X tiene distribución normal *tipificada* o *estándar* y se nota $X \sim n(0, 1)$. La figura B.3 muestra la función de densidad normal y la función de densidad acumulada normal.

Una propiedad importante de la distribución normal es la siguiente:

- Si $X \sim n(\mu, \sigma^2)$ entonces la variable aleatoria $Z = \frac{X-\mu}{\sigma} \sim n(0,1)$.

La transformación anterior se llama *estandarización* de la variable aleatoria X. La utilidad de la estandarización es que sirve para calcular las probabilidades asociadas a una distribución normal cualquiera, así,

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P(Z \leq z) = \Phi(z).$$

De lo anterior se tiene que $P(Z \geq z) = 1 - \Phi(z) = \Phi(-z)$.

La tabla C.5 contiene las probabilidades acumuladas hasta un cierto valor; para valores entre -3.00 y 3.00.

La función generadora de momentos asociada a una variable cuya distribución es normal, viene dada por: $M_X(t) = e^{\mu t + \frac{1}{2}t^2\sigma^2}$. De aquí se deducen el primero (μ) y segundo momento (μ_2') de la distribución normal; así,

$$\frac{d\big(M_X(t)\big)}{dt}\bigg|_{t=0} = \frac{d\big(e^{\mu t + \frac{1}{2}t^2\sigma^2}\big)}{dt}\bigg|_{t=0} = (\mu + t\sigma^2)(e^{\mu t + \frac{1}{2}t^2\sigma^2})\Big|_{t=0} = \mu$$

y

$$\begin{aligned}\frac{d^2\big(M_X(t)\big)}{dt^2}\bigg|_{t=0} &= \frac{d^2\big(e^{\mu t + \frac{1}{2}t^2\sigma^2}\big)}{dt^2}\bigg|_{t=0} \\ &= (\sigma^2)(e^{\mu t + \frac{1}{2}t^2\sigma^2}) + (\mu + t\sigma^2)^2(e^{\mu t + \frac{1}{2}t^2\sigma^2})\Big|_{t=0} = \sigma^2 + \mu^2 = \mu_2',\end{aligned}$$

respectivamente.

∘ Distribución ji-cuadrado

Si $Z_1, Z_2, \cdots, Z_p$ son variables aleatorias independientes con distribución $n(0,1)$, entonces, la variable aleatoria

$$U = Z_1^2 + Z_2^2 + \cdots + Z_p^2 = \sum_{i=1}^{p} Z_i^2, \tag{B.9}$$

tiene *distribución ji-cuadrado* con p grados de libertad; se nota $U \sim \chi^2_{(p)}$.

La función de densidad de probabilidad de U está dada por

$$f(u) = \frac{1}{2^{p/2}\Gamma\big(\frac{1}{2}p\big)} u^{\frac{p}{2}-1} e^{-u/2}; \quad 0 < u < \infty. \tag{B.10}$$

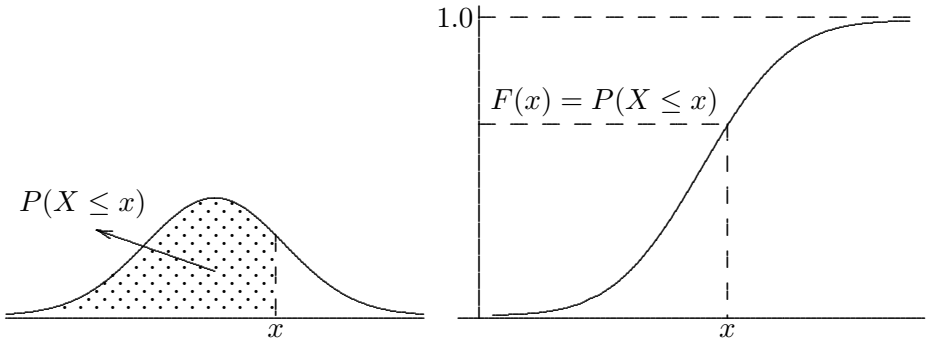


Figura B.3 *Distribución normal.*

La expresión $\Gamma(x)$ representa la *función Gama*, la cual está definida por

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du, \quad x > 0. \tag{B.11}$$

Al integrar por partes en (B.11) se obtiene $\Gamma(x+1) = x\Gamma(x)$ para $x > 0$, en particular $\Gamma(p+1) = p!$, para p número entero positivo.

Se prueba que $\mathcal{E}(U) = p$ y que $\text{var}(U) = \sigma_{\text{U}}^2 = 2p$. La figura B.4 muestra la gráfica de una función de densidad tipo ji-cuadrado, se observa que para cada valor de p está asociada una distribución ji-cuadrado, y por tanto una gráfica por cada valor de éste. En la tabla C.7 se muestran los cuantiles asociados con algunos grados de libertad y con ciertos valores de probabilidad para esta distribución.

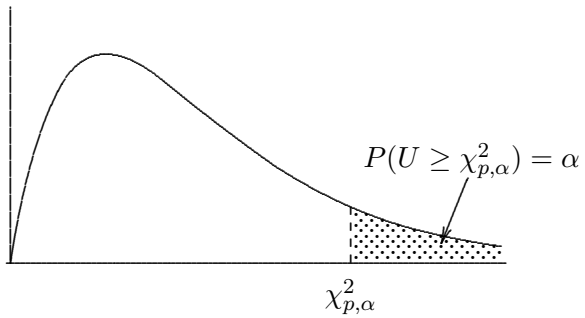


Figura B.4 *Distribución ji-cuadrado.*

◦ Distribución t-Student

Si Z y U son variables aleatorias independientes tales que $Z \sim n(0,1)$ y $U \sim \chi^2_{(p)}$, entonces, la variable aleatoria

$$t = \frac{Z}{\sqrt{U/p}}, \tag{B.12}$$

tiene *distribución t-Student*, con p grados de libertad; se nota $t \sim t_{(p)}$.

Su fdp está dada por

$$f(t) = \frac{\Gamma\left(\frac{1}{2}p + \frac{1}{2}\right)}{(p\pi)^{\frac{1}{2}}\Gamma\left(\frac{1}{2}p\right)}(1 + t^2/p)^{-\frac{(p+1)}{2}}. \tag{B.13}$$

Para $p > 1$ se tiene que $\mathcal{E}(t) = 0$ y para $p > 2$, $\text{var}(t) = \sigma_t^2 = p/(p-2)$.

La tabla C.6 contiene los cuantiles asociados con algunos grados de libertad y con varios valores de probabilidad para esta distribución.

◦ Distribución F

Si U_1 y U_2 son dos variables aleatorias independientes, con distribuciones ji-cuadrado de p_1 y p_2 grados de libertad, respectivamente, entonces, la variable aleatoria

$$F = \frac{U_1/p_1}{U_2/p_2}, \tag{B.14}$$

tiene *distribución F* con (p_1, p_2) grados de libertad y se nota $F \sim F_{(p_1,p_2)}$.

La fdp de F está dada por

$$g(f) = \frac{\Gamma\left(\frac{p_1+p_2}{2}\right)p_1^{p_1/2}p_2^{p_2/2}}{\Gamma\left(\frac{p_1}{2}\right)\Gamma\left(\frac{p_2}{2}\right)}\frac{f^{\frac{p_1}{2}-1}}{(p_1 f + p_2)^{(p_1+p_2)/2}}. \tag{B.15}$$

Se puede notar que si $t \sim t_{(p)}$ entonces $t^2 \sim F_{(1,p)}$. En la tabla C.8 se presentan los cuantiles asociados con algunos pares de grados de libertad y con ciertos valores de probabilidad para la distribución F.

◦ Distribución Gama

Se dice que la variable aleatoria X tiene *distribución Gama*, con parámetros $\alpha > 0$ y $\beta > 0$, si su función de densidad de probabilidad está dada por

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}, \quad 0 < x < \infty. \tag{B.16}$$

Se observa que para $\alpha = p/2$ y $\beta = 2$ se obtiene la distribución ji-cuadrado con p grados de libertad.

La media y la varianza de una variable aleatoria, con distribución Gama, son:

$$\mathcal{E}(X) = \mu = \alpha\beta \text{ y } \text{var}(X) = \sigma^2 = \alpha\beta^2, \tag{B.17}$$

respectivamente. Nótese que si $\alpha = 1$ se tiene la *distribución exponencial*. En la distribución exponencial del ejemplo B.1 el parámetro β es $\frac{1}{2}$.

◦ Distribución Beta

Se define la *función Beta* por la integral

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx; \quad \text{donde } \alpha > 0 \text{ y } \beta > 0. \tag{B.18}$$

Una propiedad de la función Beta es que $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$.

La variable aleatoria X tiene *distribución Beta* si su fdp se puede expresar como

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}; \quad 0 < x < 1. \tag{B.19}$$

Un caso especial de la distribución Beta es cuando $\alpha = \beta = 1$, la cual corresponde a la *distribución uniforme*.

El valor esperado y la varianza de una variable aleatoria con distribución Beta son, respectivamente,

$$\mathcal{E}(X) = \frac{\alpha}{\alpha + \beta} \quad y \quad \text{var}(X) = \sigma_2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{B.20}$$

A continuación se presentan algunas distribuciones ligadas a variables aleatorias discretas.

◦ Distribución Bernoulli

Una varaible aleatoria X tiene distribución *Bernoulli* si la función de probabilidad discreta de X es dada por

$$f(x) = \begin{cases} p^x(1-p)^{1-x}, & \text{para } x = 0, 1, \\ 0, & \text{en otra parte.} \end{cases}$$

El parámetro p satisface la siguiente relación $0 \leq p \leq 1$. Para esta distribución se demuestra que

$$\mathcal{E}(X) = p \text{ y } \text{var}(X) = p(1-p).$$

◦ Distribución Binomial

Una variable aleatoria X tiene distribución *binomial* si su función de probabilidad está dada por

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{para } x = 0, 1, \ldots, n, \\ 0, & \text{en otra parte,} \end{cases}$$

donde los dos parámetros n y p son tales que n es un entero no negativo y $0 \leq p \leq 1$. Se nota $X \sim B(n, p)$.

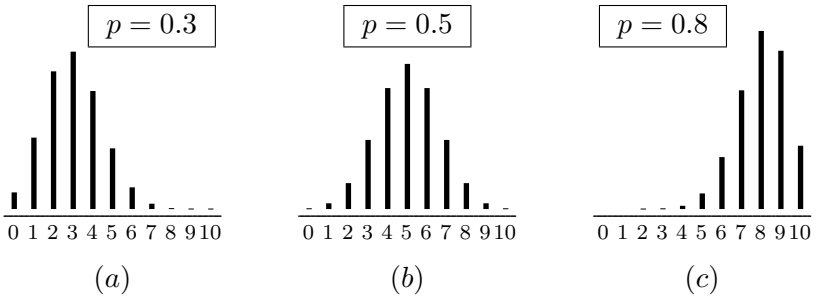


Figura B.5 *Distribución binomial.*

La figura B.5 muestra tres casos especiales de esta distribución con el mismo valor $n = 10$ y $p = 0.3,\ 0.5\ y\ 0.8$, respectivamente. La distribución de la figura B.5(a) corresponde a $p = 0.3$ la cual es sesgada a la derecha, para $p = 0.5$; la figura B.5(b) representa la distribución simétrica entorno a su media $\mu = np = 5$; finalmente, la distribución para $p = 0.8$ es sesgada hacia la izquierda como se muestra en la figura B.5(c). Nótese que el sesgo se tiene para valores de p diferentes de 0.5. Para valores de n suficientemente grandes, y cualquier valor de p, la distribución tiende a ser simétrica en torno a su media.

◦ Distribución de Poisson

Una variable aleatoria que toma los valores $0, 1, 2, \ldots$ tiene una distribución de *Poisson* si su función de probabilidad está dada por:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!};\ \text{para } x = 0, 1, 2, \ldots$$

Se escribe $X \sim P(\lambda)$ para indicar que X tiene distribución de Poisson con parámetro λ. Una característica de esta variable aleatoria es que: $\mathcal{E}(X) = \text{var}(X) = \lambda$.

B.3 Inferencia

Con un propósito didáctico, se presentan los conceptos de población y muestra; elementos ligados a la inferencia estadística, desde una óptica más hacia lo descriptivo y lo práctico.

Los valores que toman una o más variables respecto a uno o varios atributos considerados sobre un conjunto de objetos en estudio, se denomina *población*[1]. La población queda determinada por la distribución que tome la o las variables a estudiar.

Se considera *muestra* a un subconjunto de valores de una población. La bondad de una muestra radica en la cantidad de información que ella contenga o represente de la población. Estadísticamente se garantiza tal representatividad cuando cada valor tenga, independientemente de los demás, una misma probabilidad de ser seleccionado; tal muestra se llama *muestra aleatoria*.

El proceso mediante el cual se extraen conclusiones de una población, partiendo de la información contenida en una muestra, se llama *inferencia* estadística.

Los procedimientos de inferencia estadística pueden clasificarse en función de los supuestos de la inferencia y en función del tipo de información que utilicen.

a) Respecto a los supuestos: *Métodos paramétricos frente a no paramétricos.*

Los métodos paramétricos suponen que los datos provienen de una distribución conocida cuyos parámetros se desean estimar. Los métodos no paramétricos no requieren del conocimiento de la distribución y solamente introducen hipótesis muy generales respecto a ésta (continuidad, simetría, etc.), para estimar su forma o contrastar su estructura.

b) Respecto a la información utilizada: *Métodos clásicos frente a bayesianos.*

[1] Generalmente se confunde conjunto de objetos con conjunto de valores.

Los métodos clásicos suponen que los parámetros son cantidades fijas desconocidas y que la única información existente respecto a ellos está contenida en la muestra. Los métodos bayesianos consideran a los parámetros como variables aleatorias y permiten introducir información a priori sobre ellos a través de una distribución a priori.

Los métodos clásicos ofrecen una respuesta simple a una mayoría de problemas de inferencia; tal respuesta es sustancialmente análoga a la obtenida con el enfoque bayesiano suponiendo poca información a priori. El enfoque clásico es más adecuado en la etapa de crítica del modelo, donde se pretende que los datos muestren por sí solos la información que contienen, con el menor número de restricciones posibles.

Una *estadística* es una función de variables aleatorias observables, la cual es también una variable aleatoria que no contiene parámetros desconocidos.

En general, se nota a un parámetro mediante θ, donde θ puede ser un escalar, un vector de parámetros; también una función de θ, $\tau(\theta)$, la cual es nuevamente un parámetro.

Al conjunto de valores, que puede asumir θ, se llama *espacio de parámetros* y se nota por $\boldsymbol{\Omega}$. Así por ejemplo, para una variable aleatoria con distribución Poisson; es decir, $X \sim P(\lambda)$, su espacio de parámetros es $\boldsymbol{\Omega} = \mathbb{R}^+$. Para la variable aleatoria $X \sim n(\boldsymbol{\mu}; \sigma^2)$,

$$\boldsymbol{\Omega} = \{(\boldsymbol{\mu}, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ : -\infty < \boldsymbol{\mu} < \infty,\ \sigma^2 > 0\}.$$

Cualquier estadística cuyos valores se usen para estimar una función $\tau(\theta)$ del parámetro θ se define como un *estimador* de $\tau(\theta)$, donde $\tau(\theta)$ es una función del parámetro θ. Aunque, usualmente $\tau(\theta) = \theta$.

Se nota por $\widehat{\tau}(\theta)$ o $\widehat{\theta}$ a los estimadores de $\tau(\theta)$ y θ, respectivamente.

▶ Propiedades de un estimador

La pretención central de la inferencia estadística es acercarse al conocimiento de un parámetro a través de la información muestral. En este intento influyen, entre otros, los siguientes aspectos

- El muestreo.
- El diseño de muestreo.
- El tamaño de la muestra.
- Los supuestos permisibles sobre la población.

- El procedimiento de estimación.
- Los mismos datos que conforman la muestra tales como la presencia de datos atípicos (outliers), datos faltantes, entre otros.

La bondad de cualquier procedimiento de estimación generalmente se mide en términos de la distancia entre un estimador y el parámetro objetivo. Esta cantidad, que varía de una manera aleatoria en un muestreo repetitivo, se denomina *error de estimación.* Lo deseable es un error de estimación lo más pequeño posible.

Definición: El *error de estimación* (ε) es la distancia (euclidiana) entre el estimador y su parámetro objetivo

$$\varepsilon = |\widehat{\theta} - \theta|, \tag{B.21}$$

como esta cantidad es de carácter aleatorio, no se puede anticipar su valor para una estimación particular, pero en cambio, si se puede asignar una cota a sus valores en forma probabilística.

En este sentido se dice que el estimador $\widehat{\theta}_1$ del parámetro θ es más concentrado que el estimador $\widehat{\theta}_2$ del mismo parámetro, si y sólo si,

$$P_\theta(|\widehat{\theta}_1 - \theta| < \varepsilon) \geq P_\theta(|\widehat{\theta}_2 - \theta| < \varepsilon), \tag{B.22}$$

para todo $\varepsilon > 0$.

Observaciones:

- Hasta ahora se ha asumido el tamaño de muestra fijo. Se puede proponer un valor para ε y otro para P_θ y obtener el tamaño de muestra n.
- Las pruebas de bondad de ajuste miden el error de estimación entre la distribución de frecuencias observada y la distribución que se supone genera los datos.

Una medida útil de la bondad de un estimador $\widehat{\theta}$ de θ es *el cuadrado medio del error.*

Definición: El *cuadrado medio del error* (CME), de un estimador puntual $\widehat{\theta}$ es igual a la media de la desviación cuadrática del estimador respecto al parámetro; es decir:

$$CME_\theta = \mathcal{E}(\widehat{\theta} - \theta)^2. \tag{B.23a}$$

Al desarrollar el cuadrado sobre (B.23a) y aplicar valor esperado se obtiene:

$$CME_\theta = V(\widehat{\theta}) + \mathcal{E}\big[(\widehat{\theta}) - \theta\big]^2, \tag{B.23b}$$

ésta es una medida de la dispersión de $\widehat{\theta}$ respecto a θ, semejante a la varianza de una variable aleatoria, la cual es una medida de la dispersión alrededor de su media. Un estimador $\hat{\theta}_1$ es mejor, *en cuadrado medio*, que un estimador $\hat{\theta}_2$ si $CME_{\hat{\theta}_1} < CME_{\hat{\theta}_2}$

La propiedad de menor error de estimación o de error cuadrático medio no es concluyente sobre la buena calidad del estimador, es necesario reunir otras características.

Definición: Un estimador $\widehat{\theta}$ de θ es *insesgado* o *centrado* si $\mathcal{E}(\widehat{\theta}) = \theta$. A la cantidad $\mathcal{E}(\widehat{\theta}) - \theta = B$ se denomina *sesgo* de $\widehat{\theta}$. Si $\widehat{\theta}$ es insesgado, $B = 0$ y de acuerdo con la última definición $CME = V(\widehat{\theta})$.

Ejemplo B.3

1. Sea $X_1, \cdots, X_n$ una muestra aleatoria, de una población $n(\mu, \sigma^2)$ con

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}, \qquad s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n},$$

estimadores de $\boldsymbol{\mu}$ y σ^2 respectivamente, entonces, $\bar{X}$ es un estimador insesgado de μ. En efecto,

$$\mathcal{E}(\bar{X}) = \mathcal{E}\Big(\sum_{i=1}^{n} \frac{X_i}{n}\Big) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}(X_i) = \frac{1}{n} n\mu = \mu.$$

De otra parte,

$$\begin{aligned}
\mathcal{E}\left\{\sum_{i=1}^{n}(X_i - \bar{X})^2\right\} &= \mathcal{E}\left\{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu} + \boldsymbol{\mu} - \bar{X})^2\right\} \\
&= \mathcal{E}\left\{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^2 + 2\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})(\boldsymbol{\mu} - \bar{X}) + \sum_{i=1}^{n}(\boldsymbol{\mu} - \bar{X})^2\right\} \\
&= \mathcal{E}\left\{\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})^2 + 2(\boldsymbol{\mu} - \bar{X})\sum_{i=1}^{n}(X_i - \boldsymbol{\mu}) + n(\boldsymbol{\mu} - \bar{X})^2\right\}
\end{aligned}$$

$$
\begin{aligned}
&= \mathcal{E}\left\{\sum_{i=1}^{n}(X_i-\boldsymbol{\mu})^2 - 2n(\bar{X}-\boldsymbol{\mu})^2 + n(\bar{X}-\boldsymbol{\mu})^2\right\}\\
&= \mathcal{E}\left\{\sum_{i=1}^{n}(X_i-\boldsymbol{\mu})^2\right\} - \mathcal{E}\left\{n(\bar{X}-\boldsymbol{\mu})^2)\right\}\\
&= n\sigma^2 - \sigma^2\\
&= \sigma^2(n-1),
\end{aligned}
$$

de donde resulta que s^2 no es un estimador insesgado de σ^2. Si se define a s^2 como:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)},$$

resulta un estimador insesgado de σ^2.

2. Sea $X_1, \cdots, X_n$ una muestra aleatoria de una distribución uniforme en $[0;\theta]$. Un estimador de θ es $\widehat{\theta} = X_{max}$

$$\mathcal{E}(X_{max}) = \frac{n}{n+1}\theta.$$

La función de distribución de la variable aleatoria "el valor máximo de una variable aleatoria", X_{max} es $F_{X_{max}}(x) = [F_X(x)]^n$.

Entonces, $f_{X_max}(x) = n[F_X]^{n-1}f_X(x) = n\left(\frac{x}{\theta}\right)^{n-1}\left(\frac{1}{\theta}\right)$; de donde

$$
\begin{aligned}
\mathcal{E}(X_{max}) &= \int_0^\theta x\left(n\left(\frac{x}{\theta}\right)^{n-1}\left(\frac{1}{\theta}\right)\right)dx\\
&= n\int_0^\theta \frac{x^n}{\theta^n}dx\\
&= \frac{n}{n+1}\theta;
\end{aligned}
$$

así, el *máximo*, es un estimador *sesgado* de θ, con $B = \theta\left(\frac{n}{n+1} - 1\right)$.

3. En el modelo lineal de rango completo

$$Y = X\beta + \varepsilon \qquad con\ \mathcal{E}(\varepsilon) = 0 \text{ y } \ \mathrm{Cov}(\varepsilon) = \sigma^2\boldsymbol{I}_n$$

Donde Y es un vector aleatorio $(n \times 1)$ de observaciones , X es una matriz de tamaño $(n \times p)$, β un vector de parámetros con tamaño $(p \times 1)$, y ε un vector aleatorio no observable de tamaño $(n \times 1)$.

