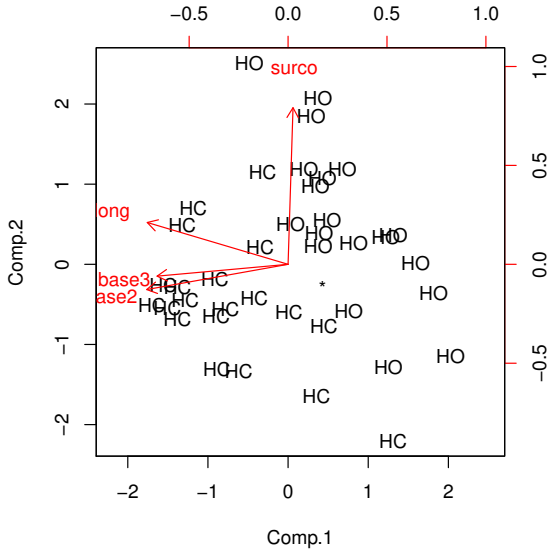
UNIVERSIDAD DE MURCIA

ANÁLISIS ESTADÍSTICO MULTIVARIANTE USANDO R



Jorge L. Navarro Camacho

2019

Contents

List of Figures

Índice de tablas

Prefacio

Este libro corresponde a los contenidos de una asignatura optativa de estadística multivariante para los últimos cursos de un grado en ciencias. Como principal novedad incluye los comandos del programa estadístico R para la resolución de algunos problemas y prácticas. Este programa es un programa estadístico de uso libre que se mejora con los procedimientos aportados por los propios usuarios. El programa R se puede descargar en

`http://www.r-project.org/`

y el programa R-studio en

`http://www.rstudio.com/ide/download/desktop`.

Es una gran verdad que cuando no está a nuestro alcance determinar lo que es verdadero, debemos aceptar aquello que sea más probable.
René Descartes.

1

Análisis de regresión múltiple

1.1 Introducción al modelo de regresión lineal.

1.2 Estimación del modelo de regresión.

1.3 Inferencia y predicción. Análisis de los residuos.

1.4 Extensiones de los modelos de regresión múltiple.

1.5 Problemas

1.

2

Análisis de componentes principales

En este capítulo mostramos cómo calcular las componentes principales asociadas a un conjunto de variables aleatorias tanto desde el punto de vista teórico como empíricamente. El Análisis de las Componentes Principales (PCA o ACP) permitirá resumir la información contenida en las variables, mostrar sus principales relaciones y analizar las características de los individuos de la población usando sus valores (puntuaciones) en las componentes principales.

2.1 Introducción.

El primer investigador que se percató de la necesidad de eliminar la información redundante en un conjunto de variables aleatorias fue Galton (Francis, Inglaterra, 1822-1911) quién criticó un intento de identificar criminales a partir de 12 medidas corporales alegando que varias de las medidas tomadas estarían altamente correlacionadas. Posteriormente, en 1901, un colaborador de Pearson (Karl, Inglaterra 1857-1936), McDonald hizo un estudio similar sobre 7 variables y 3000 criminales, publicando los resultados en una matriz de correlaciones, con la idea de encontrar algún índice que resumiera la información contenida en los datos. Pearson estaba convencido que los "índices" ideales coincidirían con los ejes del elipsoide de concentración. Pearson probó que el plano formado por estos ejes y que pasaba por la media era el que minimizaba la suma de las distancias al cuadrado con cada punto original. Finalmente, en 1933, fue Hotelling (Harold, USA 1895-1973) el que encontró un algoritmo para calcular dichos ejes lo que, en esencia, requería la obtención de los valores propios de una matriz simétrica y definida positiva. También debemos destacar las aportaciones de Rao (Calyampudi Radhakrishna, India, 1920). Con la llegada de los ordenadores la resolución de este problema para muchas variables (matrices de gran dimensión) ha supuesto uno de los principales problemas de la Programación Matemática.

Para presentar el problema usaremos varios ejemplos.

Ejemplo 2.1. *En el primer ejemplo consideramos un conjunto de datos denominado* `LifeCycleSavings` *incluido en el programa R. Los ficheros de datos incluidos en R se pueden ver con* `data()`*. Para ver los datos de ese fichero haremos:*

`LifeCycleSavings`

y para guardarlos en `d`

`d<-LifeCycleSavings`

El fichero contiene 5 variables medidas en 50 países diferentes. Los primeros datos se pueden ver en la Tabla 2.1.

Tabla 2.1: Primeros datos del fichero `LifeCycleSavings`.

	sr	pop15	pop75	dpi	ddpi
Australia	11.43	29.35	2.87	2329.68	2.87
Austria	12.07	23.32	4.41	1507.99	3.93
Belgium	13.17	23.80	4.43	2108.47	3.82
Bolivia	5.75	41.89	1.67	189.13	0.22
Brazil	12.88	42.19	0.83	728.47	4.56

La información proporcionada en R sobre datos se puede ver con:

`help(LifeCycleSavings)`

donde se indica que:

sr: incremento de los ahorros personales 1960-1970.

pop15: % población menor de 15 años.

pop75: % población mayor de 75.

dpi: ingresos per-capita.

ddpi: crecimiento del dpi 1960-1970.

Para estudiar las relaciones entre las variables podemos hacer

`plot(d)`

La gráfica se puede ver en la Figura 2.1.

Podemos calcular las matrices de covarianza y correlación con `cov(d)` *y* `cor(d)`*, respectivamente. Las correlaciones se pueden ver la Tabla 2.2. Se aprecia que existen variables con correlaciones (lineales) positivas, negativas y casi nulas. Estas relaciones se verán reflejadas en las componentes principales que calcularemos posteriormente.*

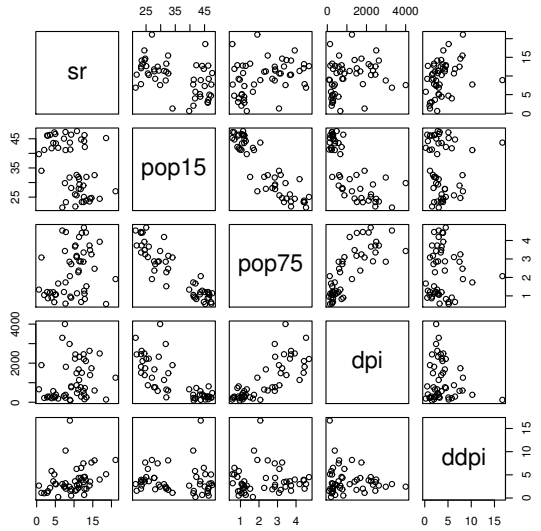


Figure 2.1: Gráficos bidimensionales para todas las variables del fichero de R `LifeCycleSavings`.

Tabla 2.2: Correlaciones entre las 5 variables del fichero `LifeCycleSavings`.

	sr	pop15	pop75	dpi	ddpi
sr	1.0000000	-0.45553809	0.31652112	0.2203589	0.30478716
pop15	-0.4555381	1.00000000	-0.90847871	-0.7561881	-0.04782569
pop75	0.3165211	-0.90847871	1.00000000	0.7869995	0.02532138
dpi	0.2203589	-0.75618810	0.78699951	1.0000000	-0.12948552
ddpi	0.3047872	-0.04782569	0.02532138	-0.1294855	1.00000000

Ejemplo 2.2. *En este ejemplo analizaremos el objeto* `d` *del fichero* `nota.rda` *(Aula virtual*[1]*) que contiene las notas (sobre 100) de 88 alumnos de matemáticas en una universidad americana (Fuente: Rencher, 1995). Los primeros datos se pueden ver en la Tabla 2.3.*

Para estudiar las relaciones entre las variables podemos hacer

```
plot(d)
```

[1]Para leer este tipo de archivos teclear `load('f:/.../name.rda')` indicando la ruta completa en donde se encuentra el archivo y reemplazando name por el nombre del archivo. Alternativamente, se puede cambiar el directorio de trabajo en Session>Set working directory y teclear `load('name.rda')`.

Tabla 2.3: Primeros datos del fichero `nota.rda`.

	Mecanica	Vectores	Algebra	Analisis	Estadist
1	77	82	67	67	81
2	63	78	80	70	81
3	75	73	71	66	81
4	55	72	63	70	68
5	63	63	65	70	63
6	53	61	72	64	73
7	51	67	65	65	68
8	59	70	68	62	56
9	62	60	58	62	70
10	64	72	60	62	45

La gráfica se puede ver en la Figura 2.2.

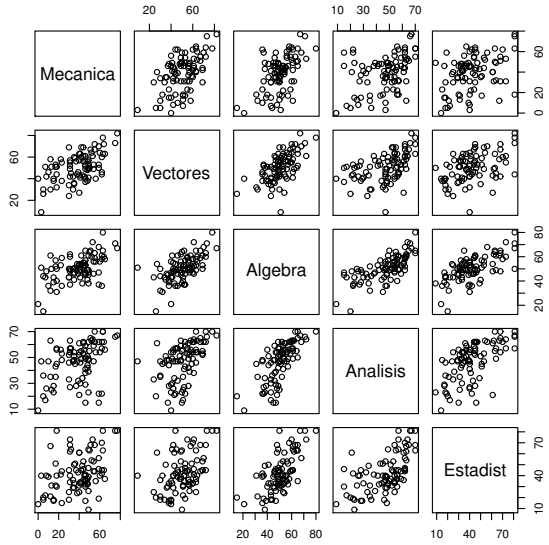


Figure 2.2: Gráficos bidimensionales para las 5 variables del fichero `nota.rda`.

Podemos calcular las matrices de covarianza y correlación con `cov(d)` *y* `cor(d)`*, respectivamente. Las correlaciones se pueden ver la Tabla 2.4. Se aprecia que todas las variables tienen correlaciones (lineales) positivas. En la matriz de covarianzas se aprecia que aunque las cinco variables se miden en las*

mismas unidades (notas sobre 100), las cuasivarianzas son bastante diferentes. Estas relaciones se verán reflejadas en las componentes principales.

Tabla 2.4: Correlaciones entre las 5 variables del fichero `nota.rda`.

	Mecanica	Vectores	Algebra	Analisis	Estadist
Mecanica	1.0000000	0.5534052	0.5467511	0.4093920	0.3890993
Vectores	0.5534052	1.0000000	0.6096447	0.4850813	0.4364487
Algebra	0.5467511	0.6096447	1.0000000	0.7108059	0.6647357
Analisis	0.4093920	0.4850813	0.7108059	1.0000000	0.6071743
Estadist	0.3890993	0.4364487	0.6647357	0.6071743	1.0000000

Ejemplo 2.3. *En el tercer ejemplo analizamos los datos del fichero* `heptathlon` *del paquete de R* `MVA`[2] *correspondientes a los resultados en la prueba femenina de heptatlon en las olimpiadas de Seul 1988 (ver Tabla 2.5). Las variables corresponden a las pruebas: hurdles (110 m. vallas, en segundos), highjump (salto de altura, en metros), shot (lanzamiento de peso, en metros), run200m (carrera de 200m., en segundos), longjump (salto de longitud, en metros), javelin (lanzamiento de jabalina, en metros), run800m (carrera de 800m., en segundos) y score (puntuación final).*

[2]Para leer este conjunto de datos hay que instalar el paquete `MVA` pinchando en el menú: `Paquetes > Instalar Paquetes` seleccionando `MVA` y tecleando en R: `library('MVA')`.

Tabla 2.5: Resultados de Heptatlon en la Olimpiada de Seul 1988.

	hurd.	HJ	shot	200m	LJ	jav.	800m	score
1	12.69	1.86	15.80	22.56	7.27	45.66	128.51	7291
2	12.85	1.80	16.23	23.65	6.71	42.56	126.12	6897
3	13.20	1.83	14.20	23.10	6.68	44.54	124.20	6858
4	13.61	1.80	15.23	23.92	6.25	42.78	132.24	6540
5	13.51	1.74	14.76	23.93	6.32	47.46	127.90	6540
6	13.75	1.83	13.50	24.65	6.33	42.82	125.79	6411
7	13.38	1.80	12.88	23.59	6.37	40.28	132.54	6351
8	13.55	1.80	14.13	24.48	6.47	38.00	133.65	6297
9	13.63	1.83	14.28	24.86	6.11	42.20	136.05	6252
10	13.25	1.77	12.62	23.59	6.28	39.06	134.74	6252
11	13.75	1.86	13.01	25.03	6.34	37.86	131.49	6205
12	13.24	1.80	12.88	23.59	6.37	40.28	132.54	6171
13	13.85	1.86	11.58	24.87	6.05	47.50	134.93	6137
14	13.71	1.83	13.16	24.78	6.12	44.58	142.82	6109
15	13.79	1.80	12.32	24.61	6.08	45.44	137.06	6101
16	13.93	1.86	14.21	25.00	6.40	38.60	146.67	6087
17	13.47	1.80	12.75	25.47	6.34	35.76	138.48	5975
18	14.07	1.83	12.69	24.83	6.13	44.34	146.43	5972
19	14.39	1.71	12.68	24.92	6.10	37.76	138.02	5746
20	14.04	1.77	11.81	25.61	5.99	35.68	133.90	5734
21	14.31	1.77	11.66	25.69	5.75	39.48	133.35	5686
22	14.23	1.71	12.95	25.50	5.50	39.64	144.02	5508
23	14.85	1.68	10.00	25.23	5.47	39.14	137.30	5290
24	14.53	1.71	10.83	26.61	5.50	39.26	139.17	5289
25	16.42	1.50	11.78	26.16	4.88	46.38	163.43	4566

Para estudiar las relaciones entre las variables podemos hacer

```
plot(heptathlon)
```

La gráfica se puede ver en la Figura 2.3.

Podemos calcular las matrices de covarianza y correlación con `cov(d)` *y* `cor(d)`*, respectivamente. Las correlaciones se pueden ver la Tabla 2.6. Se aprecia que algunas variables tienen correlaciones (lineales) positivas y otras negativas. Note que en algunas variables (las carreras) es mejor tener valores bajos (poco tiempo) mientras que en otras es mejor tener valores altos (lanzamientos y saltos). Estas relaciones se verán reflejadas en las componentes principales.*

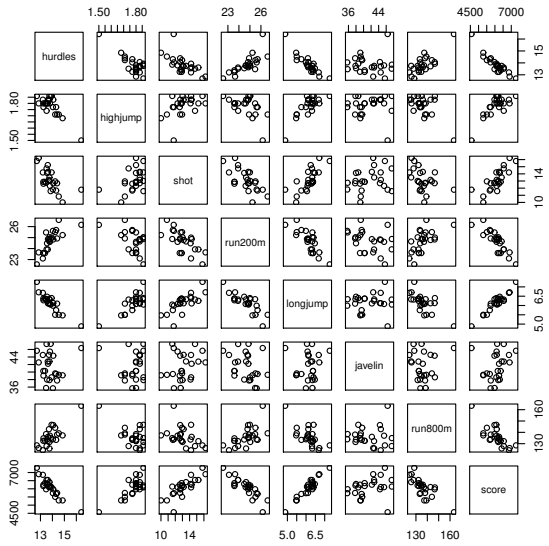


Figure 2.3: Gráficos bidimensionales para las variables del fichero `heptathlon`.

Pueden pensarse ejemplos similares como ¿cómo construir un índices que midan los cambios de precios en un país?, ¿qué relación hay entre las variables de un recién nacido, una persona o una animal? o ¿cómo resumir los indicadores económicos de diversos países?

2.2 Cálculo teórico de las componentes principales.

Desde el punto de vista teórico la idea es resumir la información de un vector aleatorio (v.a.) $k-$dimensional $X = (X_1, \ldots, X_k)'$ (A' denota la traspuesta de A, es decir, X es un vector columna) en unas "pocas" variables que proporcionen la información más relevante. Se puede dar una aproximación geométrica mediante el concepto de elipsoide de concentración.

Definición 2.1. *Si X es un vector aleatorio de dimensión k, media μ y matriz de covarianzas $V = (\sigma_{i,j})$ definida positiva, se define el* ***elipsoide de concentración*** *de X como*

$$E_k = \{x \in \mathbb{R}^k : (x-\mu)'V^{-1}(x-\mu) \leq k+2\}.$$

Puede probarse que existe un v.a. uniforme sobre E_k con media μ y matriz de covarianzas V (ver Zoroa y Zoroa 2008, p. 206). En la definición del elipsoide

Tabla 2.6: Correlaciones entre las variables del fichero `heptathlon`.

	hurd.	HJ	shot	200m
hurd.	1.000000000	-0.811402536	-0.6513347	0.7737205
HJ	-0.811402536	1.000000000	0.4407861	-0.4876637
shot	-0.651334688	0.440786140	1.0000000	-0.6826704
200m	0.773720543	-0.487663685	-0.6826704	1.0000000
LJ	-0.912133617	0.782442273	0.7430730	-0.8172053
jav.	-0.007762549	0.002153016	0.2689888	-0.3330427
800m	0.779257110	-0.591162823	-0.4196196	0.6168101
score	-0.923198458	0.767358719	0.7996987	-0.8648825
	LJ	jav.	800m	score
hurd.	-0.91213362	-0.007762549	0.77925711	-0.9231985
HJ	0.78244227	0.002153016	-0.59116282	0.7673587
shot	0.74307300	0.268988837	-0.41961957	0.7996987
200m	-0.81720530	-0.333042722	0.61681006	-0.8648825
LJ	1.00000000	0.067108409	-0.69951116	0.9504368
jav.	0.06710841	1.000000000	0.02004909	0.2531466
800m	-0.69951116	0.020049088	1.00000000	-0.7727757
score	0.95043678	0.253146604	-0.77277571	1.0000000

interviene la **distancia de Mahalanobis** basada en la matriz V entre x y la media μ dada por

$$d_V(x,\mu) = \sqrt{(x-\mu)'V^{-1}(x-\mu)}.$$

Además, si X es normal, el elipsoide se puede definir a partir de las curvas de nivel de la función de densidad ($f(x) \geq cte$) ya que

$$f(x) = \frac{1}{\sqrt{|V|(2\pi)^k}} \exp\left(-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)\right).$$

La mayor parte de los individuos (puntos) estarán dentro de este elipsoide y, si queremos distinguirlos con una única variable, parece claro que lo mejor sería proyectarlos sobre el eje mayor. Para una Normal bivariante

$$N_2\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, V = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$$

se tiene

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = \frac{4}{3}x^2 - \frac{4}{3}xy + \frac{4}{3}y^2$$

por lo que el elipsoide de concentración sería

$$\frac{4}{3}x^2 - \frac{4}{3}xy + \frac{4}{3}y^2 \leq 4$$

(ver Figura 2.4). La función de densidad puede verse en la Figura 2.5 y las curvas de nivel en la Figura 2.6.

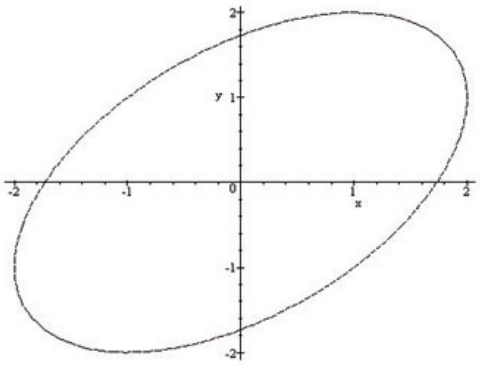


Figure 2.4: Elipsoide de concentración para una normal bidimensional con medias cero, varianzas 1 y correlación 1/2.

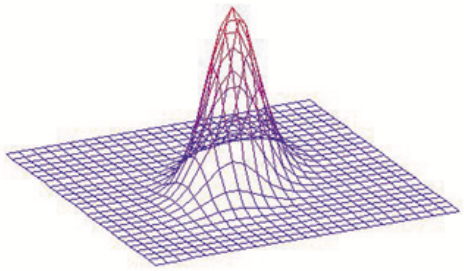


Figure 2.5: Función de densidad para una normal bidimensional con medias cero, varianzas 1 y correlación 1/2.

Supongamos que $X = (X_1, ..., X_k)'$ es un v.a. k dimensional con vector de medias μ y matriz de covarianzas V semidefinida positiva. Entonces la primera

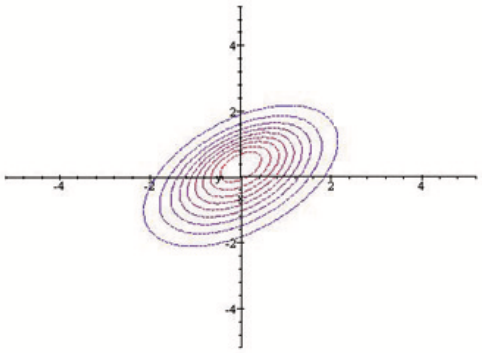


Figure 2.6: Curvas de nivel para una normal bidimensional con medias cero, varianzas 1 y correlación 1/2.

componente principal será la v.a. unidimensional $Y_1 = a_1X_1 + ... + a_kX_k$ con $a_1^2 + ... + a_k^2 = 1$ cuya varianza es máxima. Nótese que si no se normaliza la combinación lineal, la variable Y_1 puede tener varianza tan grande como queramos. Geométricamente, hacemos un cambio de variable (primer eje) para que la dispersión sea máxima y la normalización equivale a mantener la escala original (proyectar).

El problema puede expresarse de la forma siguiente:

$$\left.\begin{array}{l} \max Var(a'X) \\ s.a.: a'a = 1 \end{array}\right\}$$

donde $a = (a_1, ..., a_k)' \in \mathbb{R}^k$.

Una vez calculada una primera componente principal Y_1, la segunda componente principal Y_2 debe verificar $Cov(Y_1, Y_2) = 0$ (no debe contener información ya incluida en Y_1) y debe tener varianza máxima, es decir

$$\left.\begin{array}{ll} \max Var(a'X) & \\ s.a.: & a'a = 1 \\ & Cov(Y_1, a'X) = 0 \end{array}\right\}$$

Así, sucesivamente, por inducción, se definen las siguientes componentes principales como la (una) solución de

$$\left.\begin{array}{ll} \max Var(a'X) & \\ s.a.: & a'a = 1 \\ & Cov(Y_i, a'X) = 0, i = 1, ..., j-1 \end{array}\right\}$$

La solución general viene dada en el teorema siguiente que prueba la existencia de las (unas) componentes principales y muestra cómo calcularlas. Además, se demuestra que las componentes principales no son únicas (puede haber más soluciones).

Teorema 2.1. *Si X es un v.a. k dimensional con matriz de covarianzas V definida positiva, las (unas) componentes principales valen*

$$Y = (Y_1, ..., Y_k)' = T'X = \begin{pmatrix} t_{1,1} & ... & t_{k,1} \\ ... & ... & ... \\ t_{1,k} & ... & t_{k,k} \end{pmatrix} \begin{pmatrix} X_1 \\ ... \\ X_k \end{pmatrix}$$

donde T es una matriz ortogonal ($T'T = TT' = I$) tal que $T'VT = D = diag(\lambda_1, ..., \lambda_k)$ con $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k > 0$.

Proof. Como V es una matriz simétrica y definida positiva, existe una matriz $T = (t_{i,j})$ ortogonal ($T'T = TT' = I$) tal que $T'VT = D = diag(\lambda_1, ..., \lambda_k)$ con los valores propios verificando $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k > 0$ (ver Burgos 1994, pág. 630). De esta forma, si

$$Y = \begin{pmatrix} Y_1 \\ ... \\ Y_k \end{pmatrix} = T'X = \begin{pmatrix} t_{1,1} & ... & t_{k,1} \\ ... & ... & ... \\ t_{1,k} & ... & t_{k,k} \end{pmatrix} \begin{pmatrix} X_1 \\ ... \\ X_k \end{pmatrix}$$

entonces $Y_1, ..., Y_k$ verifican

$$Cov(Y) = Cov(T'X) = E(T'(X - \mu)(X - \mu)'T) = T'VT = D$$

lo que implica que $Cov(Y_i, Y_j) = 0$ para $i \neq j$ y $Var(Y_j) = \lambda_j$. Los vectores columnas de T (filas de T') $t_j = (t_{1,j}, ...t_{k,j})'$, serán una base ortonormal de vectores propios de V verificándose $Y_j = t_j'X$ y $Vt_j = \lambda_j t_j$ para $j = 1, \ldots, k$.

Para comprobar que Y_1 es una primera componente principal, supongamos que $a'X$ es una combinación lineal con $a'a = 1$. Entonces, como los vectores propios son una base, existirán $c_1, ..., c_k$ números reales tales que

$$a = c_1 t_1 + ... + c_k t_k, \text{ con } c = (c_1, ..., c_k)'$$

con lo que

$$\begin{aligned} Var(a'X) &= E(a'(X-\mu)(X-\mu)'a) \\ &= a'Va \\ &= \left(\sum_{i=1}^{k} c_i t_i'\right) V \left(\sum_{i=1}^{k} c_i t_i\right) \\ &= \left(\sum_{i=1}^{k} c_i t_i'\right) \left(\sum_{i=1}^{k} c_i V t_i\right) \\ &= \left(\sum_{i=1}^{k} c_i t_i'\right) \left(\sum_{j=1}^{k} c_j \lambda_j t_j\right) \\ &= \sum_{i,j} c_i c_j \lambda_j t_i' t_j \\ &= \sum_{i=1}^{k} c_i^2 \lambda_i \end{aligned}$$

y, como

$$a'a = \left(\sum_{i=1}^{k} c_i t_i'\right) \left(\sum_{j=1}^{k} c_j t_j\right) = \sum_{i,j} c_i c_j t_i' t_j = \sum_{i=1}^{k} c_i^2 = c'c = 1$$

la varianza será máxima si $c_1^2 = 1, c_2 = 0, ..., c_k = 0$ ya que

$$Var(\pm t_1' X) = \lambda_1 = \sum_{i=1}^{k} c_i^2 \lambda_1 \geq \sum_{i=1}^{k} c_i^2 \lambda_i = Var(a'X),$$

para todo a tal que $a'a = 1$, es decir, $Y_1 = \pm t_1' X$ es una primera componente principal (puede haber otras soluciones si $\lambda_1 = \lambda_2$).

Por inducción, supongamos que $Y_1 = t_1' X, ..., Y_{j-1} = t_{j-1}' X$ son las primeras $(j-1)$ componentes principales y veamos que $Y_j = t_j' X$ es la (una) solución de

$$\left.\begin{array}{ll} \max Var(a'X) & \\ s.a.: & a'a = 1 \\ & Cov(a'X, Y_i) = 0,\ i = 1, ..., j-1 \end{array}\right\}$$

Como se debe verificar

$$\begin{aligned} Cov(a'X, Y_i) = Cov(a'X, t_i'X) &= E(a'(X-\mu)(X-\mu)t_i) = a'Vt_i \\ &= \lambda_i a' t_i = \lambda_i \left(\sum_s c_s t_s'\right) t_i = \lambda_i c_i = 0 \end{aligned}$$

para $i = 1, ..., j-1$, y $\lambda_i > 0$, se tiene $c_1 = ... = c_{j-1} = 0$. Entonces, la varianza será máxima si $c_j = 1$ y $c_i = 0$ para $i > j$, ya que

$$Var(\pm t_j X) = \lambda_j = c'c\lambda_j = \sum_{i=j}^{k} c_i^2 \lambda_j \geq \sum_{i=j}^{k} c_i^2 \lambda_i = Var(a'X),$$

para todo a tal que $a'a = 1$ y $Cov(a'X, Y_i) = 0$, $i = 1, ..., j-1$, es decir, $Y_j = \pm t_j' X$, es una componente principal j-ésima (no necesariamente la única). □

Como consecuencia inmediata de la demostración anterior se tiene el corolario siguiente.

Corolario 2.1. *Si $\lambda_1 > \lambda_2 > ... > \lambda_k$, entonces las componentes principales son únicas salvo signo.*

Observación 2.1. *Nótese que la componente principal j-ésima se obtiene multiplicando la fila j-ésima de T' (la columna j-ésima de T) por X, es decir, $Y_j = t_j' X$ donde $t_j' = (t_{1,j}, ..., t_{k,j})$ es un vector propio unitario correspondiente al j-ésimo valor propio (vectores columna de T). Además, $Var(Y_j) = \lambda_j$ y*

$$traza(V) = \sum_{j=1}^{k} \sigma_{j,j} = \sum_{j=1}^{k} Var(X_j) = \sum_{j=1}^{k} Var(Y_j) = \sum_{j=1}^{k} \lambda_j$$

(las matrices semejantes tienen las trazas iguales), es decir, la variabilidad (información) de las variables originales es igual a la suma de las variabilidades de las componentes principales. La ***cantidad de información*** *(%) contenida en cada componente será $I_j = 100\lambda_j / (\sum_{i=1}^{k} \lambda_i)\%$. Por esto, la traza se usa como una medida unidimensional de la dispersión de una variable k-dimensional. La otra medida es el determinante de V para el que también se verifica:*

$$|V| = \lambda_1 ... \lambda_k = |Cov(Y)|$$

(es decir, la variabilidad se mide calculando el área encerrada en el paralelogramo de lados iguales a los valores propios).

Observación 2.2. *Otros autores llaman componentes principales a $Y = T'(X - \mu)$ con lo que, además, se consigue que sean centradas ($E(Y_j) = 0$). También se pueden definir las componentes principales estandarizadas $Z_j = t_j'(X - \mu)\lambda_j^{-1/2}$ ($Z = D^{-1/2}T'(X - \mu)$) que además de ser centradas tendrán varianza 1.*

Observación 2.3. *Cuando hay valores propios iguales a cero (V es semidefinida positiva) no suelen considerarse sus correspondientes componentes principales (degeneradas) y se puede conservar toda la información en las componentes principales de valores propios distintos de cero. En este caso hay variables que pueden obtenerse como combinación lineal de las restantes (aunque no siempre pueden eliminarse del análisis).*

Observación 2.4. *Como comentamos al inicio, geométricamente, las componentes principales se corresponden con los ejes principales del elipsoide de concentración. Como $Y = T'X$, podemos interpretar las componentes en función de los pesos que tengan en ellas las variables originales. Si ponemos X en función de Y como $X = TY$, entonces las variables originales se pueden interpretar en función de las componentes principales e incluso, podemos representar aproximadamente, las variables originales usando las dos (tres) primeras componentes.*

Si la población X es normal, entonces las componentes principales son normales e independientes entre sí, ya que en éstas poblaciones equivalen independencia e incorrelación (independencia lineal) y Z será una normal estándar multivariante ($N_k(0, I)$).

Proposición 2.1. *Si Y son las componentes principales obtenidas a partir de X, entonces X es normal multivariante si, y solo si $Y_1, \ldots, Y_k$ son independientes y normales univariantes para todo $j = 1, ..., k$.*

La demostración es inmediata. Esta propiedad puede ser utilizada para estudiar la normalidad multivariante a partir de un test de normalidad univariante sobre las componentes principales. Incluso si la normal multivariante no es de rango completo (V no es definida positiva), puede utilizarse con las m primeras componentes con valores propios distintos de cero (las otras serán degeneradas) coincidiendo m con el rango de V.