Un estimador de β está dado por $\widehat{\beta} = (X'X)^{-1}X'Y$, y resulta ser insesgado. En efecto,

$$
\begin{aligned}
\mathcal{E}(\widehat{\beta}) &= \mathcal{E}\{(X'X)^{-1}X'(X\beta+\varepsilon)\}\\
&= \mathcal{E}\{(X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon\} = \beta. \quad \boxtimes
\end{aligned}
$$

◦ *Eficiencia*

Entre todos los estimadores de un parámetro θ aparece el problema de escoger aquel cuyos valores posibles sean muy cercanos al del parámetro θ, una forma de medir la "bondad" del estimador $\widehat{\theta}$, o bien de medir el "riesgo" de obtener valores muy distantes de θ, consiste en considerar su varianza

$$\text{var}(\widehat{\theta}) = \sigma^2_{\widehat{\theta}} = \mathcal{E}(\widehat{\theta} - \mathcal{E}(\widehat{\theta}))^2.$$

Definición: La *eficiencia (EF)* o precisión de un estimador $\widehat{\theta}$ de θ está dada por

$$EF(\widehat{\theta}) = \frac{1}{\sigma^2_{\widehat{\theta}}}. \tag{B.25}$$

Sean $\widehat{\theta}_1$ y $\widehat{\theta}_2$ dos estimadores de θ, se dice que $\widehat{\theta}_1$ es *más eficiente* que $\widehat{\theta}_2$ si $\sigma^2_{\widehat{\theta}_1} < \sigma^2_{\widehat{\theta}_2}$. El recíproco de esta proposición también es cierto, por tanto,

$$\sigma^2_{\widehat{\theta}_2} \geq \sigma^2_{\widehat{\theta}_1}, \text{ si y solo si, } EF(\widehat{\theta}_1) \geq EF(\widehat{\theta}_2).$$

Se llama *eficiencia relativa (ER)* de $\widehat{\theta}_1$ respecto a $\widehat{\theta}_2$ al cociente entre ellas; es decir,

$$ER = \frac{\sigma^2_{\widehat{\theta}_1}}{\sigma^2_{\widehat{\theta}_2}}. \tag{B.26}$$

La eficiencia es especialmente útil para decidir sobre estimadores insesgados; pues entre ellos se prefiere el más eficiente.

Ejemplo B.4 Supongase que $\widehat{\theta}_1$ y $\widehat{\theta}_2$ son la media muestral y la mediana muestral, respectivamente; las cuales se consideran como estimadores de la media θ. Se puede comprobar que

$$\text{var}(\widehat{\theta}_2) = (1.2533)^2 \frac{\sigma^2}{n}.$$

Por lo tanto, la eficiencia de la mediana repecto a la media muestral (ER) es:

$$ER = \frac{\sigma^2_{\widehat{\theta}_1}}{\sigma^2_{\widehat{\theta}_2}} = \frac{\sigma^2/n}{(1.2533)^2 \dfrac{\sigma^2}{n}} = 0.6366,$$

así, la variabilidad aproximada de la media es el 64% de la variabilidad asociada con la mediana de la muestra; así, la estadística $\hat{\theta}_1$ es más eficiente que la estadística $\hat{\theta}_2$. ☑

Definición: un estimador *insesgado de mínima varianza y uniforme* (en inglés UMVUE) $\widehat{\theta}$ de θ, es aquel para el cual se satisfacen las siguientes propiedades:

$i)$ $\mathcal{E}(\widehat{\theta}) = \theta,$ insesgado

$ii)$ $\text{var}(\widehat{\theta}) \leq \text{var}(\widehat{\theta}^*),$ para todo $\widehat{\theta}^*$ estimador insesgado de θ.

A veces es infructuoso buscar la varianza de cada uno de los estimadores insesgados de θ, en cambio puede ser útil conocer el valor mínimo que su varianza puede tomar; este valor se conoce como *la cota inferior de Cramer-Rao*, la cual se define como sigue:

Sea $X_1, \cdots, X_n$ una muestra aleatoria de una distribución con una función (densidad) de probabilidad $f(X, \theta)$. Si $\widehat{\theta}$ es un estimador insesgado de θ, entonces, bajo ciertas condiciones de regularidad (Mood y colaboradores 1982, pág 315), la varianza de un estimador $\widehat{\theta}$ del parḿetro θ debe satisfacer la siguiente desigualdad,

$$\text{var}(\widehat{\theta}) \geq \frac{1}{n\mathcal{E}\left(\left(\dfrac{\partial \ln f(X,\theta)}{\partial \theta}\right)^2\right)}. \tag{B.27}$$

Observaciones:

- La desigualdad establece un límite inferior para la varianza de un estimador de θ.
- Lo anterior no implica que la varianza de un UMVUE de θ tenga que ser igual a la cota inferior de Cramer-Rao. Es decir, es posible encontrar un estimador insesgado de θ que tenga la varianza más pequeña entre todos los estimadores insesgados de θ, pero cuya varianza sea más grande que el límite inferior de Cramer-Rao.
- Para conseguir estimadores UMVUE se puede acudir al teorema *Lehmann-Scheffé* (Hogg y Craig, 1978 pág. 355).

Definición: Si $\widehat{\theta}$ es un estimador insesgado del parámetro θ tal que

$$\text{var}(\widehat{\theta}) = \frac{1}{n\mathcal{E}\left(\left(\dfrac{\partial \ln f(X,\theta)}{\partial \theta}\right)^2\right)}, \tag{B.27a}$$

entonces $\widehat{\theta}$ es un estimador *eficiente*.

Ejemplo B.5

1. Sea $X_1, \cdots, X_n$ una muestra aleatoria de una población $f(X, \theta) = \theta e^{-\theta x}$, entonces

$$\frac{\partial}{\partial \theta}\big(\ln f(x,\theta)\big) = \frac{\partial}{\partial \theta}\big(\ln \theta - \theta x\big) = \Big(\frac{1}{\theta} - x\Big)$$

luego

$$\mathcal{E}_\theta\left(\left(\frac{\partial \ln f(x,\theta)}{\partial \theta}\right)^2\right) = \mathcal{E}_\theta\big((1/\theta - x)^2\big) = \text{var}(X) = \frac{1}{\theta^2}$$

puesto que $\mathcal{E}(X) = \dfrac{1}{\theta}$, la cota de Cramer-Rao para la varianza del estimador insesgado de θ es

$$\text{var}(\widehat{\theta}) \geq \frac{1}{n\Big(\dfrac{1}{\theta^2}\Big)} = \frac{\theta^2}{n};$$

es decir, que $\bar{X}$ es un UMVUE de $1/\theta$, pues su varianza es igual a la cota inferior de Cramer-Rao.

2. Sea $X_1, \cdots, X_n$ una muestra aleatoria de una distribución de Poisson cuya función de probabilidad es

$$p(x, \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Para encontrar la cota de Cramer-Rao se procede a determinar el denominador de (B.27). A continuación se muestra el proceso.

Al aplicar logaritmos en los dos miembros de la igualdad anterior se obtiene

$$\ln\{p(x,\lambda)\} = -\lambda + x\ln(\lambda) - \ln(x!).$$

La derivada parcial respecto a λ es

$$\frac{\partial}{\partial \lambda}\ln\{p(x,\lambda)\} = \frac{x}{\lambda} - 1 = \frac{x-\lambda}{\lambda}.$$

El valor esperado del cuadrado del resultado anterior es

$$\mathcal{E}\left(\Big(\frac{x-\lambda}{\lambda}\Big)^2\right) = \frac{1}{\lambda^2}\,\text{var}(X) = \frac{1}{\lambda}.$$

De esta manera, la cota inferior de Cramer-Rao es

$$\text{var}(\widehat{\theta}) = \frac{1}{n\dfrac{1}{\lambda}} = \frac{\lambda}{n} = \frac{\sigma^2}{n},$$

luego, como ésta es la varianza de la media muestral $\bar{X}$, se concluye que un estimador eficiente de λ es $\hat{\lambda} = \bar{X}$. ☑

∘ *Consistencia*

Hasta ahora las propiedades han sido consideradas teniendo en cuenta una muestra de tamaño fijo, veamos cual es el comportamiento que se "espera" tenga un estimador $\widehat{\theta}$ de θ, teniendo en cuenta el tamaño de la muestra. Se escribe $\widehat{\theta}_n$ para señalar su dependencia con el tamaño de muestra.

La escritura anterior indica una sucesión de estimadores, por tanto se observará su comportamiento límite (convergencia). Se definen dos tipos de consistencia, así:

- *Consistencia en error cuadrático medio (ECM).* Sea $\widehat{\theta}_n$ una sucesión de estimadores de θ, donde $\widehat{\theta}_n$ se basa en una muestra aleatoria de tamaño n. La sucesión de estimadores se dice *consistente en error cuadrático medio*, si y sólo si,

$$\lim_{n\to\infty} \mathcal{E}_\theta\big((\widehat{\theta}_n - \theta)^2\big) = 0. \tag{B.28}$$

Observación:

De acuerdo con (B.23), la consistencia en error cuadrático medio implica que tanto la varianza como el sesgo de $\widehat{\theta}$ tienden a cero cuando el tamaño de muestra es suficientemente grande.

Propiedad

Si $\lim_{n\to\infty}\mathcal{E}(\widehat{\theta}_n) = \theta$ y $\lim_{n\to\infty}\text{var}(\widehat{\theta}_n) = 0$ entonces $\widehat{\theta}_n$ es consistente en error cuadrático medio. Recuérdese que

$$\begin{aligned} ECM &= \mathcal{E}(\widehat{\theta} - \theta)^2 \\ &= \text{var}(\widehat{\theta}) + (\mathcal{E}(\widehat{\theta}) - \theta)^2. \end{aligned}$$

- *Consistencia simple*

Sea $\widehat{\theta}_n$ una sucesión de estimadores de θ. La sucesión $\widehat{\theta}_n$ es consistente si satisface

$$\lim_{n\to\infty} P_\theta(|\widehat{\theta}_n - \theta| < \varepsilon) = 1, \text{ para todo } \varepsilon > 0. \tag{B.29}$$

Propiedad

Un estimador consistente en error cuadrático medio, es consistente; el recíproco no siempre es cierto.

Lo anterior es una consecuencia de la desigualdad de Chebyshev

$$P_\theta(|\widehat{\theta}_n - \theta| < \varepsilon) = P_\theta(|\widehat{\theta}_n - \theta| < \varepsilon^2) \geq 1 - \frac{\mathcal{E}(\theta_n - \theta)^2}{\varepsilon^2}.$$

El segundo término del lado izquierdo de la desigualdad tiende a 0 cuando n es suficientemente grande, ésto demuestra que la consistencia en ECM implica la consistencia simple.

Ejemplo B.6 Si $X_1, \cdots, X_n$ es una muestra aleatoria con $\mathcal{E}(X_i) = \boldsymbol{\mu}$ y $\text{var}(X_i) = \sigma^2$ finitas, para $i = 1, \cdots, n$, entonces, $\bar{X}_n$ es un estimador consistente de $\boldsymbol{\mu}$.

Se debe probar que $\lim_{n\to\infty} P(|\bar{X}_n - \boldsymbol{\mu}| < \varepsilon) = 1$ para todo $\varepsilon > 0$.

De la desigualdad de Chebyshev, como $\mathcal{E}(\bar{X}_n) = \boldsymbol{\mu}$ y $\sigma^2_{\bar{X}_n} = \dfrac{\sigma^2}{n}$, entonces,

$$P(|\bar{X}_n - \boldsymbol{\mu}| > k\frac{\sigma}{\sqrt{n}}) \leq \frac{1}{k^2},$$

sea $k = \dfrac{\varepsilon}{\sigma}\sqrt{n}$, entonces,

$$P(|\bar{X}_n - \boldsymbol{\mu}| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2},$$

como σ^2 es finito, entonces,

$$\lim_{n\to\infty} P(|\bar{X}_n - \boldsymbol{\mu}| > \varepsilon) = 0,$$

luego

$$\lim_{n\to\infty} P(|\bar{X}_n - \boldsymbol{\mu}| < \varepsilon) = 1.$$

• *Otro método*

Como $\mathcal{E}(\bar{X}_n) = \boldsymbol{\mu}$ y $\lim_{n\to\infty}\sigma^2_{\bar{X}_n} = \lim_{n\to\infty} \dfrac{\sigma^2}{n} = 0$, entonces, $\bar{X}_n$ es consistente en error cuadrático medio. ☑

∘ *Suficiencia*

De manera intuitiva, una estadística es *suficiente* para un parámetro θ si aquella utiliza toda la información contenida en la muestra aleatoria con respecto a θ.

Definición: Sea $X_1, \cdots, X_n$ una muestra aleatoria de una población cuya función de densidad es $f(X, \theta)$. Una estadística S es *suficiente*, si y sólo si, la distribución condicional de $X_1, \cdots, X_n$ dado $S = s$ no depende de θ para cualquier valor s de S.

De otra forma, si se afirma S, entonces, $X_1, \cdots, X_n$ no tiene más que decir respecto a θ

• *Otra definición*

Considere la estadística $S = T(x_1, \cdots, x_n)$ con fdp $k(s, \theta)$. La estadística S es una *estadística suficiente* para todo θ, si y sólo si,

$$\frac{f(x_1, \theta) \cdots f(x_n, \theta)}{k[T(x_1, \cdots, x_n); \theta]} = h(x_1, \cdots, x_n)$$

donde $h(x_1, \cdots, x_n)$ no depende de θ para cada valor s de S.

El siguiente criterio es útil para determinar si una estadística es suficiente.

► Teorema de factorización

Sea $X_1, \cdots, X_n$ una muestra aleatoria de una población $f(X, \theta)$. Una estadística $\mathcal{S}$ es suficiente, si y sólo si, la función de densidad conjunta de $X_1, \cdots, X_n$ se puede descomponer como

$$f(x_1, \cdots, x_n; \theta) = g(s, \theta) h(x, \cdots, x_n), \tag{B.30}$$

donde g sólo depende de la estadística $\mathcal{S}$ y del parámetro θ y h es independiente de θ.

► Propiedad

La transformación uno a uno de estadísticas suficientes, es suficiente.

Ejemplo B.7 Sea $X_1, \cdots, X_n$ una muestra aleatoria de una población con distribución de Poisson. Probar que $\widehat{\lambda} = \bar{X}$ es una estadística suficiente.

El procedimiento consiste en probar si la función de probabilidad conjunta se puede escribir conforme a la igualdad (B.30).

$$\begin{aligned} f(x_1, \cdots, x_n) &= \lambda^{\sum x_i} \exp(-n\lambda) \frac{1}{x_1! \cdots, x_n!} \\ &= g(\Sigma x_i, \lambda) \cdot h(x_1, \cdots, x_n) \end{aligned}$$

como $\bar{X}$ es una función uno a uno de Σx_i, que es suficiente, entonces $\bar{X}$ es suficiente.

▶ Estimación puntual y por intervalo

Un *estimador puntual* de un parámetro es cualquier función de las variables aleatorias cuyos valores observados son usados para estimar el verdadero valor del parámetro. De esta manera, si $X_1, \cdots, X_n$ es una muestra aleatoria de una población $f(x, \theta)$, entonces $\widehat{\theta} = T(X_1, \cdots, X_n)$ es un estimador puntual de θ.

Un método útil para encontrar estimadores puntuales de parámetros asociados a una distribución particular $f(x, \theta)$, es el de *máxima verosimilitud*. La función de verosimilitud de una muestra aleatoria de la población $f(\cdot; \theta)$ es la función de probabilidad conjunta de las variables muestrales en función de θ

$$L(x_1, \cdots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta). \tag{B.31}$$

La función de verosimilitud suministra la probabilidad de que una muestra aleatoria tome un valor particular $x_1, \cdots, x_n$. Para una muestra aleatoria dada, el problema se reduce a determinar el valor de θ ligado a la densidad $f(\cdot; \theta)$, de donde *muy probablemente* proviene la muestra.

El *estimador de máxima verosimilitud* (MV) de θ, es un valor de $\widehat{\theta}$ tal que

$$L(x_1, \cdots, x_n; \widehat{\theta}) \geq L(x_1, \cdots, x_n; \theta), \text{ para todo } \theta.$$

En muchos casos el estimador MV se obtiene por diferenciación de la función de verosimilitud, o la de su logaritmo, hallando los puntos donde estas derivadas se anulen. Hay casos en los que el máximo ocurre en puntos donde la derivada no existe, y por lo tanto, otros procedimientos deben ser desarrollados. Algunos métodos numéricos como el de Newton-Raphson, se emplean para encontrar estimadores máximo verosímiles.

Un estimador de MV (o un conjunto de estimadores MV) depende de una muestra a través de estadísticas suficientes; de otra manera, si existe un estimador suficiente, todo estimador MV es función de si mismo.

∘ *Propiedades de los estimadores máximo verosímiles*

Bajo condiciones generales respecto al modelo de distribución de probabilidad, el método de máxima verosimilitud proporciona estimadores que son:

a) Asintóticamente centrados.

b) Con distribución asintóticamente normal.

c) Asintóticamente de varianza mínima (eficientes).

d) Si existe una estadística suficiente para el parámetro, el estimador máximo verosímil es suficiente .

e) Invariantes; si $\widehat{\theta}$ es un estimador MV del parámetro θ y g es una función uno a uno, entonces $g(\widehat{\theta})$ es el estimador MV de $g(\theta)$.

∘ Un estimador del parámetro θ, *por intervalo*, con un nivel de confianza $(1-\alpha)\%$, es una expresión de la forma:

$$L \leq \theta \leq U, \tag{B.32}$$

donde los límites L y U dependen de la muestra. Se interpretan de tal forma que si se construyen muchos de ellos (uno por muestra), el $(1-\alpha)\%$ de ellos contienen el verdadero valor del parámetro.

De manera más general, sea $g(X;\theta)$ una variable aleatoria cuya función de probabilidad es conocida, la cual se asume continua y monótona sobre θ; en consecuencia, dado α, se pueden encontrar valores l_1 y l_2 tales que

$$P(l_1 \leq g(X;\theta) \leq l_2) = 1-\alpha. \tag{B.33}$$

Como g es continua y monótona sobre θ, la última expresión se puede escribir como:

$$P(g^{-1}(l_1;X) \leq \theta \leq g^{-1}(l_2;X)) = 1-\alpha$$

Llamando $L = g^{-1}(l_1;X)$ y $U = g^{-1}(l_2;X)$, el intervalo de confianza del $(1-\alpha)\%$ para θ es:

$$L \leq \theta \leq U.$$

Ejemplo B.8 Si $X_1, \ldots, X_n$ es una muestra aleatoria de una población $n(\mu, \sigma^2)$, la función de verosimilitud de la muestra es

$$L(x_1, \ldots, x_n, \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\Big(\frac{x_i-\mu}{\sigma}\Big)^2\right\}.$$

El logaritmo de la función de verosimilitud es

$$\begin{aligned} l(x_1, \ldots, x_n, \mu, \sigma^2) &= \ln L(x_1, \ldots, x_n, \mu, \sigma^2) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2. \end{aligned}$$

Para encontrar la posición del máximo, se calcula

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)$$
$$\frac{\partial \boldsymbol{l}}{\partial \boldsymbol{\sigma}^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2;$$

igualando a cero estas derivadas y resolviendo las ecuaciones que resultan respecto a μ y a σ^2, se obtienen los estimadores

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X},$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Como se observó arriba, el estimador $\hat{\mu}$ es insesgado, pero $\hat{\sigma}^2$ no lo es.

Una estimación de μ mediante un intervalo de confianza se logra a través de la función

$$g(X, \mu, \sigma^2) = Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n},$$

la cual, por ser una función lineal, cumple los requerimientos anotados arriba. Como $Z \sim n(0,1)$, entonces,

$$P\Big(-Z_{\alpha/2} \leq g(X, \mu, \sigma^2) \leq Z_{\alpha/2}\Big) = 1 - \alpha; \quad \text{con } l_1 = -Z_{\alpha/2} \text{ y } l_2 = Z_{\alpha/2}$$
$$P\big(-Z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \leq Z_{\alpha/2}\Big) = 1 - \alpha$$
$$P\left(\bar{X} - \frac{Z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

De esta forma $L = \bar{X} - \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}$ y $U = \bar{X} + \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}$, son los extremos del intervalo del $(1 - \alpha)$ de confianza para estimar μ. ☑

▶ Contraste de hipótesis

Una hipótesis estadística es una afirmación sobre la distribución de una o más variables aleatorias. También se puede considerar como los supuestos acerca de una o más poblaciones; por ejemplo, la forma de la distribución, el valor de los parámetros, etc.. Una hipótesis se llama *simple* si el supuesto

define completamente la población, de otra manera se denomina *compuesta*. La hipótesis *nula* (H_0) es la hipótesis bajo contraste[2], la hipótesis *alterna* (H_1), es la conclusión alcanzada si la hipótesis nula se rechaza.

Un contraste de hipótesis estadística es una regla con la que se decide rechazar o no la hipótesis nula H_0, de acuerdo con el valor observado en una *estadística de prueba*; la cual es función de un conjunto de variables aleatorias.

En el caso unidimensional, se define *la región crítica* como un intervalo C de $\mathbb{R}$, constituido por los valores del contraste estadístico que permitan rechazar H_0. Las cotas de la región crítica se denominan *puntos críticos*.

Se incurre en *Error Tipo I* cuando se rechaza la hipótesis nula, siendo ésta cierta. El *Error Tipo II* se comete al aceptar la hipótesis nula, siendo ésta falsa.

Para un contraste estadístico T de $H_0: \ \theta \in \Omega_0$ frente a $H_1: \ \theta \in \boldsymbol{\Omega} - \boldsymbol{\Omega}_0$, las probabilidades de estos errores son, respectivamente,

$$\alpha(\theta) = P(\lambda \in C | \theta \in \boldsymbol{\Omega}_0), \text{ y } \beta(\theta) = P(\lambda \notin C | \theta \in \boldsymbol{\Omega} - \boldsymbol{\Omega}_0). \tag{B.34}$$

Observaciones

- $\boldsymbol{\Omega}$ corresponde al espacio de parámetros y Ω_0 el subconjunto determinado por H_0.
- Recuérdese que como $\Omega_0 \subseteq \boldsymbol{\Omega}$, entonces $P(\Omega_0) \leq P(\boldsymbol{\Omega})$. Esta propiedad debe tenerse presente cuando se calcule la razón de verosimilitud, pues tal cociente será menor o igual ó mayor o igual que 1, dependiendo de si el numerador o el denominador se asocia con Ω_0 o con $\boldsymbol{\Omega}$.

El valor máximo de $\alpha(\theta)$ se llama el *tamaño de la prueba*. El nivel de significación es una cota preseleccionada para $\alpha(\theta)$. La *potencia* de la prueba es la probabilidad de que la estadística de prueba permita rechazar H_0; se nota por

$$\Pi_{\Omega_0}(\theta) = P(\lambda \in C). \tag{B.35}$$

Un método para construir pruebas es la *razón de verosimilitud*; que bajo ciertos supuestos tiene algunas "buenas" propiedades. Resumidamente se construye asi:

[2]En lo posible, se prefiere este término al de "prueba".

Sea $X_1, \cdots, X_n$ una muestra aleatoria de una población $f(\cdot;\theta)$ con función de verosimilitud $L(x_1, \cdots, x_n; \theta)$. Supóngase que $f(\cdot;\theta)$ es una familia específica de funciones, una por cada $\theta \in \boldsymbol{\Omega}$ y sea $\boldsymbol{\Omega}_0$ un subconjunto de $\boldsymbol{\Omega}$. La prueba de razón de verosimilitud de

$$H_0 : \ \theta \in \boldsymbol{\Omega}_0, \text{ frente a } \ H_1 : \ \theta \in \boldsymbol{\Omega} - \boldsymbol{\Omega}_0$$

tiene como región de rechazo al conjunto de puntos $\lambda \in C$, tales que $\lambda \leq \lambda_0, \ 0 \leq \lambda_0 \leq 1$; donde λ es la razón

$$\lambda = \frac{L(\widehat{\boldsymbol{\Omega}_0})}{L(\widehat{\boldsymbol{\Omega}})}. \tag{B.36}$$

$L(\widehat{\boldsymbol{\Omega}_0})$ y $L(\widehat{\boldsymbol{\Omega}})$ son los máximos de la función de verosimilitud con respecto a θ en Ω_0 y $\boldsymbol{\Omega}$, respectivamente.

De manera intuitiva, se rechaza H_0 cuando el cociente (B.36) sea pequeño, su tamaño se mide por el valor de λ_0, en sentido probabilístico.

Si el conocimiento de la distribución de H_0 permite determinar la distribución de λ, entonces, para un valor fijo de α se puede tomar como región crítica C al conjunto

$$\lambda > \lambda_c \quad \textit{donde} \quad P(\lambda > \lambda_c | H_0) = \alpha.$$

Como en el caso de la estimación vía máxima verosimilitud, cualquier función monótona $g(\lambda)$, puede emplearse como estadística de prueba con la región crítica especificada por valores apropiados de $g(\lambda)$. Un resultado asintótico muy importante es que, bajo ciertas condiciones de regularidad, la distribución de probabilidad de $-2\ln\lambda$ es aproximadamente *ji-cuadrado* con $(k_1 - k_2)$ grados de libertad, en tanto $n \to \infty$; donde k_1 y k_2 son las dimensiones de $\boldsymbol{\Omega} - \boldsymbol{\Omega}_0$ y $\boldsymbol{\Omega}_0$ respectivamente, y $k_1 > k_2$.

Ejemplo B.9 Supóngase una muestra aleatoria $X_1, \cdots, X_n$ de una población $n(\boldsymbol{\mu}; \sigma^2)$.

La función de verosimilitud de la muestra es

$$L(x_1, \cdots, x_n; \boldsymbol{\mu}, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \boldsymbol{\mu})^2 \right).$$

Un contraste de razón de verosimilitud para la hipótesis

$$H_0 : \ \boldsymbol{\mu} = \boldsymbol{\mu}_0 \text{ frente a } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

sobre la población anterior, se obtiene enseguida. El espacio de parámetros y el espacio inducido por la hipótesis nula, para este caso, respectivamente, son:

$$\boldsymbol{\Omega} = \{(\boldsymbol{\mu}, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ : -\infty < \boldsymbol{\mu} < \infty;\ \sigma^2 > 0\}$$

y

$$\Omega_0 = \{(\boldsymbol{\mu}, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ : \boldsymbol{\mu} = \boldsymbol{\mu}_0;\ \sigma^2 > 0\}.$$

Los valores de $\boldsymbol{\mu}$ y de σ^2 que maximizan $L(x_1, \cdots, x_n; \boldsymbol{\mu}, \sigma^2)$ en $\boldsymbol{\Omega}$ son $\widehat{\boldsymbol{\mu}} = \bar{x}$ y

$\widehat{\sigma}^2 = \sum_{i=1}^n \dfrac{1}{n}(x_i - \bar{x})^2$; esto es

$$L(\widehat{\boldsymbol{\Omega}}) = \left[\frac{n}{2\pi \sum (x_i - \bar{x})^2}\right]^{n/2} e^{-n/2}.$$

Para maximizar L sobre $\boldsymbol{\Omega}_0$, se hace $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ y $\sigma^2 = \sum_{i=1}^n \dfrac{1}{n}(x_i - \boldsymbol{\mu}_0)^2$.

$$L(\widehat{\boldsymbol{\Omega}}_0) = \left[\frac{n}{2\pi \sum (x_i - \boldsymbol{\mu}_0)^2}\right]^{n/2} e^{-n/2}.$$

Con todo esto, la razón de verosimilitud es

$$\lambda = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \boldsymbol{\mu}_0)^2}\right)^{n/2};$$

de la identidad

$$\sum_{i=1}^n (x_i - \boldsymbol{\mu}_0)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \boldsymbol{\mu}_0)^2,$$

resulta

$$\lambda^{2/n} = \left\{1 + \frac{1}{n-1} \frac{n(\bar{x} - \boldsymbol{\mu}_0)^2}{\dfrac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2}\right\}^{-1} = \left(1 + \frac{t^2}{n-1}\right)^{-1},$$

donde

$$t = \frac{\sqrt{n}(\bar{x} - \boldsymbol{\mu}_0)}{\sqrt{\dfrac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Entonces $\lambda < \lambda_0$ es equivalente a $t^2 > k$, para una determinada constante k. La región crítica C está determinada por

$$P(|t_{n-1}| > k) = \alpha,$$

que corresponde a la conocida estadística *t-Student*.

B.4 Distribuciones conjuntas

Una *variable aleatoria* p-dimensional, es un vector en el que cada una de sus componentes es una variable aleatoria. Así,

$$X' = (X_1, \ldots, X_p), \tag{B.37}$$

Similar al caso unidimensional, se define la *función de distribución conjunta* para el vector X mediante:

$$F(x_1, \ldots, x_p) = P(X_1 \leq x_1, \ldots, X_p \leq x_p). \tag{B.38}$$

Si el vector aleatorio X es continuo y F es absolutamente continua, entonces la *función de densidad conjunta* es:

$$\frac{\partial^p F(x_1, \ldots, x_p)}{\partial x_1 \ldots \partial x_p} = f(x_1, \ldots, x_p). \tag{B.39}$$

Si las p variables aleatorias que conforman el vector X son variables aleatorias independientes, entonces,

$$F(x_1, \ldots, x_p) = F_1(x_1) \cdots F_p(x_p), \tag{B.40a}$$

y

$$f(x_1, \ldots, x_p) = f_1(x_1) \ldots f_p(x_p). \tag{B.40b}$$

De manera recíproca, si la función de distribución conjunta (o la densidad) se puede expresar como en (B.40), las variables aleatorias que conforman a X son independientes.

La propiedad anterior es importante, pues muchas de las metodologías estadísticas se sustentan en el supuesto de independencia estocástica. No obstante, se debe tener cuidado con el tipo de independencia (o dependencia) que un conjunto de datos exhiba; pues de una parte está la posible independencia estocástica o estadística entre las variables, y de otra, la independencia (o dependencia) que puedan tener las observaciones. La independencia estocástica entre variables (columnas de $\mathbb{X}$) es la que se aprovecha a través de los métodos factoriales tales como el análisis por componentes principales, análisis factorial, análisis de correspondencias, la correlación canónica, entre otros. La dependencia entre observaciones (filas de $\mathbb{X}$) es utilizada por métodos tales como las series de tiempo, el análisis espacial; y una buena parte de los métodos multivariados presuponen independencia entre las observaciones.

Las anteriores son razones suficientes para estar atentos al tipo de independencia que se requiere y dispone en las diferentes aplicaciones de la estadística multivariada.

▶ Distribuciones marginales

Dada una variable aleatoria p-dimensional X, con función de distribución $F(x_1, \ldots, x_p)$, se define la *función de ditribución marginal* para algún subconjunto de variables $X_1, \ldots, X_r$ con $(r \leq p)$ como:

$$\begin{aligned} P(X_1 \leq x_1, \ldots, X_r \leq x_r) &= P(X_1 \leq x_1, \ldots, X_r \leq x_r, X_{r+1} \leq \infty, \ldots, X_p \leq \infty) \\ &= F(x_1, \ldots, x_r, \infty, \ldots, \infty) \\ &= F(x_1, \ldots, x_r). \end{aligned} \tag{B.41}$$

La función de densidad marginal de $X_1, \ldots, X_r$ es[3]

$$f(x_1, \ldots, x_r) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_p)\, dx_{r+1}, \ldots, dx_p. \tag{B.42}$$

▶ Distribuciones condicionales

En anlogícon la definición de probabilidad condicional entre eventos, se define la función de distribución condicional de un subconjunto de variables aleatorias $X_1, \ldots, X_r$, dada las variables $X_{r+1} = x_{r+1}, \ldots, X_p = x_p$, como:

$$F(x_1, \ldots, x_r | x_{r+1}, \ldots, x_p) = \frac{F(x_1, \ldots, x_p)}{F(x_{r+1}, \ldots, x_p)}. \tag{B.43a}$$

Las función de densidad condicional está definida en forma semejante a como se muestra enseguida

$$f(x_1, \ldots, x_r | x_{r+1}, \ldots, x_p) = \frac{f(x_1, \ldots, x_p)}{f(x_{r+1}, \ldots, x_p)}. \tag{B.43b}$$

▶ Transformación de variables

Hay situaciones donde las variables aleatorias deben ser transformadas a otras variables aleatorias. En tales circunstancias es necesario conocer la distribución de las "nuevas" variables. La siguiente expresión es una herramienta matemática útil para encontrar la distribución de las variables

[3] Para el caso discreto la integral corresponde a la sumatoria.

resultantes al aplicar transformaciones uno a uno, sobre un conjunto de variables aleatorias.

Sea $f(x_1, \ldots, x_p)$ la función de densidad conjunta de $X_1, \ldots, X_p$. La función de valor real

$$Y_i = Y_i(x_1, \ldots, x_p), \text{ con } i = 1, \ldots, p$$

es una transformación del X-espacio en el Y-espacio; la cual se asume uno a uno. La transformación inversa es:

$$X_i = X_i(y_1, \ldots, y_p) \text{ para } i = 1, \ldots, p.$$

Las variables aleatorias $Y_1, \ldots, Y_p$ definidas por

$$Y_i = Y_i(x_1, \ldots, x_p) \text{ para } i = 1, \ldots, p$$

tienen como función de densidad conjunta a:

$$g(y_1, \ldots, y_p) = f[x_1(y_1, \ldots, y_p), \ldots, x_p(y_1, \ldots, y_p)].|J(y_1, \ldots, y_p)|, \quad \text{(B.44)}$$

donde $J(y_1, \ldots, y_p)$ es el jacobiano de la transformación

$$J(y_1, \ldots, y_p) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_p}{\partial y_1} & \cdots & \frac{\partial x_p}{\partial y_p} \end{vmatrix};$$

se asume que estas derivadas parciales existen.

Através de (B.44), y bajo las condiciones exigidas, se obtiene la función de densidad de las "nuevas" variables. Ésta también es una herramienta útil cuando se desea obtener la distribución para un número $q \leq p$ de "nuevas" variables, basta con definir sobre las restantes (p-q) variables, funciones de tal forma que se tenga una transformación uno a uno de todo el X-espacio en el Y-espacio; la función de distribución de las q-variables se obtiene como la distribución marginal de las "nuevas" variables.

Ejemplo B.10 Supóngase que la variable aleatoria bidimensional (X_1, X_2) tiene función de densidad conjunta

$$f(x_1, x_2) = \begin{cases} x_1^2 + \frac{x_1 x_2}{3}, & 0 < x_1 < 1,\ 0 < x_2 < 2 \\ 0, & \text{en otra parte.} \end{cases}$$

La prueba de que f corresponde a una auténtica fdp es inmediata, pues $f(x_1, x_2) \geq 0$ y $\int_0^2 \int_0^1 f(x_1, x_2) dx_1 dx_2 = 1$. La función de densidad marginal para X_1 es

$$f_{X_1}(x_1) = \int_0^2 f(x_1, x_2) dx_2 = \int_0^2 (x_1^2 + \frac{x_1 x_2}{3}) dx_2 = 2x_1^2 + \frac{2}{3} x_1$$

La fdp condicional de X_2 para $X_1 = x_1$ es

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)} = \frac{x_1^2 + \frac{x_1 x_2}{3}}{2x_1^2 + \frac{2}{3} x_1} \qquad \boxtimes$$

Ejemplo B.11 Considere el vector aleatorio $X' = (X_1, X_2)$, cuya fdp conjunta es definida mediante la siguiente expresión

$$f(x_1, x_2) = \begin{cases} k, & 0 \leq x_1 \leq x_2 \leq 1, \\ 0, & \text{en otra parte.} \end{cases}$$

a) Halle el valor de k

b) Encuentre la fdp conjunta para la transformación $Y' = (Y_1, Y_2)$, definida por:

$$Y_1 = \frac{X_1 + X_2}{2} \quad \text{y} \quad Y_2 = X_2 - X_1;$$

es decir,

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

Solución

a) Como f es una fdp de X_1 y X_2, entonces se satisface $\iint\limits_R f(x_1, x_2) dx_1 dx_2 = 1$, donde el conjunto $R = \{(x_1, x_2) \in \mathbb{R}^2 : \ 0 \leq x_1 \leq x_2 \leq 1\}$. Así:

$$\begin{aligned} \int_0^1 \int_0^{x_2} f(x_1, x_2) dx_1 dx_2 &= \int_0^1 \int_0^{x_2} k dx_1 dx_2 = 1 \\ &= \int_0^1 k x_2 dx_2 = 1 \\ &= k \left. \frac{x_2^2}{2} \right|_0^1 = 1, \quad \text{de donde} \quad k = 2. \end{aligned}$$

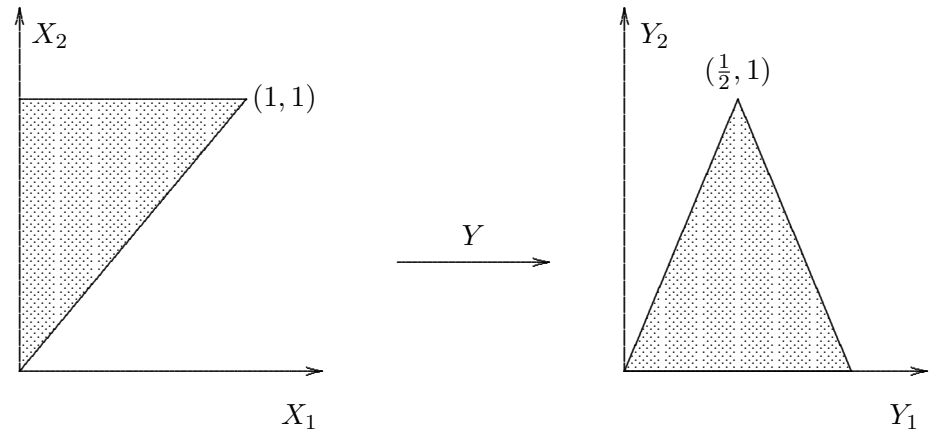


Figura B.6 *Tansformación Y.*

b) La figura B.6 muestra la transformación generada por Y.

Nótese que

$$Y = AX = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

De manera que la transformación inversa viene dada por

$$X = A^{-1}Y = \begin{pmatrix} 1 & -\frac{1}{2} \\ 1 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

El jacobiano de esta transformación es

$$J = \begin{vmatrix} 1 & -\frac{1}{2} \\ 1 & \frac{1}{2} \end{vmatrix} = 1$$

y de acuerdo con (B.44) la función de densidad conjunta de $Y' = (Y_1, Y_2)$ es

$$\begin{aligned} g_Y(y_1, y_2) &= f(x_1(y_1, y_2), x_2(y_1, y_2))J(y_1, y_2) \\ &= \begin{cases} 2, & y_1, y_2 \geq 0 \text{ tal que } |2y_1 - 1| \leq 1 - y_2 \\ 0, & \text{en otra parte} \end{cases} \end{aligned}$$ ☑

► Función generadora de momentos

La *función generadora de momentos* (*fgm*) del vector aleatorio $\mathbf{X}$ se define por

$$M_X(t) = \mathcal{E}(e^{t'X}),$$

con $t' = (t_1, \ldots, t_p)$, si y sólo si, este valor esperado existe para todo t_i tal que $|t_i| < a$; donde $i = 1, \ldots, p$ y $a > 0$. Una definición más general es la *función característica*[4] $\phi_X = \mathcal{E}(e^{it'X})$. Para los propósitos de estas notas es suficiente con la *fgm*.

Algunas utilidades de la *fgm* son la identificación de la distribución de una variable o vector aleatorio; el cálculo de los momentos asociados con una variable o vector aleatorio. Éstas se pueden observar de acuerdo con las siguientes propiedades. Las demostraciones de éstas se dejan como ejercicio. Se pueden consultar en Mood y Colaboradores (1982) o en Roussas (1972)

∘ *Propiedades de la fgm*

1. Si $\mathbf{X}$ y $\mathbf{Y}$ son vectores aleatorios con la misma función generadora de momentos en algún rectángulo abierto que contenga al origen, entonces ellos tienen la misma función de distribución.

2. Sea $Y = AX + b$. Entonces $M_Y = e^{(t'b)} M_X(t'A)$.

3. Sea $X' = (X_{(1)}, X_{(2)})$. Los vectores aleatorios $X_{(1)}$ y $X_{(2)}$ son independientes si y sólo si

$$M_X(t) = M_{X_{(1)}}(t_1) M_{X_{(2)}}(t_2); \text{ con } t = (t_1, t_2).$$

4. La *fgm* de un vector aleatorio, si existe, es única.

5. Con esta propiedad se justifica el nombre de generadora de momentos,

$$\left. \frac{\partial^{k_1 + \ldots + k_p}}{\partial t_1^{k_1} \ldots \partial t_p^{k_p}} M_{X_1, \ldots, X_p}(t_1, \ldots, t_p) \right|_{t_1 = \cdots = t_p = 0} = \mathcal{E}(X_1^{k_1} \cdots X_p^{k_p}),$$

en particular,

$$\left. \frac{\partial^k}{\partial t_j^k} M_{X_1, \ldots, X_p}(t_1, \ldots, t_p) \right|_{t_1 = \cdots = t_p = 0} = \mathcal{E}(X_j^k), \quad j = 1, \ldots, p,$$

que es el momento de orden k, centrado en 0, para la variable aleatoria X_j.

Con intención meramente ilustrativa se desarrolla la fgm para el caso bidimensional (tenga paciencia y disfrute con los cálculos...). Así, para el vector $X' = (X_1, X_2)$ con fdp conjunta $f(x_1, x_2)$ se tiene:

$$M_X(t) = \mathcal{E}\left(e^{\{t_1 X_1 + t_2 X_2\}}\right).$$

[4] Es una extensión al campo de los complejos $\mathbb{C}$,

Una función en dos variables $g(x_1, x_2)$, bajo algunas condiciones de regularidad, se puede aproximar mediante el desarrollo del polinomio de Taylor, alrededor de un punto (a, b); es decir,

$$g(x_1, x_2) = g(a, b) + \sum_{1 \leq r+s \leq n} \frac{\partial^{r+s} g(a, b)}{\partial x_1^r \partial x_2^s} \cdot \frac{(x_1 - a)^r}{r!} \frac{(x_2 - b)^s}{s!} + R_n.$$

Para este caso, se desarrolla en torno al punto $(0, 0)$,

$$\begin{aligned} g(x_1, x_2) &= e^{\{t_1 x_1 + t_2 x_2\}}, \text{ luego } g(0, 0) = 1. \\ \frac{\partial^r (x_1, x_2)}{\partial x_1^2} &= t_1^r e^{\{t_1 x_1 + t_2 x_2\}} \\ \frac{\partial^s (x_1, x_2)}{\partial x_2^2} &= t_2^s e^{\{t_1 x_1 + t_2 x_2\}} \\ \frac{\partial^{r+s} (x_1, x_2)}{\partial x_1^r \partial x_2^s} &= t_1^r t_2^s e^{\{t_1 x_1 + t_2 x_2\}}, \quad \text{con} \quad 1 \leq r + s \leq n, \end{aligned}$$

por la definición de fgm y del desarrollo del polinomio de Taylor, para $g(x_1, x_2)$ en torno al punto $(0, 0)$ resulta

$$\begin{aligned} M_X(t) = & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{\{t_1 X_1 + t_2 X_2\}} f(x_1, x_2) dx_1 dx_2 \\ = & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big[1 + t_1 x_1 + t_2 x_2 + \frac{(t_1 x_1)^2}{2!} + \frac{(t_2 x_2)^2}{2!} + (t_1 x_1)(t_2 x_2) + \\ & \cdots + \frac{(t_1 x_1)^r}{r!} + \frac{(t_2 x_2)^s}{s!} + (t_1 x_1)^r (t_2 x_2)^s + \cdots \Big] f(x_1, x_2) dx_1 dx_2. \end{aligned}$$

La integral de esta suma es la suma de las integrales, de manera que la expresión anterior es equivalente a

$$
\begin{aligned}
M_X(t) = &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2)dx_1dx_2 + \\
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (t_1x_1)f(x_1,x_2)dx_1dx_2 + \\
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (t_2x_2)f(x_1,x_2)dx_1dx_2 + \\
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{(t_1x_1)^2}{2!}f(x_1,x_2)dx_1dx_2 + \\
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{(t_2x_2)^2}{2!}f(x_1,x_2)dx_1dx_2 + \\
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (t_1x_1)(t_2x_2)f(x_1,x_2)dx_1dx_2 + \cdots + \\
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{(t_1x_1)^r}{r!}f(x_1,x_2)dx_1dx_2 + \\
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{(t_2x_2)^s}{s!}f(x_1,x_2)dx_1dx_2 + \\
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (t_1x_1)^r(t_2x_2)^s f(x_1,x_2)dx_1dx_2 + \cdots
\end{aligned}
$$

De la propiedad 5, después de derivar $M_X(t)$ respecto a t_1 y evaluar en $t_1 = t_2 = 0$ se obtiene

$$
\begin{aligned}
\frac{\partial M_X(t)}{\partial t_1}|_{t_1=t_2=0} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 f(x_1,x_2)dx_1dx_2 \\
&= \int_{-\infty}^{\infty} x_1\Big\{\int_{-\infty}^{\infty} f(x_1,x_2)dx_2\Big\}dx_1 \\
&= \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1)dx_1 \\
&= \mathcal{E}(X_1),
\end{aligned}
$$

dado que los términos a partir del tercero se anulan.