Ejemplo 2.4. *Para la v.a. normal de media $\mu = (0, 0)'$ y matriz de covarianzas*

$$V = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

(ver Figura 2.5), sus componentes principales se calcularán diagonalizando V mediante

$$|V - \lambda I| = \begin{vmatrix} 1-\lambda & 0.5 \\ 0.5 & 1-\lambda \end{vmatrix} = 1 - 2\lambda + \lambda^2 - 1/4 = 0$$

que tiene soluciones

$$\lambda = \frac{2 \pm \sqrt{4 - 4(1 - 1/4)}}{2} = 1 \pm 0.5,$$

y la primera componente se obtendrá resolviendo

$$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = (1 + 0.5) \begin{pmatrix} x \\ y \end{pmatrix},$$
$$\begin{pmatrix} -0.5x + 0.5y \\ 0.5x - 0.5y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

lo que da $x = y$, es decir, sus vectores propios son $v = \lambda(1,1)'$. Como usamos vectores normalizados (de norma 1), una primera componente valdrá

$$Y_1 = \left(1/\sqrt{2}, 1/\sqrt{2}\right) \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = (X_1 + X_2)/\sqrt{2}$$

y su varianza es $\lambda_1 = 1.5$. Análogamente, la segunda valdrá $Y_2 = (X_1 - X_2)/\sqrt{2}$ (ya que tiene que ser perpendicular a la primera) y tendrá varianza $\lambda_2 = 0.5$. Es decir, tenemos

$$Y_1 = (X_1 + X_2)/\sqrt{2}$$
$$Y_2 = (X_1 - X_2)/\sqrt{2}$$

por lo que

$$Y = T'X = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

La primera componente explicará un $I_1 = 100\lambda_1/(\lambda_1 + \lambda_2)\% = 75\%$ de la varianza total y la segunda un $I_2 = 100\lambda_2/(\lambda_1 + \lambda_2)\% = 25\%$. Como las varianzas iniciales son iguales, ambas tienen igual peso en las componentes con distinto signo en el caso de la segunda de ellas. Nótese que aunque las varianzas iniciales sean todas iguales (1) las componentes principales tienen varianzas (en general) distintas. Si X_1 fuese el peso de una persona y X_2 su altura (estandarizadas), la primera componente se podría interpretar como lo "grande" que es dicha persona, mientras que la segunda estaría relacionada con su "constitución" (Y_2 grande significaría gran peso y poca altura, es decir, complexión fuerte). Despejando, se tiene

$$X_1 = (Y_1 + Y_2)/\sqrt{2}$$
$$X_2 = (Y_1 - Y_2)/\sqrt{2},$$

lo que nos permite representar las variables X_1, X_2 en función de las componentes Y_1, Y_2 (ver Figura 2.7). Nótese que Y_1 aumenta si lo hacen X_1 y X_2 mientras que Y_2 aumenta si aumenta X_1 y disminuye X_2. Estas relaciones servirán para interpretar (dar significado) a las componentes principales.

Para realizar estos cálculos en R introduciremos los comandos siguientes. El primer lugar definimos e introducimos V con:

```
V<-matrix(nrow=2,ncol=2)
V[1,1]<-1
V[2,2]<-1
V[2,1]<-1/2
V[1,2]<-1/2
```

Tecleando `V` *podemos comprobar que hemos introducido los datos bien. Para calcular los valores y vectores propios haremos:*

`eigen(V)`

Si queremos guardar la matriz T de vectores propios haremos

`eigen(V)$vectors->T`

Recuerde que los vectores normalizados aparecen en las columnas de T. Para comprobar que T es una matriz ortogonal haremos

`t(T)%*%T`

donde `t(A)` *es la traspuesta de A y* `A%*%B` *es el producto de las matrices A y B en R. De esta forma, si queremos comprobar que T diagonaliza a V haremos*

`t(T)%*%V%*%T`

lo que nos dará (aproximadamente) la matriz diagonal con los valores 1.5 y 0.5 en la diagonal. Como $Y = T'X$, la primera componente principal será $Y_1 = 0.7071068X_1 + 0.7071068X_2$ y la segunda $Y_1 = -0.7071068X_1 + 0.7071068X_2$. Para calcular las informaciones contenidas en cada una (en tanto por 100) haremos:

`100*eigen(V)$values/sum(eigen(V)$values)`

obteniendo 75 y 25.

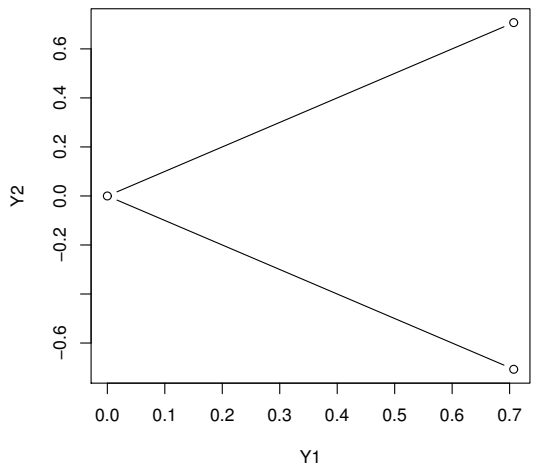


Figure 2.7: Variables del Ejemplo 2.4 en función de las componentes principales.

La desigualdad de Chebyshev para variables aleatorias da una cota inferior en función de la varianza para el porcentaje de valores a una distancia dada de la media. Se puede obtener de la desigualdad de Markov siguiente. Si Z es una variable aleatoria no negativa con media finita $E(Z)$ y $\varepsilon > 0$, entonces

$$\varepsilon \Pr(Z \geq \varepsilon) = \varepsilon \int_{[\varepsilon,\infty)} dF_Z(x) \leq \int_{[\varepsilon,\infty)} x dF_Z(x) \leq \int_{[0,\infty)} x dF_Z(x) = E(Z)$$

(donde $F_Z(x) = \Pr(Z \leq x)$ es su función de distribución), es decir

$$\Pr(Z \geq \varepsilon) \leq \frac{E(Z)}{\varepsilon}. \tag{2.1}$$

La desigualdad de Chebyshev se obtiene como sigue. Si X es una variable aleatoria con media finita $\mu = E(X)$ y varianza $\sigma^2 = Var(X) > 0$, entonces tomando $Z = (X-\mu)^2/\sigma^2 \geq 0$ en (2.1), tenemos

$$\Pr\left(\frac{(X-\mu)^2}{\sigma^2} \geq \varepsilon\right) \leq \frac{1}{\varepsilon} \tag{2.2}$$

para todo $\varepsilon > 0$. También se puede escribir como

$$\Pr((X-\mu)^2 < \varepsilon\sigma^2) \geq 1 - \frac{1}{\varepsilon}$$

o como

$$\Pr(|X-\mu| < r) \leq 1 - \frac{\sigma^2}{r^2}$$

para todo $r > 0$. De forma análoga, la desigualdad de Chebyshev multivariante se obtiene usando las componentes principales como sigue.

Teorema 2.2. *Sea $X = (X_1, \ldots, X_k)'$ un vector aleatorio con vector de medias finito $\mu = E(X)$ y matriz de covarianzas definida positiva V, entonces*

$$\Pr((X-\mu)'V^{-1}(X-\mu) \geq \varepsilon) \leq \frac{k}{\varepsilon} \tag{2.3}$$

para todo $\varepsilon > 0$.

Proof. Como V es definida positiva y simétrica existe T tal que $TT' = T'T = I$ y $T'VT = D$, donde $D = diag(\lambda_1, \ldots, \lambda_k)$ es la matriz diagonal con los valores propios ordenados $\lambda_1 \geq \cdots \geq \lambda_k > 0$. Entonces $V = TDT'$ y $V^{-1} = TD^{-1}T'$. Entonces la variable aleatoria no negativa

$$Z = (X-\mu)'V^{-1}(X-\mu)$$

se puede escribir como

$$Z = (X-\mu)'TD^{-1}T'(X-\mu) = [D^{-1/2}T'(X-\mu)]'[D^{-1/2}T'(X-\mu)] = \mathbf{Z}'\mathbf{Z},$$

donde

$$\mathbf{Z} = D^{-1/2}T'(X-\mu)$$

y $D^{-1/2} = diag(\lambda_1^{-1/2}, \ldots, \lambda_n^{-1/2})$. El vector aleatorio $\mathbf{Z} = (Z_1, \ldots, Z_k)$ (con las componentes principales estandarizadas) verifica

$$E(\mathbf{Z}) = E(D^{-1/2}T'(X-\mu)) = D^{-1/2}T'E(X-\mu) = \mathbf{0}_k$$

y

$$Cov(\mathbf{Z}) = Cov(D^{-1/2}T'(X-\mu)) = D^{-1/2}T'VTD^{-1/2} = D^{-1/2}DD^{-1/2} = I_k.$$

Por lo tanto

$$E(Z) = E(\mathbf{Z}'\mathbf{Z}) = E\left(\sum_{i=1}^{k} Z_i^2\right) = \sum_{i=1}^{k} E(Z_i^2) = \sum_{i=1}^{k} Var(Z_i) = k.$$

Entonces, usando la desigualdad de Markov (2.1), tenemos

$$\Pr(Z \geq \varepsilon) = \Pr((X-\mu)'V^{-1}(X-\mu) \geq \varepsilon) \leq \frac{E(Z)}{\varepsilon} = \frac{k}{\varepsilon}$$

para todo $\varepsilon > 0$, lo que finaliza la demostración. □

La desigualdad (2.3) también se puede escribir como

$$\Pr((X-\mu)'V^{-1}(X-\mu) < \varepsilon) \geq 1 - \frac{k}{\varepsilon} \tag{2.4}$$

para todo $\varepsilon > 0$. En particular, para el elipsoide de concentración

$$E_k = \{x \in \mathbb{R}^k : (x-\mu)'V^{-1}(x-\mu) \leq k+2\},$$

obtenemos

$$\Pr(X \in E_k) \geq 1 - \frac{k}{k+2} = \frac{2}{k+2}.$$

Para obtener regiones con más datos podemos tomar $\varepsilon = ck$, resultando

$$\Pr((X-\mu)'V^{-1}(X-\mu) < ck) \geq 1 - \frac{k}{\varepsilon} = 1 - \frac{1}{c} = \frac{c-1}{c}. \tag{2.5}$$

Si X es normal, entonces $Z_1, \ldots, Z_n$ son normales independientes y $Z = \sum_{i=1}^{k} Z_i^2$ sigue una distribución chi-cuadrado con k grados de libertad (ya que es la suma de k normales $N(0,1)$ independientes).

2.3 Propiedades.

En primer lugar estudiaremos las relaciones entre la nuevas variables y las componentes principales obtenidas mediante la matriz de covarianzas V.

Proposición 2.2. *Si Y son las componentes principales obtenidas a partir de X, entonces*

$$\begin{aligned} Cov(X,Y) &= TD \\ Corr(X,Y) &= diag(V)^{-1/2}TD^{1/2} \end{aligned} \tag{2.6}$$

donde $diag(V) = diag(\sigma_1^2, ..., \sigma_k^2)$.

Proof. En primer lugar señalaremos que

$$Cov(X, Y) = Cov(X, T'X) = VT$$

y, como $T'VT = D$ y T es ortogonal, entonces $VT = TD$ y $Cov(X, Y) = TD$.

Por otro lado se tiene que como

$$Corr(X_i, Y_j) = \frac{Cov(X_i, Y_j)}{\sigma_i \lambda_j^{1/2}},$$

entonces $Corr(X, Y) = diag(V)^{-1/2} Cov(X, Y) D^{-1/2}$ y

$$Corr(X, Y) = diag(V)^{-1/2} TDD^{-1/2} = diag(V)^{-1/2} TD^{1/2}.$$

□

Corolario 2.2. *En las condiciones de la proposición anterior se tiene:*

$$Cov(X_i, Y_j) = t_{i,j} \lambda_j$$

$$Corr(X_i, Y_j) = \frac{t_{i,j}}{\sigma_i} \lambda_j^{1/2}$$

para todo i, j.

Definición 2.2. *Llamaremos* ***matriz de saturaciones*** *a* $A = Corr(X, Y)$.

Ejemplo 2.5. *Para el vector aleatorio* $(X_1, X_2)'$ *con normal de media* $\mu = (0, 0)'$ *y matriz de covarianzas*

$$V = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

se obtiene

$$T = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}, D = \begin{pmatrix} 1.5 & 0 \\ 0 & 0.5 \end{pmatrix}$$

por lo que la matriz de saturaciones valdrá:

$$A = TD^{1/2} = \frac{1}{2} \begin{pmatrix} \sqrt{3} & 1 \\ \sqrt{3} & -1 \end{pmatrix} = \begin{pmatrix} 0.86603 & 0.5 \\ 0.86603 & -0.5 \end{pmatrix}.$$

Nótese que la primera componente explica un 75% ($0.86603^2 100$*) de las variables* X_1 *y* X_2*, mientras que la segunda solo un 25%. Las saturaciones y sus cuadrados suelen representarse en tablas de la forma siguiente:*

$a_{i,j}$	Y_1	Y_2
X_1	*0.866*	*0.5*
X_2	*0.866*	*-0.5*

$a_{i,j}^2$	Y_1	Y_2	*total*
X_1	*0.75*	*0.25*	*1*
X_2	*0.75*	*0.25*	*1*

lo que nos puede ayudar a "interpretar" las componentes principales. Las saturaciones también se pueden representar gráficamente igual que hacíamos con los coeficientes en la Figura 2.7. Aunque en este ejemplo, las saturaciones con las distintas variables coincidan, esto no siempre es así, y tendremos variables mejor explicadas por las componentes elegidas que otras.

Proposición 2.3. *Si A es la matriz de saturaciones, entonces*

$$AA' = Corr(X)$$

Proof. Si multiplicamos se obtiene

$$\begin{aligned} AA' &= diag(V)^{-1/2}TD^{1/2}(diag(V)^{-1/2}TD^{1/2})' \\ &= diag(V)^{-1/2}TD^{1/2}D^{1/2}T'diag(V)^{-1/2} \\ &= diag(V)^{-1/2}TDT'diag(V)^{-1/2} \\ &= diag(V)^{-1/2}Vdiag(V)^{-1/2} \\ &= Corr(X). \end{aligned}$$

□

También se pueden obtener las correlaciones múltiples entre cada variable original y el grupo de componentes principales elegidas lo que nos dará una idea de lo bueno que es el modelo formado por las componentes principales (y los correspondientes coeficientes) para predecir cada variable original. Recordaremos la definición de coeficiente de correlación múltiple.

Definición 2.3. *Si $X = (X_1, ..., X_k)'$ es un vector aleatorio se llama* ***coeficiente de correlación múltiple****(al cuadrado) de X_1 respecto de $Z = (X_2, ..., X_k)'$ a*

$$Corr^2(X_1, Z) = \rho^2_{1(2,...,k)} = \frac{v'_{1,2}V^{-1}_{2,2}v_{1,2}}{\sigma_{1,1}}$$

donde $Cov(X) = \begin{pmatrix} \sigma_{1,1} & v'_{1,2} \\ v_{1,2} & V_{2,2} \end{pmatrix}$, $Cov(Z) = V_{2,2}$, $\sigma_{1,1} = Var(X_1)$ *y* $v'_{1,2} = (\sigma_{1,2}, ..., \sigma_{1,k}) = Cov(X_1, Z)$.

Nótese que si $k = 2$, entonces $\rho^2_{1(2)} = \sigma_{1,2}\sigma^{-1}_{2,2}\sigma_{1,2}/\sigma_{1,1} = \rho^2_{1,2}$. La interpretación de este coeficiente se obtiene del resultado siguiente.

Proposición 2.4. *El coeficiente de correlación múltiple es el máximo de las correlaciones (al cuadrado) de X_1 con combinaciones lineales de $Z = (X_2, ..., X_k)'$, es decir*

$$\max_{\alpha} Corr^2(X_1, \alpha' Z) = \rho^2_{1(2,...,k)}$$

Proof. De la definición se tiene

$$Corr^2(X_1, \alpha' Z) = \frac{(Cov(X_1, \alpha' Z))^2}{\sigma_{1,1} Var(\alpha' Z)} = \frac{(Cov(X_1, Z)\alpha)^2}{\sigma_{1,1} Cov(\alpha' Z, \alpha' Z)}$$
$$= \frac{(\alpha' v_{1,2})^2}{\sigma_{1,1} \alpha' V_{2,2} \alpha} = \frac{(\alpha' V_{2,2}^{1/2} V_{2,2}^{-1/2} v_{1,2})^2}{\sigma_{1,1} \alpha' V_{2,2} \alpha}$$

y, usando la desigualdad de Cauchy-Schwarz ($(x'y)^2 \leq (x'x)(y'y)$) para $x' = \alpha' V_{2,2}^{1/2}$ e $y = V_{2,2}^{-1} v_{1,2}$, se tiene

$$Corr^2(X_1, \alpha' Z) \leq \frac{\alpha' V_{2,2} \alpha v_{1,2}' V_{2,2}^{-1} v_{1,2}}{\sigma_{1,1} \alpha' V_{2,2} \alpha} = \frac{v_{1,2}' V_{2,2}^{-1} v_{1,2}}{\sigma_{1,1}},$$

es decir $\rho^2_{1(2,...,k)}$ es una cota superior. Además, la igualdad se obtiene si x e y tienen la misma dirección

$$x = V_{2,2}^{1/2} \alpha = \lambda y = \lambda V_{2,2}^{-1/2} v_{1,2},$$

es decir, si $\alpha = \lambda V_{2,2}^{-1} v_{1,2}$. □

Proposición 2.5. *Si las variables de $Z = (X_2, ..., X_k)'$ son independientes (incorreladas) entre sí, entonces*

$$Corr^2(X_1, Z) = \sum_{j=2}^{k} Corr^2(X_1, X_j).$$

Proof. La demostración es inmediata ya que al ser $V_{2,2}$ diagonal, se tiene

$$Corr^2(X_1, Z) = \frac{v_{1,2}' V_{2,2}^{-1} v_{1,2}}{\sigma_{1,1}} = \sum_{j=2}^{k} \frac{\sigma_{1,j}^2}{\sigma_{1,1} \sigma_{j,j}} = \sum_{j=2}^{k} \rho_{1,j}^2.$$

□

Por lo tanto, es interesante calcular las correlaciones múltiples de cada variable original con el grupo de las p primeras componentes principales elegidas ($p \leq k$) para medir el máximo que podemos explicar de cada variable original a partir de combinaciones lineales esas componentes principales.

Proposición 2.6. *Si Y son las componentes principales obtenidas a partir de X, entonces*

$$Corr^2(X_i, (Y_1, ..., Y_p)) = \sum_{j=1}^{p} Corr^2(X_i, Y_j) = \frac{1}{\sigma_{i,i}} \sum_{j=1}^{p} t_{i,j}^2 \lambda_j = \sum_{j=1}^{p} a_{i,j}^2.$$

La demostración es inmediata ya que las componentes son incorreladas entre sí. A estas correlaciones se las suele denominar **comunalidades**

$$c_i = Corr^2(X_i, (Y_1, ..., Y_p))$$

y se suelen representar en la tabla de las saturaciones al cuadrado (como totales de las filas). Además, el máximo de la correlación se obtiene con la combinación lineal $\alpha_i'(Y_1, ..., Y_p)'$ con

$$\alpha_i = \lambda V_{2,2}^{-1} v_{1,2} = \lambda \begin{pmatrix} \lambda_1^{-1} & ... & 0 \\ ... & ... & ... \\ 0 & ... & \lambda_p^{-1} \end{pmatrix} \begin{pmatrix} t_{i,1}\lambda_1 \\ ... \\ t_{i,p}\lambda_p \end{pmatrix} = \lambda \begin{pmatrix} t_{i,1} \\ ... \\ t_{i,p} \end{pmatrix}.$$

Es decir, si tenemos que obtener X en función de las p primeras componentes principales, lo haremos a partir de la relación $X = TY$ eliminando el resto de las componentes. Lógicamente, si $p = k$, se obtiene $\alpha_i'(Y_1, ..., Y_k)' = \lambda X_i$ y $Corr^2(X_i, (Y_1, ..., Y_k)) = 1$ y la igualdad

$$Corr^2(X_i, (Y_1, ..., Y_k)) = \sum_{j=1}^{k} Corr^2(X_i, Y_j) = \frac{1}{\sigma_{i,i}} \sum_{j=1}^{k} t_{i,j}^2 \lambda_j = 1.$$

Recíprocamente, la información contenida en la componente principal j-ésima vale:

$$\lambda_j = \lambda_j \sum_{i=1}^{k} t_{i,j}^2 = \sum_{i=1}^{k} \sigma_{i,i} \frac{1}{\sigma_{i,i}} t_{i,j}^2 \lambda_j = \sum_{i=1}^{k} \sigma_{i,i} Corr^2(X_i, Y_j),$$

ya que $1 = \sum_{i=1}^{k} t_{i,j}^2$ es el módulo al cuadrado del vector propio t_j (columnas de T) y la información (variación) total contenida en las p primeras componentes principales vale:

$$\sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \sum_{i=1}^{k} \sigma_{i,i} Corr^2(X_i, Y_j) = \sum_{i=1}^{k} \sigma_{i,i} Corr^2(X_i, (Y_1, ..., Y_p)) = \sum_{i=1}^{k} c_i \sigma_i^2.$$

Si todas las varianzas son 1, la información total $\sum_{j=1}^{p} \lambda_j$ será la suma de la comunalidades, es decir, la suma de la información que se tiene de cada variable original. Si $p = k$, entonces $c_i = 1$ y se tiene $sum_{j=1}^{k} \lambda_j = \sum_{i=1}^{k} \sigma_i^2$.

En el ejemplo anterior tenemos

$a_{i,j}^2$	Y_1	Y_2	Total
X_1	0.75	0.25	1
X_2	0.75	0.25	1
Total	1.5	0.5	2

donde si $p = 1$ se tiene $\lambda_1 = 3/2 = 0.75 + 0.75$ y si $p = 2$, se tiene $\lambda_1 + \lambda_2 = 3/2 + 1/2 = 2 = 1 + 1 = \sigma_1^2 + \sigma_2^2$.

2.4 Cálculo de las componentes a partir de la matriz de correlaciones.

Cuando se estudian variables en las que se usan unidades diferentes o queremos que éstas no sean significativas (todas las variables sean iguales a priori), las componentes principales suelen calcularse a partir de la matriz de correlaciones $\Pi = (\rho_{i,j})$ con $\rho_{i,j} = \sigma_{i,j}/(\sigma_i\sigma_j)$, lo que equivale a considerar desde el principio las variables estandarizadas $Z_i = (X_i - \mu_i)/\sigma_i$ (se igualan las varianzas a 1). De esta forma, usando el teorema principal, se obtienen las componentes

$$\widetilde{Y} = \widetilde{T}'Z = \widetilde{T}'diag(V)^{-1/2}(X-\mu)$$
$$\widetilde{Y}_j = \widetilde{t}_j'Z = \sum_{i=1}^{k}\widetilde{t}_{i,j}Z_i = \sum_{i=1}^{k}\widetilde{t}_{i,j}\frac{X_i-\mu_i}{\sigma_i},$$

donde $\widetilde{T}$ es la matriz ortogonal que diagonaliza $\Pi = Corr(X) = Cov(Z)$

$$\widetilde{T}'\Pi\widetilde{T} = diag(\widetilde{\lambda}_1, ..., \widetilde{\lambda}_k) = \widetilde{D}$$

$\Pi\widetilde{t}_j = \lambda_j\widetilde{t}_j$ y $Z = (Z_1, \ldots, Z_k)'$. De esta forma, se obtiene

$$Cov(\widetilde{Y}) = Cov(\widetilde{T}'Z) = \widetilde{T}'\Pi\widetilde{T} = \widetilde{D}.$$

Es decir, las componentes principales obtenidas a partir de la matriz de correlaciones serán las variables incorreladas con varianza máxima que se pueden obtener a partir de combinaciones lineales de las variables estandarizadas $Z = diag(V)^{-1/2}(X-\mu)$. Sin embargo, los resultados que se obtienen son (en general) diferentes de los que se obtienen a partir de V.

Proposición 2.7. *Si $\widetilde{Y}$ son las componentes principales obtenidas a partir de la matriz de correlaciones de X, entonces*

$$Corr(X, \widetilde{Y}) = \widetilde{T}\widetilde{D}^{1/2}.$$

En efecto, si $\widetilde{Y} = \widetilde{T}'Z = \widetilde{T}'diag(V)^{-1/2}(X-\mu)$, entonces

$$Cov(Z, \widetilde{Y}) = Cov(Z, \widetilde{T}'Z) = \Pi\widetilde{T} = \widetilde{T}\widetilde{D}.$$
$$Corr(X, \widetilde{Y}) = Corr(Z, \widetilde{Y}) = Cov(Z, \widetilde{Y})\widetilde{D}^{-1/2} = \widetilde{T}\widetilde{D}\widetilde{D}^{-1/2} = \widetilde{T}\widetilde{D}^{1/2}.$$

Observación 2.5. *Nótese que las correlaciones con la componente $\widetilde{Y}_j$ son proporcionales al vector propio $\widetilde{t}_j$ (columnas de $\widetilde{T}$) con constante de proporcionalidad $\widetilde{\lambda}_j^{1/2}$ ($Corr(X_i, \widetilde{Y}_j) = \widetilde{t}_{i,j}\widetilde{\lambda}_j^{1/2}$) y que*

$$\sum_{i=1}^{k} Corr^2(X_i, \widetilde{Y}_j) = \sum_{i=1}^{k}\widetilde{t}_{i,j}\widetilde{\lambda}_j = \widetilde{\lambda}_j.$$

De forma similar, se define la matriz de saturaciones $\widetilde{A} = Corr(X, \widetilde{Y})$, que verifica

$$\widetilde{A}\widetilde{A}' = \widetilde{T}\widetilde{D}^{1/2}\widetilde{D}^{1/2}\widetilde{T}' = Cov(Z) = Corr(X)$$

(como vimos en la sección anterior) y

$$\widetilde{A}'\widetilde{A} = \widetilde{D}^{1/2}\widetilde{T}'\widetilde{T}\widetilde{D}^{1/2} = \widetilde{D},$$

es decir, la matriz de saturación es una matriz que factoriza Π junto a su traspuesta de forma que si las multiplicamos al revés nos da una matriz diagonal.

2.5 Cálculo práctico de las componentes principales.

Si estudiamos k variables (numéricas) en una determinada población usando una muestra de n individuos, tendremos una tabla de datos de la forma siguiente:

Datos	X_1	...	X_k
O_1'	$X_{1,1}$	...	$X_{1,k}$
...	...	...	...
O_n'	$X_{n,1}$	...	$X_{n,k}$

Esta tabla será una m.a.s. $O_1, ..., O_n$ (formada por n vectores aleatorios columna independientes e idénticamente distribuidos) de la variable aleatoria k dimensional $X = (X_1, ..., X_k)'$ que, en muchas ocasiones, podremos suponer normal. Sin embargo, otras veces prescindiremos de estas hipótesis y únicamente analizaremos una tabla de datos, tratando de condensar la información contenida en la misma y de analizar (de forma descriptiva) las relaciones entre las variables y los individuos.

Así, en la práctica, tendremos que la matriz de covarianzas V es desconocida, por lo que tendremos que estimarla y, una vez estimada, procederemos al cálculo de las componentes principales. De esta forma, las componentes principales (y los valores de la matriz T) dependerán de los valores muestrales y, por lo tanto serán v.a. (con individuos distintos, obtendremos componentes distintas) y lo mismo les ocurrirá a los valores propios (serán estimaciones de los verdaderos valores propios).

Para estimar V podemos utilizar la matriz de cuasicovarianzas muestrales

S calculada como:

$$\begin{aligned} O_l &= (X_{l,1}, ..., X_{l,k})' \\ \overline{X}_j &= \frac{1}{n}\sum_{l=1}^{n} X_{l,j} \\ \overline{O} &= (\overline{X}_1, ..., \overline{X}_k)' = \frac{1}{n}\sum_{l=1}^{n} O_l \\ S &= \frac{1}{n-1}\sum_{l=1}^{n}(O_l - \overline{O})(O_l - \overline{O})' = (S_{i,j}) \\ S_{i,j} &= \frac{1}{n-1}\sum_{l=1}^{n}(X_{l,i} - \overline{X}_i)(X_{l,j} - \overline{X}_j). \end{aligned}$$

También podemos usar la matriz de covarianzas muestrales

$$\widehat{V} = \frac{n-1}{n} S.$$

Ambas tendrán los mismos vectores propios y, si n es grande, casi los mismos valores propios.

2.5.1 Cálculo a partir de una muestra

Como no conocemos V, la aproximaremos mediante S o $\widehat{V}$, las diagonalizaremos (calcularemos los ejes de sus elipsoides) y podremos calcular las componentes principales definidas como sigue.

Definición 2.4. *Llamaremos* ***componentes principales muestrales*** *a las variables $\widehat{Y} = \widehat{T}'X$, donde $\widehat{T}$ es la matriz ortogonal que diagonaliza S ($\widehat{V}$) y llamaremos* ***valores propios muestrales*** *$\widehat{\lambda}_i$ a los valores propios de S ($\widehat{V}$). Los valores de $\widehat{T}$ serán las* ***cargas*** *("loadings") o* ***coeficientes muestrales****. Llamaremos* ***puntuaciones muestrales*** *("scores") a los valores que obtendríamos para cada individuo en la componentes muestrales $P_{i,j} = Y_j(O_i) = \widehat{t}_j' O_i$.*

Si optamos por calcular las componentes principales a partir de la matriz de correlaciones, como también es desconocida, en su lugar se usará la matriz de correlaciones (de Pearson) muestrales

$$\begin{aligned} R &= diag(S)^{-1/2} S diag(S)^{-1/2} = (R_{i,j}) \\ R_{i,j} &= S_{i,j}(S_{i,i}S_{j,j})^{-1/2} = \widehat{V}_{i,j}(\widehat{V}_{i,i}\widehat{V}_{j,j})^{-1/2}. \end{aligned}$$

Esto equivaldría a estandarizar las variables iniciales restándoles sus medias muestrales y dividiéndolas por sus cuasivarianzas (es decir, hacer que todas

tengan la misma variabilidad). En este caso, las puntuaciones se calcularán como:

$$P_{i,j} = Y_j(O_i) = \widehat{t}'_j O_i^*$$

donde $\widehat{t}_j$ es el vector propio j-ésimo de R y los datos estandarizados se obtienen (estiman) como

$$O_i^* = \left(\frac{X_{i,1} - \overline{X}_1}{S_1}, \ldots, \frac{X_{i,k} - \overline{X}_k}{S_k} \right)$$

siendo $S_j = \sqrt{S_{j,j}}$ la cuasidesviación típica de la variable X_j. La cuasidesviación típica S_j puede ser reemplazada por la desviación típica muestra $\widehat{V}_j = \sqrt{\widehat{V}_{j,j}}$ (como hace el programa R).

Si n es grande, ambas matrices ($\widehat{V}$ y S) son prácticamente iguales. Si X es normal, $\widehat{V}$ es máximo verosímil y S es insesgado para V, teniendo $(n-1)S$ una distribución (en el muestreo) Wishart $W_k(n-1, V)$. A partir de este resultado, se puede obtener la distribución exacta de los estimadores de los valores propios, pero ésta es bastante complicada. Si usamos $\widehat{V}$ y todos sus valores propios son distintos, se obtendrán estimadores máximo verosímiles para $t_{i,j}$ y λ_j ya que se verifica el resultado siguiente:

Proposición 2.8. *Si $\widehat{\theta}$ es máximo verosímil para θ, entonces $g(\widehat{\theta})$ es máximo verosímil para $g(\theta)$.*

Si X es normal, puede probarse que asintóticamente, la distribución conjunta de los estimadores de los valores propios es normal multivariante y que la de los estimadores de los valores $t_{i,j}$, también lo es, siendo ambas independientes entre sí.

Los valores asintóticos que se obtienen son los siguientes:

Teorema 2.3 (Anderson, 1974)**.** *Si X es normal con matriz de covarianzas con valores propios distintos y $O_1, \ldots, O_n$ es una m.a.s., entonces:*

1) $E(\widehat{\lambda}_j) = \lambda_j + \frac{1}{n} \sum_{i \neq j} \lambda_j \lambda_i (\lambda_j - \lambda_i)^{-1} + O(n^{-2})$
2) $Var(\widehat{\lambda}_j) = \frac{1}{n} 2\lambda_j^2 (1 - \frac{1}{n} \sum_{i \neq j} \lambda_i^2 (\lambda_j - \lambda_i)^{-2}) + O(n^{-3})$
3) $Cov(\widehat{\lambda}_i, \widehat{\lambda}_j) = O(n^{-2})$ si $i \neq j$
4) $\widehat{\lambda} \to_{n\to\infty} N_k(\lambda, 2D^2/n)$
5) $E(\widehat{t}_j) = t_j + O_k(n^{-1})$
6) $Cov(\widehat{t}_j) = \frac{1}{n} \sum_{i \neq j} \lambda_j \lambda_i (\lambda_j - \lambda_i)^{-2} t_j t'_j + O_{k,k}(n^{-2})$
7) $\widehat{t}_j \to_{n\to\infty} N_k(t_j, \frac{1}{n} \sum_{i \neq j} \lambda_j \lambda_i (\lambda_j - \lambda_i)^{-2} t_j t'_j)$
8) $Cov(\widehat{t}_i, \widehat{t}_j) = -\frac{1}{n} \lambda_j \lambda_i (\lambda_j - \lambda_i)^{-2} t_i t'_j + O_{k,k}(n^{-2})$
9) $Cov(\widehat{\lambda}, \widehat{t}_i) \to_{n\to\infty} 0$

Las convergencias de variables aleatorias son convergencias en ley,

$$\lim_{n\to\infty} O(a_n)/a_n = cte$$

y $O_k(a_n)$ y $O_{k,k}(a_n)$ representan a un vector de dimensión k y a una matriz de dimensión $k \times k$ cuyos términos son $O(a_n)$, respectivamente.

Nótese que todos los estimadores son asintóticamente centrados y sus varianzas tienden a cero, siendo además $\widehat{\lambda}_i$ y $\widehat{\lambda}_j$ asintóticamente independientes (incorrelados). No ocurrirá esto si hay dos valores propios iguales ya que, entonces $(\lambda_j - \lambda_i)^{-1} \to \infty$.

2.5.2 Cálculo maximizando la varianza muestral

El cálculo de las componentes principales muestrales se puede enfocar de otra forma buscando la variable $a'X$ (combinación lineal de las originales) con $a'a = 1$ que aplicada a los individuos de la muestra nos de una variable con varianza (o cuasivarianza) muestral máxima. La **puntuación** o contador ("scores") del individuo j en esta nueva variable sería $a'O_j$, su media muestral sería

$$\frac{1}{n}\sum_{j=1}^{n} a'O_j = a'\frac{1}{n}\sum_{j=1}^{n} O_j = a'\overline{O}$$

y su cuasivarianza sería

$$\frac{1}{n-1}\sum_{j=1}^{n}(a'O_j - a'\overline{O})^2 = \frac{1}{n-1}\sum_{j=1}^{n} a'(O_j - \overline{O})(O_j - \overline{O})'a = a'Sa \qquad (2.7)$$

cuyo máximo se alcanza si a es un vector propio del mayor de los valores propios de S. De forma análoga, se procedería para el cálculo de las restantes componentes principales muestrales. Si, por inducción, suponemos que los primeros $i-1$ vectores propios $\hat{t}_j$ de S nos dan las variables incorreladas con mayor varianza y buscamos maximizar la varianza muestral de $a'O_j$ (es decir $a'Sa$) para $a'a = 0$ haciendo que la covarianza muestral

$$\frac{1}{n-1}\sum_{j=1}^{n}(a'O_j - a'\overline{O})(\hat{t}_j'O_j - \hat{t}_j'\overline{O}) = a'S\hat{t}_j = \hat{\lambda}_j a'\hat{t}_j$$

sea cero para $j = 1, \ldots, i-1$. Escribiendo a en función de la base de vectores propios y procediendo como en el teorema principal se obtiene que el óptimo es $a = \hat{t}_i$.

De esta forma, podemos representar a los individuos mediante sus puntuaciones en las dos o tres primeras componentes manteniendo de ellos la mayor información (variabilidad o dispersión) posible (aunque $=_1, \ldots, O_n$ no sea una m.a.s.).

2.5.3 Cálculo minimizando las distancias cuadráticas

Geométricamente, el espacio formado por las m primeras componentes y que pasa por el punto $\overline{O}$ sería el espacio de dimensión m que minimiza la suma de las distancias al cuadrado de los individuos a dicho espacio (regresión perpendicular). De esta forma, el ACP sería como realizar una regresión mínimo cuadrática usando las distancias mínimas (regresión ortogonal) en lugar de las distancias "verticales" de la regresión clásica (para predecir Y en función de X).

Veamos que es cierto para $m = 1$. Para la primera componente, la suma de las distancias al cuadrado de los puntos O_j a la recta $\overline{O} + \lambda a$ vale

$$\sum_{j=1}^{n} d_j^2 = \sum_{j=1}^{n} d^2(O_j, \overline{O}) - p_j^2 = \sum_{j=1}^{n} (O_j - \overline{O})'(O_j - \overline{O}) - \sum_{j=1}^{n} a'(O_j - \overline{O})(O_j - \overline{O})'a,$$

donde p_j es la proyección del vector $O_j - \overline{O}$ sobre la recta $\overline{O} + \lambda a$. El mínimo para $a'a = 1$ coincide con el máximo de (2.7), es decir, se alcanza haciendo que a sea el primer vector propio muestral. Si consideramos cualquier otra recta paralela $P + \lambda a$, se tendrá

$$\sum_{j=1}^{n} d_j^2 = \sum_{j=1}^{n} d^2(O_j, P) - p_j^2 = \sum_{j=1}^{n} (O_j - P)'(O_j - P) - \sum_{j=1}^{n} a'(O_j - \overline{O})(O_j - \overline{O})'a,$$

ya que eligiendo P de forma adecuada podemos conseguir que las proyecciones sean las mismas. Entonces, de nuevo su mínimo para $a'a = 1$ coincide con el máximo de (2.7). Por último, si queremos minimizar

$$\sum_{j=1}^{n} (O_j - P)'(O_j - P)$$

en P tenemos que

$$\begin{aligned}
\sum_{j=1}^{n} d_j^2 &= \sum_{j=1}^{n} (O_j - P)'(O_j - P) \\
&= \sum_{j=1}^{n} (O_j - \overline{O} + \overline{O} - P)'(O_j - \overline{O} + \overline{O} - P) \\
&= n(\overline{O} - P)'(\overline{O} - P) + \sum_{j=1}^{n} (O_j - \overline{O})'(O_j - \overline{O}) + 2\sum_{j=1}^{n} (\overline{O} - P)'(O_j - \overline{O}) \\
&= n(\overline{O} - P)'(\overline{O} - P) + \sum_{j=1}^{n} (O_j - \overline{O})'(O_j - \overline{O}) + 2(\overline{O} - P)'\sum_{j=1}^{n} (O_j - \overline{O}) \\
&= n(\overline{O} - P)'(\overline{O} - P) + \sum_{j=1}^{n} (O_j - \overline{O})'(O_j - \overline{O}),
\end{aligned}$$

donde el segundo sumando es constante y el mínimo del primer sumando se alcanza con $P = \overline{O}$ (ya que es positivo).

2.6 Análisis de componentes principales en R

Veamos como realizar una ACP (PCA en inglés) en R sobre un conjunto de datos. Los conjuntos de datos disponibles en R en el "paquete" *Datasets* se pueden ver mediante:

```
data()
```

Otros "paquetes" incluyen otros ficheros de datos. Consideraremos los datos del Ejemplo 2.1 denominados `LifeCycleSavings`. Recordemos que para ver estos datos en R debemos hacer:

```
LifeCycleSavings
```

y para guardarlos en `d`

```
d<-LifeCycleSavings
```

Como ya hemos comentado el fichero contiene 5 variables medidas en 50 países diferentes (ver Tabla 2.7).

Tabla 2.7: Primeros datos del fichero `LifeCycleSavings`.

	sr	pop15	pop75	dpi	ddpi
Australia	11.43	29.35	2.87	2329.68	2.87
Austria	12.07	23.32	4.41	1507.99	3.93
Belgium	13.17	23.80	4.43	2108.47	3.82
Bolivia	5.75	41.89	1.67	189.13	0.22
Brazil	12.88	42.19	0.83	728.47	4.56

Los detalles sobre estos datos se pueden ver tecleando:

```
help(LifeCycleSavings)
```

donde se indica que:

sr: incremento de los ahorros personales 1960-1970 (crecimiento ahorros dividido por ingresos)

pop15: % población menor de 15 años.

pop75: % población mayor de 75.

dpi: ingresos per-capita.

ddpi: crecimiento del dpi 1960-1970.

2.6.1 Análisis inicial de los datos

Antes de aplicar cualquier técnica estadística es conveniente hacer un estudio preliminar de los datos. Podemos empezar estudiando las variables por separado para detectar valores atípicos, falta de simetría o normalidad, etc. En realidad podemos usar las técnicas que consideremos oportunas. Para resumir estas variables podemos hacer:

```
summary(d)
```

obteniendo medias, medianas, cuartiles, mínimos y máximos (ver Tabla 2.8).

Tabla 2.8: Principales características (media, mediana, etc.) de todas las variables del fichero de R `LifeCycleSavings`.

	sr	pop15	pop75	dpi	ddpi
Min.	0.600	21.44	0.560	88.94	0.220
1st Qu.	6.970	26.21	1.125	288.21	2.002
Median	10.510	32.58	2.175	695.66	3.000
Mean	9.671	35.09	2.293	1106.76	3.758
3rd Qu.	12.617	44.06	3.325	1795.62	4.478
Max.	21.100	47.64	4.700	4001.89	16.710

Para obtener los valores de la primera variable `sr` podemos hacer: `d$sr` o `d[,1]`. Podemos dibujarlos con:

```
plot(d[,1])
```

(ver Figura 2.8) o con:

```
boxplot(d[,1])
```

(ver Figura 2.9). En este caso no se observan ni datos atípicos, ni falta de simetrías y tendencias (falta de aleatoriedad).

Otra opción interesante son los histogramas que se pueden realizar con `hist(d$sr)` o con `hist(d[,1])` (ver Figura 2.10). Estos gráficos nos permiten estudiar simetrías y normalidad. En este caso se observa asimetría y falta de normalidad (el gráfico no se parece a la campana de Gauss).

Para estudiar las relaciones entre las variables podemos hacer todos los gráficos por parejas mediante

```
boxplot(d)
```

o

```
plot(d)
```

Las gráficas se pueden ver en las Figuras 2.11 y 2.12. En la primera se observa que una variable (dpi) tiene mucha mas dispersión que las demás (esto es muy importante para decidir si hacemos un PCA con la matriz de covarianzas o la de correlaciones). En la segunda se aprecia que hay variables poco relacionadas con las demás (por ejemplo, sr) y que en algunas variables hay valores atípicos.

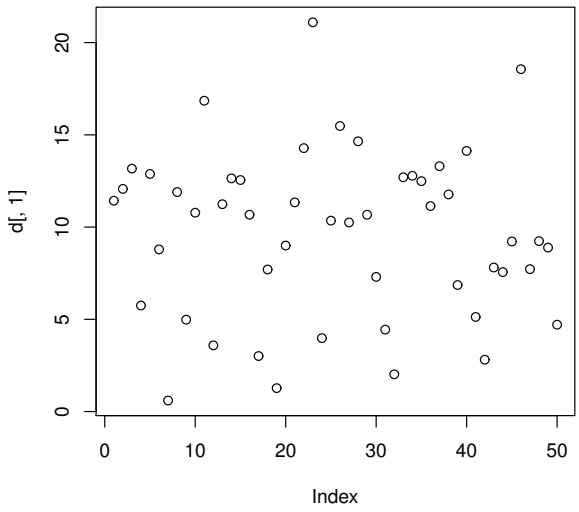


Figure 2.8: Datos en la primera variable del fichero `LifeCycleSavings`.

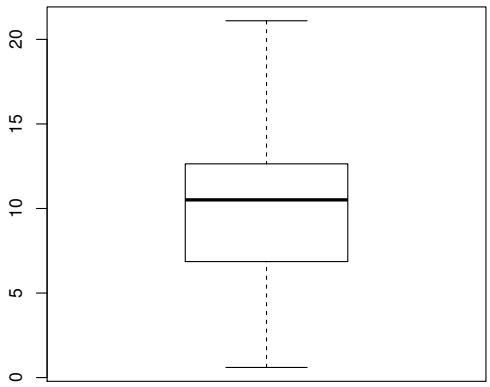


Figure 2.9: Gráfico caja-bigote los datos de la primera variable del fichero de R `LifeCycleSavings`.

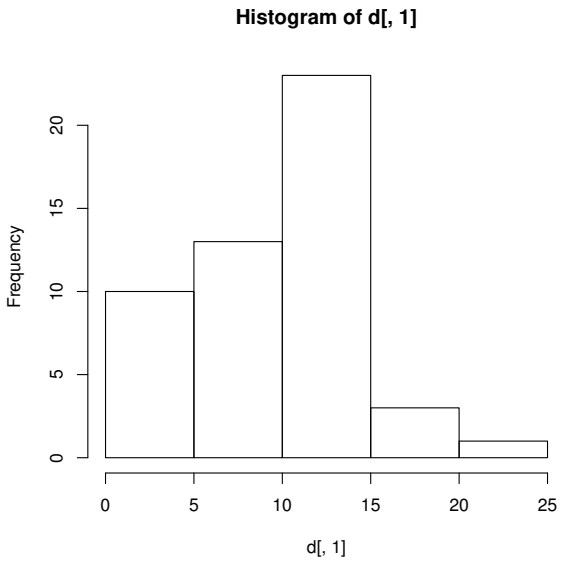


Figure 2.10: Histograma de la primera variable del fichero `LifeCycleSavings`.

Por ejemplo, en `ddpi` hay dos países con un valores muy altos (Jamaica 10.23 y Libya 16.71). Podemos usar los comandos `which.max(d[,5])`, `sort(d[,5])` o `order(d[,5])` para detectar los valores extremos.

Podemos calcular las matrices de covarianza y correlación con `cov(d)` y `cor(d)`, respectivamente. Las correlaciones se pueden ver con `View(cor(d))` (ver Tabla 2.9). Se aprecia que existen variables con correlaciones (lineales) positivas, negativas y casi nulas. Estas relaciones se verán reflejadas en las componentes principales.

Tabla 2.9: Correlaciones entre las 5 variables del fichero `LifeCycleSavings`.

	sr	pop15	pop75	dpi	ddpi
sr	1.0000000	-0.45553809	0.31652112	0.2203589	0.30478716
pop15	-0.4555381	1.00000000	-0.90847871	-0.7561881	-0.04782569
pop75	0.3165211	-0.90847871	1.00000000	0.7869995	0.02532138
dpi	0.2203589	-0.75618810	0.78699951	1.0000000	-0.12948552
ddpi	0.3047872	-0.04782569	0.02532138	-0.1294855	1.00000000

En las gráficas anteriores y en la matriz de covarianzas se aprecia que las variables se miden en unidades muy diferentes (recuerde que en la diagonal

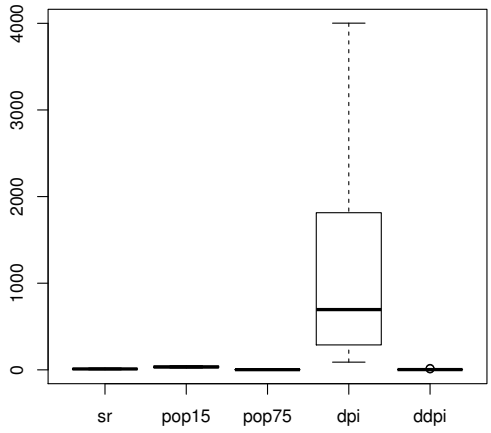


Figure 2.11: Gráficos caja-bigote de los datos de todas las variables del fichero de R `LifeCycleSavings`.

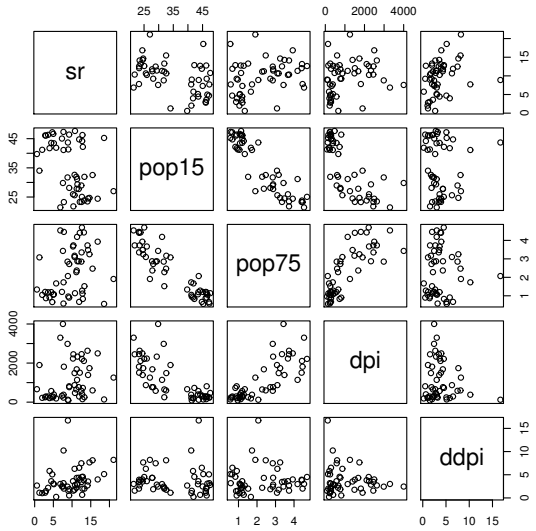


Figure 2.12: Gráficos bidimensionales para todas las variables del fichero de R `LifeCycleSavings`.

aparecen las cuasivarianzas) por lo que el ACP se deberá realizar sobre la matriz de correlaciones (es decir, el programa lo hará sobre las v.a. estandarizadas).

2.6.2 Cálculo de las componentes principales

En el programa R disponemos de dos comandos diferentes para calcular las componentes principales: `princomp` y `prcomp`. Como las componentes son únicas (salvo cambio de signo) cuando los valores propios estimados son distintos, los resultados serán muy similares, pero puede haber pequeñas diferencias debidas a los métodos numéricos usados para su cálculo.

Para hacer un PCA con `princomp` usando la matriz de correlaciones de los datos guardados en `d` basta teclear:

```
PCA<-princomp(d,cor=TRUE)
```

Para ver las características principales haremos:

```
summary(PCA,loadings=TRUE)
```

obteniendo:

Importance:	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Stan. dev.	1.6799041	1.1207437	0.777512	0.4895354	0.278721
Prop. Var.	0.5644156	0.2512133	0.120905	0.0479299	0.015537
Cum. Prop.	0.5644156	0.8156289	0.936534	0.984463	1

Loadings:	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
sr	0.308	0.554	0.750	-0.130	-0.134
pop15	-0.571			-0.416	-0.707
pop75	0.560	-0.101	-0.212	0.390	-0.692
dpi	0.514	-0.266	-0.145	-0.801	
ddpi		0.782	-0.609	-0.123	

La importancia de las componentes se mide con sus desviaciones estándar (raíces cuadradas de los valores propios de la matriz de correlaciones ordenados de mayor a menor) estimadas, la proporción en tanto por uno de sus varianzas estimadas y las proporciones acumuladas. En este caso, las varianzas iniciales suman 5 (la traza de la matriz de correlaciones muestral) por lo que el primer valor de las proporciones 0.5644156, se calcula como:

```
(1.6799041 ^ 2)/5
```

Las proporciones acumuladas se calculan sumando las de las componentes anteriores. Por ejemplo, $0.8156289 = 0.5644156 + 0.2512133$. Estos valores nos indican que la primera componente mantiene un 56.44156% de la información inicial, la segunda un 25.12133% y las dos juntas un 81.56289%.

Las cargas (*loadings*) son los vectores propios unitarios de los valores propios anteriores. Los valores ausentes son números pequeños (pero no necesariamente cero). Si queremos guardarlos podemos hacer:

```
PCA$loadings->T
```

De esta forma, tecleando `T[,1]` obtenemos el primer vector propio que se utiliza para calcular la primera componente principal. Su último coeficiente es 0.03787232 (pequeño pero no cero). Se puede comprobar que este vector tiene norma 1. Por lo tanto la primera componente principal (muestral) se calcularía como:

$$\hat{Y}_1 = 0.308 * X_1^* - 0.571 * X_2^* + 0.560 * X_3^* + 0.514 * X_4^* + 0.0379 * X_5^* \quad (2.8)$$

(los coeficientes se han redondeado), donde $X_i^* = (X_i - mean(X_i))/sd(X_i)$ es la variable i-ésima estandarizada (muestral). Si usamos la matriz de covarianzas para el PCA, en esta fórmula se usarán las variables originales.

Análogamente, las puntuaciones (*scores*), es decir, los valores que obtendrían los individuos de la muestra (países en este caso) en las componentes principales (usando esas fórmulas), se pueden calcular mediante:

```
PCA$scores->S
```

Tecleando `S` comprobamos que, por ejemplo, la puntuación en la primera componente de Australia es 1.36528994. El significado de estos valores se verá posteriormente.

Para comprobar que los valores obtenidos con `princomp` son los correctos podemos hacer:

```
eigen(cor(d))
```

con lo que obtenemos los valores propios ordenados (varianzas de las componentes) y los vectores propios (cargas o coeficientes). Para calcular las puntuaciones debemos primero estandarizar (por columnas) los datos iniciales (este paso no será necesario cuando usemos la matriz de covarianzas). Para ello usaremos:

```
z<-scale(d)
```

y para calcular las puntuaciones de la primera componente haremos:

```
y1<-0.30846174*z[,1]-0.57065322*z[,2]+0.56043119*z[,3]+
+0.51350640*z[,4]+0.03787232*z[,5]
```

De esta forma, para Australia obtenemos 1.35156808, que es similar al valor obtenido antes.

Las componentes principales se pueden calcular aunque no se dispongan de los datos completos usando únicamente la matriz de correlaciones (o covarianza). Para ello haremos:

```
princomp(covmat=cor(d))
```

sustituyendo `cor(d)` por la matriz de correlación (o covarianzas) de los datos. Las cargas se calcularán como antes pero, en este caso, no podremos calcular las

puntuaciones y no podremos hacer los gráficos de las componentes principales (objetos).

Para calcular las componentes principales con `prcomp` usando la matriz de correlaciones debemos hacer:

```
PCAbis<-prcomp(d,scale=TRUE)
```

Haciendo `summary(PCAbis)` se obtiene la importancia de las componentes, con `PCAbis` las cargas y con `PCAbis$x` las puntuaciones (usa las cuasivarianzas). Nótese que los valores son similares pero que se han cambiado de signo las dos primeras componentes (su interpretación será opuesta). Compruebe que se obtienen resultados totalmente diferentes (erróneos en este caso) si usamos la matriz de covarianzas tecleando `princomp(d)` (procure no alterar los objetos usados anteriormente ya que se usarán en las secciones siguientes).

2.6.3 Análisis de las componentes principales

Para analizar las componentes principales calculadas en la sección anterior primero debemos fijarnos en la importancia de cada una. Hablaremos posteriormente sobre el número adecuado de componentes pero, antes de analizarlas, debemos tener en cuenta que la primera tiene un 56.44156% de la información inicial, la segunda un 25.12133% y las dos juntas un 81.56289%. Por lo tanto, en este caso, la información proporcionada por la primera será en general el doble de importante que la que proporciona la segunda, etc. Esto es un cómputo global por lo que puede haber variables que estén mejor representadas en Y_2 que en Y_1 (por ejemplo `ddpi`).

En segundo lugar miraremos las cargas (loadings) o coeficientes de las componentes que queremos analizar para poder dar un significado a estas variables nuevas denominadas componentes principales. Si miramos las cargas de Y_1 dadas en (2.8) y guardadas en `T[,1]`, teniendo en cuenta que las variables están estandarizadas (y tendrán valores similares), podemos afirmar que las variables que más influyen en $Y1$ son (por orden de influencia): pop15 (negativa), pop75 (positiva), dpi (positiva) y sr (positiva). Por lo tanto, Y_1 tomará valores grandes en los países con valores pequeños en pop15 y grandes en las otras tres. Por lo tanto, Y_1 nos indicará los países que tienen poblaciones envejecidas (alta pop75 y baja pop15) y ricos (altos valores en dpi y sr). Estas suelen ser características de países muy desarrollados.

Una vez que ya hemos interpretado una componente, podemos analizar sus puntuaciones para decir cómo serán (aproximadamente) los individuos de la muestra según los valores que toman en esa componente. Usando `summary(S[,1])` y/o `plot(S[,1])` (ver Figura 2.13) observamos que los valores de la muestra en Y_1 están entre -2.258755 y 2.787708 (su media es cero ya que hemos usado variables estandarizadas). Haciendo `which.max(S[,1])` comprobamos que el valor mayor en Y_1 corresponde a Suecia (2.787708) y el menor a Malasia

(−2.258755). Por lo tanto, Australia con 1.36528994 sería, en aquella época, un país bastante desarrollado y España con 0.69294913 estaría un poco por encima de la media (ver Figura 2.13). En esta gráfica se observa que casi no hay valores entre 0 y -1, por lo que la mayoría de los países se podrían clasificar como del tercer o del primer mundo (en esa época). Analice la segunda componente y estudie las puntuaciones de estos países en esa componente.

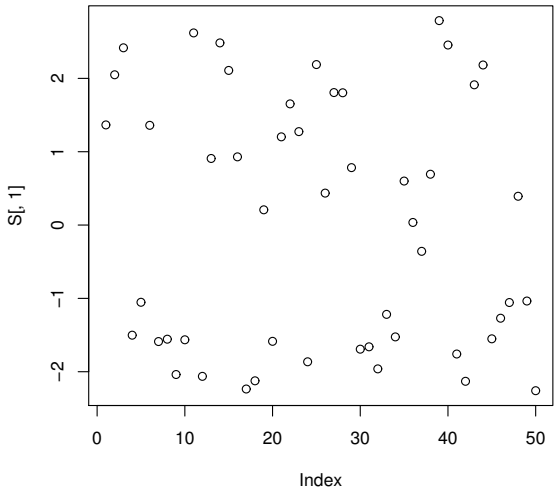


Figure 2.13: Gráfico de puntuaciones para la primera componente principal.

Como las componentes son incorreladas (independientes si las variables iniciales son normales), las podemos estudiar por separado. Sin embargo, muchas veces resulta conveniente representarlas por parejas en gráficos bidimensionales. Para representar las cargas y las puntuaciones de las dos primeras componentes haremos:

```
biplot(PCA,pc.biplot=TRUE)
```

El resultado puede verse en la Figura 2.14. Las cargas aparecen como vectores en rojo con las escalas en la derecha y arriba y las puntuaciones en negro con las etiquetas de los datos (nombres de los países) con las escalas abajo (Y_1) y en la izquierda (Y_2). Las puntuaciones en el gráfico se reescalan para tener varianza 1 (componentes principales estandarizadas).

Este gráfico es la ('mejor') proyección bidimensional ('foto') de los ejes iniciales de las cinco variables estandarizadas y de las puntuaciones de los individuos (países) de la muestra. Las cargas de este gráfico se pueden usar (igual que antes) para interpretar las componentes. Las variables con vectores largos (norma cercana a 1) estarán bien representadas por las dos primeras compo-

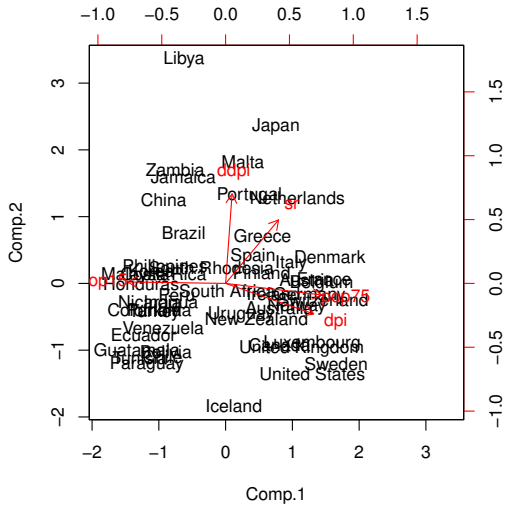


Figure 2.14: Gráfico de las dos primeras componentes principales estandarizadas.

nentes, mientras que las que tengan vectores cortos estarán mal representadas (se pierden al proyectar por ser casi perpendiculares). En este ejemplo todas las variables están bien representadas en este gráfico. Además, se aprecia que `pop75`, `dpi` y, en menor medida, `sr`, hacen crecer la primera componente, mientras que `pop15` la hace disminuir y `ddpi` no influye en ella. Lógicamente, la interpretación de Y_1 es la misma que antes. También podemos observar que la segunda componente crece cuando crece `ddpi` (incremento ingresos per-capita) y en menor medida `sr` (incremento de los ahorros personales), decrece un poco si crece `dpi` y casi no se ve afectada por el envejecimiento de la población. Por lo tanto se puede interpretar como un índice del crecimiento de los países en esa década (los valores grandes corresponderán a los países que más han crecido).

Las puntuaciones se usarán para decir cómo serán (aproximadamente) los individuos de la muestra (países) en esas características. A la derecha tendremos a los países más desarrollados y con poblaciones envejecidas (Suecia, US, etc.) a la izquierda lo contrario (Honduras, Guatemala, etc.), arriba a los países que más se desarrollaron durante esa época (Líbia, Japón, etc.) y debajo los que menos (Islandia, US, Paraguay, etc.). También nos podemos fijar en una variable concreta. Por ejemplo, con respecto a `sr` podríamos decir que los países con mayores incrementos de los ahorros personales (valores sr) deberían ser Japón, Malta y Holanda. Si vemos los datos de `sr` (con `d` y `sort(d[,1])`) podemos comprobar que efectivamente Japón es el que tienen un valor mayor

(21.10) pero que el segundo es Zambia (18.56). Es lógico que al proyectar las variables originales, se pierda algo de la información contenida en ellas.