Similarmente

$$
\frac{\partial^2 M_X(t)}{\partial t_1 \partial t_2}\Big|_{t_1=t_2=0} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 \cdot x_2 f(x_1,x_2)dx_1dx_2 = \mathcal{E}(X_1 \cdot X_2).
$$

B.5 Matriz de información de Fisher

La función de puntaje se define como

$$s(X;\theta) = \frac{\partial l(X,\theta)}{\partial \theta} = \frac{1}{L(X,\theta)} \frac{\partial}{\partial \theta} L(X,\theta)$$

con $l(X,\theta) = \ln L(X,\theta)$. Se demuestra, bajo ciertas condiciones de regularidad, que

$$\mathcal{E}\left(\frac{\partial l(X,\theta)}{\partial \theta}\right) = 0$$

La matriz de covarianzas de $s(X;\theta)$ R. A. Fisher la denominó la *matriz de información esperada* de θ, se nota $\mathcal{I}(\theta)$. Se demuestra (Mardia y colaboradores, 1992, pág. 98) que

$$\mathcal{I}(\theta) = \mathcal{E}(ss') = \mathcal{E}\left(-\frac{\partial^2 l(X,\theta)}{\partial \theta \partial \theta'}\right)$$

Por analogía con la cota de Cramer-Rao, se demuestra que $Cov(\hat{\theta})$ es "mayor" que $\mathcal{I}(\theta)$ en el sentido de que $Cov(\hat{\theta}) - \mathcal{I}(\theta)$ es una matriz definida positiva.

Se define la *matriz de información observada* como

$$\mathcal{IO}(\hat{\theta}) == \left(-\frac{\partial^2 l(X,\hat{\theta}_{MV})}{\partial \hat{\theta} \partial \hat{\theta}'}\right)$$

con $\hat{\theta}$ el estimador de máxima verosimilitud del vector de parámetros θ.

Una de las aplicaciones de esta matriz consiste en determinar la varianza de los estimadores, con la cual se puede hacer inferencia sobre el vector de parámetros θ.

B.6 Funciones en SAS para calcular probabilidades en algunas distribuciones

A continuación se listan las distribuciones de probabilidad y de densidad de aplicación más frecuente, las cuales están incorporadas al paquete SAS.

`POISSON(lamda,k);` /* Calcula $Pr(X \leq k)$ distribución de Poisson con parámetro lamda */

`PROBBETA(x,a,b);` /∗ Calcula $Pr(X \leq x)$ distribución Beta de parámetros a y b ∗/
`PROBBNML(p,n,k);` /∗ Calcula $Pr(X \leq k)$ distribución Binomial de parámetros p y n ∗/
`PROBCHI(x,n);` /∗ Calcula $Pr(X \leq x)$ distribución ji-cuadrado de n grados de libertad ∗/
`PROBF(x,m, n);` /∗ Calcula $Pr(X \leq x)$ distribución F con m y n grados de libertad ∗/
`PROBGAM(x,a);` /∗ Calcula $Pr(X \leq x)$ distribución Gama con parámetro a ∗/
`PROBNORM(x);` /∗ Calcula $Pr(X \leq x) = \Phi(x)$ distribución normal estándar ∗/
`PROBT(x,n);` /∗ Calcula $Pr(X \leq x)$ distribución t-Student de n grados de libertad ∗/

B.7 Procesamiento de datos con R

R permite calcular la función de distribución, y sus afines.

Distribución	Nombre en R	Argumentos adicionales
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
ji cuadrado	chisq	df, ncp
exponencial	exp	rate
F de Snedecor	f	df1, df1, ncp
gamma	gamma	shape, scale
geométrica	geom	prob
hipergeométrica	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logística	logis	location, scale
binomial negativa	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
t de Student	t	df, ncp
uniforme	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

La tabla anterior contiene un conjunto de distribuciones (discretas y continuas).

Para construir el nombre de cada función, utilice el nombre de la distribución precedido de la letra "d" para la función de densidad, la letra "p" para la función de distribución, la letra "q" para la función de distribución inversa, y la letra "r" para la generación de números pseudoaleatorios (muestras aleatorias). El primer argumento es x para la función de densidad, q para la función de distribución, p para la función de distribución inversa, y n para la función de generación de números pseudoaleatorios (excepto en el caso de rhyper y rwilcox, en los cuales es nn).

Los siguientes ejemplos clarificarán estos conceptos:

```
> ## P valor a dos colas de la distribución t_13
> 2*pt(-2.43, df = 13)
> ## Percentil 1 superior de una distribución F(2, 7)
> qf(0.99, 2, 7)
> ## Genera una muestra aleatioria de tamaño n=60 de una dis-
tribución normal con media
mean=5.0 y desviación estándar sd=1.5
> rnorm(n=60; mean=5.0; sd=1.5)
```

Apéndice C

Tablas Estadísticas

Tabla C.1 Percentiles superiores de la distribución T^2 de Hotelling

G.L. ν	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$	$p=9$	$p=10$
					$\alpha = 0.05$					
2	18.513									
3	10.128	57.000								
4	7.709	25.472	114.986							
5	6.608	17.361	46.383	192.468						
6	5.987	13.887	29.661	72.937	289.446					
7	5.591	12.001	22.720	44.718	105.157	405.920				
8	5.318	10.828	19.028	33.230	62.561	143.050	541.890			
9	5.117	10.033	16.766	27.202	45.453	83.202	186.622	697.356		
10	4.965	9.459	15.248	23.545	36.561	59.403	106.649	235.873	872.317	
11	4.844	9.026	14.163	21.108	31.205	47.123	75.088	132.903	290.806	1066.774
12	4.747	8.689	13.350	19.376	27.656	39.764	58.893	92.512	161.967	351.421
13	4.667	8.418	12.719	18.086	25.145	34.911	49.232	71.878	111.676	193.842
14	4.600	8.197	12.216	17.089	23.281	31.488	42.881	59.612	86.079	132.582
15	4.543	8.012	11.806	16.296	21.845	28.955	38.415	51.572	70.907	101.499
16	4.494	7.856	11.465	15.651	20.706	27.008	35.117	45.932	60.986	83.121
17	4.451	7.722	11.177	15.117	19.782	25.467	32.588	41.775	54.041	71.127
18	4.414	7.606	10.931	14.667	19.017	24.219	30.590	38.592	48.930	62.746
19	4.381	7.504	10.719	14.283	18.375	23.189	28.975	36.082	45.023	56.587
20	4.351	7.415	10.533	13.952	17.828	22.324	27.642	34.054	41.946	51.884
21	4.325	7.335	10.370	13.663	17.356	21.588	26.525	32.384	39.463	48.184
22	4.301	7.264	10.225	13.409	16.945	20.954	25.576	30.985	37.419	45.202
23	4.279	7.200	10.095	13.184	16.585	20.403	24.759	29.798	35.709	42.750
24	4.260	7.142	9.979	12.983	16.265	19.920	24.049	28.777	34.258	40.699
25	4.242	7.089	9.874	12.803	15.981	19.492	23.427	27.891	33.013	38.961
26	4.225	7.041	9.779	12.641	15.726	19.112	22.878	27.114	31.932	37.469
27	4.210	6.997	9.692	12.493	15.496	18.770	22.388	26.428	30.985	36.176
28	4.196	6.957	9.612	12.359	15.287	18.463	21.950	25.818	30.149	35.043
29	4.183	6.919	9.539	12.236	15.097	18.184	21.555	25.272	29.407	34.044
30	4.171	6.885	9.471	12.123	14.924	17.931	21.198	24.781	28.742	33.156
35	4.121	6.744	9.200	11.674	14.240	16.944	19.823	22.913	26.252	29.881
40	4.085	6.642	9.005	11.356	13.762	16.264	18.890	21.668	24.624	27.783
45	4.057	6.564	8.859	11.118	13.409	15.767	18.217	20.781	23.477	26.326
50	4.034	6.503	8.744	10.934	13.138	15.388	17.709	20.117	22.627	25.256
55	4.016	6.454	8.652	10.787	12.923	15.090	17.311	19.600	21.972	24.437
60	4.001	6.413	8.577	10.668	12.748	14.850	16.992	19.188	21.451	23.790
70	3.978	6.350	8.460	10.484	12.482	14.485	16.510	18.571	20.676	22.834
80	3.960	6.303	8.375	10.350	12.289	14.222	16.165	18.130	20.127	22.162
90	3.947	6.267	8.309	10.248	12.142	14.022	15.905	17.801	19.718	21.663
100	3.936	6.239	8.257	10.167	12.027	13.867	15.702	17.544	19.401	21.279
110	3.927	6.216	8.215	10.102	11.934	13.741	15.540	17.340	19.149	20.973
120	3.920	6.196	8.181	10.048	11.858	13.639	15.407	17.172	18.943	20.725
150	3.904	6.155	8.105	9.931	11.693	13.417	15.121	16.814	18.504	20.196
200	3.888	6.113	8.031	9.817	11.531	13.202	14.845	16.469	18.083	19.692
400	3.865	6.052	7.922	9.650	11.297	12.890	14.447	15.975	17.484	18.976
1000	3.851	6.015	7.857	9.552	11.160	12.710	14.217	15.692	14.141	18.570
∞	3.841	5.991	7.815	9.488	11.070	12.592	14.067	15.507	16.919	18.307

(Continuación Tabla C.1)

G. L. ν	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$	$p=9$	$p=10$
					$\alpha=0.01$					
2	98.503									
3	34.116	297.000								
4	21.198	82.177	594.997							
5	16.258	45.000	147.283	992.494						
6	13.745	31.857	75.125	229.679	1489.489					
7	12.246	25.491	50.652	111.839	329.433	2085.984				
8	11.259	21.821	39.118	72.908	155.219	446.571	2781.978			
9	10.561	19.460	32.598	54.890	98.703	205.293	581.106	3577.472		
10	10.044	17.826	28.466	44.838	72.882	128.067	262.076	733.045	4472.464	
11	9.646	16.631	25.637	38.533	58.618	93.127	161.015	325.576	902.392	5466.956
12	9.330	15.722	23.588	34.251	49.739	73.969	115.640	197.555	395.797	1089.149
13	9.074	15.008	22.041	31.171	43.745	62.114	90.907	140.429	237.692	472.742
14	8.862	14.433	20.834	28.857	39.464	54.150	75.676	109.441	167.449	281.428
15	8.683	13.960	19.867	27.060	36.246	48.472	65.483	90.433	129.576	196.853
16	8.531	13.566	19.076	25.626	33.672	44.240	58.241	77.755	106.391	151.316
17	8.400	13.231	18.418	24.458	31.788	40.975	52.858	68.771	90.969	123.554
18	8.285	12.943	17.861	23.487	30.182	38.385	48.715	62.109	80.067	105.131
19	8.185	12.694	17.385	22.670	28.852	36.283	45.435	56.992	71.999	92.134
20	8.096	12.476	16.973	21.972	27.734	34.546	42.779	52.948	65.813	82.532
21	8.017	12.283	16.613	21.369	26.781	33.088	40.587	49.679	60.932	75.181
22	7.945	12.111	16.296	20.843	25.959	31.847	38.750	46.986	56.991	69.389
23	7.881	11.958	16.015	20.381	25.244	30.779	37.188	44.730	53.748	64.719
24	7.823	11.820	15.763	19.972	24.616	29.850	35.846	42.816	51.036	60.879
25	7.770	11.695	15.538	19.606	24.060	29.036	34.680	41.171	48.736	57.671
26	7.721	11.581	15.334	19.279	23.565	28.316	33.659	39.745	46.762	54.953
27	7.677	11.478	15.149	18.983	23.121	27.675	32.756	38.496	45.051	52.622
28	7.636	11.383	14.980	18.715	22.721	27.101	31.954	37.393	43.554	50.604
29	7.598	11.295	14.825	18.471	22.359	26.584	31.236	36.414	42.234	48.839
30	7.562	11.215	14.683	18.247	22.029	26.116	30.589	35.538	41.062	47.283
35	7.419	10.890	14.117	17.366	20.743	24.314	28.135	32.259	36.743	41.651
40	7.314	10.655	13.715	16.750	19.858	23.094	26.502	30.120	33.984	38.135
45	7.234	10.478	13.414	16.295	19.211	22.214	25.340	28.617	32.073	35.737
50	7.171	10.340	13.181	15.945	18.718	21.550	24.470	27.504	30.673	33.998
55	7.119	10.228	12.995	15.667	18.331	21.030	23.795	26.647	29.603	32.682
60	7.077	10.137	12.843	15.442	18.018	20.613	23.257	25.967	28.760	31.650
70	7.011	9.996	12.611	15.098	17.543	19.986	22.451	24.957	27.515	30.139
80	6.963	9.892	12.440	14.849	17.201	19.536	21.877	24.242	26.642	29.085
90	6.925	9.813	12.310	14.660	16.942	19.197	21.448	23.710	25.995	28.310
100	6.895	9.750	12.208	14.511	16.740	18.934	21.115	23.299	25.496	27.714
110	6.871	9.699	12.125	14.391	16.577	18.722	20.849	22.972	25.101	27.243
120	6.851	9.567	12.057	14.292	16.444	18.549	20.632	22.705	24.779	26.862
150	6.807	9.565	11.909	14.079	16.156	18.178	20.167	22.137	24.096	26.054
200	6.763	9.474	11.764	13.871	15.877	17.819	19.720	21.592	23.446	25.287
400	6.699	9.341	11.551	13.569	15.473	17.303	19.080	20.818	22.525	24.209
1000	6.660	9.262	11.426	13.392	15.239	17.006	18.743	20.376	22.003	23.600
∞	6.635	9.210	11.345	13.277	15.086	16.812	18.475	20.090	21.666	23.209

Tabla C.2 Valores críticos inferiores de lambda de Wilks $\mathbf{\Lambda}$

ν_E	$\nu_H = 1$	$\nu_H = 2$	$\nu_H = 3$	$\nu_H = 4$	$\nu_H = 5$	$\nu_H = 6$	$\nu_H = 7$	$\nu_H = 8$	$\nu_H = 9$	$\nu_H = 10$
	$p = 1$									
1	.006157	.002501	.001543	.001112	.000868	.000712	.000603	.000523	.000462	.000413
2	.097504	.050003	.033615	.025322	.020309	.016953	.014549	.012741	.011333	.010208
3	.228516	.135712	.097321	.076019	.062408	.052963	.046005	.040672	.036446	.033020
4	.341614	.223602	.168243	.135345	.113373	.097610	.085724	.076447	.068985	.062851
5	.430725	.301697	.235535	.194031	.165283	.144073	.127777	.114822	.104279	.095505
6	.500549	.368408	.295990	.248596	.214783	.189255	.169266	.153168	.139893	.128754
7	.555908	.424896	.349304	.298096	.260620	.231812	.208893	.190186	.174606	.161423
8	.600708	.472870	.396057	.342590	.302612	.271332	.246124	.225311	.207825	.192902
9	.637512	.513916	.437164	.382446	.340790	.307770	.280823	.258362	.239288	.222931
10	.668243	.549286	.473389	.418213	.375519	.341248	.313019	.289246	.268936	.251373
11	.694275	.580017	.505463	.450317	.407104	.372040	.342834	.318054	.296768	.278229
12	.716553	.606964	.534027	.479309	.435913	.400299	.370453	.344940	.322876	.303528
13	.735840	.630737	.559570	.505524	.462189	.426361	.396057	.369995	.347321	.327362
14	.752686	.651825	.582581	.529327	.486267	.450348	.419800	.393372	.370239	.349823
15	.767548	.670715	.603333	.551025	.508362	.472534	.441864	.415222	.391754	.370941
16	.780701	.687653	.622162	.570862	.528717	.493103	.462433	.435638	.411957	.390869
17	.792480	.702972	.639343	.589081	.547516	.512177	.481598	.454742	.430939	.409637
18	.803070	.716858	.655029	.605835	.564911	.529907	.499481	.472687	.448807	.427368
19	.812622	.729553	.669434	.621307	.581024	.546448	.516235	.489502	.465637	.444138
20	.821320	.741135	.682709	.635651	.596039	.561890	.531952	.505341	.481506	.459991
21	.829224	.751770	.694977	.648941	.610046	.576355	.546692	.520264	.496521	.475006
22	.836472	.761597	.706329	.661316	.623108	.589905	.560562	.534332	.510712	.489258
23	.843140	.770660	.716858	.672867	.635361	.602631	.573639	.547638	.524139	.502762
24	.849274	.779083	.726685	.683655	.646851	.614609	.585968	.560211	.536896	.515594
25	.854950	.786896	.735870	.693771	.657639	.625900	.597626	.572128	.547817	.527817
26	.860199	.794189	.744446	.703278	.667786	.636566	.608643	.583435	.560486	.539459
27	.865112	.800995	.752487	.712189	.677383	.646637	.619080	.594147	.571411	.550537
28	.869675	.807373	.760040	.720612	.686432	.656174	.628998	.604370	.581833	.561127
29	.873947	.813339	.767151	.785846	.694992	.665222	.638428	.614075	.591766	.571228
30	.877945	.818970	.773865	.736053	.703110	.673798	.647385	.623322	.601242	.580872
40	.907349	.860886	.824463	.793274	.765594	.740540	.717575	.696365	.676636	.651888
60	.937485	.904968	.878807	.855911	.835175	.816055	.798233	.781494	.765686	.750702
80	.952527	.927841	.907471	.889450	.872940	.857590	.843124	.829437	.816391	.803925
100	.962128	.941845	.925179	.910324	.896637	.883835	.871696	.860153	.849083	.838455
120	.968363	.951297	.937200	.924578	.912894	.901916	.891475	.881501	.871901	.862660
140	.972836	.958107	.945890	.934921	.924731	.915131	.905971	.897200	.888734	.880563
170	.977588	.965370	.955195	.946025	.937478	.929401	.921669	.914245	.907057	.900101
200	.980926	.970487	.961768	.953893	.946532	.939564	.932877	.926443	.920200	.914149
240	.984086	.975345	.968024	.961396	.955187	.949296	.943631	.938171	.932861	.927705
320	.988046	.981451	.975907	.970876	.966145	.961649	.957311	.953121	.949035	.945058
440	.991295	.986475	.982411	.978715	.975232	.971914	.968704	.965599	.962561	.959605
600	.993610	.990064	.987067	.984337	.981759	.979301	.976917	.974611	.972349	.970144
800	.995204	.992539	.990282	.988225	.986279	.984422	.982619	.980873	.979158	.977487
1000	.996161	.994026	.992216	.990566	.989003	.987512	.986062	.984658	.983276	.981931

(Continuación Tabla C.2)

ν_E	$\nu_H = 1$	$\nu_H = 2$	$\nu_H = 3$	$\nu_H = 4$	$\nu_H = 5$	$\nu_H = 6$	$\nu_H = 7$	$\nu_H = 8$	$\nu_H = 9$	$\nu_H = 10$
	$p = 2$									
1	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
2	.002500	.000641	.000287	.000162	.000104	.000072	.000053	.000041	.000032	.000026
3	.049998	.018318	.009528	.005844	.003950	.002849	.002152	.001683	.001352	.001110
4	.135725	.061800	.035817	.023460	.016578	.012346	.009555	.007615	.006212	.005165
5	.223606	.117368	.073621	.050765	.037211	.028476	.022507	.018244	.015092	.012695
6	.301715	.174902	.116450	.083663	.063188	.049481	.039834	.032772	.027440	.023320
7	.368405	.229737	.160239	.118984	.092129	.073571	.060172	.050155	.042465	.036426
8	.424876	.280187	.202813	.154741	.122376	.099380	.082397	.069475	.059404	.051386
9	.472866	.325883	.243151	.189781	.152779	.125881	.105643	.089993	.077615	.067661
10	.513885	.367036	.280802	.223433	.182644	.152421	.129282	.111138	.096610	.084797
11	.549281	.404052	.315720	.255369	.211592	.178545	.152598	.135206	.116013	.102453
12	.580029	.437339	.347988	.285511	.239373	.203997	.176155	.153782	.135511	.120356
13	.606971	.467384	.377744	.313837	.265838	.228568	.198874	.174774	.154909	.138311
14	.630737	.494599	.405216	.340396	.291016	.252171	.220930	.195325	.174061	.156149
15	.651851	.519281	.430564	.365263	.314863	.274786	.242249	.215357	.192837	.173755
16	.677011	.541775	.454003	.388530	.337412	.296391	.262763	.234782	.211185	.191059
17	.687662	.562317	.475724	.410322	.358763	.316990	.282502	.253583	.226039	.208000
18	.702982	.581146	.495888	.430784	.378964	.336632	.301430	.271723	.246366	.224530
19	.716866	.598489	.514629	.449961	.398041	.355335	.319573	.289225	.263169	.240614
20	.729531	.614483	.532092	.467968	.416109	.373163	.336951	.306072	.279429	.256249
21	.741124	.629283	.548399	.484925	.433211	.390129	.353609	.322287	.295147	.271431
22	.751776	.643011	.563622	.500886	.449429	.406286	.369555	.337873	.310325	.286147
23	.761598	.655775	.577893	.515922	.464800	.421699	.384810	.352883	.324978	.300409
24	.770680	.667666	.591286	.530135	.479373	.436391	.399429	.367295	.339116	.314213
25	.779088	.678783	.603884	.543551	.493227	.450412	.413436	.381165	.352775	.327593
26	.786893	.689182	.615752	.556269	.506409	.463802	.426867	.394506	.365946	.340539
27	.794192	.698945	.626937	.568306	.518951	.476588	.439744	.407337	.378645	.353047
28	.800992	.708108	.637517	.579727	.530891	.488822	.452093	.419700	.390911	.365171
29	.807354	.716737	.647497	.590582	.542291	.500519	.463948	.431586	.402753	.376900
30	.813343	.724899	.656962	.600899	.553155	.511722	.475325	.443028	.414182	.388244
40	.857594	.786433	.729818	.681627	.639419	.601870	.568076	.537426	.509476	.483873
60	.903437	.852599	.810662	.773804	.740586	.710190	.682157	.656096	.631804	.609029
80	.926967	.887496	.854347	.824736	.797636	.772490	.748974	.726849	.405927	.686107
100	.941272	.909051	.881684	.856993	.834186	.812834	.792697	.773596	.755405	.738034
120	.950898	.923673	.900382	.879233	.859569	.841056	.823491	.806739	.790700	.775302
140	.957812	.934247	.913983	.895493	.878224	.861896	.846339	.831442	.817125	.803326
170	.965169	.945562	.928606	.913057	.898465	.884603	.871338	.858581	.844267	.834352
200	.970341	.953554	.938982	.925569	.912940	.900904	.889349	.878202	.867412	.856939
240	.975243	.961158	.948887	.937554	.926848	.916613	.906758	.897224	.887968	.878959
320	.981393	.970741	.961415	.952766	.944563	.936691	.929082	.921692	.914493	.907461
440	.986449	.978644	.971788	.965408	.959337	.953491	.947824	.942303	.936908	.931623
600	.990047	.984298	.979233	.974507	.969998	.965648	.961420	.957293	.953251	.949283
800	.992529	.988203	.984384	.980814	.977404	.974108	.970900	.967763	.944687	.961662
1000	.994021	.990552	.987487	.984620	.981877	.979224	.976640	.974110	.971627	.969184