Como las etiquetas de los datos (nombres de los países) son muy grandes, algunos de ellos no se aprecian bien en el gráfico. Para sustituirlos por sus números de línea podemos hacer:

```
biplot(PCA,pc.biplot=TRUE,xlabs=1:50)
```

Si queremos hacer un gráfico de las componentes tercera y cuarta haremos:

```
biplot(PCA,pc.biplot=TRUE,choices=c(3,4),xlabs=1:50)
```

También podemos hacer un gráfico solo de las puntuaciones de las dos primeras componentes (sin estandarizar) con:

```
plot(S[,1],S[,2],xlab='Y1',ylab='Y2')
```

Para encontrar un individuo (país) bastar mirar sus puntuaciones. La mayor dispersión (varianza) de la primera componente indica que ésta es más importante (tiene más información) a la hora de distinguir los datos. Además podemos localizar cualquier país usando sus puntuaciones. Por ejemplo, encuentre España y diga como serán sus medidas según su posición en el gráfico. Para poner una etiqueta al dato $i = 38$ haremos:

```
text(S[38,1],S[38,2],labels='Esp')
```

Las cargas se pueden representar de forma similar. Aunque no son muy habituales, las tres primeras componentes se podrían representar en gráficos 3D. Es mucho mejor realizar el gráfico siguiente que contiene las tres primeras componentes (sin estandarizar):

```
pairs(PCA$scores[,1:3])
```

2.6.4 Saturaciones.

Para medir las relaciones lineales entre las variables iniciales y las componentes principales, se puede calcular la matriz de correlaciones conocida como matriz de saturaciones mediante:

$$Corr(X_i, Y_j) = \frac{t_{i,j}}{\sigma_i}\lambda_j^{1/2}.$$

En la práctica trabajaremos con sus estimaciones. Si el PCA se ha realizado con la matriz de correlaciones (o si todas las varianzas son 1) bastará con multiplicar la columna de los coeficientes (cargas) de cada componente principal por la raíz cuadrada de su valor propio (su desviación estándar). Las saturaciones al cuadrado nos indicarán cuánta información (en tanto por 1) tendrá cada componente de cada variable. En nuestro ejemplo, las saturaciones de la primera componente principal se calcularán con:

```
S1<-T[,1]*1.6799041
```

y las saturaciones al cuadrado con `S1 ^ 2` obteniendo los valores de la Tabla 2.10.

Tabla 2.10: Saturaciones de la primera componente principal.

	sr	pop15	pop75	dpi	ddpi
Sat.	0.5181861	-0.9586427	0.9414706	0.8626415	0.0636219
Inf.	0.26851688	0.91899580	0.88636698	0.74415038	0.00404774

De esta forma comprobamos que la variable mejor representada en Y_1 es *pop15* con un 91.89958% y que de la última variable Y_1 prácticamente no tiene información.

Para calcular todas las saturaciones (usemos o no la matriz de correlaciones) podemos hacer:

```
SAT<-cor(d,S)
```

obteniendo:

Sat.	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
sr	0.5181861	0.6211673	0.5832461	-0.0637125	-0.03740324
pop15	-0.9586427	0.0142035	0.0206425	-0.2037487	-0.19713655
pop75	0.9414706	-0.1131882	-0.1647000	0.1911276	-0.19278378
dpi	0.8626415	-0.2984564	-0.1127514	-0.3922864	0.01310945
ddpi	0.0636219	0.8764293	-0.4733754	-0.0604464	0.00927110

Como las componentes son incorreladas, las correlaciones múltiples al cuadrado o comunalidades serán la suma de las correlaciones al cuadrado:

$$Corr^2(X_i,(Y_1,\ldots,Y_P)) = \sum_{j=1}^{p} Corr^2(X_i,Y_j).$$

Estos valores nos indicarán la información (en tanto por 1) que matienen las p primeras componentes sobre cada variable. Por ejemplo, si decidimos usar las dos primeras componentes, las correlaciones multiples se calcularán mediante:

```
SAT[,1] ^ 2+ SAT[,2] ^ 2
```

obteniendo:

Inf.	Comp.1	Comp.2	Comunalidad
sr	0.268516885	0.38584877	0.6543657
pop15	0.918995810	0.00020174	0.9191976
pop75	0.886366987	0.01281156	0.8991785
dpi	0.744150387	0.08907624	0.8332266
ddpi	0.004047742	0.76812824	0.7721760

Se observa que la variable mejor representada por las dos primeras componentes (gráfico *biplot*) es `pop15` de la que se mantiene un 91.91976% de su información y la peor representada es `sr` con un 65.43657% (no es necesario usar tantos decimales). Se puede comprobar que la media de los valores de la última columna es 0.8156289 que coincide con la información que (en promedio) mantienen Y_1 y Y_2 (calculada anteriormente con `summary(PCA)`). Compruebe que ocurre lo mismo con las informaciones individuales de cada componente. También se puede comprobar que la suma de los valores de cada columna nos dan los valores propios (informaciones) de cada componente principal (solo si usamos la matriz de correlaciones). Por ejemplo, compruebe que sumando los valores de la primera obtenemos 2.822078, es decir, el mayor valor propio de la matriz de correlaciones.

La correlación múltiple al cuadrado $Corr^2(X_i, (Y_1, \ldots, Y_p))$ es el máximo de las correlaciones que se pueden obtener con combinaciones lineales de las componentes $Y_1, \ldots, Y_p$. Además, el máximo de esas correlaciones se obtiene con los coeficientes incluidos en la matriz T. Por ejemplo, la mejor combinación lineal de las dos primeras componentes para aproximar *sr* es la que se obtiene cortando el vector fila `T[1,]`, es decir, $Z_1 = 0.3084617 * Y_1 + 0.5542456 * Y_2$. De esta forma, si calculamos Z_1 con:

```
Z1<-0.3084617*S[,1]+ 0.5542456*S[,2]
```

y calculamos `cor(d[,1],Z1) ^ 2`, se obtiene 0.6543657 que coincide con la información que mantienen esas dos componentes sobre `sr` (correlación al cuadrado máxima). La variable Z_1 se podría usar para predecir `sr` usando las técnicas de regresión lineal vistas en las prácticas anteriores.

2.7 Número de componentes

Una vez realizado un PCA podemos preguntarnos con cuántas componentes principales debemos quedarnos. Las respuesta no es única y puede depender de factores subjetivos. Todas las soluciones serán correctas ya que lo que estamos haciendo es perder algo de información (la menor posible) a cambio de reducir la dimensión (número de variables) inicial. A continuación comentamos algunas

de las técnicas más usadas. En todos ellas el número de componentes elegidas se representará por m y, lógicamente, se tomarán siempre las m primeras componentes principales (ya que son las que más información tienen).

2.7.1 Fijar un número concreto de componentes.

Una opción válida es fijar un número de componentes concreto. Por ejemplo, si queremos hacer una única gráfica bidimensional, evidentemente debemos tomar $m = 2$, con lo que únicamente analizaremos Y_1 e Y_2. En esta opción es fundamental informar de la información total mantenida por las componentes elegidas y advertir si ese número es bajo. Se suelen tomar números pares de componentes para poder realizar gráficas bidimensionales y el valor más usual es $m = 2$.

Tecleando `summary(PCA)` comprobamos que, en nuestro ejemplo, si tomamos $m = 2$, mantendríamos $p = 81.56\%$ de la información inicial lo que podemos considerar como aceptable al reducir la dimensión de 5 a 2. También podríamos informar sobre las comunalidades, es decir, sobre la información mantenida por esas componentes de cada variable (ver sección anterior). En nuestro ejemplo, para $m = 2$, la variable peor representada es *sr* de la que mantienen un 65.44%. Por lo tanto, todas las variables están bien representadas. En otros ejemplos nos podremos encontrar con variables que no están representadas en las componentes elegidas. En estos casos es importante señalarlo.

2.7.2 Fijar un porcentaje mínimo de información mantenida.

Si queremos mantener un porcentaje p de la variabilidad inicial deberemos quedarnos con las primeras m componentes que verifiquen

$$\frac{\widehat{\lambda}_1 + ... + \widehat{\lambda}_m}{\widehat{\lambda}_1 + ... + \widehat{\lambda}_k} \geq \frac{p}{100},$$

donde $\widehat{\lambda}_i$ representan las estimaciones de los valores propios. En nuestro ejemplo, si queremos mantener más de un $p = 90\%$, debemos tomar $m = 3$ con lo que mantendríamos un 93.65%.

Otra regla (diferente) podría ser el fijar un porcentaje mínimo para las comunalidades. De esta forma nos aseguramos de que todas las variables originales (sean importantes o no), estén representadas en las componentes. En nuestro ejemplo, si queremos que las comunalidades sean mayores que 0.5 (es decir queremos mantener al menos un 50% de todas las variables), debemos tomar $m = 2$. Nótese que con esta regla, en nuestro ejemplo, nunca obtendríamos $m = 1$ a pesar de que Y_1 tiene un 56% de la información total.

2.7.3 Regla de Rao.

Esta regla establece que solo serán relevantes las componentes que tengan una variabilidad (varianza o valor propio) mayor que la variabilidad mínima de las variables originales. De esta forma, se tiene

$$\max m : \widehat{\lambda}_m \geq \min\{S_j^2\},$$

donde S_j^2 representan a las cuasivarianzas muestrales de las variables originales. Si las componentes se calculan usando la matriz de correlaciones, como esto es equivalente a usar las variables estandarizadas, se entiende que las varianzas son 1 y, por lo tanto, se toman solo las componentes con valores propios (varianzas o desviaciones típicas) mayores que uno. En nuestro ejemplo, esta regla nos conduce a $m = 2$ ya que

$$\widehat{\lambda}_2 = 1.256066 > 1 > \widehat{\lambda}_3 = 0.6045255.$$

Si calculásemos las componentes con la matriz de covarianzas (aunque ya hemos comentado que esto no sería correcto), el mínimo de las cuasivarianzas muestrales corresponde a la variable `pop75` y vale 1.66609082 (hacer `var(d$pop75)` o `cov(d)`) y los valores propios de la matriz de covarianzas valen: 981871.2, 43.14338, 13.68328, 6.629537 y 0.2351568 por lo que con este criterio tomaríamos $m = 4$.

2.7.4 Regla de Kaiser.

Esta regla es similar a la anterior y establece que solo serán relevantes las componentes que tengan una variabilidad mayor que la variabilidad media de las variables originales. De esta forma, se tiene

$$\max m : \widehat{\lambda}_m \geq \frac{1}{k}\sum_{j=1}^{k} S_j^2.$$

Si usamos la matriz de correlaciones para calcular las componentes, como las varianzas iniciales son 1, su media es 1 y este criterio coincide con el de Rao, por lo que, en nuestro ejemplo, obtenemos el mismo resultado $m = 2$. Si calculásemos las componentes con la matriz de covarianzas (aunque ya hemos comentado que esto no sería correcto con estos datos), la media de las cuasivarianzas muestrales es 196387 y los valores propios de la matriz de covarianzas valen: 981871.2, 43.14338, 13.68328, 6.629537 y 0.2351568 por lo que con este criterio tomaríamos $m = 1$.

2.7.5 Regla del codo o del gráfico de sedimentación.

Es uno de los métodos más usados y suele ir incluido en casi todos los programas de estadística. El método consiste en representar j (eje x) frente a los valores propios estimados $\widehat{\lambda}_j$ obteniéndose el denominado gráfico de sedimentación o desmoronamiento (*scree graph*). El gráfico será similar a la acumulación de sedimentos en la ladera de una montaña (cono de desmoronamiento). Se trataría de separar "la montaña" de los "sedimentos". La regla establece que serán representativas las componentes hasta el primer "codo" (sin incluirlo) de la gráfica o hasta que comience la línea recta aproximada final. Para realizar este gráfico en R haremos:

```
screeplot(PCA)
```

con lo que se obtiene el gráfico de la Figura 2.15. Se puede obtener un gráfico similar mediante:

```
plot(eigen(cor(d))$values,type='l',ylab='valores propios')
```

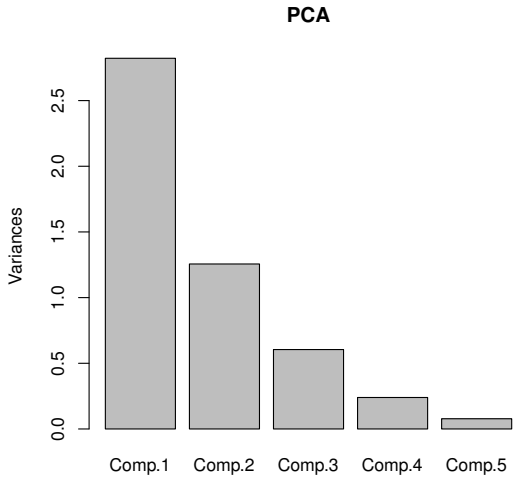


Figure 2.15: Gráfico de sedimentación (screeplot).

En estos gráficos, aunque no está muy claro, parece que el codo (los sedimentos) se encuentra en $j = 3$, por lo que tomaríamos las dos primeras componentes ($m = 2$). Las soluciones $m = 1$ y $m = 3$ también serían aceptables. En otras ocasiones el "codo" aparece más claro y solo hay una opción para m con esta regla.

2.7.6 Prueba de esfericidad.

Esta regla se basa en la contrastación de la hipótesis $H_0 : \lambda_{m+1} = ... = \lambda_k$, cuyo significado es que para un m dado, las componentes restantes tienen igual variabilidad teórica (las diferencias en las varianzas muestrales se deben al azar) y, por lo tanto, no debemos dar preferencia a una sobre otra (también se dice que hay "esfericidad" en $Y_{m+1}, \ldots, Y_k$). El test de esfericidad de Bartlett se basa en el test de razón de verosimilitudes que da el estadístico:

$$T = \left(n - \frac{2k+11}{6}\right)(k-m)\ln\left(\frac{m_a}{m_g}\right),$$

donde $m_a = \frac{1}{k-m}\sum_{i=m+1}^{k}\widehat{\lambda}_i$ y $m_g = (\prod_{i=m+1}^{k}\widehat{\lambda}_i)^{1/(k-m)}$ (medias aritmética y geométrica de los últimos valores propios). En condiciones de normalidad de los datos iniciales y cuando H_0 es cierta, T sigue una distribución chi-cuadrado χ^2_{gl} con $gl = 0.5(k-m-1)(k-m+2)$ grados de libertad. Si H_0 no es cierta, T tiende a tomar valores mayores por lo que la región de rechazo sería de la forma $T > \chi^2_{1-\alpha,gl}$, donde $\chi^2_{1-\alpha,gl}$ es el cuantil $1-\alpha$ de esa distribución chi-cuadrado.

Para aplicar este test a nuestro ejemplo con $m = 2$ calcularemos

```
eigen(cor(d))
(0.60452546+0.23964496+0.07768522)/3->ma
(0.60452546*0.23964496*0.07768522) ^ (1/3)->mg
(50-(2*5+11)/6)*(5-2)*log(ma/mg)->T
0.5*(5-2-1)*(5-2+2)->gl
1-pchisq(T,gl)
```

obteniendo:

$$\begin{aligned}
m_a &= \frac{1}{k-m}\sum_{i=m+1}^{k}\widehat{\lambda}_i = 0.3072852\\
m_g &= \left(\prod_{i=m+1}^{k}\widehat{\lambda}_i\right)^{1/(k-m)} = 0.2240993\\
T &= \left(n - \frac{2k+11}{6}\right)(k-m)\ln\left(\frac{m_a}{m_g}\right) = 44.03837\\
gl &= 0.5(k-m-1)(k-m+2) = 5\\
P-valor &= \Pr(\chi^2_5 > 44.03837) = 2.27505 * 10^{-8}
\end{aligned}$$

y, como el P-valor obtenido es muy pequeño (menor que 0.05), rechazaremos la esfericidad de las tres últimas componentes (H_0) por lo que, si queremos, podemos calcular más componentes (y éstas no serán al azar). La región crítica (de rechazo) para este test con $\alpha = 0.05$ es $(11.0705, \infty)$ donde $11.0705 = \chi^2_{0.95,5}$ se calcula en R mediante `qchisq(0.95,gl)`. La gráfica de la función de densidad de una χ^2_5 se puede obtener con:

```
curve(dchisq(x,5),0,50)
```
obteniéndose la gráfica de la Figura 2.16.

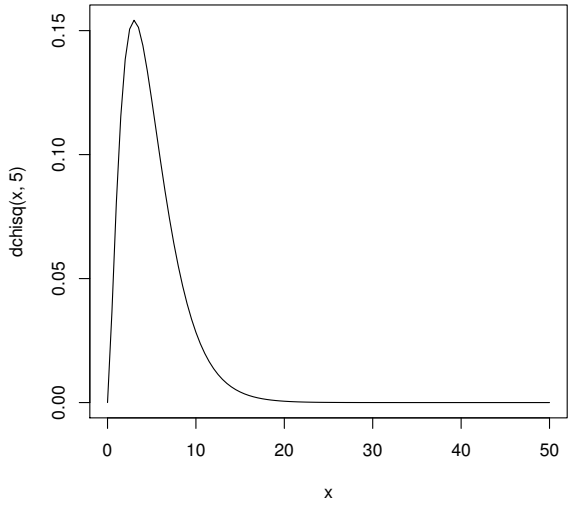


Figure 2.16: Gráfico de la función de densidad Chi-cuadrado con 5 grados de libertad.

Note que es muy posible que no haya esfericidad para ningún m ($m = 1, 2, 3$) (los valores propios teóricos son todos diferentes), pero esto no implica que tengamos que tomar todas las componentes principales. Si para algún m se acepta la esfericidad, no sería conveniente aumentar las componentes (ya que estás podrían obtenerse por azar) y sí podríamos intentar disminuir m.

2.8 Problemas.

1. Calcular las componentes principales para una variable bidimensional con matriz de covarianzas
$$V = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$
Calcular la información que tiene cada componente.

2. Calcular las componentes principales para una variable bidimensional con

matriz de correlaciones

$$\Pi = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}.$$

¿Qué condiciones debe verificar r? Calcular la información que tiene cada componente.

3. Calcular las componentes principales para una variable bidimensional con matriz de covarianzas

$$\begin{pmatrix} 10 & -3 \\ -3 & 2 \end{pmatrix}.$$

4. Calcular la primera componente principal para una variable tridimensional con media cero y matriz de correlaciones

$$\Pi = \begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix}.$$

5. Calcular las componentes principales para una variable tridimensional con media cero y matriz de covarianzas

$$\Sigma = \begin{pmatrix} \beta^2 + \delta & \beta & \beta \\ \beta & 1 + \delta & 1 \\ \beta & 1 & 1 + \delta \end{pmatrix}.$$

(Indicación: $\Sigma - \delta I = (\beta, 1, 1)'(\beta, 1, 1)$).

6. Demostrar que si las varianzas iniciales son iguales entonces las componentes principales que se obtienen con la matriz de covarianzas son iguales a las que se obtienen con la matriz de correlaciones.

7. Calcular las componentes principales de k variables con media cero, varianza uno y correlaciones iguales a r. ¿Qué condiciones debe verificar r? Calcular la información que tiene cada componente.

8. Demostrar que las componentes principales no son invariantes por cambio de escala.

9. (Teorema Perron-Frobenius). Si A es una matriz simétrica con todos sus elementos positivos, entonces todos los coeficientes del vector propio del mayor valor propio de A se pueden tomar todos positivos.

10. Aplicar un PCA a los datos del fichero: `USArrests` que contiene datos sobre los arrestos por cada 100000 residentes por asesinato, asalto o violación en cada uno de los 50 estados de USA en 1973. También se incluye el porcentaje de población que vive en las áreas urbanas. Fuente: `help(USArrests)`.

11. Aplicar un PCA a los datos del fichero de R: `USJudgeRatings`. Fuente: `help(USJudgeRatings)`.

12. Aplicar un PCA a la matriz de covarianza incluida en el fichero: `ability.cov` sobre diversos tests de inteligencia. Fuente: `help(ability.cov)`.

13. Aplicar un PCA a los datos de las columnas 5-10 del objeto `d` del fichero `bears.rda` ([3]). Esas columnas contienen diversas medidas de 143 osos (Head.L= longitud de la cabeza (pulgadas), Head.W=anchura de la cabeza (pulgadas), Neck.G=perímetro cuello (pulgadas), Length=altura (pulgadas), Chest.G=perímetro pecho (pulgadas), Weight=peso (libras)). Fuente: Minitab15. (Indicación: Para aplicar un PCA a las columnas 5-10 del objeto `d` debemos teclear: `PCA<-princomp(d[,5:10])`).

14. Aplicar un PCA a los datos del fichero `heptathlon` del paquete `MVA`[4] correspondientes a los resultados en la prueba femenina de heptatlon en las olimpiadas de Seul 1988.

15. Aplicar un PCA a los datos del fichero `pottery` del paquete `MVA`[4] que contiene resultados de análisis químicos de cerámica británica de la época romana de diversas regiones y hornos (kiln). La región 1 corresponde al horno 1, la región 2 a los hornos 2 y 3, y la región 3 a los hornos 4 y 5. ¿Podemos usar estas medidas para determinar el origen de la cerámica?

16. Aplicar un PCA a los datos del objeto `d` del fichero `nota.rda`[3] que contiene las notas (sobre 100) de alumnos de matemáticas en una universidad americana. Fuente: Rencher (1995, Methods of Multivariate Analysis, Wiley).

17. Aplicar un PCA a los datos del objeto `d` del fichero `madres.rda`[3] que contiene las medidas de madres y sus bebés recién nacidos. Las variables son: PESOM (peso madre), TALLAM (altura de la madre), SEM (semanas de gestación), PASM (presión sanguínea sistólica de la madre), PADM (presión sanguínea diastólica de la madre), PESOR (peso del recién nacido), TALLAR (altura recién del nacido), PTR (perímetro torácico del recién nacido), PCR (perímetro craneal del recién nacido).

18. Aplicar un PCA a los datos del objeto `d` del fichero `decatlon.rda`[3] que contiene los resultados obtenidos por 24 atletas en las 10 pruebas de

[3]Para este tipo de archivos teclear `load('f:/name.rda')` indicando la ruta completa en donde se encuentra el archivo y reemplazando name por el nombre del archivo.

[4]Para leer este conjunto de datos hay que instalar el paquete `MVA` pinchando en el menú: `Paquete > Instalar Paquete` seleccionando `MVA` y tecleando en R: `library('MVA')` (o indicando en el menú que se carge este paquete).

décatlon en los Juegos Olímpicos de Seul 1988. Las variables corresponden a las pruebas siguientes: X1 100 metros lisos (en segundos), X2 salto de longitud (metros), X3 lanzamiento de peso (metros), X4 salto de altura (metros), X5 400 metros lisos (segundos), X6 110 metros vallas (segundos), X7 lanzamiento de disco (metros), X8 salto con pértiga (metros), X9 lanzamiento de jabalina (metros), X10 1500 metros lisos (segundos) y X11 puntuación.

3

Análisis discriminante

En este capítulo mostramos cómo clasificar individuos entre varios grupos a partir de sus medidas en diversas variables aleatorias. Para ello necesitaremos disponer de una muestra de las variables en estudio en individuos de cada grupo (al menos dos individuos por cada grupo) y de las medidas de los elementos a clasificar en esas variables.

3.1 Introducción.

El análisis discriminante trata de decidir si uno (o varios) "individuos" sobre el que se han medido una serie de características (variables) pertenece a una de las poblaciones existentes "a priori". Para ello construiremos *funciones discriminantes* que servirán para decidir en qué población incluimos a cada sujeto. Los ejemplo típicos son la diagnosis de enfermedades, las clasificación de individuos de diferentes especies, diagnosis de autoría en obras de arte, clasificación de perfiles de clientes (por ejemplo en la concesión de créditos), diseño de máquinas de clasificación automática en ingeniería, etc., aunque esta técnica se puede aplicar a muy diferentes situaciones (Economía, Psicología, Meteorología, Genética, etc.). Las variables estudiadas sobre los individuos deben ser numéricas (en muchos casos normales multivariantes) y, lógicamente, cuando no se conozcan las características de las poblaciones en las que se pueden clasificar los individuos, necesitaremos una muestra de individuos de cada una de ellas. El ACP (ver capítulo anterior) y otras técnicas de estadística descriptiva pueden servir de ayuda a la hora de visualizar las diferencias entre los individuos de distintas poblaciones, así como de los que están por clasificar, aunque veremos que pueden existir otras direcciones de proyección que permitan separar mejor a los grupos. Usaremos la técnica denominada *validación cruzada* para dar una estimación de las probabilidades de cada uno de los errores posibles (clasificar a un individuo de la población 1 en la población 2, etc...). La principal diferen-

cia del Análisis Discriminante con el Análisis Cluster (de grupos) es que, en el primer caso, las poblaciones están establecidas de antemano, mientras que en el segundo la clasificación se realiza a posteriori (pudiéndose elegir el número de grupos deseados o el índice de "afinidad" deseada para los individuos de un determinado grupo). Cuando algunas de las variables de clasificación sean de tipo discreto es preferible utilizar otras técnicas de clasificación como la regresión logística (ver Peña 2002).

El primero que claramente estudió un problema de Análisis Discriminante (AD o DA) fue Fisher (Ronald Aylmer, Inglaterra 1890-1962) quién fue consultado por Barnard (1935) para clasificar restos de esqueletos. Fisher (1936) introdujo la función discriminante para clasificar a un individuo en una de dos poblaciones normales con una matriz de covarianzas común. Básicamente veremos que esta clasificación se basa en la distancia de Mahalanobis del individuo a cada una de las poblaciones (sus medias). La utilización de esta distancia es equivalente bajo normalidad a la utilización del criterio de máxima verosimilitud que clasificará a un individuo en donde sus medidas sean "más probables" (verosímiles), es decir, donde la función de densidad sea más grande. Este segundo criterio permitirá la extensión de dicha clasificación a más de dos poblaciones con diferentes matrices de covarianzas incluso sin la necesidad de la normalidad de las mismas. Esta extensión fue llevada a cabo entre otros, por Welch (1939), Anderson (1951) y Okamoto (1963).

Recordemos que si X es una v.a. de dimensión k, media μ y matriz de covarianzas $V = (\sigma_{i,j})$ definida positiva, se define la **distancia de Mahalanobis** (Prasanta Chandra Mahalanobis, India 1893-1972) de $x, y \in \mathbb{R}^k$ como

$$\triangle(x,y) = \sqrt{(x-y)'V^{-1}(x-y).}$$

Obviamente, si V es la matriz identidad, obtenemos la distancia Euclídea. En la Figura 3.1 puede verse una circunferencia para la distancia de Mahalanobis con centro en la media para una Normal bivariante

$$N_2\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, V = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right).$$

Como la función de densidad de la normal es

$$f(x) = \frac{1}{\sqrt{|V|(2\pi)^k}} \exp\left(-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)\right)$$

las circunferencias para la distancia de Mahalanobis con centro en μ coincidirán con las curvas de nivel de la función de densidad ($f(x) = cte$).

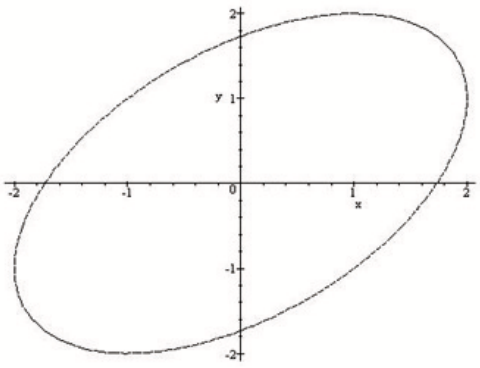


Figure 3.1: Circunferencia para la distancia de Mahalanobis en una población normal bidimensional con medias 0, varianzas 1 y correlación 1/2.

3.2 Clasificación teórica.

3.2.1 Dos poblaciones normales con la misma matriz de covarianza.

Supongamos que $X = (X_1, ..., X_k)'$ e $Y = (Y_1, ..., Y_k)'$ son dos v.a. normales k dimensionales con vectores de medias μ_X y μ_Y y matriz de covarianzas común V definida positiva. Supongamos que $Z = (Z_1, ..., Z_k)$ representa las medidas obtenidas para el individuo que se quiere clasificar y que Z proviene de X o de Y. Es decir, Z será una v.a. normal k dimensional con media igual a μ_X o μ_Y y matriz de covarianzas V. En la práctica z será un punto de $\mathbb{R}^k$ que debemos clasificar en X o en Y.

La idea de Fisher es usar una función discriminante D unidimensional lineal basada en Z. De esta forma,

$$D = a'Z = a_1Z_1 + \cdots + a_kZ_k$$

donde $a \in \mathbb{R}^k$ y si $Z \equiv N_k(\mu, V)$, entonces

$$D = a'Z \equiv N_1\left(a'\mu, \sqrt{a'Va}\right)$$

ya que $E(a'Z) = a'E(Z)$ y

$$Var(a'Z) = Cov(a'Z) = a'Cov(Z)a = a'Va,$$

donde $\mu = E(Z) = \mu_X$ ó μ_Y.

Esta función debe elegirse de forma que discrimine (aleje) a los individuos de X de los de Y, es decir, debemos resolver el problema siguiente:

$$\max_a \frac{(a'\mu_X - a'\mu_Y)^2}{a'Va}. \tag{3.1}$$

Nótese que el objetivo es alejar las "proyecciones" de las medias $a'\mu_X$ y $a'\mu_Y$ y disminuir la varianza común $\sigma^2 = a'Va$ (ver Figura 3.2). La solución se obtiene en el teorema siguiente.

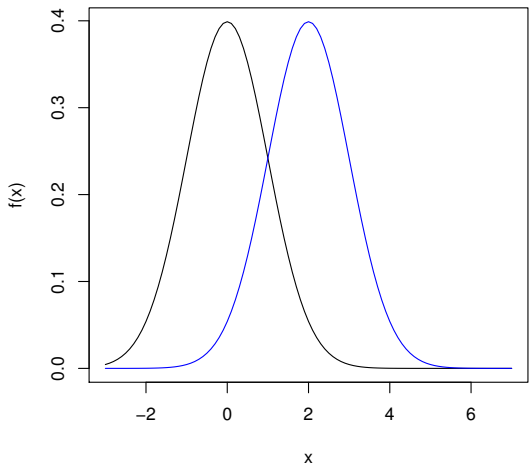


Figure 3.2: Funciones de densidad de las proyecciones en cada grupo.

Teorema 3.1. *Si V es definida positiva, la solución general de (3.1) es*

$$a = \lambda V^{-1}(\mu_X - \mu_y)$$

para $\lambda \neq 0$, y el máximo vale $\triangle^2(\mu_X, \mu_Y)$.

Proof. La demostración se basa en la desigualdad de Cauchy-Schwarz:

$$(x'y)^2 \leq (x'x)(y'y),$$

donde se da la igualdad si, y solo si, $x = \lambda y$. Como V es definida positiva,

existe su inversa V^{-1} y $a'Va > 0$ para todo vector $a \neq 0$. Entonces, tenemos

$$\begin{aligned}\frac{(a'\mu_X - a'\mu_Y)^2}{a'Va} &= \frac{(a'V^{1/2}V^{-1/2}(\mu_X - \mu_Y))^2}{a'Va} \\ &\leq \frac{a'Va(\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y)}{a'Va} \\ &= (\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y) \\ &= \triangle^2(\mu_X, \mu_Y),\end{aligned}$$

donde $x' = a'V^{1/2}$ e $y = V^{-1/2}(\mu_X - \mu_Y)$. Además, se verifica la igualdad si, y solo si $x = \lambda y$, es decir, si

$$V^{1/2}a = \lambda V^{-1/2}(\mu_X - \mu_Y),$$

lo que implica

$$a = \lambda V^{-1}(\mu_X - \mu_Y).$$

□

Definición 3.1. *Llamaremos* ***función discriminante de Fisher*** *a la v.a.*

$$D = L(Z) = a'Z = (\mu_X - \mu_Y)'V^{-1}Z.$$

Nótese que en la proposición anterior no es necesario que las variables sean normales. Si las variables X e Y son normales, entonces la nueva variable D será normal

$$D \sim N_1((\mu_X - \mu_Y)'V^{-1}\mu, \triangle(\mu_X, \mu_Y)),$$

donde $\mu = E(Z)$ es igual a μ_X o μ_Y (ver Figura 3.2). Nótese que hemos tomado $\lambda = 1$, pero que esto no influye en la clasificación ya que podemos tomar cualquier otro λ no nulo. Por ejemplo, si tomamos $\lambda = 1/ \mid a \mid$ obtenemos una proyección en la dirección de a.

Con la función discriminante de Fisher, la regla de discriminación será:

- Si $L(Z) > K$, entonces Z es clasificado en X;
- Si $L(Z) < K$, entonces Z es clasificado en Y;

donde $K = L((\mu_X + \mu_Y)/2)$. En realidad clasificamos a un individuo con características z según $a'z$ esté más cerca de $a'\mu_X$ o de $a'\mu_Y$, ya que, como

$$(\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y) \geq 0,$$

entonces

$$a'\mu_X = (\mu_X - \mu_Y)'V^{-1}\mu_X \geq (\mu_X - \mu_Y)'V^{-1}\mu_Y = a'\mu_Y,$$

es decir, con esta función discriminante, la proyección de la media de X será siempre mayor que la proyección de la media de Y. Ocurrirá lo mismo si tomamos $\lambda > 0$ y lo contrario si tomamos $\lambda < 0$.

De esta forma, se crean dos regiones en el conjunto de posibles valores de Z, la región de individuos que serán clasificados en X y la de los que lo serán en Y:

$$R_X = \{z \in \mathbb{R}^k : L(z) > K\},$$
$$R_Y = \{z \in \mathbb{R}^k : L(z) < K\}.$$

Lógicamente, debemos dar una medida de lo "buena" que es la función discriminante obtenida. Está claro que será mejor cuanto más alejadas estén las medias $a'\mu_X$ y $a'\mu_Y$, y cuanto más pequeña sea la varianza $a'Va$. Así, el cociente

$$\frac{(a'\mu_X - a'\mu_Y)^2}{a'Va} = (\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y) = \triangle^2(\mu_X, \mu_Y)$$

(que no depende de λ) puede servir para comparar una función de discriminación con otra. Nótese que la discriminación será buena si las medias poblacionales (las poblaciones) están alejadas según la distancia de Mahalanobis asociada a V.

Otra forma de medir la bondad de un criterio de clasificación es la que calcula las probabilidades de malas (buenas) clasificaciones. Si llamamos error tipo 1, e_1, al que clasifica a un individuo de la población X en la población Y, entonces

$$\begin{aligned}
\Pr(e_1) &= \Pr(Z \in R_Y \mid Z \equiv X) \\
&= \Pr(L(X) < K) \\
&= \Pr\left(a'X < a'\frac{\mu_X + \mu_Y}{2}\right) \\
&= \Pr\left(\frac{a'X - a'\mu_X}{\sqrt{a'Va}} < \frac{a'(\mu_Y - \mu_X)}{2\sqrt{a'Va}}\right) \\
&= \Pr\left(U < \frac{(\mu_X - \mu_Y)'V^{-1}(\mu_Y - \mu_X)}{2\sqrt{(\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y)}}\right) \\
&= \Pr\left(U < -\frac{1}{2} \triangle (\mu_X, \mu_Y)\right),
\end{aligned}$$

donde $U \equiv N_1(0, 1)$. De forma análoga, puede comprobarse que

$$\Pr(e_2) = \Pr(Z \in R_X \mid Z \equiv Y) = \Pr\left(U > \frac{1}{2} \triangle (\mu_X, \mu_Y)\right) = \Pr(e_1).$$

Por lo tanto, las probabilidades de clasificaciones erróneas son iguales y solo dependen de la distancia de Mahalanobis entre las poblaciones. Lógicamente las probabilidades de clasificaciones correctas vienen dadas por:

$$\Pr(c_1) = \Pr(Z \in R_X \mid Z \equiv X) = 1 - \Pr(e_1),$$

$$\Pr(c_2) = \Pr(Z \in R_Y \mid Z \equiv Y) = 1 - \Pr(e_2),$$

y también son iguales.

Ejemplo 3.1. *Supongamos que tenemos que decidir si un individuo con medidas* $z = (z_1, z_2)' = (2, 0.9)'$ *se clasifica en una población normal bivariante de media* $\mu_X = (0, 0)'$ *o en una de media* $\mu_Y = (1, 2)'$ *siendo la matriz de covarianzas común*

$$V = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

Entonces la función discriminante será:

$$\begin{aligned} D &= L(Z) = a'Z = (\mu_X - \mu_Y)'V^{-1}Z \\ &= -(1,2)\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}^{-1} Z \\ &= -(1,2)\begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix} Z \\ &= -(0,2)\,Z \\ &= -2Z_2, \end{aligned}$$

es decir, $L(z_1, z_2) = -z_2$*. La distancia de Mahalanobis entre las dos poblaciones vale*

$$\begin{aligned} \triangle^2(\mu_X, \mu_Y) &= (\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y) \\ &= (0,2)(1,2)' = 4. \end{aligned}$$

Un individuo Z *será clasificado en la primera población si*

$$-2Z_2 > K = a'\frac{\mu_X + \mu_Y}{2} = -(0,2)(0.5,1)' = -2,$$

es decir, si $Z_2 < 1$*. En este caso,* $z = (1, 0.9)$ *será clasificado en* X*, con una probabilidad de error global*

$$\begin{aligned} \Pr(e_2) &= \Pr(Z \in R_X \mid Z \equiv Y) \\ &= \Pr(U > \frac{1}{2} \triangle (\mu_X, \mu_Y)) \\ &= \Pr(U > 1) \\ &= 1 - F_U(1) \\ &= 1 - 0.8413 = 0.1587, \end{aligned}$$

donde la función de distribución normal estándar $F_U(1)$ se calcula en la Tabla 4.3 o en R haciendo: `pnorm(1)`.

Note que otra función discriminate equivalente sería

$$L^*(z_1, z_2) = z_2,$$

(proyección sobre el eje y) con la que obtendríamos

$$L^*(\mu_X) = L^*(0,0) = 0,$$

$$L^*(\mu_Y) = L^*(1,2) = 2,$$

$$K^* = (L^*(\mu_X) + L^*(\mu_X))/2 = 1$$

y

$$L^*(z) = L^*(1, 0.9) = 0.9,$$

con lo que z se clasificaría en X. Las proyecciones con esta función serán $N(0,1)$ ($L^*(X)$) y $N(2,1)$ ($L^*(Y)$). Las densidades de las proyecciones (sobre el eje y) pueden verse en la Figura 3.3. La probabilidad de error 0.1587 *corresponde a las áreas menores determinadas por la recta vertical en el punto de corte de las densidades en $K = 1$. Para dibujarlas en R podemos hacer:*

```
curve(dnorm(x,0,1),-3,7,ylab='f(x)')
curve(dnorm(x,2,1),add=TRUE,col='blue').
```

Si hacemos cualquier otra proyección, los grupos aparecerán más mezclados. Por ejemplo, si proyectamos sobre el eje x, obtendremos las densidades de la Figura 3.4. ¿Cuál sería la probabilidad de error si usáramos estas proyecciones sobre el eje x?

Welch (1939) probó que, si las poblaciones son normales, el procedimiento de clasificación mediante la función discriminante de Fisher es máximo verosímil, es decir, que se clasifica a un individuo con características z en X si y solo si $f_X(z) > f_Y(z)$. Además, se comprueba que también es equivalente al criterio de clasificación basado en la distancia de Mahalanovis mínima.

Teorema 3.2. *Si las variables X e Y son normales multivariantes con matriz de covarianzas común V y función discriminante de Fisher L, entonces equivalen:*

1) $L(z) > K$.

2) $\triangle_V(z, \mu_X) < \triangle_V(z, \mu_Y)$.

3) $f_X(z) > f_Y(z)$.

Proof. La primera condición $L(z) > K$ es

$$a'z > a'\frac{\mu_X + \mu_Y}{2},$$

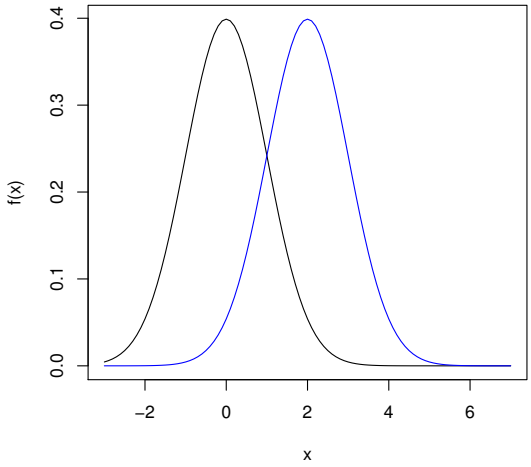


Figure 3.3: Funciones de densidad de las proyecciones sobre el eje y en cada grupo para las poblaciones del Ejemplo 3.1.

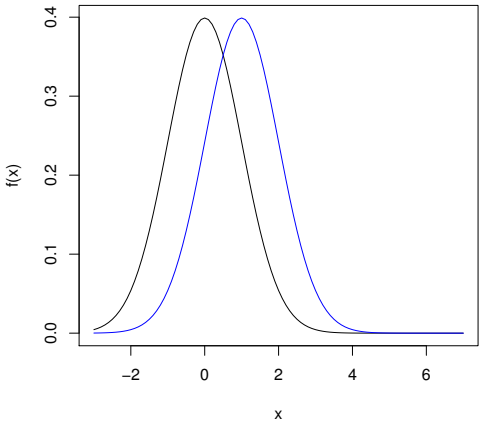


Figure 3.4: Funciones de densidad de las proyecciones sobre el eje x en cada grupo para las poblaciones del Ejemplo 3.1.

con $a' = (\mu_X - \mu_Y)'V^{-1}$, es decir,

$$\begin{aligned}(\mu_X - \mu_Y)'V^{-1}z &> \frac{1}{2}(\mu_X - \mu_Y)'V^{-1}(\mu_X + \mu_Y) \\ &= \frac{1}{2}\mu_X'V^{-1}\mu_X - \frac{1}{2}\mu_Y'V^{-1}\mu_Y\end{aligned}$$

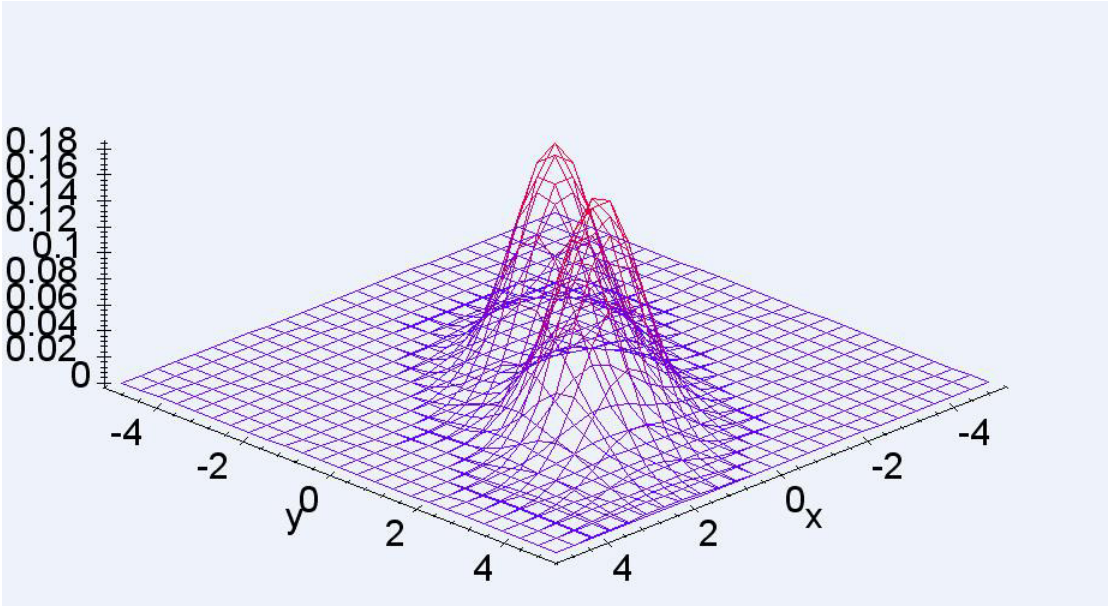


Figure 3.5: Funciones de densidad bivariantes para las poblaciones del Ejemplo 3.1.

lo que equivale a

$$2\mu_X' V^{-1} z - 2\mu_Y' V^{-1} z > \mu_X' V^{-1} \mu_X - \mu_Y' V^{-1} \mu_Y.$$

La segunda condición es equivalente a

$$(z - \mu_X)' V^{-1} (z - \mu_X) < (z - \mu_Y)' V^{-1} (z - \mu_Y) \tag{3.2}$$

y, desarrollando, se obtiene

$$z' V^{-1} z - 2\mu_X' V^{-1} z + \mu_X' V^{-1} \mu_X < z' V^{-1} z - 2\mu_Y' V^{-1} z + \mu_Y' V^{-1} \mu_Y,$$

es decir,

$$2\mu_X' V^{-1} z - 2\mu_Y' V^{-1} z > \mu_X' V^{-1} \mu_X - \mu_Y' V^{-1} \mu_Y$$

y, por tanto, las dos primeras condiciones son equivalentes.

Si las poblaciones son normales, la tercera condición es

$$c \exp\left(-\frac{1}{2}(z - \mu_X)' V^{-1} (z - \mu_X)\right) > c \exp\left(-\frac{1}{2}(z - \mu_Y)' V^{-1} (z - \mu_Y)\right),$$

es decir,

$$(z - \mu_X)' V^{-1} (z - \mu_X) < (z - \mu_Y)' V^{-1} (z - \mu_Y)$$

lo que es equivalente a la condición segunda. □

Observación 3.1. *Note que en la demostración anterior la hipótesis de normalidad no es necesaria para demostrar la equivalencia entre las dos primeras condiciones.*

Observación 3.2. *En ocasiones no es conveniente dar la misma importancia a los dos tipos de errores. Así, por ejemplo, podemos usar el criterio utilizado en los contrastes de hipótesis (Neyman-Pearson) que fija un máximo para uno de los errores* $\Pr(e_1) \leq \alpha$ *e intenta reducir la probabilidad del otro error. Usando este criterio sobre la función discriminate de Fisher, cambiaría la constante* K*, que ahora se calcularía a partir de la relación*

$$\Pr(e_1) = \Pr(Z \in R_Y | Z \equiv X) = \Pr(L(X) < K_\alpha) = \alpha,$$

donde $L(X) \equiv N_1((\mu_X - \mu_Y)'V^{-1}\mu_X, \sigma)$ *y*

$$\sigma^2 = \triangle^2(\mu_X, \mu_Y) = (\mu_X - \mu_Y)'V^{-1}(\mu_X - \mu_Y).$$

De esta forma, la probabilidad del otro error valdrá

$$\Pr(e_2) = \Pr(L(Y) > K_\alpha)$$

donde $L(Y) \equiv N_1((\mu_X - \mu_Y)'V^{-1}\mu_Y, \sigma)$.

Observación 3.3. Criterio de mínimo coste (probabilidad de error) *Otras veces se da un coste a cada uno de los posibles errores* $(c_1, c_2 > 0)$ *y, si se conocen las probabilidades "a priori" de pertenencia a cada una de las poblaciones* $q_1 = \Pr(Z \equiv X)$ *y* $q_2 = \Pr(Z \equiv Y)$*, entonces usando el teorema de la probabilidad total, se tiene*

$$\begin{aligned}\Pr(error) &= \Pr(Z \in R_Y \mid Z \equiv X)\Pr(Z \equiv X) + \Pr(Z \in R_X \mid Z \equiv Y)\Pr(Z \equiv Y) \\ &= \Pr(e_1)q_1 + \Pr(e_2)q_2\end{aligned}$$

y el coste esperado asociado para una constante k *será*

$$c(k) = c_1 \Pr(e_1)q_1 + c_2 \Pr(e_2)q_2,$$

donde

$$\Pr(e_1) = \Pr(L(X) < k) = G\left(\frac{k - L(\mu_X)}{\sigma}\right)$$

y

$$\Pr(e_2) = \Pr(L(Y) > k) = 1 - G\left(\frac{k - L(\mu_Y)}{\sigma}\right).$$

Por lo tanto,

$$c(k) = c_1 q_1 G\left(\frac{k - L(\mu_X)}{\sigma}\right) + c_2 q_2 - c_2 q_2 G\left(\frac{k - L(\mu_Y)}{\sigma}\right)$$

y, derivando, tenemos

$$c'(k) = \frac{c_1 q_1}{\sigma} g\left(\frac{k - L(\mu_X)}{\sigma}\right) - \frac{c_2 q_2}{\sigma} g\left(\frac{k - L(\mu_Y)}{\sigma}\right)$$

donde $g(u) = c\exp(-u^2/2)$ *es la función de densidad normal estándar. Igualando a cero, obtenemos*

$$c_1 q_1 g\left(\frac{k - L(\mu_X)}{\sigma}\right) = c_2 q_2 g\left(\frac{k - L(\mu_Y)}{\sigma}\right).$$

Despejando, se obtiene

$$\frac{(k - L(\mu_Y))^2}{\sigma} - \frac{(k - L(\mu_X))^2}{\sigma} = 2\log\left(\frac{c_2 q_2}{c_1 q_1}\right),$$

$$(L(\mu_Y))^2 - (L(\mu_X))^2 - 2k(L(\mu_Y) - L(\mu_X)) = 2\sigma^2 \log\left(\frac{c_2 q_2}{c_1 q_1}\right)$$

y, finalmente,

$$k = \frac{L(\mu_Y) + L(\mu_X)}{2} + \frac{\sigma^2}{L(\mu_X) - L(\mu_Y)} \log\left(\frac{c_2 q_2}{c_1 q_1}\right),$$

donde

$$L(\mu_X) - L(\mu_Y) = a'(\mu_X - \mu_Y) = (\mu_X - \mu_Y)' V^{-1} (\mu_X - \mu_Y) = \sigma^2$$

por lo que, para minimizar el coste esperado, debe tomarse

$$k = a' \frac{\mu_X + \mu_Y}{2} + \log\left(\frac{c_2 q_2}{c_1 q_1}\right).$$

Cuando $c_1 = c_2$ *lo que se hace es minimizar la probabilidad total de error (mala clasificación). Si* $c_1 q_1 = c_2 q_2$*, la constante* k *coincide con la del criterio de Fisher* $K = (L(\mu_X) + L(\mu_Y))/2$*. Así demostramos que este criterio es el que minimiza la suma de las probabilidades de error.*

Observación 3.4. Criterio de máxima probabilidad a posteriori. *Cuando se conozcan las probabilidades "a priori"* q_1 *y* q_2*, también se pueden calcular las probabilidades "a posteriori" (es decir, cuando conocemos sus los valores de* Z*) para un individuo con medidas* z *mediante el Teorema de Bayes como:*

$$\Pr(Z \equiv X \mid Z = z) = \frac{\Pr(Z = z \mid Z \equiv X) \Pr(Z \equiv X)}{\Pr(Z = z)},$$

con

$$\Pr(Z = z) = \Pr(Z = z \mid Z \equiv X) \Pr(Z \equiv X) + \Pr(Z = z \mid Z \equiv Y) \Pr(Z \equiv Y).$$

Si las variables son discretas, podremos calcular esas probabilidades. Si son continuas, reemplazaremos las probabilidades puntuales por las respectivas funciones de densidad obteniendo:

$$\Pr(Z \equiv X \mid Z = z) = \frac{q_1 f_X(z)}{q_1 f_X(z) + q_2 f_Y(z)}$$

y

$$\Pr(Z \equiv Y \mid Z = z) = \frac{q_2 f_Y(z)}{f_X(z) q_1 + q_2 f_Y(z)},$$

clasificándose a un individuo con características z en la población en la que tenga mayor "probabilidad a posteriori" (pero note que esos valores no son probabilidades sino verosimilitudes ponderadas para que sumen 1). En cualquier caso, esos valores nos indicarán la fiabilidad de la clasificación de un individuo z con esas medidas.

En la práctica, las probabilidades "a priori" q_1 y q_2 suelen ser desconocidas por lo que se tendrán que estimar (si es posible) o suponer iguales (si es razonable). Obviamente, si $q_1 = q_2$, entonces el criterio coincide con el de máxima verosimilitud. En general, se puede probar el resultado siguiente.

Proposición 3.1. *Los criterios de mínima probabilidad de error y de máxima probabilidad a posteriori son equivalentes.*

Proof. Un individuo con medidas z se clasifica en X usando el criterio de máxima probabilidad a posteriori si y solo si

$$q_1 f_X(z) > q_2 f_Y(z).$$

Esta condición es equivalente a

$$\exp(-\Delta^2(z, \mu_X)/2) > \frac{q_2}{q_1} \exp(-\Delta^2(z, \mu_Y)/2),$$

es decir, a

$$\Delta^2(z, \mu_Y) - \Delta^2(z, \mu_X)/2) > 2 \log \frac{q_2}{q_1}.$$

Operando se obtiene que esta condición es equivalente a

$$2(\mu_X - \mu_Y)' V^{-1} z > 2 \log \frac{q_2}{q_1} + \mu_X' V^{-1} \mu_X - \mu_Y' V^{-1} \mu_Y \tag{3.3}$$

donde

$$\mu_X' V^{-1} \mu_X - \mu_Y' V^{-1} \mu_Y = (\mu_X - \mu_Y)' V^{-1} \mu_X + (\mu_X - \mu_Y)' V^{-1} \mu_Y.$$

Finalmente, si $a' = (\mu_X - \mu_Y)' V^{-1}$, la condición (3.3) puede escribirse como

$$a' z > \log \frac{q_2}{q_1} + \frac{a' \mu_X + a' \mu_Y}{2}$$

donde la expresión de la derecha coincide con la constante k obtenida en el criterio de mínimo coste cuando $c_1 = c_2 = 1$ (es decir, en el criterio de mínima probabilidad de error total). □

Observación 3.5. *Cuando el análisis discriminante se aplica a datos biomédicos (tests) siendo X la población de los individuos que tienen una determinada enfermedad e Y los que no la tienen, suelen utilizarse los términos "sensibilidad" o "especificidad" siendo*

$$sensibilidad = 100(1 - \Pr(e_1)) = 100 \Pr(Z \in R_X \mid Z \equiv X) \approx \frac{TP}{TP + FN} 100$$

$$especificidad = 100(1 - \Pr(e_2)) = 100 \Pr(Z \in R_Y \mid Z \equiv Y) \approx \frac{TN}{TN + FP} 100,$$

donde TP=verdaderos positivos, FN=falsos negativos, TN=verdaderos de negativos y FP=falsos positivos. Es decir, la sensibilidad es el porcentaje de individuos con la enfermedad detectados y la especificidad es el porcentaje de individuos sin la enfermedad descartados. También se habla del "valor predictivo" del test como el porcentaje de individuos bien clasificados entre los que han sido clasificados dentro de una de las poblaciones:

$$valor\ predictivo\ positivos = 100 \Pr(Z \equiv X \mid Z \in R_X) \approx 100 \frac{TP}{TP + FP}$$

$$valor\ predictivo\ negativos = 100 \Pr(Z \equiv Y \mid Z \in R_Y) \approx 100 \frac{TN}{TN + FN},$$

es decir, nos dice cuántos de los positivos (negativos) son realmente positivos (negativos). La "eficiencia" de una prueba indica el porcentaje de individuos bien clasificados

$$eficiencia = q_1(1 - \Pr(e_1)) + q_2(1 - \Pr(e_2)) \approx \frac{TP + TN}{total}.$$

3.2.2 Varias poblaciones con la misma matriz de covarianza.

Cuando tengamos más de dos poblaciones con matriz de covarianzas común V, podemos usar el criterio de mínima distancia de Mahalanobis a las medias de los grupos. Para ello, si queremos clasificar a Z entre diversos grupos con variables $X^{(i)} = (Z \mid G = i)$ de medias $\mu^{(i)} = E(X^{(i)})$, para $i = 1, \ldots, m$, calcularemos

$$\begin{aligned} \Delta^2(z, \mu^{(i)}) &= (z - \mu^{(i)})' V^{-1} (z - \mu^{(i)}) \\ &= z' V^{-1} z - 2(\mu^{(i)})' V^{-1} z + (\mu^{(i)})' V^{-1} \mu^{(i)}. \end{aligned}$$

Como la parte cuadrática $z' V^{-1} z$ es común, podemos quedarnos solo con la parte lineal (en realidad, su opuesta) dada por

$$L_i(z) = (\mu^{(i)})' V^{-1} z - (\mu^{(i)})' V^{-1} \mu^{(i)}/2$$

conocida como **función discriminante lineal** (FDL), clasificándose un individuo con características z en el grupo en el que tenga un valor máximo dicha función discriminante. Como consecuencia se obtiene el teorema siguiente.

Teorema 3.3. *Si $X^{(i)}$ tienen medias $\mu^{(i)} = E(X^{(i)})$ y matriz de covarianzas común V para $i = 1, \ldots, m$, entonces equivalen:*
1) $L_i(z) \geq L_j(z)$ para todo j.
2) $\Delta^2(z, \mu^{(i)}) \leq \Delta^2(z, \mu^{(j)})$ para todo j.

Corolario 3.1. *Si solo hay dos grupos, este criterio de clasificación es equivalente a usar la función discriminante de Fisher.*

La demostración del corolario es inmediata ya que demostramos en la sección anterior que el criterio de mínima distancia de Mahalanobis era equivalente a usar la función discriminante de Fisher.

En este caso también podemos aplicar el criterio de máxima verosimilitud clasificando a Z en el grupo para el que $f_i(z)$ sea máxima, donde f_i representa la densidad del grupo i. Bajo normalidad, tenemos el resultado siguiente.

Teorema 3.4. *Si $X^{(j)} \sim N(\mu^{(j)}, V)$ para $j = 1, \ldots, m$, entonces equivalen:*
1) $L_i(z) \geq L_j(z)$ para todo j.
2) $\Delta^2(z, \mu^{(i)}) \leq \Delta^2(z, \mu^{(j)})$ para todo j.
3) $f_i(z) \geq f_j(z)$ para todo j.

La demostración es inmediata ya que la densidad normal multivariante del grupo i vale

$$f_i(z) = \frac{1}{\sqrt{\mid V \mid (2\pi)^k}} \exp\{-\frac{1}{2}(z - \mu^{(i)})' V^{-1} (z - \mu^{(i)})\}$$

y será máxima cuando la distancia de Mahalanobis al cuadrado

$$\Delta^2(z, \mu^{(i)}) = (x - \mu^{(i)})' V^{-1} (x - \mu^{(i)})$$

sea mínima.

Nótese que esto no será, en general, cierto si las poblaciones no son normales o si tienen distintas matrices de covarianzas. Como consecuencia inmediata, se obtiene el resultado siguiente.

Proposición 3.2. *Si todas las poblaciones son normales con matriz de covarianzas común V, entonces los criterios de clasificación de máxima verosimilitud y de mínima distancia de Mahalanobis son equivalentes a aplicar el criterio de discriminación de Fisher paso a paso tomando las poblaciones de dos en dos.*

De esta forma, podríamos estudiar primero si z se clasifica en la población 1 o en la 2. En el segundo paso discriminaríamos entre la 3 y la ganadora del

primer paso y así, sucesivamente. Sin embargo, este método no se puede aplicar en la práctica ya que al discriminar entre las poblaciones 1 y 2 solo se utilizarían los individuos de estas poblaciones para estimar V (ver siguiente sección).

El método de proyección de Fisher puede generalizarse para más de dos grupos obteniéndose las denominadas **proyecciones canónicas** que permiten separar (discriminar) mejor a los grupos i y j. La idea se basa en la representación de la función discriminante de Fisher siguiente:

$$L(Z) = (\mu_X - \mu_Y)'V^{-1}Z \tag{3.4}$$
$$= (\mu_X - \mu_Y)'V^{-1/2}V^{-1/2}Z \tag{3.5}$$
$$= (V^{-1/2}\mu_X - V^{-1/2}\mu_Y)'V^{-1/2}Z, \tag{3.6}$$

donde $Z^* = V^{-1/2}Z$ es una variable con $Cov(V^{-1/2}Z) = V^{-1/2}VV^{-1/2} = I$, $V^{-1/2} = T'D^{-1/2}T$ y donde T es la matriz ortonormal que diagonaliza a V ($T'VT = D$, $T'T = TT' = I$). Es decir, el criterio basado en la función discriminante de Fisher equivale a, primero estandarizar las variables haciendo que tengan covarianza I y medias $\mu_X^* = V^{-1/2}\mu_X'$ y $\mu_Y^* = V^{-1/2}\mu_Y'$, y luego proyectarlas en la dirección de la recta que une ambas medias $\mu_X^* - \mu_Y^*$. Esto corresponde con la solución óptima para la distancia de Mahalanobis cuando $V = I$, es decir, cuando usamos la distancia Euclídea. Así, una vez proyectados el punto z y las medias sobre el espacio canónico (Euclídeo), z se clasificará en el grupo cuya proyección de la media esté más cercana a su proyección en la distancia euclídea (unimos las medias y trazamos la mediatriz para separar los grupos). Este criterio también es equivalente al criterio de mínima distancia de Mahalanobis ya que las distancias euclídeas de la proyección de z a los grupos verificarán:

$$d_i^2(z) = d^2(V^{-1/2}z, V^{-1/2}\mu^{(i)}) = (z - \mu^{(i)})'V^{-1/2}V^{-1/2}(z - \mu^{(i)}) = \Delta^2(z, \mu^{(i)})$$

para todo i.