(Continuación Tabla C.2)

ν_E	$\nu_H = 1$	$\nu_H = 2$	$\nu_H = 3$	$\nu_H = 4$	$\nu_H = 5$	$\nu_H = 6$	$\nu_H = 7$	$\nu_H = 8$	$\nu_H = 9$	$\nu_H = 10$
	$p = 3$									
1	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
2	.000000	.000000	.000000	.000000	.000000	.000001	.000002	.000004	.000005	.000008
3	.001698	.000354	.000179	.000127	.000105	.000095	.000091	.000090	.000091	.000092
4	.033740	.009612	.004205	.002314	.001479	.001052	.000809	.000659	.000562	.000496
5	.097355	.035855	.017521	.010010	.006357	.004369	.003195	.002458	.001971	.001636
6	.168271	.073634	.039672	.024047	.015792	.011018	.008067	.006148	.004849	.003939
7	.235525	.116476	.067711	.043226	.029433	.021043	.015642	.012012	.009485	.007674
8	.295976	.190244	.098932	.065947	.046378	.033966	.025706	.019990	.015911	.012927
9	.349277	.202814	.131378	.090794	.065660	.049161	.037855	.029838	.023995	.019637
10	.396084	.243139	.163846	.116701	.086448	.066012	.051643	.041238	.033514	.027654
11	.437147	.280808	.195556	.142927	.108110	.083979	.066659	.053876	.044225	.036801
12	.473377	.315719	.226090	.168939	.130131	.102644	.082534	.067443	.055894	.046882
13	.505452	.347981	.255220	.194414	.152160	.121656	.098973	.081704	.068298	.057724
14	.534018	.377735	.282849	.219113	.173959	.140775	.115736	.096413	.081246	.069166
15	.559570	.405221	.308951	.249244	.195322	.159796	.132619	.111416	.094593	.081052
16	.582577	.430566	.333588	.265812	.216138	.178574	.149493	.126564	.108178	.093264
17	.603338	.454006	.356777	.287689	.236338	.197017	.166236	.141728	.121917	.105704
18	.622168	.475728	.378631	.308599	.255858	.215044	.182762	.156827	.135694	.118273
19	.639337	.495908	.399223	.328552	.274710	.232604	.199009	.171789	.149446	.130904
20	.655028	.514622	.418629	.347546	.292843	.249666	.214918	.186544	.163097	.143521
21	.669437	.532101	.436898	.365676	.310304	.266216	.230467	.201077	.176620	.156088
22	.682712	.548393	.454182	.322934	.327083	.282253	.245626	.215325	.189969	.168561
23	.694960	.563637	.470473	.399402	.343191	.297740	.260397	.229291	.203123	.180907
24	.706310	.577895	.485889	.415077	.358665	.312738	.274743	.242939	.216044	.193091
25	.716875	.591311	.500491	.430041	.373523	.327222	.288709	.256276	.228718	.205103
26	.726681	.603899	.514336	.444332	.387790	.341199	.302238	.269280	.241137	.216929
27	.735837	.615757	.527435	.457946	.401488	.354711	.315386	.281968	.253300	.228535
28	.744404	.626944	.539914	.470981	.414658	.367742	.328131	.294313	.265188	.239935
29	.752437	.637514	.551741	.483431	.427307	.380334	.340477	.306326	.276805	.251110
30	.759984	.647501	.563023	.495347	.439475	.392490	.352461	.318033	.288158	.262062
40	.816139	.723938	.651356	.590773	.538846	.493686	.453976	.418785	.387401	.359271
60	.874843	.807778	.752424	.704238	.661334	.622640	.587440	.555224	.525598	.498272
80	.905160	.852653	.808266	.768805	.732964	.700027	.669520	.641124	.614572	.589678
100	.923660	.880557	.843610	.810333	.779746	.751296	.724666	.699598	.675935	.653520
120	.936178	.899588	.867973	.839253	.812632	.787686	.764150	.741841	.720623	.700389
140	.945137	.913391	.885776	.860534	.836998	.814820	.793780	.773732	.754565	.736197
170	.954680	.928199	.904999	.883652	.863624	.844636	.826518	.809156	.792465	.776383
200	.961395	.938685	.918687	.900202	.882782	.866197	.850307	.835018	.820262	.805990
240	.967765	.948679	.931793	.916116	.901281	.887100	.873459	.860284	.847521	.835131
320	.975762	.961296	.948422	.936405	.924972	.913987	.903369	.893064	.883033	.873250
440	.922336	.971725	.962235	.953337	.944835	.936632	.928671	.920913	.913333	.905910
600	.987028	.979198	.972173	.965563	.959229	.953099	.947133	.941302	.935589	.929978
800	.990261	.984364	.979060	.974060	.969257	.964600	.960057	.955610	.951243	.946947
1000	.992204	.987475	.983215	.979193	.975326	.971571	.967905	.964310	.960776	.957296

(Continuación Tabla C.2)

ν_E	$\nu_H = 1$	$\nu_H = 2$	$\nu_H = 3$	$\nu_H = 4$	$\nu_H = 5$	$\nu_H = 6$	$\nu_H = 7$	$\nu_H = 8$	$\nu_H = 9$	$\nu_H = 10$
	$p = 4$									
1	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
2	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
3	.000000	.000000	.000000	.000000	.000000	.000001	.000001	.000001	.000002	.000002
4	.001378	.000292	.000127	.000075	.000052	.000040	.000033	.000029	.000026	.000025
5	.025529	.006091	.002314	.001128	.000647	.000416	.000292	.000218	.000172	.000141
6	.076071	.023604	.010010	.005073	.002903	.001818	.001223	.000872	.000652	.000508
7	.135374	.050839	.024047	.013014	.007737	.004938	.003338	.002365	.001745	.001333
8	.194043	.083695	.043226	.024857	.015415	.010129	.006975	.004994	.003698	.002819
9	.248619	.118995	.065947	.039919	.025729	.017408	.012249	.008907	.006664	.005112
10	.298130	.154758	.090794	.057378	.038260	.026586	.019107	.014130	.010706	.008288
11	.342596	.189778	.116701	.076502	.052524	.037385	.027402	.020589	.015806	.012365
12	.382448	.223411	.142927	.096664	.068077	.049495	.036933	.028170	.021899	.017314
13	.418181	.255376	.168939	.117377	.084546	.062632	.047493	.036731	.028895	.023075
14	.450335	.285511	.194414	.138286	.101586	.076537	.058886	.046115	.036676	.029572
15	.479286	.313829	.219113	.159131	.118954	.090983	.070925	.056188	.045140	.036722
16	.505512	.340400	.242944	.179688	.136434	.105779	.083443	.066806	.054181	.044440
17	.529312	.365253	.265812	.199832	.153891	.120780	.096316	.077856	.063688	.052645
18	.551035	.388530	.287689	.219490	.171171	.135856	.109411	.089236	.073577	.061263
19	.570858	.410325	.308599	.238570	.188209	.150905	.122643	.100843	.083764	.070213
20	.589077	.430766	.388552	.257052	.204926	.165853	.135926	.112607	.094180	.079441
21	.605832	.449947	.347546	.274909	.221288	.180626	.149180	.124462	.104757	.088877
22	.621318	.467988	.365676	.292142	.237242	.195197	.162364	.136342	.115440	.098474
23	.635634	.484922	.382934	.308765	.252783	.209511	.175434	.148204	.126185	.108191
24	.648934	.500883	.399402	.324767	.267896	.223535	.188341	.160009	.136950	.117977
25	.661320	.515918	.415077	.340175	.282568	.237277	.201067	.171726	.147695	.127818
26	.672864	.530124	.430041	.355004	.296810	.250710	.213597	.183333	.158399	.137656
27	.683663	.543561	.444332	.369254	.310608	.263809	.225900	.194794	.169017	.147483
28	.693769	.556262	.457946	.382979	.323980	.276602	.237971	.206105	.179569	.157274
29	.703259	.568303	.470981	.396197	.336947	.289051	.249798	.217241	.189991	.167006
30	.712188	.579734	.483431	.408914	.349488	.301188	.261373	.228198	.200311	.176673
40	.778877	.668158	.582817	.513297	.455181	.405867	.363565	.326959	.295085	.267163
60	.849044	.767047	.700066	.642556	.592126	.547349	.507256	.471148	.438462	.408771
80	.885442	.820705	.766251	.718260	.675124	.635912	.600023	.566986	.536460	.508176
100	.907714	.854312	.808614	.767700	.730354	.695928	.663968	.634166	.606280	.580112
120	.922736	.877325	.838018	.802443	.769650	.739118	.710513	.683595	.658183	.634132
140	.933554	.894066	.859605	.828176	.798994	.771635	.745829	.721386	.698162	.676045
170	.945088	.912072	.883006	.856283	.831279	.807662	.785224	.763821	.743347	.723717
200	.953211	.924848	.899727	.876499	.854647	.833900	.814087	.795095	.776838	.759251
240	.960919	.937047	.915781	.896012	.877319	.859482	.842366	.825881	.809961	.794554
320	.970605	.952477	.936212	.920990	.906503	.892593	.879164	.866153	.853513	.841211
440	.978571	.965253	.953233	.941922	.931100	.920655	.910522	.900654	.891022	.881602
600	.984259	.974422	.965507	.957084	.948995	.941160	.933530	.926075	.918772	.911606
800	.988181	.980767	.974028	.967644	.961498	.955529	.949702	.943994	.938390	.932877
1000	.990538	.984589	.979173	.974034	.969078	.964257	.959545	.954922	.950376	.945898

(Continuación Tabla C.2)

ν_E	$\nu_H = 1$	$\nu_H = 2$	$\nu_H = 3$	$\nu_H = 4$	$\nu_H = 5$	$\nu_H = 6$	$\nu_H = 7$	$\nu_H = 8$	$\nu_H = 9$	$\nu_H = 10$
	$p = 5$									
1	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
2	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
3	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
4	.000000	.000000	.000000	.000000	.000001	.000001	.000001	.000001	.000001	.000001
5	.001598	.000291	.000105	.000052	.000031	.000021	.000015	.000012	.000010	.000008
6	.021145	.004391	.001479	.000647	.000335	.000197	.000126	.000087	.000064	.000049
7	.062771	.016898	.006357	.002903	.001514	.000872	.000544	.000361	.000253	.000185
8	.113526	.037390	.015792	.007737	.004208	.002479	.001557	.001032	.000716	.000516
9	.165351	.063279	.029433	.015415	.008787	.005348	.003433	.002304	.001607	.001159
10	.214794	.092191	.046378	.025729	.015321	.009639	.006343	.004335	.003062	.002225
11	.260635	.122403	.065660	.038260	.023674	.015360	.010358	.007216	.005173	.003802
12	.302608	.152793	.086448	.052524	.033618	.022418	.015467	.010980	.007991	.005946
13	.340813	.182662	.108110	.068077	.044878	.030680	.021607	.015611	.011530	.008685
14	.375528	.211602	.130131	.084546	.057198	.039965	.028683	.021031	.015774	.012024
15	.407128	.239373	.152160	.101586	.070324	.050117	.036584	.027266	.020687	.015949
16	.435899	.265851	.173959	.118954	.084048	.060965	.045199	.034145	.026219	.020428
17	.462173	.291015	.195322	.136434	.098187	.072367	.054409	.041618	.032312	.025427
18	.486266	.314859	.216138	.153891	.112582	.084178	.064111	.049602	.038909	.030904
19	.508362	.337418	.236338	.171171	.127108	.096308	.074209	.058024	.045951	.036810
20	.528714	.358776	.255858	.188209	.141662	.108634	.084619	.066805	.053373	.043100
21	.547516	.378956	.274710	.204926	.156176	.121083	.095254	.075885	.061122	.049724
22	.564905	.398038	.292843	.221288	.170563	.133590	.106063	.085203	.069149	.056652
23	.581036	.416105	.310304	.237242	.184782	.146095	.116974	.094699	.077408	.063832
24	.596032	.433216	.327083	.252783	.198795	.158544	.127948	.104337	.085849	.071231
25	.610030	.449429	.343191	.267896	.212568	.170898	.138945	.114058	.094444	.078809
26	.623126	.464800	.358665	.282568	.226071	.183129	.149909	.123843	.103144	.086536
27	.635368	.479382	.373523	.296810	.239294	.195207	.160826	.133657	.111931	.094385
28	.646832	.493247	.387790	.130608	.252224	.207116	.171667	.143454	.120766	.102328
29	.657645	.506421	.401488	.323980	.264873	.218828	.182408	.153240	.129630	.110336
30	.667803	.518945	.414658	.336947	.277200	.230347	.193043	.162971	.138499	.118393
40	.744010	.617178	.521747	.446045	.384424	.333492	.290896	.254963	.224433	.198322
60	.824764	.729155	.652037	.586878	.530670	.481578	.438367	.400085	.365997	.335520
80	.866847	.790730	.727186	.671775	.622536	.578316	.538319	.501966	.468774	.438392
100	.892643	.829563	.775817	.728040	.684827	.645343	.609037	.575509	.544420	.515540
120	.910071	.856268	.809790	.767957	.729656	.694256	.661341	.630608	.601822	.574793
140	.922634	.875748	.834850	.797705	.763400	.731431	.701466	.673268	.646653	.621477
170	.936039	.896748	.862122	.830370	.800777	.772953	.746649	.721687	.697934	.675284
200	.945486	.911680	.881674	.853973	.827989	.803406	.780024	.757705	.736343	.715856
240	.954455	.925960	.900496	.876838	.854512	.833264	.812938	.793426	.774647	.756540
320	.965732	.944055	.924519	.906224	.888827	.872146	.856074	.840535	.825476	.810855
440	.975013	.959064	.944590	.930949	.917894	.905302	.893096	.881226	.866955	.858357
600	.981642	.969850	.959096	.948913	.939124	.929642	.920411	.911369	.902572	.893921
800	.986214	.977320	.969181	.961450	.953996	.946753	.939682	.932756	.925957	.919273
1000	.988963	.981823	.975277	.969047	.963029	.957171	.951441	.945820	.940292	.934848

(Continuación Tabla C.2)

ν_E	$\nu_H = 1$	$\nu_H = 2$	$\nu_H = 3$	$\nu_H = 4$	$\nu_H = 5$	$\nu_H = 6$	$\nu_H = 7$	$\nu_H = 8$	$\nu_H = 9$	$\nu_H = 10$
	$p = 6$									
1	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
2	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
3	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
4	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
5	.000007	.000002	.000001	.000001	.000001	.000000	.000000	.000000	.000000	.000000
6	.002045	.000315	.000095	.000040	.000021	.000012	.000008	.000006	.000004	.000003
7	.018804	.003479	.001052	.000416	.000197	.000106	.000063	.000040	.000027	.000020
8	.053911	.012883	.004369	.001818	.000872	.000465	.000270	.000168	.000111	.000076
9	.098038	.028824	.011018	.004938	.002479	.001358	.000798	.000497	.000325	.000222
10	.144274	.049685	.021043	.010129	.005348	.003035	.001826	.001155	.000762	.000521
11	.189355	.073697	.033966	.017408	.009639	.005672	.003507	.002263	.001514	.001046
12	.231866	.099450	.049161	.026586	.015360	.009348	.005940	.003915	.002664	.001865
13	.271356	.125933	.066012	.037385	.022418	.014071	.009172	.006173	.004273	.003033
14	.307797	.152453	.083979	.049495	.030680	.019795	.013205	.009066	.006381	.004592
15	.341285	.178581	.102644	.062632	.039965	.026433	.018012	.012593	.009005	.006568
16	.372033	.204010	.121656	.076537	.050117	.033893	.023544	.016741	.012147	.008974
17	.400304	.228568	.140775	.090983	.060965	.042061	.029737	.021472	.015794	.011811
18	.426364	.252176	.159796	.105779	.072367	.050834	.036522	.026746	.019924	.015070
19	.450349	.274785	.178574	.120780	.084178	.060119	.043825	.032520	.024510	.018734
20	.472532	.296393	.197017	.135856	.096308	.069818	.051576	.038739	.029518	.022785
21	.493091	.316990	.215044	.150905	.108634	.079840	.059715	.045350	.034906	.027193
22	.512182	.336628	.232604	.165853	.121083	.090122	.068178	.052311	.040646	.031936
23	.529913	.355328	.249666	.180626	.133590	.100596	.076899	.059574	.046695	.036988
24	.546452	.373143	.266216	.195197	.146095	.111189	.085836	.060790	.053016	.042316
25	.561889	.390109	.282253	.209511	.158544	.121873	.094944	.074824	.059586	.047895
26	.576348	.406285	.297740	.223535	.170898	.132587	.104168	.082735	.066362	.053696
27	.589899	.421688	.312738	.237277	.183129	.143309	.113485	.090793	.073318	.059697
28	.602633	.436379	.327222	.250710	.195207	.153998	.122849	.098970	.080420	.065867
29	.614602	.450416	.341199	.263809	.207116	.164629	.132250	.107224	.087654	.072196
30	.625896	.463794	.354711	.276602	.218828	.175171	.141648	.115539	.094994	.078649
40	.710937	.569976	.466792	.387183	.324162	.273470	.232192	.198251	.170132	.146678
60	.801604	.693451	.607528	.536153	.475641	.423707	.378774	.339636	.305361	.275238
80	.849063	.762264	.690479	.628610	.574313	.526153	.483144	.444543	.409736	.378269
100	.878218	.805945	.744748	.690824	.642495	.598763	.558956	.522538	.489125	.458377
120	.897944	.836112	.782919	.735354	.692128	.652489	.615927	.582063	.550602	.521300
140	.912172	.858176	.811198	.768751	.729786	.693709	.660119	.628724	.599296	.571649
170	.927365	.882016	.842092	.805615	.771776	.740119	.710350	.682254	.655667	.630455
200	.938078	.899001	.864314	.832375	.802523	.774395	.747758	.722444	.698328	.675308
240	.948255	.915270	.885761	.858391	.832628	.808187	.784886	.762599	.741229	.720701
320	.961056	.935919	.913212	.891956	.871772	.852459	.833892	.815985	.798676	.781916
440	.971597	.953076	.936212	.920308	.905097	.890438	.876249	.862471	.849063	.835996
600	.979129	.965422	.952870	.940969	.929529	.918448	.907669	.897152	.886868	.876798
800	.984325	.973979	.964469	.955420	.946689	.938203	.929921	.921812	.913858	.906042
1000	.987450	.979142	.971487	.964187	.957129	.950256	.943532	.936937	.930455	.924073

(Continuación Tabla C.2)

ν_E	$\nu_H = 1$	$\nu_H = 2$	$\nu_H = 3$	$\nu_H = 4$	$\nu_H = 5$	$\nu_H = 6$	$\nu_H = 7$	$\nu_H = 8$	$\nu_H = 9$	$\nu_H = 10$
	$p = 7$									
1	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
2	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
3	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
4	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
5	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
6	.000043	.000006	.000002	.000001	.000001	.000000	.000000	.000000	.000000	.000000
7	.002625	.000350	.000091	.000033	.000015	.000008	.000005	.000003	.000002	.000002
8	.017612	.002953	.000809	.000292	.000126	.000063	.000034	.000020	.000013	.000009
9	.047835	.010329	.003195	.001223	.000543	.000270	.000147	.000086	.000053	.000035
10	.086645	.023060	.008067	.003338	.001558	.000798	.000440	.000259	.000160	.000104
11	.128234	.040186	.015642	.006974	.003433	.001836	.001035	.000619	.000387	.000252
12	.169506	.060396	.025707	.012249	.006343	.003508	.002048	.001252	.000796	.000525
13	.209026	.082538	.037857	.019109	.010357	.005940	.003571	.002234	.001448	.000967
14	.246203	.105734	.051646	.027402	.015466	.009172	.005668	.003628	.002395	.001625
15	.280861	.129346	.066659	.036933	.021607	.013206	.008371	.005476	.003682	.002537
16	.313032	.152929	.082533	.047494	.028684	.018013	.011688	.007801	.005337	.003733
17	.342842	.176179	.098971	.058884	.035686	.023544	.015606	.010611	.007379	.005235
18	.370455	.198894	.115731	.070921	.045199	.029736	.020096	.013900	.009814	.007057
19	.396050	.220944	.132623	.083445	.054409	.036520	.025122	.017653	.012640	.009204
20	.419802	.242252	.149498	.096315	.064111	.043824	.030640	.021845	.015847	.011676
21	.441876	.262777	.166240	.109415	.074209	.051579	.036603	.026450	.019422	.014469
22	.462425	.282503	.182765	.122645	.084616	.059717	.042965	.031435	.023345	.017571
23	.481587	.301432	.199007	.135923	.095257	.068177	.049678	.036769	.027595	.020971
24	.499486	.319577	.214919	.149181	.106063	.076901	.056697	.042416	.032148	.024653
25	.516238	.336959	.230467	.162364	.116978	.085838	.063980	.048346	.036980	.028599
26	.531942	.353606	.245631	.175429	.127951	.094941	.071488	.054525	.042067	.032794
27	.546689	.369546	.260395	.188340	.138940	.104168	.079183	.060924	.047385	.037217
28	.560561	.384810	.274752	.201068	.149909	.113482	.087032	.067514	.052911	.041851
29	.573629	.399430	.288701	.213591	.160826	.122851	.095005	.095005	.074268	.058622
30	.585961	.413438	.302243	.225894	.141667	.132247	.103073	.081161	.064496	.051680
40	.679228	.525996	.417050	.335433	.272668	.223571	.184671	.153533	.128393	.107941
60	.779306	.659576	.566032	.489695	.426135	.372561	.327012	.288026	.254476	.225471
80	.831906	.735024	.655779	.588321	.529875	.478709	.433602	.393626	.358051	.326284
100	.864288	.783251	.715144	.655689	.602930	.555673	.513081	.474521	.439488	.407570
120	.886219	.816680	.757179	.704361	.656738	.613420	.573796	.537400	.503866	.472893
140	.902052	.841199	.788462	.741086	.697881	.658148	.621410	.587314	.555578	.525974
170	.918970	.867751	.822764	.781839	.744063	.708913	.676042	.645194	.616167	.588800
200	.930906	.886705	.847518	.811553	.778074	.746666	.717058	.689053	.662499	.637274
240	.942249	.904887	.871471	.840546	.811527	.784091	.758031	.733198	.709478	.686784
320	.956525	.928004	.902213	.878097	.855239	.833417	.812491	.792362	.772959	.754224
440	.968286	.947243	.928043	.909937	.892635	.875985	.859892	.844294	.829142	.814403
600	.976693	.961103	.946788	.933208	.920155	.907522	.895244	.883276	.871588	.860157
800	.982494	.970720	.959861	.949517	.939535	.929836	.920373	.911114	.902038	.893128
1000	.985983	.976524	.967778	.959426	.951346	.943478	.935782	.928236	.920822	.913527