Si solo hay dos grupos, podemos representar los puntos proyectados usando la función discriminante de Fisher, es decir, las proyecciones de los puntos transformados $z^* = V^{-1/2}z$ sobre la recta que une las dos medias transformadas $\mu_X^* = V^{-1/2}\mu_X$ y $\mu_Y^* = V^{-1/2}\mu_Y$. Si hay tres grupos, podemos representar las proyecciones de los puntos transformados $z^* = V^{-1/2}z$ sobre el plano que forman las tres medias transformadas $\mu_X^* = V^{-1/2}\mu_X$, $\mu_Y^* = V^{-1/2}\mu_Y$ y $\mu_Z^* = V^{-1/2}\mu_Z$. Si estos puntos forman un triángulo, sus mediatrices y su circuncentro determinarán las regiones de clasificación (ya que, en este espacio, podemos usar la distancia Euclídea). Si los puntos están alineados, las regiones de clasificación vendrán dadas por las dos mediatrices obtenidas con el punto central y los dos extremos. En este último caso, en realidad bastaría con proyectar sobre la recta formada por estos puntos. Se procede de forma análoga si hay más

grupos. Finalmente mencionar que la matriz $V^{-1/2}$ se puede reemplazar por cualquier matriz U' no singular (invertible) tal que $UU' = V^{-1}$ ya que entonces $V = (UU')^{-1} = (U')^{-1}U^{-1}$ y

$$Cov(U'Z) = U'VU = U'(U')^{-1}U^{-1}U = I.$$

Ejemplo 3.2. *Supongamos que tenemos que decidir si un individuo con medidas* $z = (x, y)' = (1, 0.9)'$ *se clasifica en una población normal bivariante de media* $(0, 0)'$, *en una de media* $(1, 2)'$ *o en una de media* $(-1/2, 1)$ *siendo la matriz de covarianzas común*

$$V = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}.$$

Entonces las funciones discriminantes lineales (LDF) para $z = (x, y)'$ *serán:*

$$L_1(x, y) = (\mu^{(1)})'V^{-1}z - (\mu^{(1)})'V^{-1}\mu^{(1)}/2 = 0,$$

$$\begin{aligned} L_2(x, y) &= (\mu^{(2)})'V^{-1}z - (\mu^{(2)})'V^{-1}\mu^{(2)}/2 \\ &= (1, 2)\begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}^{-1}\begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2}(1, 2)\begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ 2 \end{pmatrix} \\ &= (1, 2)\begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \frac{1}{2}(1, 2)\begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \end{pmatrix} \\ &= 2y - 2 \end{aligned}$$

y

$$\begin{aligned} L_3(x, y) &= (\mu^{(3)})'V^{-1}z - \frac{1}{2}(\mu^{(3)})'V^{-1}\mu^{(3)} \\ &= (-1/2, 1)\begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \\ &\quad - \frac{1}{2}(-1/2, 1)\begin{pmatrix} 4/3 & -2/3 \\ -2/3 & 4/3 \end{pmatrix}\begin{pmatrix} -1/2 \\ 1 \end{pmatrix} \\ &= -\frac{4}{3}x + \frac{5}{3}y - \frac{7}{6}. \end{aligned}$$

En particular, para $z = (1, 0.9)'$ *obtenemos*

$$\begin{aligned} L_1(1, 0.9) &= 0 \\ L_2(1, 0.9) &= 2(0.9) - 2 = -0.2 \\ L_3(1, 0.9) &= -\frac{4}{3} + \frac{5}{3}0.9 - \frac{7}{6} = -1, \end{aligned}$$

por lo que z se clasificará en la primera población (donde L_i es máxima).

Para calcular las regiones de clasificación para los grupos 1 y 2 haremos $L_1 = L_2$, es decir,

$$0 = 2y - 2$$

obteniendo $y = 1$ (como ya vimos anteriormente usando la función discriminante de Fisher). Para los grupos 1 y 3, obtenemos

$$0 = -\frac{4}{3}x + \frac{5}{3}y - \frac{7}{6},$$

es decir, $y = \frac{4}{5}x + \frac{7}{10}$ y para los grupos 2 y 3, obtenemos

$$2y - 2 = -\frac{4}{3}x + \frac{5}{3}y - \frac{7}{6},$$

es decir, $y = -4x + \frac{5}{2}$. Las regiones de clasificación se pueden ver en la Figura 3.6. Las tres rectas se cortan en el punto $(3/8, 1)$. Para representarlas en R haremos:

```
curve(4*x/5+7/10,-3,5,ylab='y',axes=TRUE)
curve(1+x-x,add=TRUE)
curve(-4*x+5/2,add=TRUE)
text(0,0,labels='mu1')
text(1,2,labels='mu2')
text(-0.5,1,labels='mu3')
text(1,0.9,labels='z')
```

Para calcular las proyecciones canónicas $z^* = V^{-1/2}z$ al espacio Euclídeo, debemos calcular la matriz ortogonal T tal que $T'VT = D$, donde D es diagonal con lo que

$$V^{-1/2} = TD^{-1/2}T'.$$

Las matrices T y D se calcularon en el Ejemplo 2.4, obteniéndose

$$T = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

y

$$D = \begin{pmatrix} 3/2 & 0 \\ 0 & 1/2 \end{pmatrix},$$

con lo que

$$\begin{aligned} V^{-1/2} &= TD^{-1/2}T' \\ &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} \sqrt{2/3} & 0 \\ 0 & \sqrt{2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{6}\sqrt{6} + \frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{6} - \frac{1}{2}\sqrt{2} \\ \frac{1}{6}\sqrt{6} - \frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{6} + \frac{1}{2}\sqrt{2} \end{pmatrix}. \end{aligned}$$

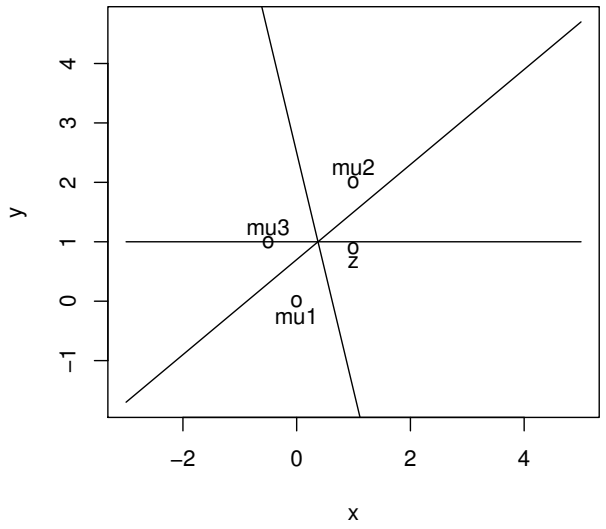


Figure 3.6: Regiones de clasificación para las poblaciones del Ejemplo 3.2.

De esta forma, se tiene:

$$V^{-1/2}z = \begin{pmatrix} \frac{1}{6}\sqrt{6}+\frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{6}-\frac{1}{2}\sqrt{2} \\ \frac{1}{6}\sqrt{6}-\frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{6}+\frac{1}{2}\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 \\ 0.9 \end{pmatrix} = \begin{pmatrix} 0.846382 \\ 0.704961 \end{pmatrix}$$

$$V^{-1/2}\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$V^{-1/2}\mu_2 = \begin{pmatrix} \frac{1}{6}\sqrt{6}+\frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{6}-\frac{1}{2}\sqrt{2} \\ \frac{1}{6}\sqrt{6}-\frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{6}+\frac{1}{2}\sqrt{2} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0.517638 \\ 1.938516 \end{pmatrix}$$

$$V^{-1/2}\mu_3 = \begin{pmatrix} \frac{1}{6}\sqrt{6}+\frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{6}-\frac{1}{2}\sqrt{2} \\ \frac{1}{6}\sqrt{6}-\frac{1}{2}\sqrt{2} & \frac{1}{6}\sqrt{6}+\frac{1}{2}\sqrt{2} \end{pmatrix} \begin{pmatrix} -1/2 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.856536 \\ 1.264784 \end{pmatrix}.$$

Por último, calculamos las distancias euclídeas entre la proyección de z y las proyecciones de las medias obteniendo

$$d_1 = \sqrt{0.846382^2 + 0.704961^2} \approx 1.101514$$
$$d_2 = \sqrt{(0.846382 - 0.517638)^2 + (0.704961 - 1.938516)^2} \approx 1.276609$$
$$d_3 = \sqrt{(0.846382 + 0.856536)^2 + (0.704961 - 1.264784)^2} \approx 1.792577,$$

con lo que (como ya sabíamos) z se clasifica (por poco) en el grupo 1

3.2.3 Varias poblaciones con distintas matrices de covarianza.

Los criterios de clasificación por máxima verosimilitud o por mínima distancia de Mahalanobis a las medias de los grupos pueden utilizarse aunque las poblaciones no tengan la misma matriz de covarianzas. De hecho, tampoco es necesario que las poblaciones sean normales, pudiéndose aplicar incluso a poblaciones de tipo discreto (siempre que se conozcan las densidades o las funciones puntuales de probabilidad). Cuando las poblaciones sean normales suele hablarse de **Análisis Discriminante Cuadrático** (ADC o QDA) ya que las funciones que determinan las regiones de clasificación son polinomios de grado 2. Sin embargo, en este caso, las funciones discriminantes para mínima distancia o máxima verosimilitud no coinciden. El criterio de máxima verosimilitud buscaría el máximo de

$$f_i(z) = \frac{1}{\sqrt{\mid V_i \mid (2\pi)^k}} \exp\left(-\frac{1}{2}(z-\mu^{(i)})' V_i^{-1}(z-\mu^{(i)})\right)$$

o, equivalentemente, el mínimo de

$$Q_i(z) = c - 2\log f_i(z) = (z-\mu^{(i)})' V_i^{-1}(z-\mu^{(i)}) + \log \mid V_i \mid,$$

conocida como **función discriminante cuadrática (QDF)** para $i = 1, \ldots, m$. Un individuo con medidas z se clasificará en el grupo donde la función discriminante cuadrática sea mínima (máxima verosimilitud).

Sin embargo, el criterio basado en la distancia de Mahalanobis, usará la función discriminante cuadrática

$$Q_i^*(z) = \Delta^2(z, \mu^{(i)}) = (z-\mu^{(i)})' V_i^{-1}(z-\mu^{(i)}).$$

Por lo tanto, los resultados pueden ser diferentes (cuando los determinantes de las matrices de covarianzas sean diferentes).

En general, las funciones discriminantes cuadráticas son muy "sensibles" cuando las poblaciones no son normales, por lo que no es muy recomendable su uso en este caso, siendo preferible usar funciones discriminantes lineales. Como veremos posteriormente, en la práctica podemos usar las técnicas de *validación cruzada* para elegir el mejor método posible con los datos disponibles.

Veamos un ejemplo.

Ejemplo 3.3. *Sean dos poblaciones normales bidimensionales con medias* $\mu_1 = (2,0)'$ *y* $\mu_2 = (0,0)'$ *y matrices de covarianzas*

$$V_1 = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$

y

$$V_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

Se pide: Calcular las funciones discriminates cuadráticas, dar el criterio de clasificación, dibujar las regiones de clasificación en R^2 y clasificar a $z = (1,1)'$.

Como

$$V_1^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix},$$

la primera QDF es

$$\begin{aligned} Q_1(x,y) &= \begin{pmatrix} x-2 & y \end{pmatrix} V_1^{-1} \begin{pmatrix} x-2 \\ y \end{pmatrix} \\ &= (x-2+y)(x-2) + (x-2+2y)\,y \\ &= x^2 - 4x + 4 + 2yx - 4y + 2y^2. \end{aligned}$$

Análogamente, para el segundo grupo tenemos:

$$\begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}$$

con lo que la segunda QDF será

$$Q_2 = \begin{pmatrix} x & y \end{pmatrix} V_2^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = 0.5x^2 + 2y^2,$$

es decir,

$$Q_2(x,y) = 0.5x^2 + 2y^2.$$

Entonces $z = (x,y)$ se clasifica en 1 si y solo si $Q_1 < Q_2$, es decir, si

$$(2x-4)y < 4x - 4 - x^2/2.$$

Como:

$$Q_1(1,1) = 1$$

y

$$Q_2(1,1) = 5/2,$$

se clasifica en 1.

Para calcular las regiones de clasificación notamos que, en la ecuación anterior, tenemos tres casos:

1) Si $x > 2$, $Q_1 < Q_2$, si y solo si, $y < (4x - 4 - x^2/2)/(2x-4)$.

2) Si $x < 2$, $Q_1 < Q_2$, si y solo si, $y < (4x - 4 - x^2/2)/(2x-4)$.

3) Si $x = 2$, $0 < 8 - 4 - 4/2 = 2$, por lo que $Q_1 < Q_2$ y se clasifica en 1.

Las regiones de clasificación se pueden ver en la Figura 3.7.

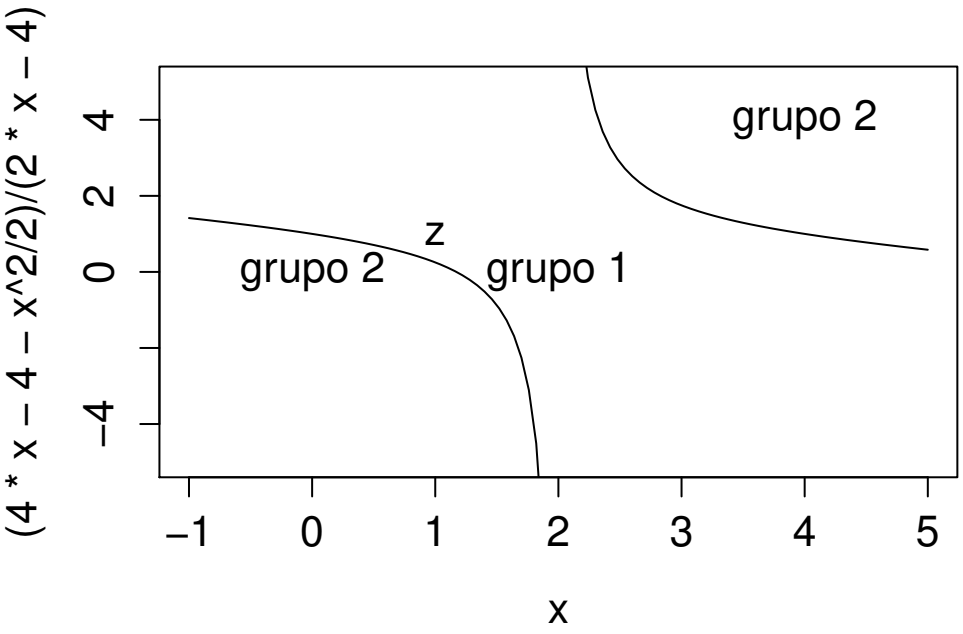


Figure 3.7: Regiones de clasificación Ejemplo 3.3 (1 centro, 2 laterales).

3.3 Clasificación a partir de una muestra.

En la práctica, los valores de las medias y las matrices de covarianzas teóricos usados en los criterios de clasificación dados en la sección anterior serán desconocidos, por lo que tendrán que ser estimados. Estas estimaciones dependerán de las hipótesis de partida (normalidad, igualdad de matrices de covarianzas, etc...), hipótesis que en muchos casos deberán ser corroboradas mediante algún procedimiento cuando no se esté muy seguro de su validez.

En general, si estudiamos k variables numéricas $(Z_1, \ldots, Z_k)$ en m distintas poblaciones indicadas por una variable discreta G (grupo) para una muestra de n individuos, tendremos una tabla de datos de la forma

	Z_1	...	Z_k	G
ω_1	$z_{1,1}$	...	$z_{1,k}$	g_1
...	...	...	...	...
ω_n	$z_{n,1}$	...	$z_{n,k}$	g_n

donde $g_i \in \{1, \ldots, m\}$ para todo i.

Para cada valor de $G = j$, esta tabla proporcionará una m.a.s. de la variable aleatoria k dimensional $Z = (Z_1, ..., Z_k)'$ condicionada por $G = j$ que, en muchas ocasiones, podremos suponer normal. Es decir, en realidad tendremos m muestras de m poblaciones normales k dimensionales $N(\mu_j, V_j)$.

Así, en la práctica, tendremos m medias muestrales teóricas y m matrices de covarianzas desconocidas, por lo que tendremos que estimarlas. Para dichas estimaciones usaremos:

$$\widehat{\mu}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n} \omega_j 1(G(\omega_i) = j)$$

$$\widehat{V}_j = \frac{1}{n_j - 1} \sum_{i=1}^{n} 1(G(\omega_i) = j)(\omega_i - \widehat{\mu}^{(j)})(\omega_i - \widehat{\mu}^{(j)})'$$

$$\omega_i = (Z_{1,i}, ..., Z_{k,i})'$$

$$n_j = \sum_{i=1}^{n} 1(G(\omega_i) = j)$$

$$n = n_1 + \cdots + n_m,$$

donde $1(G(\omega_i) = j)$ indica si el individuo i pertenece (1) o no (0) a la población j-ésima y, por lo tanto, n_j es el número de individuos de la muestra pertenecientes a la población j-ésima. Note que necesitamos $n_j > 1$ para todo j. Si $(Z|G = j)$ es normal, $\widehat{V}_j$ es insesgado para V_j, teniendo $(n_j - 1)\widehat{V}_j$ una distribución (en el muestreo) Wishart $W_k(n_j - 1, V_j)$.

Si suponemos que las matrices de covarianzas teóricas son todas iguales $(V_1 = \cdots = V_m = V)$, entonces la matriz de covarianzas común V se aproximará

mediante la matriz de covarianzas ponderada (pooled)

$$\widehat{V} = \frac{1}{n-m}\sum_{j=1}^{m}(n_j - 1)\widehat{V}_j$$

que será un estimador insesgado para V.

A partir de estos estimadores se pueden obtener estimaciones de las distintas funciones discriminantes y con ellas, obtener clasificaciones (empíricas) para nuevos individuos. Como los estimadores se aproximan a los verdaderos valores de los parámetros, las clasificaciones "se parecerán" (si n es grande) a las que se obtendrían usando los verdaderos parámetros. Por ejemplo, para el caso de dos poblaciones con la misma matriz de covarianzas, la función discriminante de Fisher se estimará mediante

$$\widehat{D} = \widehat{L}(Z) = \widehat{a}'Z = (\widehat{\mu}_X - \widehat{\mu}_Y)'\widehat{V}^{-1}Z.$$

De esta forma, la probabilidad del error tipo 1 se estimará mediante

$$\Pr(e_1) = \Pr\left(U < -\frac{1}{2}\widehat{\triangle}\right),$$

donde $U \equiv N_1(0,1)$ y

$$\widehat{\triangle} = \sqrt{(\widehat{\mu}_X - \widehat{\mu}_Y)'\widehat{V}^{-1}(\widehat{\mu}_X - \widehat{\mu}_Y)}$$

es la distancia de Mahalanobis muestral entre la medias (muestrales) de los grupos.

Bajo la hipótesis de normalidad, estos estimadores son asintóticamente insesgados. En Srivastava y Carter (1983, pag. 238) pueden verse otros estimadores basados en las distribuciones obtenidas para los distintos errores a partir de la hipótesis de normalidad.

Se procederá de forma similar en los otros casos. Por ejemplo, las Funciones Discriminantes Lineales (FDL) muestrales serán

$$\widehat{L}_i(z) = (\widehat{\mu}^{(i)})'\widehat{V}^{-1}z - (\widehat{\mu}^{(i)})'\widehat{V}^{-1}\widehat{\mu}^{(i)}/2,$$

clasificándose z en $G_i : \widehat{L}_i(z) \geq \widehat{L}_j(z)$ para todo j. Análogamente, las proyecciones canónicas muestrales serán $Z^* = \widehat{V}^{-1/2}Z$ y las funciones discriminantes cuadráticas (FDC o QDF) muestrales serán

$$\widehat{Q}_i(z) = c - 2\log\widehat{f}_i(z) = (z - \widehat{\mu}^{(i)})'\widehat{V}_i^{-1}(z - \widehat{\mu}^{(i)}) + \log \mid \widehat{V}_i \mid,$$

clasificándose z en $G_i : \widehat{Q}_i(z) \leq \widehat{Q}_j(z)$ para todo j.

Veamos otra forma de aproximar los errores de clasificación que no precisa de la hipótesis de normalidad.

3.3.1 Validación cruzada

Una forma de estimar las probabilidades de los posibles errores consiste en clasificar a los individuos de la muestra (de los que se conoce su verdadero grupo) usando las funciones discriminantes obtenidas y hacer un recuento de los individuos mal (o bien) clasificados según el grupo al que pertenecen y/o el grupo en el que son clasificados por error.

Cuando se clasifica a los individuos de los que se conoce su grupo se está usando la propia información que ellos han proporcionado a la función discriminante. Esto puede influir en las proporciones de individuos bien clasificados ya que si en la muestra hay un individuo con las mismas características que las del que se quiere clasificar, es bastante probable que coincidan los grupos. Para evitar esto suele utilizarse la técnica denominada **validación cruzada** (CV=cross validation, también llamada leave-one-out o jackknife) que deja fuera del análisis (se tacha) al individuo que se quiere clasificar.

Es decir, para aplicar esta técnica a un individuo deberemos recalcular las funciones discriminantes excluyéndolo de la muestra y aplicárselas, observando si se clasifica o no en su verdadero grupo. Repitiendo esto con cada individuo (o con un número suficiente de ellos) obtendremos estimaciones de las probabilidades de clasificación errónea (o correcta).

Lógicamente este proceso conlleva numerosos cálculos por lo que solo se puede realizar ayudándonos de un ordenador e incluso, si la muestra es muy grande, reduciendo el número de individuos a los que se le aplica (o haciendo grupos). Los errores de estimación de las distintas proporciones (probabilidades) son fáciles de calcular ya que se trata de variables Binomiales y, por lo tanto, incluso podemos calcular los intervalos de confianza para las distintas proporciones. La ventaja de usar validación cruzada es que al ser una prueba empírica, no es necesaria ninguna comprobación estadística (no se necesitan hipótesis iniciales).

También se puede utilizar esta técnica para ver qué variables son las que más influyen a la hora de clasificar correctamente a los individuos. Para ello podemos eliminar variables y comprobar cómo afecta esta eliminación al porcentaje de correctos. También se puede realizar una análisis discriminante lineal con las variables estandarizadas $(X_i - \mu_i)/\sigma_i$ (sustituyendo las medias y las varianzas por sus estimaciones) y, en este caso, la variable que más influya será aquella que tenga un coeficiente mayor en valor absoluto en la función discriminante lineal de Fisher. La estandarización no influirá en la clasificación (se obtienen los mismos resultados) pero sí nos permitirá el poder comparar los coeficientes al tener las variables el mismo rango de variación. Si no estandarizamos, los coeficientes dependerán de las unidades usadas en cada variable.

3.4 Ejemplos

3.4.1 Ejemplo con dos grupos

En el primer ejemplo consideramos el objeto d del fichero `escarabajos.rda` (Aula virtual) que contiene una muestra de 40 escarabajos de dos especies diferentes (Haltica oleracea y Haltica carduorum) a los que se les han medido 4 variables. Para leer este archivo debemos teclear

```
load('f:/escarabajos.rda')
```

indicando la ruta completa en donde se encuentra el archivo. Para ver los datos basta teclear `d` (o `View(d)`). Las variables son: distancia desde el tórax al surco transversal X_1 (micras), longitud X_2 (0.01mm.), longitud de la base de las antenas secundarias X_3 y terciarias X_4 (en micras). La variable código (X_6) indica la especie a la que pertenece cada individuo (HO=1, HC=2). Puede observarse que hay un individuo (40) del que se desconoce la especie lo que en R se escribe como `NA`.

Podemos comenzar estudiando las variables por separado. Si queremos ver solo los datos de la variable `surco`, haremos: `d$surco` o `d[,1]`. Por ejemplo, para estudiar esta variable podemos comenzar calculando sus estadísticos básicos (medias, cuartiles y valores extremos) en cada grupo haciendo:

```
tapply(d$surco,d$especie,summary)
```

De esta forma observamos que la media de la variable `surco` es más grande en la especie HO (194.5) que en la HC (179.6) y que su valor en el escarabajo 40 (182.2) está más cerca de la media de la especie HC. También podemos representarla gráficamente tecleando:

```
plot(d$surco,d$codigo)
```

Si queremos que aparezca el escarabajo 40 podemos hacer:

```
text(d$surco[40],1.5,labels='e40')
```

obteniendo el gráfico de la Figura 3.8. En esta gráfica podemos observar que la variable `surco` parece un poco mayor en el grupo 1 (HO) pero que no discrimina (separa) bien a los grupos. Con esta variable no es sencillo clasificar al escarabajo 40 pero, si tenemos que elegir un grupo, lo incluiríamos en el grupo 2 (HC) ya que está más cerca de su media. Se obtiene una gráfica similar haciendo `plot(d$surco,d$especie)`. En este caso, R etiqueta los datos por orden alfabético (ASCII) con *=1, HC=2 y HO=3. Estudie las restantes variables.

También se pueden hacer los gráficos caja-bigote por grupos con:

```
boxplot(d$surco~d$especie)
```

(el símbolo ~ se puede escribir pulsando simultáneamente Alt y 126) obteniendo el gráfico de la Figura 3.9. Las cajas contienen al 50% de los datos, los "bigotes" al 25% y los puntos señalan valores atípicos (para una distribución normal). En este gráfico apreciamos que si usáramos solo la variable surco para clasificar, como las cajas no se solapan, más del 75% de los individuos se clasificarían bien.

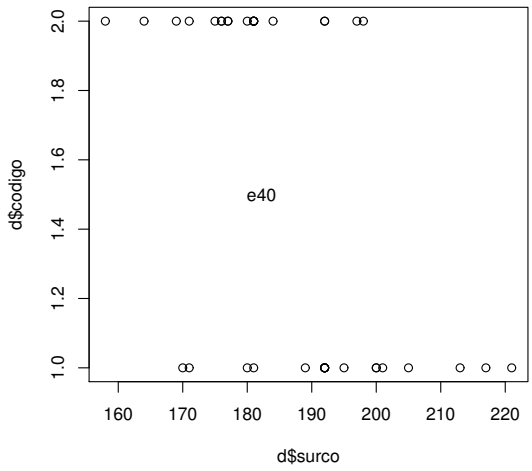


Figure 3.8: Gráfico de la variable `surco` por grupos.

También observamos que el escarabajo 40 estaría en la caja de la especie HC (por poco) pero que no sería un valor atípico en la HO. Estudie las restantes variables.

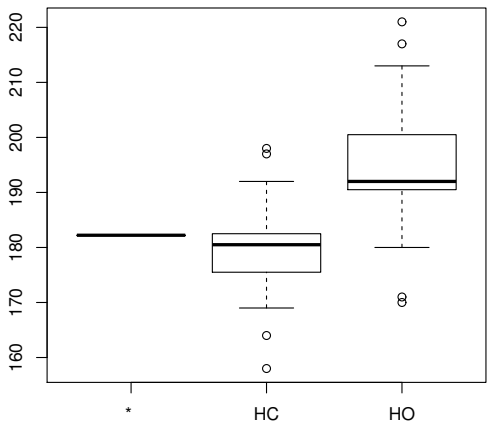


Figure 3.9: Gráficos caja-bigote de la variable `surco` por grupos.

En segundo lugar podemos estudiar las variables por parejas. Por ejemplo, para analizar `surco` y `long`, podemos hacer:

```
plot(d$surco,d$long,pch=as.integer(d$especie))
legend('topright',legend=c('e40','HC','HO'),pch=1:3)
```

obteniendo el gráfico de la Figura 3.10 en el que se observa que, con estas dos variables, los dos grupos están bastante separados, pero que el escarabajo 40 estaría entre ambos grupos por lo que no es sencillo clasificarlo. Estudie las otras dos variables. Se obtiene un gráfico similar haciendo

```
plot(d$surco, d$long)
text(d$surco,d$long,d$especie,cex=0.7,pos=4,col='red')
```

(`cex` indica el tamaño y `pos` la posición de la etiqueta).

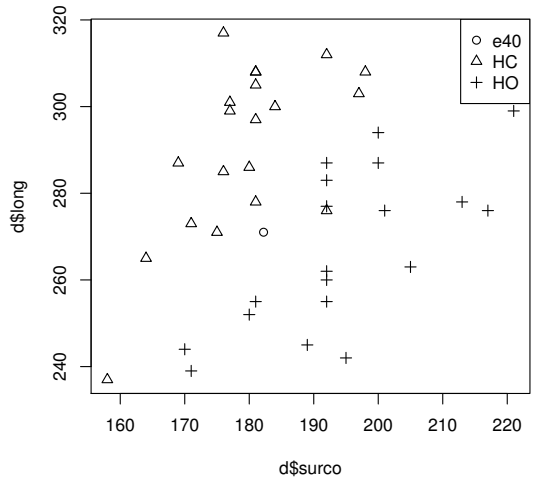


Figure 3.10: Gráfico conjunto de las variables `surco` y `long` por grupos.

Finalmente podemos hacer un gráfico similar al de la Figura 3.10 pero usando las dos primeras componentes principales (ver tema anterior) que contienen información sobre todas las variables. Recordemos que las componentes principales se calculan con:

```
pca<-princomp(d[,1:4],cor=TRUE)
```

y que se pueden representar las dos primeras componentes por grupos haciendo:

```
biplot(pca,pc.biplot=TRUE,xlabs=d$especie)
```

El resultado puede verse en la Figura 3.11. En este gráfico también se aprecia que los grupos se pueden separar bastante bien. Señalar no obstante que las dos primeras componentes principales no son necesariamente las mejores variables para clasificar a estos individuos.

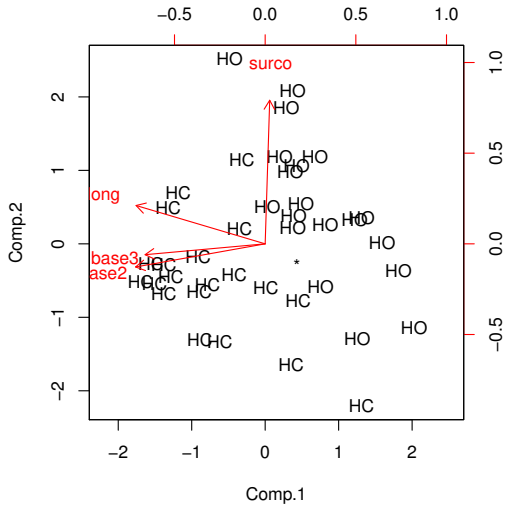


Figure 3.11: Gráfico de las dos primeras componentes principales por grupos.

Comenzaremos realizando un **Análisis Discriminante Lineal** (LDA). Para calcular la función discriminate lineal (FDL) de Fisher para distinguir entre dos grupos debemos suponer que sus matrices de covarianzas (teóricas) son iguales. Entonces, la FDL valdrá $L = L(Z) = \mathbf{a}'Z$ donde $(Z_1, \ldots, Z_k)'$ son las medidas del individuo a clasificar y los coeficientes teóricos se calculan como

$$\mathbf{a}' = \lambda(\mu_X - \mu_Y)'V^{-1}, \tag{3.7}$$

donde λ es un número real cualquiera distinto de cero, V es la matriz de covarianzas común y μ_X y μ_Y son los vectores de medias en cada grupo de las variables usadas para clasificar. En la práctica estas medias teóricas se sustituyen por sus estimaciones $\overline{X}$ e $\overline{Y}$ y V se estima mediante:

$$S = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)S_1 + (n_2 - 1)S_2]$$

siendo n_1 y n_2 los tamaños muestrales de cada grupo y S_1 y S_2 las matrices de cuasicovarianzas muestrales de cada grupo.

Para calcular (estimar) **a** en R debemos cargar primero el "paquete" denominado **MASS**. Una vez cargado, debemos hacer:

```
LDA<-lda(d[1:39,1:4],d[1:39,6],prior=c(0.5,0.5))
```

Tecleando `LDA` comprobamos que las probabilidades de pertenencia a priori asignadas a cada grupo valen 0.5 (si no se especifica nada se computan como

si los individuos fuesen una muestra, es decir, como $19/39 = 0.4871795$ (HO) y $20/39 = 0.5128205$ (HC) en este ejemplo), los vectores de medias de los grupos son:

$$\overline{X} = (194.4737, 267.0526, 137.3684, 185.9474)$$

y

$$\overline{Y} = (179.5500, 290.8000, 157.2000, 209.2500)$$

y que los (unos) coeficientes estimados de la FDL son

$$\mathbf{a} = (-0.09327642, 0.03522706, 0.02875538, 0.03872998).$$

Si queremos guardar estos coeficientes en el objeto `a` haremos:

```
a<-LDA$scaling
```

Para clasificar a un individuo con medidas z calcularemos su proyección $L(z) = \mathbf{a}'z$ y las proyecciones de las medias de los grupos $L(\overline{X})$ y $L(\overline{Y})$, clasificándolo en el grupo que tenga la media más cerca a su proyección. La frontera de las regiones de clasificación vendrá dada por la media de las proyecciones de las medias: $K = (L(\overline{X}) + L(\overline{Y}))/2$. Para calcular L podemos definir la función:

```
L<-function(z) sum(a*z)
```

De esta forma, podemos calcular la proyección de la media $L(\overline{X})$ de la especie HO haciendo:

```
mHO<-L(LDA$means[1,])
```

obteniendo $L(\overline{X}) = 2.419488$. Análogamente, podemos calcular `mHC` obteniendo $L(\overline{Y}) = 6.120841$. De esta forma, haciendo `K<-(mHC+mHO)/2`, obtenemos $K = 4.270164$. Por lo tanto, la regla de decisión óptima según este criterio sería: Si $L(z) > K$, se clasifica como HC (grupo 2) y si no como HO (grupo 1).

Podemos calcular las proyecciones de los 40 escarabajos haciendo:

```
z<-d[,1:4]
D<-1:40
for (i in 1:40) D[i]<-L(z[i,])
```

Tecleando `D` comprobamos que para el escarabajo 1 se obtiene $D[1] = 1.253859$ que, como es menor que $K = 4.270164$, nos conduciría a clasificarlo como del grupo HO (correctamente). Análogamente, para el escarabajo 40, obtenemos $D[40] = 3.968782$ que, como es menor que K, nos conduciría a clasificarlo como del grupo HO (con un margen pequeño). Podemos representar estas "puntuaciones discriminates" haciendo:

```
plot(D,d$codigo)
text(D,d$codigo,cex=0.7,pos=3,col='red')
```

Podemos incluir la puntuación del escarabajo 40 y la constante K en el gráfico haciendo:

```
text(D[40],1.5,labels='*')
text(D[40],1.5,labels='e40',cex=0.7, pos=3,col='red')
text(K,1.5,labels='|')
```

```
text(K,1.5,labels='K',cex=0.7,pos=3,col='red')
```

De esta forma se obtiene el gráfico de la Figura 3.12. En este gráfico se observa que el escarabajo 27 se clasificaría erróneamente y que el 40 se clasificaría en el grupo 1 (HO) pero con un margen pequeño (cerca de K).

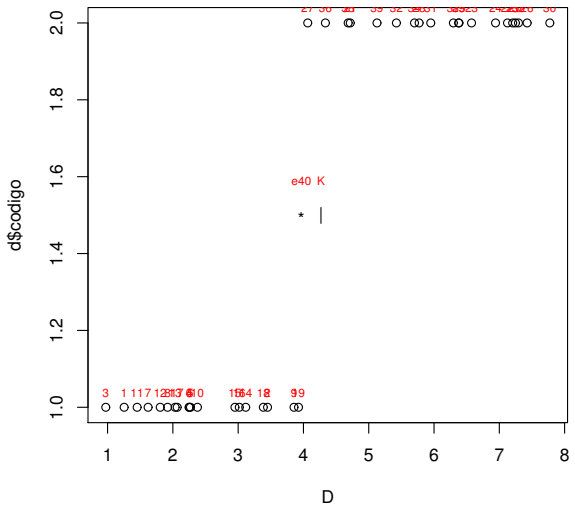


Figure 3.12: Gráfico de las puntuaciones discriminantes.

Otros autores prefieren calcular las puntuaciones como $D - K$ con lo que la regla de decisión dependerá de si las puntuaciones son positivas o negativas. La puntuación $D - K$ se puede obtener de forma automática haciendo:

```
predict(LDA,d[,1:4])->P
```

Las puntuaciones se obtienen haciendo `P` o `P$x`. Compruebe que coinciden con los valores de `D-K`. Estos valores se pueden representar como en la Figura 3.12 o en forma de histograma haciendo:

```
ldahist(P$x,g=d$especie)
```

Haciendo `P$class` podemos ver en qué grupo se clasifican los 40 escarabajos. Tecleando

```
P$class==d[,6]->Resumen
```

podemos ver cuando la clasificación es correcta (para los 39 escarabajos de los que se conoce su grupo). Podemos hacer un recuento de estos resultados con:

```
table(P$class,d[,6])
```

Estos valores se pueden resumir con los valores de la Tabla 3.1. Esta tabla sirve para comprobar si este procedimiento de clasificación es adecuado. En este caso, obtenemos buenos resultados ya que todos los individuos del primer grupo se

clasifican correctamente y solo uno del grupo 2 (el escarabajo 27) se clasifica erróneamente como del grupo 1. Análogamente, comprobamos que todos los individuos clasificados como del grupo 2 se han clasificado correctamente pero que uno clasificado como del grupo 1, en realidad pertenecía al grupo 2 (de nuevo el 27).

Tabla 3.1: Resumen de los resultados de clasificación usando LDA.

Grupo verdadero=	1 (HO)	2 (HC)	Total
Clasificados en el grupo 1	19	1	20
Clasificados en el grupo 2	0	19	19
Total	19	20	39

Finalmente, haciendo:

```
P$posterior
```

podemos ver las "probabilidades" a posteriori (verosimilitudes normalizadas) de pertenencia a cada grupo bajo normalidad dadas por:

$$\Pr(i|z) = \frac{\pi_i f_i(z)}{\pi_1 f_1(z) + \pi_2 f_2(z)},$$

donde π_1 y π_2 son las probabilidades a priori (0.5 en este caso) y f_1 y f_2 son las funciones de densidad normales estimadas de cada grupo. Aquí podemos ver que las probabilidades de pertenencia para el escarabajo 40 valen $\Pr(1|z = e40) = 0.7531572$ y $\Pr(2|z = e40) = 0.2468428$, que nos muestran que para un individuo de estas medidas la clasificación no es muy fiable. Evidentemente, los individuos se clasifican usando LDA en el grupo en el que resultan más verosímiles (ambos métodos son equivalentes).

La función `predict` también se puede usar para clasificar a nuevos individuos. Por ejemplo, para clasificar a un escarabajo con las medidas siguientes $z = (185, 280, 150, 200)$, haremos:

```
z<-c(185,280,150,200)
predict(LDA,z)
```

con lo que z se obtiene que se clasifica en el grupo 2, con una puntuación 0.3965766 y una probabilidad a posteriori de pertenencia al grupo 2 de 0.8127334. Compruebe que la puntuación coincide con $L(z) - K$.

Los valores de la Tabla 3.1 se pueden usar para estimar las proporciones de acierto en cada caso. Por ejemplo, la probabilidad de acierto global estimada es $38/39 = 0.974359$. Estas estimaciones suelen dar valores ligeramente mayores que los reales ya que al clasificar a un individuo, se ha usado la información proporcionada por el propio individuo. Sin embargo, cuando se clasifica a un individuo nuevo (e40), éste no se usa en el procedimiento de clasificación. Para

evitar esto, podemos usar la técnica denominada *validación cruzada* (*cross validation* o CV) que consiste en que, al clasificar a los individuos de los que se conoce su grupo, el individuo a clasificar no se usa en el procedimiento de clasificación (se tacha). Para hacer esto en R debemos teclear:

```
LDACV<-lda(d[1:39,1:4],d[1:39,6],prior=c(0.5,0.5),CV=TRUE)
table(LDACV$class,d[1:39,6])
```

De esta forma, podemos comprobar que hay 3 escarabajos del grupo 2 que se clasifican mal (21, 27 y 36) con validación cruzada y el resumen correcto de clasificación sería el dado en la Tabla 3.2. En ella comprobamos, por ejemplo, que la verdadera (no sesgada) estimación de la probabilidad global de acierto es: $p_{LDA} = (19 + 17)/39 = 0.9230769$ (ligeramente menor que la calculada anteriormente sin CV). Al usar validación cruzada las probabilidades a posteriori de los individuos con grupos conocidos también cambian (ya que no se usan). Por ejemplo, para el escarabajo 21 obtenemos 0.5291374 y 0.4708626, mientras que antes eran 0.1606631 y 0.8393369. La validación cruzada no afecta a la clasificación de los individuos de los que se desconoce el grupo.

Tabla 3.2: Resumen de los resultados de clasificación usando LDA con validación cruzada.

Grupo verdadero	1 (HO)	2 (HC)	Total
Clasificados en el grupo 1	19	3	22
Clasificados en el grupo 2	0	17	17
Total	19	20	39

Tanto las probabilidades de pertenencia, como las puntuaciones (la constante K) y las clasificaciones finales se verán influenciadas por las probabilidades a priori. Por ejemplo, si no indicamos las probabilidades a priori (es decir, asumimos que éstas se calculen a partir de la muestra), para el escarabajo 40 se obtiene una puntuación en $D - K$ de -0.34883551 y probabilidades a posteriori de $\Pr(1|e40) = 0.7434978$ y $\Pr(2|e40) = 0.2565022$, por lo que se sigue clasificando en el grupo 1. Los aciertos con estas probabilidades a priori son los mismos. Sin embargo, compruebe que con las probabilidades a priori 0.2 y 0.8, existe un escarabajo (19) del grupo 1 que se clasifica en el 2 y que el escarabajo 40 se clasifica en el grupo 2. La clasificación será óptima cuando se usen las probabilidades de pertenencia reales en cada grupo (que suelen ser desconocidas).

Cuando las variables usadas para clasificar sean normales (multivariantes) en cada grupo pero sus matrices de covarianzas (teóricas) no sean iguales, el procedimiento óptimo de clasificación será el proporcionado por el **Análisis**

Discriminante Cuadrático (QDA) que consiste en comparar sus funciones de densidad (verosimilitudes o probabilidades a posteriori) estimadas mediante:

$$f_1(z) = c \mid S_1 \mid^{-1/2} \exp\left(-\frac{1}{2}(z-\overline{X})' S_1^{-1}(z-\overline{X})\right)$$
$$f_2(z) = c \mid S_2 \mid^{-1/2} \exp\left(-\frac{1}{2}(z-\overline{Y})' S_2^{-1}(z-\overline{Y})\right).$$

En el LDA estas funciones se estimaban usando la estimación de la matriz de varianzas común S. Ahora note que las matrices de covarianzas de cada grupo se estiman usando solo los datos de ese grupo. Esto es equivalente a comparar las funciones discriminantes cuadráticas:

$$QDF_1(z) = (z-\overline{X})' S_1^{-1}(z-\overline{X}) + \log \mid S_1 \mid \tag{3.8}$$
$$QDF_2(z) = (z-\overline{Y})' S_2^{-1}(z-\overline{Y}) + \log \mid S_2 \mid, \tag{3.9}$$

clasificando a un individuo en donde QDF sea mínima. Note que las funciones QDF son iguales a las distancias de Mahalanobis al cuadrado de cada grupo mas una constante que depende del grupo. Cuando los determinantes sean iguales, el método será equivalente al de distancia de Mahalanobis mínima.

Para realizar un QDA en R con los datos de los escarabajos incluidos en el objeto `d` debemos hacer:

```
QDA<-qda(d[1:39, 1:4], d[1:39, 6],prior=c(0.5,0.5))
```

Tecleando QDA comprobamos que en este procedimiento no aparecen los coeficientes de las QDF. Para obtener los coeficientes que convierten a los datos en esféricos y las constantes debemos teclear:

```
QDA$scaling
QDA$ldet
```

respectivamente. Compruebe que con la segunda opción se obtiene $\log \mid S_1 \mid = 19.41635$. La primera opción nos proporciona matrices triangulares U_i tales que $U_i U_i' = S_i^{-1}$. De esta forma, las funciones discriminantes cuadráticas se pueden calcular como:

$$QDF_1(z) = (U_1'z - U_1'\overline{X})'(U_1'z - U_1'\overline{X}) + \log \mid S_1 \mid \tag{3.10}$$
$$QDF_2(z) = (U_2'z - U_2'\overline{Y})'(U_2'z - U_2'\overline{Y}) + \log \mid S_2 \mid, \tag{3.11}$$

es decir, la transformación $U_i'z$ convierte a los datos del grupo i en esféricos ya que $Cov(U_i'z) = U_i' S_i U_i$ y como $U_i U_i' = S_i^{-1}$, entonces $S_i = (U_i')^{-1} U_i^{-1}$ y

$$Cov(U_i'Z) = U_i' S_i U_i = U_i (U_i')^{-1} U_i^{-1} U_i = I$$

cuando Z pertenece al grupo i.

Para obtener las predicciones basadas en las probabilidades a posteriori podemos hacer:

```
predict(QDA,d[,1:4])->P
```

Tecleando `P` comprobamos que solo hay un escarabajo mal clasificado (el 27) y que el escarabajo 40 se clasifica en el grupo 1 (como en el LDA). En este caso, las probabilidades de pertenencia valen 0.5817418 y 0.4182582 por lo que esta clasificación no es fiable.

De nuevo podemos obtener una tabla resumen de las clasificaciones con:

```
table(P$class,d$codigo)
```

Para que esta tabla sea más realista debemos usar validación cruzada haciendo:

```
QDACV<-qda(d[1:39,1:4],d[1:39,6],prior=c(0.5,0.5),CV=TRUE)
table(QDACV$class,d[1:39,6])
```

obteniendo los resultados de la Tabla 3.3. Los resultados son similares a los obtenidos con el LDA con una probabilidad global de acierto estimada de $p_{QDA} = 35/39 = 0.8974359$.

Tabla 3.3: Resumen de los resultados de clasificación usando QDA con validación cruzada.

Grupo verdadero	1 (HO)	2 (HC)	Total
Clasificados en el grupo 1	17	2	19
Clasificados en el grupo 2	2	18	20
Total	19	20	39

La función `predict` también se puede usar para clasificar a nuevos individuos. Por ejemplo, para clasificar a un escarabajo con medidas

$$z = (185, 280, 150, 200),$$

haremos:

```
z<-c(185,280,150,200)
predict(QDA,z)
```

con lo que se obtiene que se clasifica en el grupo 2, con probabilidad a posteriori de pertenencia al grupo 2 de 0.9636754 con lo que esta clasificación sí es fiable (bajo la hipótesis de normalidad). En este ejemplo, los dos procedimientos dan buenos resultados y proporcionan las mismas clasificaciones para $e40$ y este z.

Finalmente, podemos realizar algunas comprobaciones sobre los modelos usados. En primer lugar, podemos tener la duda de si es mejor aplicar LDA o QDA. El primer método funciona bien si las matrices de covarianzas teóricas son iguales y el segundo si los datos son normales en cada grupo. Se cumplan o no esas hipótesis, el método de validación cruzada nos proporciona estimaciones de las probabilidades de acierto en cada caso y nos permite la comparación de

las técnicas LDA y QDA. También tenemos la opción de usar ambas técnicas y comprobar si los resultados coinciden.

Si queremos estudiar las hipótesis del LDA, la matriz de cuasicovarianzas del primer grupo se puede calcular con:

```
S1<-cov(d[1:19,1:4])
```

También se pueden separar los datos del grupo 1 con:

```
d1<-d[d$especie=='HO',1:4]
```

y así, su matriz de cuasicovarianzas se calcula con `cov(d1)`. Análogamente, se calcula la del segundo grupo obteniéndose:

$$S_1 = \begin{pmatrix} 187.596 & 176.863 & 48.371 & 113.582 \\ 176.863 & 345.386 & 75.980 & 118.781 \\ 48.371 & 75.980 & 66.357 & 16.243 \\ 113.582 & 118.781 & 16.243 & 239.942 \end{pmatrix}$$

y

$$S_2 = \begin{pmatrix} 101.839 & 128.063 & 36.989 & 32.592 \\ 128.063 & 389.011 & 165.358 & 94.368 \\ 36.989 & 165.358 & 167.537 & 66.526 \\ 32.592 & 94.368 & 66.526 & 177.882 \end{pmatrix}.$$

De esta forma, comprobamos que las matrices de covarianzas de los grupos son bastante diferentes.

Para comprobar que las computaciones de R para los coeficientes del LDA dados en (3.7) son correctas podemos calcular la estimación de la matriz de covarianzas común V con $S = \frac{1}{n+m-2}[(n-1)S_1 + (m-1)S_2]$. Haciendo:

```
S<-(18*S1+19*S2)/37
```

obtenemos

$$S = \begin{pmatrix} 143.559 & 151.803 & 42.527 & 71.993 \\ 151.803 & 367.788 & 121.877 & 106.245 \\ 42.527 & 121.877 & 118.314 & 42.064 \\ 71.993 & 106.245 & 42.064 & 208.073 \end{pmatrix}.$$

Su inversa se calcula con:

```
solve(S)->In
```

Las medias de los grupos se calculan con:

```
LDA$means[1,]->m1
LDA$means[2,]->m2
```

(o con `mean(d1)`) y los coeficientes como

```
(m1-m2)%*%In->a
```

(donde $\%*\%$ denota el producto de matrices en R) obteniendo

$$\mathbf{a} = (0.345249, -0.1303878, -0.1064338, -0.1433533).$$

Para comprobar que son proporcionales a los obtenidos por R haremos:

```
LDA$scaling/t(a)
```

donde `t(a)` denota el vector traspuesto de `a`. Note que la constante de proporcionalidad es negativa (-0.2701715) y, por eso, la media del grupo 2 es mayor que la del 1.

Si queremos estudiar qué variables influyen más en los procedimientos de clasificación LDA, como las variables originales pueden tener escalas diferentes (como ocurre en nuestro ejemplo), no podemos comparar los coeficientes obtenidos con ellas. Sin embargo, si estandarizamos las variables originales, como éstas tendrán valores similares, los coeficientes obtenidos con ellas en el LDA sí se podrán usar estudiar la influencia de las variables en la clasificación. Al contrario de lo que ocurría en el PCA, los procedimientos de clasificación LDA y QDA dan el mismo resultado si se usan las variables estandarizadas (no se ven afectados por cambios de escala y/o localización). Para estandarizar los datos haremos:

```
ds<-scale(d[,1:4])
```

y calculando los coeficientes con:

```
lda(ds[1:39,1:4],d[1:39,6],prior=c(0.5,0.5))
```

obtenemos: -1.2937164 (surco), 0.7809833 (long), 0.4182667 (base2) y 0.7084167 (base3). Por lo tanto, la variable que más influye (mejor discrimina) es surco y la que menos base2.

También nos podemos plantear si queremos eliminar alguna variable cuál sería la más adecuada. Para esto podemos usar los procedimientos de validación cruzada y estudiar qué opción proporciona los mejores resultados teniendo claro que la mejor opción es siempre usarlas todas. Por ejemplo, si eliminamos *surco* haciendo:

```
lda(d[1:39,2:4],d[1:39,6],prior=c(0.5,0.5),CV=TRUE)
```

comprobamos que hay 7 escarabajos que se clasifican mal. Eliminado las otras variables comprobamos que las mejores opciones son eliminar la variable *long* o la variable *base*2 (en ambos casos solo hay 2 escarabajos que se clasifiquen mal). Análogamente, podemos estudiar cuál es la mejor pareja de variables (o la variable individual) que mejor discriminan. Se puede aplicar un procedimiento similar en el QDA.

También podemos comprobar cómo se calculan las probabilidades a posteriori. Para ello debemos cargar el "paquete" **mvtnorm** y teclear:

```
dmvnorm(d[40,1:4],m1,S)->f1
dmvnorm(d[40,1:4],m2,S)->f2
f1/(f1+f2)
```

De esta forma se obtiene que la probabilidad a posteriori de pertenencia del escarabajo 40 en el grupo 1 es $\Pr(1|z = e40) = 0.7531572$ en el caso de probabilidades a priori iguales. Para obtener la que se obtiene con las probabilidades

a priori proporcionadas por los grupos debemos hacer:

```
19*f1/(19*f1+20*f2)
```

obteniendo $\Pr(1|e40) = 0.743497$ (como anteriormente). Compruebe usando un procedimiento análogo (pero sustituyendo S por $S1$ y $S2$) las probabilidades a posteriori calculadas en el QDA.

Por último, señalar que para que estas "probabilidades" (verosimilitudes) sean correctas, las variables deben ser normales en cada grupo. Esta hipótesis también se usa en el QDA. Para hacer un test de normalidad multivariante (Shapiro-Wilk) debemos cargar el paquete: **mvnormtest** y hacer:

```
mshapiro.test(t(d[1:19,1:4]))
```

obteniendo un p-valor de 0.2013 por lo que el primer grupo pasaría el test de normalidad. Análogamente, para el segundo se obtiene un p-valor de 0.05769 que nos conduciría a aceptar la normalidad con $\alpha = 0.05$ por muy poco. Esto se puede deber al escarabajo 27 que, como hemos visto durante toda la práctica tiene unas medidas raras para ser del grupo 2. Los datos para este grupo se pueden ver haciendo `plot(d[20:39,1:4])`.

Cuando en un LDA hay más de dos grupos, algunos autores prefieren calcular las funciones discriminantes lineales por grupos dadas por:

$$L_i(z) = z'S^{-1}m_i - m_i'S^{-1}m_i/2,$$

donde S es la matriz de covarianzas ponderada (calculada anteriormente) y m_i son las medias muestrales de los grupos. Para calcularlas en R haremos:

```
solve(S)%*%m1
solve(S)%*%m2
-0.5*t(m1)%*%solve(S)%*%m1
-0.5*t(m2)%*%solve(S)%*%m2
```

obteniendo:

$$L_1(z) = 0.9557217z_1 - 0.0208622z_2 + 0.6842504z_3 + 0.4353125z_4 - 177.6155,$$

$$L_2(z) = 0.6104728z_1 + 0.1095255z_2 + 0.7906842z_3 + 0.5786658z_4 - 193.4209.$$

Los individuos se clasificarán en el grupo con valor máximo de estas funciones. Este método es equivalente al de máxima verosimilitud (probabilidad a posteriori) con matrices de covarianzas iguales por lo que se obtendrán los mismos resultados de clasificación que antes. También es equivalente a usar las funciones discriminantes de Fisher paso a paso. De hecho, éstas se obtienen restando las funciones discriminates de los grupos, es decir: $L_1(z) - L_2(z) = a'z - K$ (aunque en la estimación de V se usan los individuos de todos los grupos). Por ejemplo, para el escarabajo 40 obtenemos:

$$L_1(182.22, 271.01, 140.99, 190.15) = 170.1294,$$

$$L_2(182.22, 271.01, 140.99, 190.15) = 169.0138,$$

por lo que se clasificaría en el grupo 1 (HO) por un pequeño margen.

De forma análoga, en el QDA se pueden calcular las funciones cuadráticas definidas por (3.9). En este caso, los individuos se incluyen en el grupo con el valor mínimo para esas funciones. Esto es equivalente a usar el método se máxima verosimilitud (probabilidad a posteriori) con matrices de covarianzas distintas por lo que se obtendrán las mismas clasificaciones que en la sección 3. Para el escarabajo 40 se obtiene:

$$QDF_1(z) = 22.76789,$$

$$QDF_2(z) = 23.42774,$$

por lo que se clasificaría (de nuevo) en el grupo 1.

De forma similar se pueden calcular las distancias de Mahalanobis (al cuadrado) dadas en el QDA por:

$$D_1^2(z) = (z - \overline{X})' S_1^{-1} (z - \overline{X}),$$
$$D_2^2(z) = (z - \overline{Y})' S_2^{-1} (z - \overline{Y}),$$

obteniendo para el escarabajo 40: $D_1^2(z) = 3.351539$ y $D_2^2(z) = 3.860477$, por lo que se clasificaría en el grupo 1 (en el más cercano). En este caso, los métodos solo son equivalentes si los determinantes de las matrices de covarianzas de los grupos son iguales. Por lo tanto, se pueden obtener resultados diferentes de los obtenidos con las QDF. Estas distancias también se pueden calcular usando las transformaciones proporcionadas por las matrices U_i incluidas en `QDA$scaling`. Por ejemplo, los transformados en el grupo 2 de la media del grupo 2 y el escarabajo 40 son

$$U_2' \overline{Y} = (-17.792105, 4.306052, -6.051622, 9.862908)$$

y

$$U_2' z = (-18.056683, 2.772973, -5.519009, 8.787518),$$

respectivamente, y su distancia Euclídea al cuadrado es 3.860477.

Para calcular estas distancias en el LDA debemos reemplazar S_1 y S_2 por S obteniendo para el escarabajo 40: $D_1^2(z) = 2.801345$ y $D_2^2(z) = 5.03239$ por lo que se clasificaría en el grupo 1 (en el más cercano). En este caso, los métodos son equivalentes por lo que se obtendrán los mismos resultados de antes (con probabilidades a priori iguales). Cuando hay más de dos grupos, `LDA$scaling` proporciona la matriz U tal que $UU' = S^{-1}$, es decir, la transformación $U'z$ es esférica en todos los grupos. Con `predict(LDA)` podemos ver los transformados de los individuos que se pueden representar con `plot`. Si solo hay dos grupos, los transformados esféricos se proyectan sobre la recta formada por los transformados de las dos medias (función de Fisher). Veamos un ejemplo con tres grupos.

3.4.2 Ejemplo con tres grupos

Vamos a aplicar un DA a los datos del fichero `wine.R` (Aula virtual) que se pueden leer en R con:

```
source('F:/Ruta/wine.R')
```

o con

```
wine<-read.table('http://archive.ics.uci.edu
/ml/machine-learning-databases/wine/wine.data',sep=',')
```

Los datos contienen resultados de 13 diferentes análisis químicos en vinos de la misma región de Italia producidos tres cultivos diferentes (indicadas en la primera columna). Fuente:

http://little-book-of-r-for-multivariate-analysis.readthedocs.org.

Como en el ejemplo anterior comenzaremos haciendo un estudio por separado de cada variable. Por ejemplo, para calcular las principales característica de la segunda variable por grupos haremos:

```
tapply(wine$V2, wine$V1, summary)
```

Estos datos se pueden representar con

```
plot(wine$V2,wine$V1)
```

obteniéndose la gráfica de la Figura 3.13. En la gráfica se observa que esta variable discrimina bien a los grupos 1 y 2 pero que los grupos 1 y 3 aparecerían muy mezclados. Estudiando las otras variables observamos que para separar a estos dos últimos grupos podemos usar la variable V8 o V13.

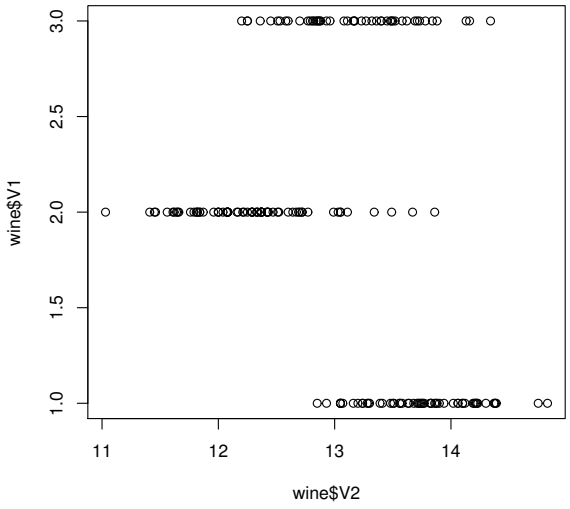


Figure 3.13: Gráfico de la segunda variable del fichero `wine` por grupos.

Otra opción sería realizar los gráficos caja-bigote por grupos. Por ejemplo, los de la Figura 3.14, se obtienen haciendo:

```
boxplot(wine$V2∼wine$V1)
```

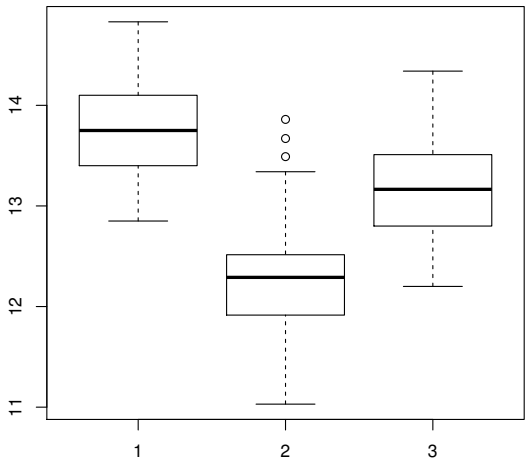


Figure 3.14: Gráficos caja-bigote de la segunda variable del fichero `wine` por grupos.

Para estudiar las variables por parejas en los grupos podemos hacer los gráficos bidimensionales con

```
plot(wine$V2,wine$V8,pch=as.integer(wine$V1))
legend('topright',legend=c('1','2','3'),pch=1:3)
```

obteniendo el gráfico de la Figura 3.15. En este gráfico se aprecia que, con estas dos variables, se pueden separar bastante bien a los tres grupos (aunque hay algunos elementos mezclados). Se obtiene un gráfico similar haciendo:

```
plot(wine$V2,wine$V8)
text(wine$V2,wine$V8,wine$V1,cex=0.7,pos=4,col='red')
```

Otra opción es calcular las componentes principales (ver tema anterior) y representar los puntos de cada grupo. Para realizar el gráfico de las dos primeras componentes haremos:

```
pca<-princomp(wine[,2:14],cor=TRUE)
biplot(pca,pc.biplot=TRUE,xlabs=wine$V1)
```

obteniendo el gráfico de la Figura 3.16.