(Continuación Tabla C.2)

ν_E	$\nu_H = 1$	$\nu_H = 2$	$\nu_H = 3$	$\nu_H = 4$	$\nu_H = 5$	$\nu_H = 6$	$\nu_H = 7$	$\nu_H = 8$	$\nu_H = 9$	$\nu_H = 10$
	$p = 8$									
1	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
2	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
3	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
4	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
5	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
6	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000	.000000
7	.000138	.000015	.000004	.000001	.000001	.000000	.000000	.000000	.000000	.000000
8	.003295	.000393	.000090	.000029	.000012	.000006	.000003	.000002	.000001	.000001
9	.017079	.002632	.000659	.000218	.000087	.000040	.000020	.000011	.000007	.000004
10	.043574	.008626	.002458	.000872	.000361	.000168	.000086	.000047	.000028	.000017
11	.078039	.019031	.006148	.002365	.001032	.000497	.000259	.000144	.000085	.000052
12	.115676	.033314	.012011	.004993	.002304	.001155	.000619	.000351	.000209	.000130
13	.153630	.050518	.019990	.008908	.004335	.002263	.001252	.000727	.000441	.000278
14	.190453	.069716	.029839	.014129	.007216	.003915	.002234	.001331	.000824	.000527
15	.225477	.090151	.041241	.020590	.010980	.006173	.003628	.002215	.001399	.000910
16	.258443	.111245	.053875	.028171	.015610	.009065	.005476	.003422	.002203	.001457
17	.289300	.132575	.067447	.036729	.021061	.012594	.007801	.004982	.003269	.002197
18	.318105	.153836	.081699	.046115	.027265	.016740	.010611	.006915	.004617	.003151
19	.344966	.174814	.096415	.056185	.034144	.021472	.013900	.009228	.006265	.004339
20	.370015	.195359	.111416	.066805	.041616	.026747	.017653	.011923	.008219	.005771
21	.393387	.215374	.126559	.077857	.049601	.032519	.021845	.014991	.010483	.007456
22	.415217	.234796	.141726	.089233	.058021	.038737	.026450	.018419	.013053	.009397
23	.435632	.253588	.156826	.100843	.066804	.045350	.031435	.022192	.015923	.011593
24	.454749	.271732	.171785	.112606	.075884	.052311	.036769	.026287	.019081	.014041
25	.472677	.289225	.186549	.124457	.085199	.059573	.042416	.030685	.022515	.016733
26	.489514	.306072	.201075	.136338	.094698	.067091	.048346	.035361	.026210	.019663
27	.505352	.322285	.215331	.148203	.104332	.074826	.054525	.040293	.030150	.022818
28	.520271	.337880	.229293	.160010	.114060	.082739	.060924	.045457	.034319	.026189
29	.534345	.352879	.242945	.171728	.123844	.090796	.067514	.050831	.038700	.029764
30	.547639	.367302	.256277	.183330	.133653	.098967	.074268	.056394	.043276	.033529
40	.648630	.484826	.311902	.289857	.228618	.182082	.146235	.118316	.096365	.078964
60	.757690	.627279	.527185	.447009	.381482	.327255	.281978	.243910	.211718	.184362
80	.815243	.708843	.622840	.550577	.488795	.435425	.388992	.348380	.312704	.281253
100	.850742	.761330	.686819	.622411	.565838	.515687	.470954	.430871	.394827	.632322
120	.874811	.797857	.732425	.674791	.623251	.576764	.534599	.496197	.461114	.428382
140	.892201	.824719	.766516	.714559	.667497	.624521	.585067	.548712	.515117	.484002
170	.910793	.853874	.804039	.758920	.717494	.679163	.643522	.610267	.579158	.549999
200	.923918	.874725	.831204	.791410	.754525	.720081	.687764	.657345	.628642	.601508
240	.936396	.894758	.857556	.823223	.791114	.760867	.732246	.705079	.679234	.654605
320	.952108	.920269	.891472	.864586	.839159	.814944	.791784	.769570	.748216	.727659
440	.965057	.941534	.920045	.899793	.880463	.861889	.843968	.826629	.809821	.793502
600	.974316	.956873	.940825	.925599	.910972	.896826	.883093	.869724	.856684	.843948
800	.980707	.967524	.955338	.943721	.932512	.921624	.911008	.900630	.890464	.880494
1000	.984551	.973955	.964134	.954746	.945661	.936815	.928167	.919691	.911367	.903183

Tabla C.3 Estadística $D^2_{(n)}$ para detectar un outlier en una normal multivariada
Percentiles superiores de la estadídistica $D^2_{(n)}$

n	$p=2$		$p=3$		$p=4$		$p=5$	
	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$
5	3.17	3.19						
6	4.00	4.11	4.14	4.16				
7	4.71	4.95	5.01	5.10	5.12	5.14		
8	5.32	5.70	5.77	5.97	6.01	6.09	6.11	6.12
9	5.85	6.37	6.43	6.76	6.80	6.97	7.01	7.08
10	6.32	6.97	7.01	7.47	7.50	7.79	7.82	7.98
12	7.10	8.00	7.99	8.70	8.67	9.20	9.19	9.57
14	7.74	8.84	8.78	9.71	9.61	10.37	10.29	10.90
16	8.27	9.54	9.44	10.56	10.39	11.36	11.20	12.02
18	8.73	10.15	10.00	11.28	11.06	12.20	11.96	12.98
20	9.13	10.67	10.49	11.91	11.63	12.93	12.62	13.81
25	9.94	11.73	11.48	13.18	12.78	14.40	13.94	15.47
30	10.58	12.54	12.24	14.14	13.67	15.51	14.95	16.73
35	11.10	13.20	12.85	14.92	14.37	16.40	15.75	17.73
40	11.53	13.74	13.36	15.56	14.96	17.13	16.41	18.55
45	11.90	14.20	13.80	16.10	15.46	17.74	16.97	19.24
50	12.23	14.60	14.18	16.56	15.89	18.27	17.45	19.83
100	14.22	16.95	16.45	19.26	18.43	21.30	20.26	23.17
200	15.99	18.94	18.42	21.47	20.59	23.72	22.59	25.82
500	18.12	21.22	20.75	23.95	23.06	26.37	25.21	28.62

Tabla C.4 Polinomios ortogonales

		Variable										
p	Polinomio	1	2	3	4	5	6	7	8	9	10	$c_i'c_i$
3	Lineal	-1	0	1								2
	Cuadrático	1	-2	1								6
4	Lineal	-3	-1	1	3							20
	Cuadrático	1	-1	-1	1							4
	Cúbico	-1	3	-3	1							20
5	Lineal	-2	-1	0	1	2						10
	Cuadrático	2	-1	-2	-1	2						14
	Cúbico	-1	2	0	-2	1						10
	Cuarto	1	-4	6	-4	1						70
6	Lineal	-5	-3	-1	1	3	5					70
	Cuadrático	5	-1	-4	-4	-1	5					84
	Cúbico	-5	7	4	-4	-7	5					180
	Cuarto	1	-3	2	2	-3	1					28
	Quinto	-1	5	-10	10	-5	1					252
7	Lineal	-3	-2	-1	0	1	2	3				28
	Cuadrático	5	0	-3	-4	-3	0	5				84
	Cúbico	-1	1	1	0	-1	-1	1				6
	Cuarto	3	-7	1	6	1	-7	3				154
	Quinto	-1	4	-5	0	5	-4	1				84
	Sexto	1	-6	15	-20	15	-6	1				924
8	Lineal	-7	-5	-3	-1	1	3	5	7			168
	Cuadrático	7	1	-3	-5	-5	-3	1	7			168
	Cúbico	-7	5	7	3	-3	-7	-5	7			264
	Cuarto	7	-13	-3	9	9	-3	-13	7			616
	Quinto	-7	23	-17	-15	15	17	-23	7			2,184
	Sexto	1	-5	9	-5	-5	9	-5	1			264
	Séptimo	-1	7	-21	35	-35	21	-7	1			3,432
9	Lineal	-4	-3	-2	-1	0	1	2	3	4		60
	Cuadrático	28	7	-8	-17	-20	-17	-8	7	28		2,772
	Cúbico	-14	7	13	9	0	-9	-13	-7	14		990
	Cuarto	14	-21	-11	9	18	9	-11	-21	14		2,002
	Quinto	-4	11	-4	-9	0	9	4	-11	4		468
	Sexto	4	-17	22	1	-20	1	22	-17	4		1,980
	Séptimo	-1	6	-14	14	0	-14	14	-6	1		858
	Octavo	1	-8	28	-56	70	-56	28	-8	1		12,870
10	Lineal	-9	-7	-5	-3	-1	1	3	5	7	9	330
	Cuadrático	6	2	-1	-3	-4	-4	-3	-1	2	6	132
	Cúbico	-42	14	35	31	12	-12	-31	-35	-14	42	8,580
	Cuarto	18	-22	-17	3	18	18	3	-17	-22	18	2,860
	Quinto	-6	14	-1	-11	-6	6	11	1	-14	6	780
	Sexto	3	-11	10	6	-8	-8	6	10	11	3	660
	Séptimo	-9	47	-86	92	56	-56	-42	86	-47	9	29,172
	Octavo	1	-7	20	-28	14	14	-28	20	-7	1	2,860
	Noveno	-1	9	-36	84	-126	126	-84	36	-9	1	48,620

Tabla C.5 Percentiles de la distribución normal estándar: $P(Z \leq z) = \Phi(z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
-2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
-2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
-2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
-2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
-2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
-2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
-2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
-2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
-2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
-2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
-1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
-1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
-1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
-1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
-1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
-1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
-1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
-1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
-1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
-1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
-0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
-0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
-0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
-0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
-0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
-0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
-0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
-0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
-0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465

Tabla (Continuación tabla C.5) Percentiles de la distribución normal estándar: $P(Z \leq z) = \Phi(z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900

Tabla C.6 Cuantiles de la distribución t-Student

G.L.	α									
ν	0.90	0.95	0.975	0.99	0.995	0.10	0.05	0.025	0.01	0.001
1	-3.07768	-6.31375	-12.7062	-31.8205	-63.6567	3.07768	6.31375	12.7062	31.8205	63.6567
2	-1.88562	-2.91999	-4.3027	-6.9646	-9.9248	1.88562	2.91999	4.3027	6.9646	9.9248
3	-1.63774	-2.35336	-3.1824	-4.5407	-5.8409	1.63774	2.35336	3.1824	4.5407	5.8409
4	-1.53321	-2.13185	-2.7764	-3.7469	-4.6041	1.53321	2.13185	2.7764	3.7469	4.6041
5	-1.47588	-2.01505	-2.5706	-3.3649	-4.0321	1.47588	2.01505	2.5706	3.3649	4.0321
6	-1.43976	-1.94318	-2.4469	-3.1427	-3.7074	1.43976	1.94318	2.4469	3.1427	3.7074
7	-1.41492	-1.89458	-2.3646	-2.9980	-3.4995	1.41492	1.89458	2.3646	2.9980	3.4995
8	-1.39682	-1.85955	-2.3060	-2.8965	-3.3554	1.39682	1.85955	2.3060	2.8965	3.3554
9	-1.38303	-1.83311	-2.2622	-2.8214	-3.2498	1.38303	1.83311	2.2622	2.8214	3.2498
10	-1.37218	-1.81246	-2.2281	-2.7638	-3.1693	1.37218	1.81246	2.2281	2.7638	3.1693
11	-1.36343	-1.79588	-2.2010	-2.7181	-3.1058	1.36343	1.79588	2.2010	2.7181	3.1058
12	-1.35622	-1.78229	-2.1788	-2.6810	-3.0545	1.35622	1.78229	2.1788	2.6810	3.0545
13	-1.35017	-1.77093	-2.1604	-2.6503	-3.0123	1.35017	1.77093	2.1604	2.6503	3.0123
14	-1.34503	-1.76131	-2.1448	-2.6245	-2.9768	1.34503	1.76131	2.1448	2.6245	2.9768
15	-1.34061	-1.75305	-2.1314	-2.6025	-2.9467	1.34061	1.75305	2.1314	2.6025	2.9467
16	-1.33676	-1.74588	-2.1199	-2.5835	-2.9208	1.33676	1.74588	2.1199	2.5835	2.9208
17	-1.33338	-1.73961	-2.1098	-2.5669	-2.8982	1.33338	1.73961	2.1098	2.5669	2.8982
18	-1.33039	-1.73406	-2.1009	-2.5524	-2.8784	1.33039	1.73406	2.1009	2.5524	2.8784
19	-1.32773	-1.72913	-2.0930	-2.5395	-2.8609	1.32773	1.72913	2.0930	2.5395	2.8609
20	-1.32534	-1.72472	-2.0860	-2.5280	-2.8453	1.32534	1.72472	2.0860	2.5280	2.8453
21	-1.32319	-1.72074	-2.0796	-2.5176	-2.8314	1.32319	1.72074	2.0796	2.5176	2.8314
22	-1.32124	-1.71714	-2.0739	-2.5083	-2.8188	1.32124	1.71714	2.0739	2.5083	2.8188
23	-1.31946	-1.71387	-2.0687	-2.4999	-2.8073	1.31946	1.71387	2.0687	2.4999	2.8073
24	-1.31784	-1.71088	-2.0639	-2.4922	-2.7969	1.31784	1.71088	2.0639	2.4922	2.7969
25	-1.31635	-1.70814	-2.0595	-2.4851	-2.7874	1.31635	1.70814	2.0595	2.4851	2.7874
26	-1.31497	-1.70562	-2.0555	-2.4786	-2.7787	1.31497	1.70562	2.0555	2.4786	2.7787
27	-1.31370	-1.70329	-2.0518	-2.4727	-2.7707	1.31370	1.70329	2.0518	2.4727	2.7707
28	-1.31253	-1.70113	-2.0484	-2.4671	-2.7633	1.31253	1.70113	2.0484	2.4671	2.7633
29	-1.31143	-1.69913	-2.0452	-2.4620	-2.7564	1.31143	1.69913	2.0452	2.4620	2.7564
30	-1.31042	-1.69726	-2.0423	-2.4573	-2.7500	1.31042	1.69726	2.0423	2.4573	2.7500
35	-1.30621	-1.68957	-2.0301	-2.4377	-2.7238	1.30621	1.68957	2.0301	2.4377	2.7238
40	-1.30308	-1.68385	-2.0211	-2.4233	-2.7045	1.30308	1.68385	2.0211	2.4233	2.7045
45	-1.30065	-1.67943	-2.0141	-2.4121	-2.6896	1.30065	1.67943	2.0141	2.4121	2.6896
50	-1.29871	-1.67591	-2.0086	-2.4033	-2.6778	1.29871	1.67591	2.0086	2.4033	2.6778
60	-1.29582	-1.67065	-2.0003	-2.3901	-2.6603	1.29582	1.67065	2.0003	2.3901	2.6603
80	-1.29222	-1.66412	-1.9901	-2.3739	-2.6387	1.29222	1.66412	1.9901	2.3739	2.6387
100	-1.29007	-1.66023	-1.9840	-2.3642	-2.6259	1.29007	1.66023	1.9840	2.3642	2.6259
120	-1.28865	-1.65765	-1.9799	-2.3578	-2.6174	1.28865	1.65765	1.9799	2.3578	2.6174
∞	-1.28240	-1.64638	-1.9623	-2.3301	-2.5808	1.28240	1.64638	1.9623	2.3301	2.5808

Tabla C.7 Cuantiles de la distribución Ji–cuadrado
$P(Ji \geq \text{valor tabla}) = \alpha$

G.L.	α									
ν	0.90	0.95	0.975	0.99	0.995	0.10	0.05	0.025	0.01	0.001
1	0.016	0.004	0.001	0.000	0.000	2.71	3.84	5.02	6.63	7.88
2	0.211	0.103	0.051	0.020	0.010	4.61	5.99	7.38	9.21	10.60
3	0.584	0.352	0.216	0.115	0.072	6.25	7.81	9.35	11.34	12.84
4	1.064	0.711	0.484	0.297	0.207	7.78	9.49	11.14	13.28	14.86
5	1.610	1.145	0.831	0.554	0.412	9.24	11.07	12.83	15.09	16.75
6	2.204	1.635	1.237	0.872	0.676	10.64	12.59	14.45	16.81	18.55
7	2.833	2.167	1.690	1.239	0.989	12.02	14.07	16.01	18.48	20.28
8	3.490	2.733	2.180	1.646	1.344	13.36	15.51	17.53	20.09	21.95
9	4.168	3.325	2.700	2.088	1.735	14.68	16.92	19.02	21.67	23.59
10	4.865	3.940	3.247	2.558	2.156	15.99	18.31	20.48	23.21	25.19
11	5.578	4.575	3.816	3.053	2.603	17.28	19.68	21.92	24.72	26.76
12	6.304	5.226	4.404	3.571	3.074	18.55	21.03	23.34	26.22	28.30
13	7.042	5.892	5.009	4.107	3.565	19.81	22.36	24.74	27.69	29.82
14	7.790	6.571	5.629	4.660	4.075	21.06	23.68	26.12	29.14	31.32
15	8.547	7.261	6.262	5.229	4.601	22.31	25.00	27.49	30.58	32.80
16	9.312	7.962	6.908	5.812	5.142	23.54	26.30	28.85	32.00	34.27
17	10.085	8.672	7.564	6.408	5.697	24.77	27.59	30.19	33.41	35.72
18	10.865	9.390	8.231	7.015	6.265	25.99	28.87	31.53	34.81	37.16
19	11.651	10.117	8.907	7.633	6.844	27.20	30.14	32.85	36.19	38.58
20	12.443	10.851	9.591	8.260	7.434	28.41	31.41	34.17	37.57	40.00
21	13.240	11.591	10.283	8.897	8.034	29.62	32.67	35.48	38.93	41.40
22	14.041	12.338	10.982	9.542	8.643	30.81	33.92	36.78	40.29	42.80
23	14.848	13.091	11.689	10.196	9.260	32.01	35.17	38.08	41.64	44.18
24	15.659	13.848	12.401	10.856	9.886	33.20	36.42	39.36	42.98	45.56
25	16.473	14.611	13.120	11.524	10.520	34.38	37.65	40.65	44.31	46.93
26	17.292	15.379	13.844	12.198	11.160	35.56	38.89	41.92	45.64	48.29
27	18.114	16.151	14.573	12.879	11.808	36.74	40.11	43.19	46.96	49.64
28	18.939	16.928	15.308	13.565	12.461	37.92	41.34	44.46	48.28	50.99
29	19.768	17.708	16.047	14.256	13.121	39.09	42.56	45.72	49.59	52.34
30	20.599	18.493	16.791	14.953	13.787	40.26	43.77	46.98	50.89	53.67
35	24.797	22.465	20.569	18.509	17.192	46.06	49.80	53.20	57.34	60.27
40	29.051	26.509	24.433	22.164	20.707	51.81	55.76	59.34	63.69	66.77
45	33.350	30.612	28.366	25.901	24.311	57.51	61.66	65.41	69.96	73.17
50	37.689	34.764	32.357	29.707	27.991	63.17	67.50	71.42	76.15	79.49
60	46.459	43.188	40.482	37.485	35.534	74.40	79.08	83.30	88.38	91.95
80	64.278	60.391	57.153	53.540	51.172	96.58	101.88	106.63	112.33	116.32
100	82.358	77.929	74.222	70.065	67.328	118.50	124.34	129.56	135.81	140.17
120	100.624	95.705	91.573	86.923	83.852	140.23	146.57	152.21	158.95	163.65
∞	943.133	927.594	914.257	898.912	888.564	1057.72	1074.68	1089.53	1106.97	1118.95

Tabla C.8 Cuantiles de la distribución F
$P(F \geq F\ de\ tabla) = \alpha$

G. L. Num.		Grados de libertad del denominador ν_2											
ν_1	α	1	2	3	4	5	6	7	8	9	10	11	12
1	0.100	39.86	8.526	5.538	4.544	4.060	3.775	3.589	3.457	3.360	3.285	3.225	3.176
	0.050	161.45	18.513	10.128	7.708	6.607	5.987	5.591	5.317	5.117	4.964	4.844	4.747
	0.025	647.79	38.506	17.443	12.217	10.007	8.813	8.072	7.570	7.209	6.936	6.724	6.553
	0.010	4052.18	98.503	34.116	21.197	16.258	13.745	12.246	11.258	10.561	10.044	9.646	9.330
	0.005	16210.72	198.501	55.552	31.332	22.784	18.635	16.235	14.688	13.613	12.826	12.226	11.754
2	0.100	49.50	9.000	5.462	4.324	3.779	3.463	3.257	3.113	3.006	2.924	2.859	2.806
	0.050	199.50	19.000	9.552	6.944	5.786	5.143	4.737	4.459	4.256	4.102	3.982	3.885
	0.025	799.50	39.000	16.044	10.649	8.433	7.259	6.541	6.059	5.714	5.456	5.255	5.095
	0.010	4999.50	99.000	30.816	18.000	13.273	10.924	9.546	8.649	8.021	7.559	7.205	6.926
	0.005	19999.50	199.000	49.799	26.284	18.313	14.544	12.404	11.042	10.106	9.427	8.912	8.509
3	0.100	53.59	9.162	5.390	4.190	3.619	3.288	3.074	2.923	2.812	2.727	2.660	2.605
	0.050	215.71	19.164	9.276	6.591	5.409	4.757	4.346	4.066	3.862	3.708	3.587	3.490
	0.025	864.16	39.165	15.439	9.979	7.763	6.598	5.889	5.416	5.078	4.825	4.630	4.474
	0.010	5403.35	99.166	29.456	16.694	12.060	9.779	8.451	7.591	6.991	6.552	6.216	5.952
	0.005	21614.74	199.166	47.467	24.259	16.529	12.916	10.882	9.596	8.717	8.080	7.600	7.225
4	0.100	55.83	9.243	5.342	4.107	3.520	3.180	2.960	2.806	2.692	2.605	2.536	2.480
	0.050	224.58	19.247	9.117	6.388	5.192	4.533	4.120	3.837	3.633	3.478	3.356	3.259
	0.025	899.58	39.248	15.101	9.604	7.387	6.227	5.522	5.052	4.718	4.468	4.275	4.121
	0.010	5624.58	99.249	28.709	15.977	11.391	9.148	7.846	7.006	6.422	5.994	5.668	5.412
	0.005	22499.58	199.250	46.194	23.154	15.556	12.027	10.050	8.805	7.955	7.342	6.880	6.521
5	0.100	57.24	9.293	5.309	4.050	3.453	3.107	2.883	2.726	2.610	2.521	2.451	2.394
	0.050	230.16	19.296	9.013	6.256	5.050	4.387	3.971	3.687	3.481	3.325	3.203	3.105
	0.025	921.85	39.298	14.884	9.364	7.146	5.987	5.285	4.817	4.484	4.236	4.044	3.891
	0.010	5763.65	99.299	28.237	15.521	10.967	8.745	7.460	6.631	6.056	5.636	5.316	5.064
	0.005	23055.80	199.300	45.391	22.456	14.939	11.463	9.522	8.301	7.4712	6.872	6.421	6.071
6	0.100	58.20	9.326	5.284	4.009	3.404	3.054	2.827	2.668	2.550	2.460	2.389	2.331
	0.050	233.99	19.330	8.940	6.163	4.950	4.283	3.866	3.580	3.373	3.217	3.094	2.996
	0.025	937.11	39.331	14.734	9.197	6.977	5.819	5.118	4.651	4.319	4.072	3.880	3.728
	0.010	5858.99	99.333	27.910	15.206	10.672	8.466	7.191	6.370	5.801	5.385	5.069	4.820
	0.005	23437.11	199.333	44.838	21.974	14.513	11.073	9.155	7.952	7.133	6.544	6.101	5.757
7	0.100	58.91	9.349	5.266	3.979	3.367	3.014	2.784	2.624	2.505	2.414	2.341	2.282
	0.050	236.77	19.353	8.886	6.094	4.875	4.206	3.787	3.500	3.292	3.135	3.012	2.913
	0.025	948.22	39.355	14.624	9.074	6.853	5.695	4.994	4.528	4.197	3.949	3.758	3.606
	0.010	5928.36	99.356	27.671	14.975	10.455	8.260	6.992	6.177	5.612	5.200	4.886	4.639
	0.005	23714.57	199.357	44.434	21.621	14.200	10.785	8.885	7.694	6.884	6.302	5.864	5.524
8	0.100	59.44	9.367	5.251	3.954	3.339	2.983	2.751	2.589	2.469	2.377	2.304	2.244
	0.050	238.88	19.371	8.845	6.041	4.818	4.146	3.725	3.438	3.229	3.071	2.948	2.848
	0.025	956.66	39.373	14.539	8.979	6.757	5.599	4.899	4.433	4.102	3.854	3.663	3.511
	0.010	5981.07	99.374	27.489	14.798	10.289	8.101	6.840	6.028	5.467	5.056	4.744	4.499
	0.005	23925.41	199.375	44.125	21.352	13.961	10.565	8.678	7.495	6.693	6.115	5.682	5.345

(Continuación Tabla C.8)

G. L. Num.		Grados de libertad del denominador ν_2											
ν_1	α	1	2	3	4	5	6	7	8	9	10	11	12
9	0.100	59.86	9.381	5.240	3.935	3.316	2.95	2.727	2.561	2.440	2.347	2.273	2.213
	0.050	240.54	19.385	8.812	5.998	4.772	4.099	3.676	3.388	3.178	3.020	2.896	2.796
	0.025	963.28	39.387	14.473	8.904	6.681	5.523	4.823	4.357	4.026	3.779	3.587	3.435
	0.010	6022.47	99.388	27.345	14.659	10.157	7.976	6.718	5.910	5.351	4.942	4.631	4.387
	0.005	24091.00	199.388	43.882	21.139	13.771	10.391	8.513	7.338	6.541	5.967	5.536	5.202
10	0.100	60.19	9.392	5.230	3.919	3.297	2.936	2.702	2.538	2.416	2.322	2.248	2.187
	0.050	241.88	19.396	8.785	5.964	4.735	4.060	3.636	3.347	3.137	2.978	2.853	2.753
	0.025	968.63	39.398	14.418	8.843	6.619	5.461	4.761	4.295	3.963	3.716	3.525	3.373
	0.010	6055.85	99.399	27.228	14.545	10.051	7.874	6.620	5.814	5.256	4.849	4.539	4.296
	0.005	24224.49	199.400	43.685	20.966	13.618	10.250	8.380	7.210	6.417	5.846	5.418	5.085
11	0.100	60.47	9.401	5.222	3.906	3.281	2.919	2.683	2.518	2.396	2.301	2.226	2.166
	0.050	242.98	19.405	8.763	5.935	4.704	4.027	3.603	3.313	3.102	2.943	2.817	2.717
	0.025	973.03	39.407	14.374	8.793	6.567	5.409	4.709	4.243	3.912	3.664	3.473	3.321
	0.010	6083.32	99.408	27.132	14.452	9.962	7.789	6.538	5.734	5.177	4.771	4.462	4.219
	0.005	24334.36	199.409	43.523	20.824	13.491	10.132	8.269	7.104	6.314	5.746	5.319	4.988
12	0.100	60.71	9.408	5.215	3.895	3.268	2.904	2.668	2.502	2.378	2.284	2.208	2.147
	0.050	243.91	19.413	8.744	5.911	4.677	3.999	3.5746	3.283	3.072	2.912	2.787	2.686
	0.025	976.71	39.415	14.336	8.751	6.524	5.366	4.665	4.199	3.868	3.620	3.429	3.277
	0.010	6106.32	99.416	27.051	14.373	9.888	7.718	6.469	5.666	5.111	4.705	4.397	4.155
	0.005	24426.37	199.416	43.387	20.704	13.384	10.034	8.176	7.014	6.227	5.661	5.236	4.906
13	0.100	60.90	9.415	5.209	3.885	3.256	2.892	2.654	2.487	2.364	2.268	2.192	2.131
	0.050	244.69	19.419	8.728	5.891	4.655	3.976	3.550	3.259	3.047	2.887	2.761	2.660
	0.025	979.84	39.421	14.304	8.715	6.487	5.329	4.628	4.162	3.830	3.583	3.391	3.239
	0.010	6125.86	99.422	26.983	14.306	9.824	7.657	6.410	5.608	5.054	4.649	4.341	4.099
	0.005	24504.54	199.423	43.271	20.602	13.293	9.950	8.096	6.938	6.153	5.588	5.164	4.835
14	0.100	61.07	9.420	5.204	3.877	3.246	2.880	2.642	2.475	2.351	2.255	2.179	2.117
	0.050	245.36	19.424	8.714	5.873	4.635	3.955	3.529	3.237	3.025	2.864	2.738	2.637
	0.025	982.53	39.427	14.276	8.683	6.455	5.296	4.596	4.129	3.797	3.550	3.358	3.206
	0.010	6142.67	99.428	26.923	14.248	9.770	7.604	6.358	5.558	5.005	4.600	4.293	4.051
	0.005	24571.77	199.428	43.171	20.514	13.214	9.877	8.027	6.872	6.088	5.525	5.103	4.774
15	0.100	61.22	9.425	5.200	3.870	3.238	2.871	2.632	2.464	2.339	2.243	2.167	2.104
	0.050	245.95	19.429	8.702	5.857	4.618	3.938	3.510	3.218	3.006	2.845	2.718	2.616
	0.025	984.87	39.431	14.252	8.656	6.427	5.268	4.567	4.101	3.769	3.521	3.329	3.177
	0.010	6157.28	99.433	26.872	14.198	9.722	7.559	6.3143	5.515	4.962	4.558	4.250	4.009
	0.005	24630.21	199.433	43.084	20.438	13.146	9.814	7.967	6.814	6.032	5.470	5.048	4.721
16	0.100	61.35	9.429	5.196	3.863	3.230	2.862	2.623	2.454	2.329	2.233	2.156	2.093
	0.050	246.46	19.433	8.692	5.844	4.603	3.922	3.494	3.201	2.988	2.827	2.700	2.598
	0.025	986.92	39.435	14.231	8.632	6.403	5.243	4.542	4.076	3.744	3.496	3.304	3.151
	0.010	6170.10	99.437	26.826	14.153	9.680	7.518	6.275	5.476	4.924	4.520	4.213	3.972
	0.005	24681.47	199.437	43.008	20.371	13.086	9.758	7.914	6.763	5.982	5.422	5.001	4.674

(Continuación Tabla C.8)

G. L. Num.		Grados de libertad del denominador ν_2											
ν_1	α	1	2	3	4	5	6	7	8	9	10	11	12
20	0.100	61.74	9.44	5.184	3.844	3.206	2.836	2.594	2.424	2.298	2.200	2.123	2.059
	0.050	248.01	19.446	8.660	5.802	4.558	3.874	3.444	3.150	2.936	2.774	2.646	2.543
	0.025	993.10	39.448	14.167	8.559	6.328	5.168	4.466	3.999	3.666	3.418	3.226	3.072
	0.010	6208.73	99.449	26.689	14.019	9.552	7.395	6.155	5.359	4.808	4.405	4.099	3.858
	0.005	24835.97	199.450	42.777	20.167	12.903	9.588	7.753	6.608	5.831	5.274	4.855	4.529
25	0.100	62.05	9.451	5.174	3.828	3.187	2.814	2.571	2.399	2.272	2.173	2.095	2.031
	0.050	249.26	19.456	8.634	5.768	4.520	3.834	3.403	3.108	2.893	2.729	2.601	2.497
	0.025	998.08	39.458	14.115	8.501	6.267	5.106	4.404	3.936	3.603	3.354	3.161	3.007
	0.010	6239.83	99.459	26.579	13.910	9.449	7.296	6.057	5.263	4.713	4.311	4.005	3.764
	0.005	24960.34	199.460	42.591	20.002	12.755	9.451	7.622	6.481	5.708	5.152	4.735	4.411
30	0.100	62.26	9.458	5.168	3.817	3.174	2.800	2.555	2.383	2.254	2.155	2.076	2.0114
	0.050	250.10	19.462	8.616	5.745	4.495	3.808	3.3758	3.0794	2.863	2.699	2.570	2.466
	0.025	1001.41	39.465	14.080	8.461	6.226	5.065	4.362	3.894	3.560	3.311	3.117	2.963
	0.010	6260.65	99.466	26.504	13.837	9.379	7.228	5.992	5.198	4.648	4.246	3.941	3.700
	0.005	25043.63	199.466	42.4658	19.891	12.655	9.358	7.534	6.396	5.624	5.070	4.654	4.330
40	0.100	62.53	9.466	5.159	3.803	3.157	2.781	2.535	2.361	2.231	2.131	2.051	1.986
	0.050	251.14	19.471	8.594	5.717	4.463	3.774	3.3404	3.0427	2.825	2.660	2.530	2.425
	0.025	1005.60	39.473	14.036	8.411	6.175	5.012	4.308	3.839	3.505	3.255	3.061	2.906
	0.010	6286.78	99.474	26.410	13.745	9.291	7.143	5.908	5.115	4.566	4.165	3.859	3.619
	0.005	25148.15	199.475	42.308	19.751	12.529	9.240	7.422	6.287	5.518	4.965	4.550	4.228
60	0.100	62.79	9.475	5.151	3.789	3.140	2.762	2.514	2.339	2.208	2.107	2.026	1.959
	0.050	252.20	19.479	8.572	5.687	4.431	3.739	3.304	3.005	2.787	2.621	2.490	2.384
	0.025	1009.80	39.481	13.992	8.360	6.122	4.958	4.254	3.784	3.449	3.198	3.003	2.847
	0.010	6313.03	99.482	26.316	13.652	9.202	7.056	5.823	5.031	4.483	4.081	3.776	3.535
	0.005	25253.14	199.483	42.149	19.610	12.402	9.121	7.308	6.177	5.410	4.859	4.445	4.122
80	0.100	62.93	9.479	5.146	3.782	3.131	2.752	2.503	2.327	2.196	2.094	2.013	1.946
	0.050	252.72	19.483	8.560	5.673	4.415	3.722	3.285	2.986	2.767	2.600	2.469	2.362
	0.025	1011.91	39.485	13.969	8.334	6.096	4.931	4.226	3.756	3.420	3.169	2.974	2.817
	0.010	6326.20	99.487	26.268	13.605	9.157	7.013	5.780	4.989	4.440	4.039	3.733	3.492
	0.005	25305.80	199.487	42.069	19.539	12.338	9.061	7.251	6.121	5.355	4.804	4.391	4.069
120	0.100	63.06	9.483	5.142	3.775	3.122	2.742	2.4927	2.316	2.184	2.081	1.999	1.932
	0.050	253.25	19.487	8.549	5.658	4.398	3.704	3.267	2.966	2.747	2.580	2.448	2.340
	0.025	1014.02	39.490	13.947	8.309	6.069	4.904	4.198	3.727	3.391	3.139	2.944	2.787
	0.010	6339.39	99.491	26.221	13.558	9.111	6.969	5.737	4.946	4.397	3.996	3.690	3.449
	0.005	25358.57	199.491	41.989	19.468	12.273	9.001	7.193	6.064	5.300	4.750	4.336	4.014
∞	0.100	63.33	9.491	5.1337	3.760	3.105	2.722	2.470	2.292	2.159	2.055	1.972	1.903
	0.050	254.31	19.496	8.526	5.628	4.365	3.668	3.229	2.927	2.706	2.537	2.404	2.296
	0.025	1018.25	39.498	13.902	8.257	6.015	4.849	4.142	3.670	3.332	3.079	2.882	2.725
	0.010	6365.83	99.499	26.125	13.463	9.020	6.880	5.649	4.858	4.310	3.909	3.602	3.360
	0.005	25464.33	199.500	41.828	19.324	12.143	8.879	7.076	5.950	5.187	4.638	4.225	3.904

(Continuación Tabla C.8)

G. L. Num.		Grados de libertad del denominador ν_2											
ν_1	α	13	14	15	16	20	25	30	40	60	80	120	∞
1	0.100	3.136	3.102	3.073	3.048	2.974	2.917	2.880	2.835	2.791	2.769	2.747	2.705
	0.050	4.667	4.600	4.543	4.494	4.351	4.241	4.170	4.084	4.001	3.960	3.920	3.841
	0.025	6.414	6.297	6.199	6.115	5.871	5.686	5.567	5.423	5.285	5.218	5.152	5.024
	0.010	9.073	8.861	8.683	8.531	8.095	7.769	7.562	7.314	7.077	6.962	6.850	6.635
	0.005	11.373	11.060	10.798	10.575	9.943	9.475	9.179	8.827	8.494	8.334	8.178	7.879
2	0.100	2.763	2.726	2.695	2.668	2.589	2.528	2.488	2.440	2.393	2.370	2.347	2.302
	0.050	3.805	3.738	3.682	3.633	3.492	3.385	3.315	3.231	3.150	3.110	3.071	2.995
	0.025	4.965	4.856	4.765	4.686	4.461	4.290	4.182	4.050	3.925	3.864	3.804	3.689
	0.010	6.701	6.514	6.358	6.226	5.848	5.568	5.390	5.178	4.977	4.880	4.786	4.605
	0.005	8.186	7.921	7.700	7.513	6.986	6.598	6.354	6.066	5.794	5.665	5.539	5.298
3	0.100	2.560	2.522	2.489	2.461	2.380	2.317	2.276	2.226	2.177	2.153	2.129	2.083
	0.050	3.410	3.343	3.287	3.238	3.098	2.991	2.922	2.838	2.758	2.718	2.680	2.605
	0.025	4.347	4.241	4.152	4.076	3.858	3.694	3.589	3.463	3.342	3.284	3.226	3.116
	0.010	5.739	5.563	5.417	5.292	4.938	4.675	4.509	4.312	4.125	4.036	3.949	3.781
	0.005	6.925	6.680	6.476	6.303	5.817	5.461	5.238	4.975	4.728	4.611	4.497	4.279
4	0.100	2.433	2.394	2.361	2.332	2.248	2.184	2.142	2.090	2.040	2.016	1.992	1.944
	0.050	3.179	3.112	3.055	3.006	2.866	2.758	2.689	2.605	2.525	2.485	2.447	2.372
	0.025	3.995	3.891	3.804	3.729	3.514	3.353	3.249	3.126	3.007	2.950	2.894	2.785
	0.010	5.205	5.035	4.893	4.772	4.430	4.177	4.017	3.828	3.649	3.563	3.479	3.319
	0.005	6.233	5.998	5.802	5.637	5.174	4.835	4.623	4.373	4.139	4.028	3.920	3.715
5	0.100	2.346	2.306	2.273	2.243	2.158	2.092	2.049	1.996	1.945	1.920	1.895	1.847
	0.050	3.025	2.958	2.901	2.852	2.710	2.602	2.533	2.449	2.368	2.328	2.289	2.214
	0.025	3.766	3.663	3.576	3.502	3.289	3.128	3.026	2.903	2.786	2.729	2.673	2.566
	0.010	4.861	4.695	4.555	4.437	4.102	3.854	3.699	3.513	3.338	3.255	3.173	3.017
	0.005	5.791	5.562	5.372	5.211	4.7615	4.432	4.227	3.986	3.759	3.652	3.548	3.350
6	0.100	2.283	2.242	2.208	2.178	2.091	2.024	1.980	1.926	1.874	1.849	1.823	1.774
	0.050	2.915	2.847	2.790	2.741	2.598	2.490	2.420	2.335	2.254	2.214	2.175	2.098
	0.025	3.604	3.501	3.414	3.340	3.128	2.968	2.866	2.744	2.627	2.570	2.515	2.408
	0.010	4.620	4.455	4.318	4.201	3.871	3.627	3.473	3.291	3.118	3.036	2.955	2.802
	0.005	5.481	5.257	5.070	4.913	4.472	4.149	3.949	3.712	3.491	3.386	3.284	3.091
7	0.100	2.234	2.193	2.158	2.128	2.039	1.971	1.926	1.872	1.819	1.793	1.767	1.716
	0.050	2.832	2.764	2.706	2.657	2.514	2.404	2.334	2.249	2.166	2.126	2.086	2.009
	0.025	3.482	3.379	3.293	3.219	3.007	2.847	2.746	2.623	2.506	2.450	2.394	2.287
	0.010	4.441	4.277	4.141	4.025	3.698	3.456	3.304	3.123	2.953	2.871	2.791	2.639
	0.005	5.252	5.031	4.847	4.692	4.256	3.939	3.741	3.508	3.291	3.187	3.087	2.897
8	0.100	2.195	2.153	2.118	2.088	1.998	1.929	1.884	1.828	1.774	1.748	1.721	1.670
	0.050	2.766	2.698	2.640	2.591	2.447	2.337	2.266	2.180	2.096	2.056	2.016	1.938
	0.025	3.388	3.285	3.198	3.124	2.912	2.753	2.651	2.528	2.411	2.354	2.299	2.191
	0.010	4.302	4.139	4.004	3.889	3.564	3.323	3.172	2.992	2.823	2.741	2.662	2.511
	0.005	5.0761	4.856	4.674	4.520	4.089	3.775	3.580	3.349	3.134	3.032	2.932	2.744

(Continuación Tabla C.8)

G. L. Num.		Grados de libertad del denominador ν_2											
ν_1	α	13	14	15	16	20	25	30	40	60	80	120	∞
9	0.100	2.163	2.122	2.086	2.055	1.964	1.894	1.848	1.792	1.738	1.711	1.684	1.631
	0.050	2.714	2.645	2.587	2.537	2.392	2.282	2.210	2.124	2.040	1.999	1.958	1.879
	0.025	3.312	3.209	3.122	3.048	2.836	2.676	2.574	2.451	2.334	2.277	2.221	2.113
	0.010	4.191	4.029	3.894	3.780	3.456	3.217	3.066	2.887	2.718	2.637	2.558	2.407
	0.005	4.935	4.717	4.536	4.383	3.956	3.644	3.450	3.221	3.008	2.906	2.808	2.621
10	0.100	2.137	2.095	2.059	2.028	1.936	1.865	1.819	1.762	1.707	1.679	1.652	1.598
	0.050	2.671	2.602	2.543	2.493	2.347	2.236	2.164	2.077	1.992	1.951	1.910	1.830
	0.025	3.249	3.146	3.060	2.986	2.773	2.613	2.511	2.388	2.270	2.213	2.157	2.048
	0.010	4.100	3.939	3.804	3.690	3.368	3.129	2.979	2.800	2.631	2.550	2.472	2.321
	0.005	4.819	4.603	4.423	4.271	3.847	3.537	3.343	3.116	2.904	2.803	2.705	2.519
11	0.100	2.115	2.072	2.036	2.005	1.912	1.841	1.794	1.736	1.680	1.652	1.625	1.570
	0.050	2.634	2.565	2.506	2.456	2.309	2.197	2.125	2.037	1.952	1.910	1.869	1.788
	0.025	3.197	3.094	3.007	2.933	2.720	2.560	2.457	2.334	2.215	2.158	2.102	1.992
	0.010	4.024	3.864	3.729	3.616	3.294	3.055	2.905	2.727	2.558	2.477	2.399	2.247
	0.005	4.724	4.508	4.329	4.178	3.755	3.446	3.254	3.028	2.816	2.715	2.618	2.432
12	0.100	2.096	2.053	2.017	1.985	1.892	1.820	1.772	1.714	1.657	1.629	1.601	1.545
	0.050	2.603	2.534	2.475	2.424	2.277	2.164	2.092	2.003	1.917	1.875	1.833	1.752
	0.025	3.153	3.050	2.963	2.889	2.675	2.514	2.412	2.288	2.169	2.111	2.054	1.944
	0.010	3.960	3.800	3.666	3.552	3.231	2.993	2.843	2.664	2.496	2.415	2.336	2.184
	0.005	4.642	4.428	4.249	4.099	3.677	3.370	3.178	2.953	2.741	2.641	2.543	2.358
13	0.100	2.080	2.037	2.000	1.968	1.874	1.801	1.753	1.695	1.637	1.608	1.580	1.524
	0.050	2.576	2.507	2.448	2.397	2.249	2.136	2.062	1.973	1.887	1.844	1.802	1.720
	0.025	3.115	3.011	2.924	2.850	2.636	2.475	2.372	2.248	2.128	2.070	2.013	1.902
	0.010	3.905	3.745	3.611	3.498	3.176	2.938	2.789	2.610	2.441	2.360	2.281	2.130
	0.005	4.573	4.359	4.181	4.031	3.611	3.304	3.113	2.888	2.677	2.576	2.479	2.294
14	0.100	2.065	2.022	1.985	1.953	1.858	1.785	1.737	1.677	1.619	1.590	1.561	1.504
	0.050	2.553	2.483	2.424	2.373	2.224	2.111	2.037	1.947	1.860	1.817	1.775	1.691
	0.025	3.081	2.978	2.891	2.817	2.603	2.441	2.337	2.212	2.092	2.034	1.977	1.865
	0.010	3.857	3.697	3.563	3.450	3.129	2.891	2.741	2.563	2.394	2.313	2.233	2.081
	0.005	4.512	4.299	4.121	3.972	3.553	3.246	3.056	2.831	2.620	2.520	2.422	2.237
15	0.100	2.053	2.009	1.972	1.939	1.844	1.770	1.722	1.662	1.603	1.574	1.545	1.487
	0.050	2.533	2.463	2.403	2.352	2.203	2.088	2.014	1.924	1.836	1.793	1.750	1.666
	0.025	3.052	2.949	2.862	2.787	2.573	2.410	2.307	2.181	2.061	2.002	1.944	1.832
	0.010	3.815	3.655	3.522	3.408	3.088	2.850	2.700	2.521	2.352	2.270	2.191	2.038
	0.005	4.459	4.246	4.069	3.920	3.501	3.196	3.005	2.781	2.570	2.470	2.372	2.186
16	0.100	2.041	1.998	1.960	1.928	1.832	1.757	1.709	1.648	1.589	1.559	1.529	1.471
	0.050	2.514	2.444	2.384	2.333	2.183	2.069	1.994	1.903	1.815	1.771	1.728	1.643
	0.025	3.026	2.923	2.836	2.761	2.546	2.384	2.279	2.154	2.033	1.974	1.916	1.802
	0.010	3.778	3.618	3.485	3.372	3.051	2.813	2.663	2.484	2.314	2.233	2.153	2.000
	0.005	4.413	4.200	4.023	3.874	3.456	3.151	2.961	2.736	2.525	2.425	2.327	2.141

(Continuación Tabla C.8)

G. L. Num.		Grados de libertad del denominador ν_2											
ν_1	α	13	14	15	16	20	25	30	40	60	80	120	∞
20	0.100	2.006	1.962	1.924	1.891	1.793	1.717	1.667	1.605	1.543	1.512	1.482	1.420
	0.050	2.458	2.387	2.327	2.275	2.124	2.007	1.931	1.838	1.747	1.703	1.658	1.570
	0.025	2.947	2.843	2.755	2.680	2.464	2.300	2.195	2.067	1.944	1.884	1.824	1.708
	0.010	3.664	3.505	3.371	3.258	2.937	2.699	2.548	2.368	2.197	2.115	2.034	1.878
	0.005	4.270	4.058	3.882	3.734	3.317	3.013	2.823	2.598	2.387	2.286	2.188	2.000
25	0.100	1.977	1.932	1.893	1.860	1.761	1.683	1.631	1.567	1.503	1.471	1.439	1.375
	0.050	2.412	2.340	2.279	2.227	2.073	1.955	1.878	1.783	1.690	1.643	1.597	1.506
	0.025	2.882	2.777	2.689	2.613	2.395	2.230	2.123	1.994	1.868	1.807	1.746	1.626
	0.010	3.570	3.411	3.278	3.164	2.843	2.604	2.452	2.271	2.098	2.014	1.932	1.772
	0.005	4.152	3.941	3.766	3.618	3.202	2.898	2.707	2.482	2.269	2.167	2.068	1.877
30	0.100	1.957	1.911	1.872	1.838	1.738	1.658	1.606	1.541	1.475	1.442	1.409	1.341
	0.050	2.380	2.308	2.246	2.193	2.039	1.919	1.840	1.744	1.649	1.601	1.554	1.459
	0.025	2.837	2.732	2.643	2.567	2.348	2.181	2.073	1.942	1.815	1.752	1.689	1.566
	0.010	3.507	3.347	3.214	3.100	2.778	2.538	2.385	2.203	2.028	1.943	1.860	1.696
	0.005	4.072	3.861	3.686	3.538	3.123	2.818	2.627	2.401	2.187	2.084	1.983	1.789
40	0.100	1.931	1.885	1.845	1.810	1.708	1.627	1.573	1.505	1.437	1.402	1.367	1.295
	0.050	2.339	2.266	2.204	2.150	1.993	1.871	1.791	1.692	1.594	1.544	1.495	1.394
	0.025	2.779	2.674	2.585	2.508	2.287	2.118	2.008	1.875	1.744	1.679	1.614	1.483
	0.010	3.425	3.265	3.131	3.018	2.694	2.452	2.299	2.114	1.936	1.848	1.762	1.592
	0.005	3.970	3.760	3.584	3.437	3.021	2.715	2.524	2.295	2.078	1.973	1.870	1.669
60	0.100	1.904	1.857	1.816	1.781	1.676	1.593	1.537	1.467	1.395	1.358	1.320	1.240
	0.050	2.296	2.222	2.160	2.105	1.946	1.821	1.739	1.637	1.534	1.482	1.429	1.318
	0.025	2.720	2.614	2.524	2.447	2.223	2.051	1.940	1.802	1.666	1.598	1.529	1.388
	0.010	3.341	3.181	3.047	2.933	2.607	2.363	2.207	2.019	1.836	1.745	1.655	1.473
	0.005	3.865	3.655	3.480	3.332	2.915	2.608	2.415	2.183	1.962	1.853	1.746	1.532
80	0.100	1.890	1.842	1.801	1.766	1.660	1.575	1.518	1.446	1.372	1.333	1.293	1.207
	0.050	2.274	2.200	2.137	2.082	1.921	1.795	1.712	1.607	1.501	1.447	1.392	1.273
	0.025	2.689	2.583	2.492	2.415	2.190	2.016	1.903	1.764	1.625	1.554	1.483	1.333
	0.010	3.298	3.138	3.003	2.881	2.562	2.317	2.160	1.969	1.782	1.690	1.596	1.404
	0.005	3.812	3.601	3.426	3.278	2.861	2.553	2.358	2.124	1.899	1.789	1.678	1.454
120	0.100	1.875	1.828	1.786	1.750	1.643	1.557	1.498	1.424	1.347	1.307	1.264	1.168
	0.050	2.252	2.177	2.114	2.058	1.896	1.768	1.683	1.576	1.467	1.410	1.351	1.221
	0.025	2.659	2.551	2.461	2.383	2.156	1.981	1.866	1.724	1.581	1.507	1.432	1.268
	0.010	3.254	3.094	2.959	2.844	2.516	2.269	2.110	1.917	1.726	1.630	1.532	1.324
	0.005	3.757	3.547	3.372	3.224	2.805	2.496	2.299	2.063	1.834	1.720	1.605	1.364
∞	0.100	1.846	1.797	1.755	1.718	1.607	1.517	1.456	1.376	1.291	1.244	1.192	1.008
	0.050	2.206	2.130	2.065	2.009	1.843	1.711	1.622	1.508	1.389	1.324	1.254	1.010
	0.025	2.595	2.487	2.395	2.316	2.085	1.905	1.786	1.637	1.482	1.399	1.310	1.012
	0.010	3.165	3.004	2.868	2.752	2.421	2.169	2.006	1.804	1.600	1.494	1.380	1.014
	0.005	3.646	3.435	3.260	3.111	2.690	2.376	2.176	1.931	1.688	1.563	1.431	1.016

Las tablas C.1 a C.4 se extractaron con permiso de Rencher (1995).
Las tablas C.5 a C.8 fueron generadas mediante el paquete SAS©.

Bibliografía

[1] Alfenderfer, Mark S., and Blashfield, Roger., *Cluster Analysis*, Series: Quantitative Applications in the Social Sciences, Sage Publications, Inc., Beverly Hills, (1984).

[2] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons., New York, 1984.

[3] Anderson, T. W., *Asymptotic Theory for Principal Component Analysis*, The Annals of Mathematical Statistics, Vol. 34, 122-148, 1963.

[4] Andrews, D. F., *Plots of high-dimensional data*, Biometrics, Vol. 28, 125-136, 1972.

[5] Andrews, D. F., Gnanadesikan, R., and Warner, J. L., *Methos for Assessing Multivariate Normality. In P. R. Kishnaiah (Ed.) Multivariate Analysis*, Vol. III, 95-116, Academic Press, New York, 1973.

[6] Arnold, Steven F., *The Theory of Linear Models and Multivariate Analysis*, John Wiley and Sons, 1981.

[7] Bartlett, M. S., *Properties of Sufficiency and Statistical Tests*, Proceedings of the Royal Society of London, Vol. 160, 268-282, 1937.

[8] Bartlett, M. S., *A note on test of significance in multivariate analysis*, Proceedings of the Cambridge Philosophical Society, Vol. 35, 180-185, 1939.

[9] Bartlett, M. S., *Anote on multiplying factors for various chi-squared approximations*, Journal of the Royal Statistical Society, Series B, Vol. 16, 296-298, 1954.

[10] Benzecri, J. P., *Cours de Linguistique Mathématique*, Publication multigraphiée, (Faculé des Sciences de Rennes). 1964.

[11] Benzecri, J. P., *L'Analyse des Données*, Tomo 1: La Taxinomie, Tomo 2: L'Analyse des Correspondances, Dunod, Paris, 1973.

[12] Benzecri, J. P., *Histoire et Préhistoire de l'Analyse des Données L'Analyse des Données.*, Les Cahiers de Analyse des Donées Dunod, Paris, 1976.

[13] Biscay, R., Valdes, P. and Pascual, R., *Modified Fisher's linear discriminant function with reduction of dimensionality*, Journal of Statistical Computation and simulation, Vol. 36, 1-8, 1990.

[14] Borg, Ingwer., and Groenen, Patrick, *Modern Multidimensional Scaling*, Springer, New York. 1997.

[15] Box, G. E. P., *A general distribution theory for a class of likelihood criteria*, Biometrika, Vol. 36, 317-346, 1949.

[16] Box, G. E. P. and Cox, D. R., *An analysis of transformations*, Journal of the Royal Statistical Society, Series B, Vol. 26,211-252, 1964.

[17] Buck, S. F. A., *A Method of estimation of missing values in multivariate data suitable for use with an electronic computer*, Journal of the Royal Statistics Society, Series B, Vol. 22, 302-307, 1960.

[18] Catell, R. B., *The screen test for the number of factors*, Multivariate Behavioral Research, Vol. 1, 140-161, 1966.

[19] Chatfield, C. and Collins, A. J., *Introduction to Multivariate Analysis* Chapman and Hall, New York. 1986.

[20] Cherkassky, Vladimir., Friedman, Jerome H. and Wechsler, Harry. *From Statistics to Neural Networks*, Theory and Pattern Recognition Applications, Springer, Berlin, 1993.

[21] Chernoff, Herman., *Using faces to represent points in k-dimensional space graphically*, Journal of the American Statistics Association, Vol. 68, 361-368, 1973.

[22] Clifford, H. and Stephenson, W., *Introduction to Numerical Taxonomic*, Academic Press, New York. 1975.

[23] Crisci, Jorge Victor y López, María Fernanda., *Introducción a la Teoría y Práctica de la Taxonomía Numérica*, Secretaría General de la OEA, Washington, D. C., 1983.

[24] Cox, Trevor F. and Cox, Michael A. A., *Multidimensional Scaling*, Chapman and Hall, London. 1994.

[25] Crowder, M. J. and Hand, D. J., *Analysis of Repeated Measures*, Chapman and Hall, New York. 1990.

[26] D'agostino, R. B. and Pearson, E. S., *Test for deperture from Normality. Empirical Results for the Distributions of* b_2 *and* $\sqrt{b_1}$, Biometrika, Vol. 60, 613-622; correction 61, 647, 1973.

[27] Díaz, Luis Guillermo y López, Luis Alberto, *Tamaño de muestra en diseño experimental*, Memorias III Simposio de Estadística Muestreo, Universidad Nacional de Colombia, Santafé de Bogotá, D. C., 132-154, 1992.

[28] Dillon, William R. and Goldstein, Matthew., *Multivariate Analysis, Methods and Applications* John Wiley and Sons, New York, 1984.

[29] Diday, E., *Optimisation en classification automatique et reconnnaisance des formes*, Revue Française de Recherche Opérationnelle, Vol. 3, 61-96, 1972.

[30] Diday, E., *Classification automatique séquentielle pour grands tableaux, Revue Française de Recherche Opérationnelle*, Vol. 9, 1-29, 1974.

[31] Efron, B. and Tibshirani, R., *An Introduction to the Bootstrap*, Chapman and Hall, London, 1993.

[32] Escofier, Brigitte. et Pages, Jérome., *Analyses factorielles simples et multiples*, Dunod, Paris, 1990.

[33] Everitt, Brian S., *Cluster Analysis*, Heineman Educational Books, London, 1980.

[34] Everitt, Brian S. and Dunn, Graham., *Applied Multivariate Data Analysis*, Edward Arnold Books, New York, 1991.

[35] Forgy, E. W., *Cluster analysis of multivariate data: efficiency versus interpretability of classifications*, Biometrics, 768, Vol. 21, 1965.

[36] Freund, Rudolf J., Litell, Ramon C. and Spector, Philip C., *SAS system for linear models*, SAS Institute Inc., Cary, NC., 1986.

[37] Giri, Narayan C., *Multivariate Statistical Inference*, Academic Press, New York, 1977.

[38] Gnanadesikan, R. *Methods for Statistical Analysis of Multivariate Observations*, John Wiley and Sons, New York., 1997.

[39] Gnanadesikan, R. and Kattenring, J. R., *Robust stimates, residulas and outlier detection with multiresponse data*, Biometrics, 81-124, 1972.

[40] Gordon, A. D., *A Review of hierarchical Classification*, Series A Journal of the Royal Statistical Society, 150-119, 1937.

[41] Graybill, Franklyn A., *Theory and Application of the Linear Model*, Duxbury Press, Massachusetts, 1976.

[42] Gorsuch, Richard L., *Factor Analysis*, Lawrence Erlbaum Associates, Publishers, London, 1983.

[43] Harville, David A., *Introduction to Matrix Algebra From a Statistician's Perspective*, Springer, New York, 1997.

[44] Hogg, Robert V. and Craig, Allent T., *Introduction to Mathematical Statistics*, Macmillan Publishing Co. Inc., New York, 1978.

[45] Hotelling, H., *The generalization of Student's ratio*, Annals of Mathematical Statistics, Vol. 2, 360-378, 1931.

[46] Jobson, J. D., *Applied Multivariate Data Analysis*, Volume I: Regression and Experimental Design, Springer, New York, 1992.

[47] Jobson, J. D., *Applied Multivariate Data Analysis*, Volume II: Categorical and Multivariate Methods, Springer, New York, 1992.

[48] Johnson, Richard and Wicher, Dean W., *Applied Multivariate Statistical Analysis*, Prentice Hall, Inc., New Jersey, 1998.

[49] Jöreskog, K. G., *Some contributions to maximum likelihood factor analysis*, Psychometrika, Vol. 32,443-482, 1967.

[50] Kaiser, K. G., *The varimax criteriom for analytic rotation in factor analysis*, Psychometrika, Vol. 23, 187-200,1958.

[51] Kaiser, K. G., *Some contributions to maximum likelihood factor analysis*, Psychometrika, Vol. 32, 443-482, 1967.

[52] Kruskal, J. B., and Wish, M., *Multidimensional Scaling*, Sage Publications, Beverly Hills, CA., 1978.

[53] Krzanowski, W. J. and Marriot, F. H. C., *Multivariate Analysis. Part 1 Distributions, Ordination and Inference*, Edward Arnold, London, 1994.

[54] Krzanowski, W. J. and Marriot, F. H. C., *Multivariate Analysis. Part 2 Classification, covariance structures and repeated measurements*, Edward Arnold, London, 1995.

[55] Lawley, D. N., *A generalization of Fisher's z test*, Biometrika, Vol. 30, 180-187, 1938.

[56] Lawley, D. N., *Some new results in maximum likelihood factor analysis*, Proceedings of the Royal Society of Education, Vol. 67, 256-264, 1967.

[57] Lebart, Ludovic, Morineau, Alan, Fénelon, Jean-Pierre, *Tratamiento Estadístico de Datos*, Marcombo-Boixareu Editores, Barcelona. 1985.

[58] Lebart, Ludovic, Morineau, Alan, Piron, Marie, *Statistique Exploratoire Multidimensionnelle*, Dunod, Paris, 1995.

[59] Lebart, Ludovic, Morineau, Alan, and Warwick, Kenneth M., *Multivariate Descriptive Statistical Analysis*, John Wiley and Sons, New York, 1984.

[60] Lee, Kerry L., *Multivariate Test for Cluster*, Journal of the American Statistical Association, Vol. 74, 708-714 , 1979.

[61] Little, R. J. A. and Rubin, D. B., *Statistical Analysis with Missing Data* John Wiley and Sons, New York, 1987.

[62] Lou, Sheldon., Jiang, Jiong., and Keng, Kenneth, *Clustering Objects Generated by Linear Regression Models*, Journal of the American Statistical Association, Vol. 88, 1356-1362, 1993.

[63] Maclachlan, Geoffrey J., *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley and Sons, New York, 1992.

[64] Manly, Bryan F. J., *Multivariate Statistical Methods, A primer*, Chapman and Hall, New York, 2000.

[65] Mardia, K. V., *Measures of multivariate skewness and kurtosis with applications*, Biometrika, Vol. 57, 519-530, 1970.

[66] Mardia, K. V., *Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies*, Sankhyā B, Vol. 36, 115-128, 1974.

[67] Mardia, K. V., Kent, J. T., and Bibby, J. M., *Multivariate Analysis*, Academic Press, New York, 1979.

[68] Mason, R. L., Tracy, N. D. and Young, J. C., *Decomposition of T^2 for multivariate control chart interpretation*, Journal of Quality Technology, Vol. 27 (2), 157-158, 1995.

[69] Mijares, Tito A., *The normal approximation to the Bartlett- Nanda-Pillai trace test in multivariate analysis*, Biometrika, Vol. 77, 230-233, 1990.

[70] Milligan, G. W. and Cooper, M. C., *An examination of procedures for determining the number of cluster*, Psychometrika, Vol. 50 , 159-179, 1985.

[71] Mood, Alexander M., Graybill, Franklyn A. and Boes, Duane C., *Introduction to the Theory of Statistics*, Mc Graw Hill Book Company, 1982.

[72] Morrison, Donald F., *Multivariate Statistical Methods*, Mc Graw Hill Book Company, New York, 1990.

[73] Muirhead, Robb J. *Aspects of Multivariate Statistical Theory*, John Wiley and Sons, New York, 1982.

[74] Nagarsenker, B. N. and Pillai, K. C. S., *Distribution of the likelihood ratio for testing $\Sigma = \Sigma_0$, $\boldsymbol{\mu} = \boldsymbol{\mu}_0$*, Journal of multivariate analysis,114-122, Vol. 4, 1974.

[75] Nanda, D. N., *Distribution of the sum of roots of the determinantal equation under a certain condition*, Annals of Mathematical Statistics, Vol. 21, 432-439, 1950.

[76] Pardo, Campo Elias., *Análisis de la Aplicación del Método de Ward de Clasificación Jerárquica al Caso de Variables Cualitativas*, Universidad Nacional de Colombia. Tesis de Magister en Estadística, Santafé de Bogotá, D. C., 1992.

[77] Peck, Roger., Fisher, LLoyd., and Van, John., *Approximate confidence intervals for the number of cluster*, Journal of the American Statistical Association, Vol. 84, 184-191, 1989.

[78] Peña S., Daniel, *Estadística modelos y métodos. Fundamentos*, Alianza Universitaria Textos, Madrid, 1998.

[79] Pillai, K. C. S., *Some new test criteria in multivariate analysis*, Annals of Mathematical Statistics, Vol. 26, 117-121, 1955.

[80] Pla, Laura E., *Análisis Multivariado: Método de Componentes Principales*, Secretaría General de la OEA, Washington, D. C., 1986.

[81] Rencher, Alvin C., *Methods of Multivariate Analysis*, John Wiley and Sons, New York, 1995.

[82] Rencher, Alvin C., *Multivariate Statistical Inference and Applications*, John Wiley and Sons, New York, 1998.

[83] Rohatgi, Vijay K., *Statistical Inference*, John Wiley and Sons, New York, 1984.

[84] Roussas, George G., *A First Course in Mathematical Statistics*, Addison-Wesley Publishing Company, Massachusetts, 1973.

[85] Roy, S. N., *On a heuristic method of test construction and its use in multivariate analysis*, Annals of Mathematical Statistics, Vol. 24, 220-238, 1953.

[86] Roy, S. N., *Some Aspects of multivariate Analysis*, John Wiley and Sons, New York, 1957.

[87] Ruiz-Velazco, S., *Asympototic efficiency of logistic regression relative to linear discriminant analysis*, Biometrika, Vol. 78, 235-243, 1991.

[88] Saporta, Gilbert., *Probabilités Analyse des Données et Statistique*, Technip, Paris, 1990.

[89] SAS Institute Inc., *SAS/STAT User's Guide*, SAS Institute Inc., Cary N. C., 2001.

[90] Schott, James R., *A test for a specific principal component of a correlation matrix*, Journal of the American Statistical Association, Vol. 86, 747-751, 1991.

[91] Searle, S. R., *Matrix Algebra Useful for Statistics*, John Wiley and Sons, New York. 1990.

[92] Seber, G.A.F., *Multivariate observations*, Jonhn Wiley and Sons, New York, 1984.

[93] Sharma, Subhash, *Applied Multivariate Techniques*, Jonhn Wiley and Sons, New York, 1996.

[94] Silverman, B. W. *Density Estimation for Statistics and Data Analysis,* Chapman and Hall, New York, 1986.

[95] Sokal, R. and Michener, C. D., *A statistical method for evaluating systematic relationship*, University of Kansas Scientific Bulletin, 1958.

[96] Takane, Y., Young, F. W., and Leeuw, J., *Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features*, Psychometrika, Vol. 42, 7-67, 1977.

[97] Thompson, Paul A., *Correspondence analysis in statistical package programs*, The American Statistician, Vol. 49, 310-316, 1995.

[98] Tiku, M. L., *Tables of the power of the F-test*, Journal of the American Statistical Association, Vol. 62, 525-539, 1967.

[99] Torres, Luz Gloria, Niño, Luis F. y Hernández, Germán., *Redes Neuronales*, X Coloquio Distrital de Matemáticas y Estadística, Santafé de Bogotá, 1993.

[100] Tukey, J. W., *On the Comparative Anatomy of Transformations*, Annals of Mathematical Statistics, Vol. 28, 602-632, 1957.

[101] Velilla, S., and Barrio, J. A., *A discriminant rule under transformation*, Technometrics, Vol. 36, 348-353, 1994.

[102] Ward, J., *Approximate confidence intervals for the number of cluster*, Journal of the American Statistical Association, Vol. 58, 236-224, 1963.

[103] Welch, B. L., *The significance of the difference between two means when the population variances are unequal*, Biometrika, Vol. 29, 350-360,1937.

[104] Welch, B. L., *The generalization of "Student" problem when several different population variances are involved*, Biometrika, Vol. 34, 28-35, 1947.

[105] Wilcox, Rand R. *Introduction to Robust Stimation and Hipothesis Testing*, Academic Press, New York, 1997.

[106] Yager, R. R., Ovchinnikov, S., Togn, R. M., and Nguyen, H. T., *Fuzzy Sets and Applications. Selected Papers by L. A. Zadeh*, John Wiley and Sons, New York, 1987.

[107] Yan, S. S., and Lee, Y., *Identification of a multivariate outlier*, Annual Meeting of the American Statistical Association, 1987.

[108] Yeo, In-Know and Johnson, Richard A. *A new famly of power transformations to improve normality or symemtry*, Biometrika, 2000.

[109] Zadeh, Lotfi A., *Fuzzy Sets*, Information and Control, 338-353, 1965.

Índice de materias