Tecleando `summary(pca)` podemos ver que estas dos primeras componentes mantienen un 55.4% de la información de las 13 variables numéricas. La primera

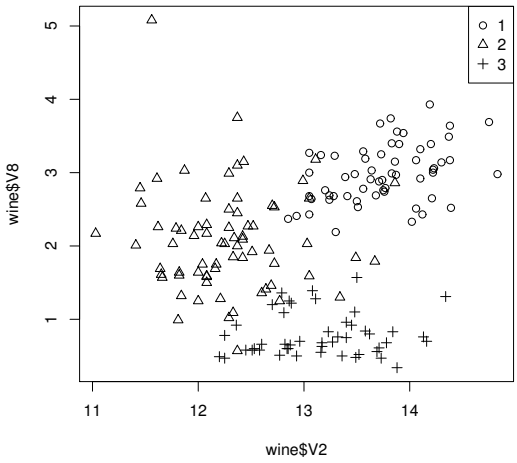


Figure 3.15: Gráficos bidimensionales de las variables V2 y V8 del fichero `wine` por grupos.

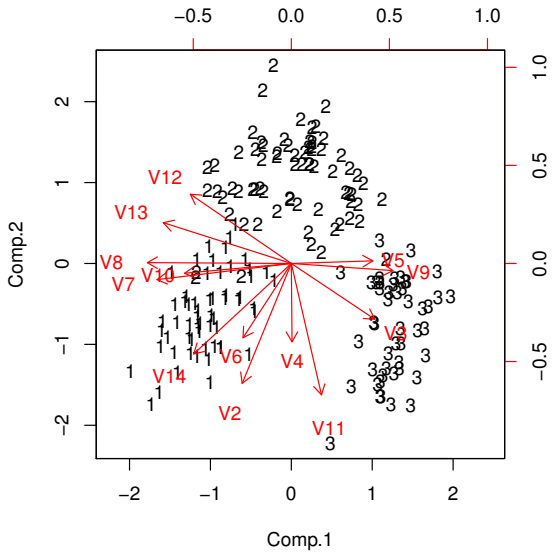


Figure 3.16: Gráficos de las dos primeras componentes principales para los datos del fichero `wine` por grupos.

componente destaca a los vinos que tienen puntuaciones altas en V3, V5 y V9 y bajas en V7, V8, V10 y V13 (derecha del gráfico). Esta primera componente distingue perfectamente a los grupos 3 (con puntuaciones altas en Y_1) y 1 (con puntuaciones bajas en Y_1). La segunda componente destaca a los vinos que tienen puntuaciones altas en V12 y bajas en V2, V4, V6, V11 y V14 (arriba en el gráfico). Esta segunda componente distingue perfectamente a los grupos 2 (con puntuaciones altas en Y_2) y 1 y 3 (con puntuaciones bajas en Y_2). Algunos elementos del grupo 2 aparecen mezclados con los otros grupos. Para detectar estos elementos podemos ver las puntuaciones en Y_2 con `pca$scores[,2]` o representarlas con:

```
plot(pca$scores[,2])
text(pca$scores[,2],cex=0.7,pos=4,col='red')
```

obteniendo el gráfico de la Figura 3.17.

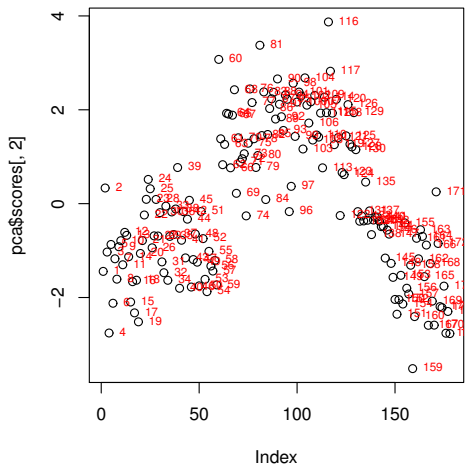


Figure 3.17: Gráfico de la segunda componente principal para los datos del fichero `wine`.

Para hacer una Análisis Discriminante (DA) debemos cargar primero el paquete **MASS** pinchando en el menú superior del programa R en *Paquetes-Cargar paquetes*. Una vez cargado, para hacer un LDA con probabilidades a priori iguales para todos los grupos debemos teclear:

```
LDA<-lda(wine[,2:14],wine[,1],prior=c(1/3,1/3,1/3))
```

Tecleando LDA podemos ver las medias de los grupos en las variables y los coeficientes para las proyecciones canónicas sobre el plano formado por las tres medias. Para guardar estos coeficientes podemos hacer

```
a<-LDA$scaling
```

Podemos calcular las puntuaciones de los vinos con estos coeficientes haciendo

```
predict(LDA,wine[,2:14])->P
```

Tecleando `P` podemos ver los grupos donde se clasificarían los 178 casos, las probabilidades de pertenencia a cada grupo (bajo normalidad) y las puntuaciones canónicas proyectadas. Para representar estas proyecciones por grupos podemos hacer:

```
plot(P$x, pch = as.integer(wine$V1))
legend('bottomright',legend=c('1','2','3'),pch=1:3)
```

obteniendo el gráfico de la Figura 3.18 donde apreciamos que los grupos se pueden separar perfectamente con estas proyecciones (mejor que con las componentes principales).

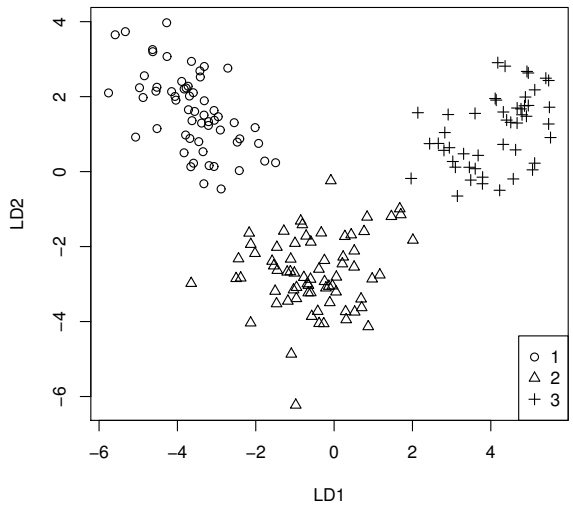


Figure 3.18: Gráfico de las puntuaciones canónicas para los datos del fichero `wine` por grupos.

Podemos comprobar que todas las observaciones de la muestra se clasificarían correctamente haciendo

```
P$class==wine[,1]
```

o

```
table(P$class,wine[,1])
```

Como comentamos en la sección anterior estas estimaciones de las proporciones de acierto son ligeramente superiores a las reales porque al clasificar a un

individuo se están usando sus propias medidas. Para evitar esto y dar estimaciones mejores (no sesgadas) podemos usar las técnicas de validación cruzada (que eliminan al individuo a clasificar) haciendo:

```
LDACV<-lda(wine[,2:14],wine[,1],prior=c(1/3,1/3,1/3),CV=TRUE)
table(LDACV$class,wine[,1])
```

observando que hay dos elementos del grupo 2 que se clasifican erróneamente en los grupos 1 y 3. Los resultados se pueden ver en la Tabla 3.4.

Tabla 3.4: Resumen de los resultados de clasificación usando LDA y validación cruzada.

Grupo verdadero=	1	2	3	Total
Clasificados en el grupo 1	59	1	0	60
Clasificados en el grupo 2	0	69	0	69
Clasificados en el grupo 3	0	1	48	49
Total	59	71	48	178

Para determinar qué individuos se clasifican mal podemos hacer:

```
LDACV$class
```

observando que corresponden a las filas 122 (se clasifica en el grupo 1) y 97 (se clasifica en el grupo 3). En todo caso, las proporciones de acierto (estimadas) son grandes en todos los casos. Por ejemplo, la proporción global de clasificación correcta es $176/178 = 0.988764$, la de clasificación correcta para individuos del grupo 2 es $69/71 = 0.971831$ y la de clasificación correcta para individuos clasificados en el primer grupo es $59/60 = 0.9833333$.

Si queremos predecir la forma de cultivo de un vino con medidas

$$z = (13, 2, 2, 19, 100, 2, 2, 0.3, 1.6, 5, 1, 3, 750)$$

teclearemos:

```
z<-c(13,2,2,19,100,2, 2, 0.3,1.6,5,1,3,750)
predict(LDA,z)
```

obteniendo que z se clasifica en el grupo 2 con una probabilidad de pertenencia a este grupo (bajo normalidad) de 0.996975. También proporciona las coordenadas para incluirlo en la Figura 3.18 haciendo

```
text(-0.8813738, -1.039406, labels ='z',col='red')
```

obteniéndose la Figura 3.19.

Análogamente, para hacer un análisis discriminante cuadrático (QDA) con probabilidades a priori iguales, debemos hacer:

```
QDA<-qda(wine[,2:14],wine[,1],prior=c(1/3,1/3,1/3))
```

Para obtener las predicciones para el punto z haremos:

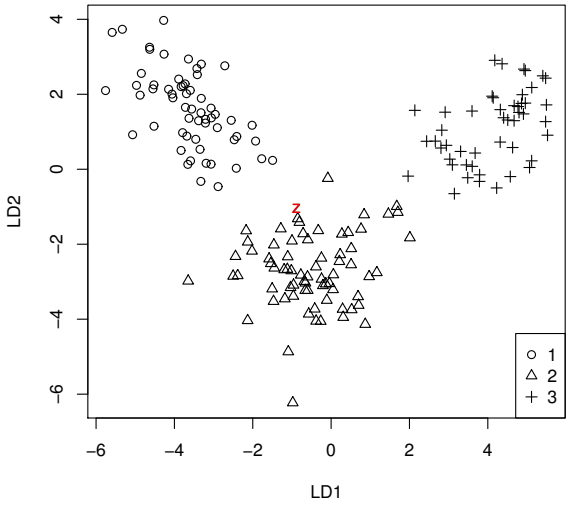


Figure 3.19: Gráfico de las puntuaciones canónicas para los datos del fichero `wine` por grupos incluyendo al proyectado del punto z.

```
predict(QDA,z)
```

obteniendo que, de nuevo, se clasifica en el grupo 2 con una probabilidad de pertenencia estimada (bajo normalidad) de 0.8616003. Para aplicar validación cruzada hacemos:

```
QDA<-qda(wine[,2:14],wine[,1],prior=c(1/3,1/3,1/3),CV=TRUE)
```

Para contar los aciertos y fallos haremos:

```
table(QDA$class,wine$V1)
```

obteniendo que solo un individuo del grupo 2 se clasifica mal en el grupo 1. Para ver cuál es podemos hacer:

```
QDA$class==wine$V1
```

obteniendo que el vino mal clasificado es el 82. De esta forma, concluimos que ambas formas de clasificación dan muy buenos resultados y que el vino a clasificar debe pertenecer al cultivo tipo 2 con una probabilidad de acierto alta.

3.5 Problemas.

1. Dadas tres poblaciones normales bidimensionales con medias $\mu_1 = (1,0)'$, $\mu_2 = (0,1)'$ y $\mu_3 = (0,0)'$ y matrices de covarianzas iguales a

$$V = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

se pide:

(i) Obtener las funciones discriminantes lineales.

(ii) Clasificar a $z = (2,2)'$.

(iii) Dibujar las regiones de clasificación para cada grupo.

2. Dadas tres poblaciones normales bidimensionales con medias $\mu_1 = (0,0)'$, $\mu_2 = (1,1)'$ y $\mu_3 = (2,0)'$ y matrices de covarianzas iguales a

$$\begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix},$$

se pide:

(i) Obtener las funciones discriminantes lineales.

(ii) Clasificar a $z = (1,1/2)'$.

(iii) Obtener la función discriminante de Fisher, la constante K y el criterio de clasificación para distinguir entre las poblaciones 2 y 3.

(iv) Dibujar las regiones de clasificación para cada grupo.

3. Dadas tres poblaciones normales bivariantes con matriz de covarianzas común

$$V = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$$

y medias $(-1,2), (0,-2)$ y $(3,5)$, respectivamente.

a) Obtener las funciones discriminantes.

b) Si $z = (2,1)$, ¿en qué población se clasificaría?

4. Dadas tres poblaciones normales bivariantes con matriz de covarianzas común

$$V = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$$

y medias $(1,2), (0,2)$ y $(3,0)$, respectivamente.

a) Obtener las funciones discriminantes.

b) Si $z = (2,1)$, ¿en qué población se clasificaría?

5. Dadas tres poblaciones normales bidimensionales con medias $\mu_1 = (0,1)'$, $\mu_2 = (1,0)'$ y $\mu_3 = (2,2)'$ y matriz de covarianzas común

$$V = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$$

obtener las funciones discriminantes y clasificar a $z = (1, 3/2)'$.

6. Dadas tres poblaciones normales con matriz de covarianzas común

$$V = \begin{pmatrix} 6 & 2 \\ 2 & 1 \end{pmatrix}$$

y medias $(0,1), (1,0)$ y $(1,1)$, respectivamente, obtener las funciones discriminantes y el criterio de clasificación.

7. Dados dos vectores aleatorios bidimensionales con medias $(0,0)$ y $(3,0)$ y matrices de covarianzas

$$V_1 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \text{ y } V_2 = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix},$$

se pide:

a) Calcular las funciones discriminantes cuadráticas.

b) Clasificar a $z = (1, -4)$ usando dichas funciones.

c) Representar gráficamente las regiones de clasificación.

Dados dos vectores aleatorios bidimensionales con medias $(0,0)$ y $(1,1)$ y matrices de covarianzas

$$V_1 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \text{ y } V_2 = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix},$$

respectivamente, se pide:

a) Calcular las funciones discriminantes cuadráticas.

b) Clasificar a $z = (1, 0)$ usando dichas funciones.

c) Representar gráficamente las regiones de clasificación para un punto cualquiera del plano $z = (x, y)$.

8. Dadas dos poblaciones normales bidimensionales con medias $\mu_1 = (1,0)'$ y $\mu_2 = (0,0)'$ y matrices de covarianzas

$$V_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix} \text{ y } V_2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

se pide:

(i) Obtener las funciones discriminantes cuadráticas y clasificar a $z = (1, 1)'$.

(ii) Clasificar a z usando el criterio de mínima distancia de Mahalanobis y representar las regiones de clasificación con este criterio para cada grupo.

9. Obtener un criterio de clasificación para dos poblaciones Exponenciales unidimensionales con medias distintas usando máxima verosimilitud. Calcular las probabilidades de error. ¿Cuál deberá ser el criterio de clasificación para que las probabilidades de ambos errores sean iguales?. Clasificar a $z = 1.5$ entre dos poblaciones exponenciales con medias 2 y 1 usando ambos criterios. (Indicación: La función de densidad de la distribución exponencial es $f(x) = (1/\mu)\exp(-x/\mu)$ para $x \geq 0$).

10. Sea $Z = (X_1, X_2)$ formada por dos variables independientes Bernoulli $(X_i = 0, 1)$ con probabilidades $\Pr(X_i = 1) = p_{i,j}$, $i = 1, 2$, según que el individuo z provenga de la población G_j, $j = 1, 2$. Sabiendo que la probabilidad de que un individuo provenga de la población G_j es $\pi_j = \Pr(z \in G_j)$ obtener la regla de clasificación.

11. Aplicar un DA a los datos de las columnas 5-10 del objeto d del fichero `bears.rda`[1]) para estudiar si esas medidas sirven para determinar el sexo del oso. Las variables son: Head.L= longitud de la cabeza (pulgadas), Head.W=anchura de la cabeza (pulgadas), Neck.G=perímetro cuello (pulgadas), Length=altura (pulgadas), Chest.G=perímetro pecho (pulgadas), Weight=peso (libras). Fuente: Minitab15.

12. Aplicar un DA a los datos del fichero `pottery` del paquete MVA[2] que contiene resultados de análisis químicos de cerámica británica de la época romana de diversas regiones y hornos (kiln). La región 1 corresponde al horno 1, la región 2 a los hornos 2 y 3, y la región 3 a los hornos 4 y 5. ¿Podemos usar estas medidas para determinar el origen de la cerámica?

13. Aplicar un DA a los datos del objeto d del fichero `pulgas.rda`[1].

[1]Para este tipo de archivos teclear `load('f:/name.rda')` indicando la ruta completa en donde se encuentra el archivo y reemplazando `name` por el nombre del archivo.

[2]Para leer este conjunto de datos hay que instalar el paquete MVA pinchando en el menú: Instalar>Paquete seleccionando MVA y tecleando en R: `library('MVA')` (o indicando en el menú que se carge este paquete).

4

Apéndice

4.1 Formulario.

Álgebra:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in M_{n\times 1}, A = (a_{i,j}) = \begin{pmatrix} a_{1,1} & a_{1,2} & ... & a_{1,m} \\ ... & ... & ... & ... \\ a_{n,1} & a_{n,2} & & a_{n,m} \end{pmatrix} \in M_{n\times m}$$

1. Traspuesta: $A' = (a_{j,i})$
2. Norma euclídea: $\| x \| = \sqrt{(x'x)} = \sqrt{x_1^2 + ... + x_n^2}$
3. Desigualdad de Cauchy-Schwarz: $(x'y)^2 \leq (x'x)(y'y)$, es decir,

$$| x'y | \leq \| x \| \| y \| .$$

   Además se da la igualdad si y solo si $y = \lambda x$.

4. Inversa generalizada A^- tal que $AA^-A = A$
5. Determinante $| A |$, si $A \in M_{n\times n}$
6. Matriz no singular si $| A | \neq 0$
7. Si A es no singular, existe A^{-1}
8. $I_n = \begin{pmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & 0 & 1 \end{pmatrix}, 1_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$
9. $diag(d_1, ..., d_n) = \begin{pmatrix} d_1 & 0 & ... & 0 \\ 0 & d_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & 0 & d_n \end{pmatrix}$
10. Una matriz simétrica es semidefinida positiva ($A \geq 0$) si $x'Ax \geq 0$, para todo x.
11. Todos los valores propios de una matriz simétrica (real) son números reales.
12. Una matriz simétrica es semidefinida positiva si todos sus valores propios son mayores o iguales que cero.
13. Teorema espectral: Si A es una matriz (real) simétrica, entonces existe una matriz ortogonal T tal que $T'AT$ es diagonal.

14. $A \leq B$ si $B - A \geq 0$ (si $B - A$ es semidefinida positiva).

15. Rango de una matriz $R(A)$ es el espacio vectorial (o su dimensión) generado por sus columnas.

16. Núcleo de una matriz $N(A)$ es el espacio vectorial $\{x : Ax = 0\}$

17. Matriz idempotente $A^2 = A$

18. $traza(ABC) = traza(BCA) = \sum valores\ propios$

19. A simétrica e idempotente, entonces sus valores propios son cero o uno y $rango = traza$.

20. Producto de Kroneker (directo), $A \otimes B = (a_{ij}B) \in M_{nm \times nm}, A \in M_{m \times m}$ y $B \in M_{n \times n}$.

Esperanza y covarianza:

X, Y, Z vectores (columna) aleatorios. A y B matrices.

1) $E(a_1 g_1(X) + a_2 g_2(X)) = a_1 E(g_1(X)) + a_2 E(g_2(X))$; $a_1, a_2 \in \mathbb{R}$

2) $X = (Y, Z), E_X(g(Y)) = E_Y(g(Y))$

3) Si (X, Y) independientes, entonces $E(g_1(X)g_2(Y)) = E(g_1(X))E(g_2(Y))$

4) $E(AX + b) = AE(X) + b; A \in M_{m,k}, b' \in \mathbb{R}^k$

5) $Cov(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$

6) $Cov(X_i, X_j) = 0$ si $X_1 ... X_k$ son independientes

7) $Var(X_i + X_j) = Var(X_i) + 2Cov(X_i, X_j) + Var(X_j)$

8) $Cov(aX_i + b, cX_j + d) = acCov(X_i, X_j)$

9) $Cov(X) = E((X - \mu)(X - \mu)') = E(XX') - \mu\mu'$

10) $Var(a'X) = a'Cov(X)a = \sum a_i a_j \sigma_{i,j}$

11) $Cov(AX + b) = ACov(X)A'$

12) $Corr(X_i, X_j) = 0$ si $X_1 ... X_k$ son independientes

13) $Corr(aX_i + b, cX_j + d) = Corr(X_i, X_j)$

14) $-1 \leq Corr(X_i, X_j) \leq 1$

15) $Corr(X_i, aX_i + b) = \pm 1$ (según el signo de a)

16) $Corr(X) = \Delta^{-1} Cov(X) \Delta^{-1}$, donde Δ es la matriz diagonal formada por las desviaciones típicas ($\Delta = diag(\sigma_1, \ldots, \sigma_k)$).

17) $Cov(X, Y) = (Cov(X_i, Y_j)) = Cov(Y, X)'$

18) $Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)$

19) Si X e Y tienen la misma dimensión, entonces

$$Cov(X + Y) = Cov(X) + Cov(X, Y) + Cov(Y, X) + Cov(Y)$$

20) $Cov(AX, BY) = ACov(X, Y)B'$

21) Si X, Y independientes, entonces $Cov(X, Y) = 0$

Modelos:

1) Distribución multinomial $M_k(n, p_1, ..., p_k)$

$$p(x_1, ..., x_k) = \frac{n!}{x_1!...x_k!} p_1^{x_1} ... p_k^{x_k}$$

2) Normal $N_k(\mu, V)$

$$f(x) = \frac{1}{\sqrt{\mid V \mid (2\pi)^k}} \exp\left(-\frac{1}{2}(x-\mu)' V^{-1}(x-\mu)\right).$$

3) Distribución Wishart $W_k(m, V) = \sum_{j=1}^{m} Y_j Y_j'$, con $Y_j \equiv N_k(0, V)$ indep.

$$f(w_{11}, ..., w_{kk}) = 2^{-mk/2} \mid V \mid^{-m/2} \Gamma_k^{-1}(\frac{m}{2}) \mid W \mid^{(m-k-1)/2} e^{tr(-V^{-1}W/2)}$$

siendo $W = (w_{ij})$, $\Gamma_k(\frac{m}{2}) = \pi^{k(k-1)/4} \prod_{j=1}^{k} \Gamma(\frac{m+1-j}{2})$ y $\Gamma(a) = \int_0^\infty y^{a-1} e^{-y} dy$.

4) Distribución T^2 de Hotelling $T^2_{k,m} = mZ'W^{-1}Z$, con $Z \equiv N_k(0, I)$ y $W \equiv W_k(m, I)$ independientes. Verifica $\frac{m+1-k}{mk} T^2_{k,m} \equiv F_{k,m-k+1}$.

5) Distribución F de Snedecord $F_{n,m}$:

$$f(x) = \frac{\sqrt{n^n m^m}}{\beta(n/2, m/2)} \sqrt{\frac{x^{m-2}}{(m+nx)^{n+m}}} \text{ si } x > 0$$

Inferencia:

1) Media: $\overline{X} = \overline{O} = (\overline{X}_j) = \frac{1}{n} \sum_{i=1}^{n} O_i$, con $\overline{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$ con $E(\overline{X}) = E(X)$, $Cov(\overline{X}) = \frac{1}{n} Cov(X)$ y $\overline{X} \equiv N_k(\mu, V/n)$.

2) Cuasivarianza muestral:

$$S = (S_{ij}) = \frac{1}{n-1} \sum_{l=1}^{n} (O_l - \overline{O})(O_l - \overline{O})',$$

con

$$S_{ij} = \frac{1}{n-1} \sum_{l=1}^{n} (X_{li} - \overline{X}_i)(X_{lj} - \overline{X}_j)',$$

$E(S) = V$ y $(n-1)S \equiv W_k(n-1, V)$.

3) Correlación: $R = (r_{ij}) = D^{-1} S D^{-1}, D = diag(S_j)$,con $r_{ij} = \frac{S_{ij}}{S_i S_j}$

4) Distancia de Mahalanobis: $\triangle(x, y) = \sqrt{(x-y)' V^{-1}(x-y)}$.

5) Distancia de Mahalanobis muestral: $D(x, y) = \sqrt{(x-y)' S^{-1}(x-y)}$ con $nD^2(\overline{X}, \mu) = n(\overline{X} - \mu)' S^{-1}(\overline{X} - \mu) \equiv T^2_{k,n-1}$ de Hotelling.

Análisis de Componentes Principales (PCA).

PCA1) Definición Y_1)

$$\left.\begin{array}{c}\max Var(a'X)\\ s.a.: \quad a'a=1\end{array}\right\}$$

PCA2) Definición Y_j: Y_1) y debe tener varianza máxima, es decir

$$\left.\begin{array}{l}\max Var(a'X)\\ s.a.: \quad a'a=1\\ \qquad\quad Cov(Y_i,a'X)=0, i=1,\ldots,j-1.\end{array}\right\}$$

PCA3) Cálculo teórico: $Y=T'X$, donde $TT'=T'T=I$ y $T'VT=D=diag(\lambda_1,...,\lambda_k)$ con $\lambda_1\geq\lambda_2\geq...\geq\lambda_k>0$.

PCA4) $Y_j=t_j'X$, $Var(Y_j)=\lambda_j$ y

$$traza(V)=\sum_{j=1}^{k}Var(X_j)=\sum_{j=1}^{k}Var(Y_j)=\sum_{j=1}^{k}\lambda_j.$$

PCA5) Información en Y_j: $I_j=100\lambda_j/(\sum_{i=1}^{k}\lambda_i)\%$.

PCA6) $|V|=\lambda_1...\lambda_k=|Cov(Y)|$.

PCA7) Relaciones entre X e $Y=T'X$:

$$\begin{aligned}Cov(X,Y)&=TD\\ Corr(X,Y)&=diag(V)^{-1/2}TD^{1/2}\end{aligned}\tag{4.1}$$

donde $diag(V)=diag(\sigma_1^2,...,\sigma_k^2)$.

PCA8) Matriz de saturaciones $A=Corr(X,Y)$ con

$$a_{i,j}=Corr(X_i,Y_j)=t_{i,j}\sqrt{\lambda_j}/\sigma_i.$$

PCA9) Información en Y_j sobre X_i:

$$100Corr^2(X_i,Y_j)\%=100t_{i,j}^2\lambda_j/\sigma_i^2\%.$$

Análisis Discriminante (DA)

DA1) Función discriminante de Fisher a la v.a.

$$D=L(Z)=a'Z=(\mu_X-\mu_Y)'V^{-1}Z.$$

DA2) Si X e Y son normales con $Cov(X)=Cov(Y)=V$,

$$D\sim N_1((\mu_X-\mu_Y)'V^{-1}\mu,\triangle(\mu_X,\mu_Y))$$

donde $\mu=E(Z)$ es igual a μ_X o μ_Y.

DA3) Regla de discriminación:

Si $L(Z) > K$, entonces Z es clasificado en X

Si $L(Z) < K$, entonces Z es clasificado en Y

donde $K = L((\mu_X + \mu_Y)/2)$.

DA4) Probabilidad de errores 1 y 2

$$\begin{aligned}\Pr(e_1) &= \Pr(Z \in R_Y \mid Z \equiv X) \\ &= \Pr\left(U < -\frac{1}{2} \triangle (\mu_X, \mu_Y)\right) \\ &= \Pr(e_2) \\ &= \Pr(Z \in R_X \mid Z \equiv Y),\end{aligned}$$

donde $U \equiv N_1(0,1)$.

DA5) Probabilidad error total

$$\begin{aligned}\Pr(error) &= \Pr(Z \in R_Y \mid Z \equiv X)\Pr(Z \equiv X) + \Pr(Z \in R_X \mid Z \equiv Y)\Pr(Z \equiv Y) \\ &= \Pr(e_1)q_1 + \Pr(e_2)q_2\end{aligned}$$

con $q_1 = \Pr(Z \equiv X)$ y $q_2 = \Pr(Z \equiv Y)$ (probabilidades a priori).

DA6) Se minimiza el coste esperado

$$c(K) = c_1 \Pr(e_1)q_1 + c_2 \Pr(e_2)q_2$$

si

$$K = a' \frac{\mu_X + \mu_Y}{2} + \log\left(\frac{c_2 q_2}{c_1 q_1}\right).$$

DA7) Probabilidades a posteriori

$$\Pr(Z \equiv X \mid Z = z) = \frac{f_X(z)q_1}{f_X(z)q_1 + f_Y(z)q_2}.$$

DA8) Función discriminante lineal (FDL)

$$L_i(z) = (\mu^{(i)})' V^{-1} z - (\mu^{(i)})' V^{-1} \mu^{(i)}/2,$$

clasificándose z en $G_i : L_i(z) \geq L_j(z)$ para todo j.

DA9) Proyecciones canónicas $Z^* = V^{-1/2}Z$ con $Cov(V^{-1/2}Z) = I$,

$$L(Z) = (V^{-1/2}\mu_X - V^{-1/2}\mu_Y)' V^{-1/2} Z$$

y

$$d_i^2(z) = d^2(V^{-1/2}z, V^{-1/2}\mu^{(i)}) = \Delta^2(z, \mu^{(i)}).$$

DA10) Función discriminante cuadrática (QDF)

$$QDF_i(z) = c - 2\log f_i(z) = (z - \mu^{(i)})' V_i^{-1} (z - \mu^{(i)}) + \log \mid V_i \mid,$$

clasificándose z en $G_i : QDF_i(z) \leq QDF_j(z)$ para todo j.

DA11) Función discriminante cuadrática basada en la distancia de Mahalanobis

$$QDF_i^*(z) = \Delta^2(z, \mu^{(i)}) = (z - \mu^{(i)})' V_i^{-1} (z - \mu^{(i)}).$$

DA12) Estimaciones por grupos

$$\begin{aligned}
\widehat{\mu}_j &= \frac{1}{n_j} \sum_{i=1}^{n} \omega_j 1(Y(\omega_i) = j) \\
\widehat{V}_j &= \frac{1}{n_j - 1} \sum_{i=1}^{n} 1(Y(\omega_i) = j)(\omega_i - \widehat{\mu}_j)(\omega_i - \widehat{\mu}_j)' \\
\omega_i &= (X_{1,i}, ..., X_{k,i})' \\
n_j &= \sum_{i=1}^{n} 1(Y(\omega_i) = j),
\end{aligned}$$

donde $1(Y(\omega_i) = j) = 1$ si $\omega_i \in G_j$ (cero si no).

DA13) Matriz de covarianzas ponderada (pooled)

$$\widehat{V} = \frac{1}{n - m} \sum_{j=1}^{m} (n_j - 1) \widehat{V}_j.$$

DA14) Función discriminante de Fisher muestral

$$\widehat{D} = \widehat{L}(Z) = \widehat{a}' Z = (\widehat{\mu}_X - \widehat{\mu}_Y)' \widehat{V}^{-1} Z.$$

DA15) Probabilidad del error tipo 1 muestral

$$\Pr(e_1) = \Pr\left(U < -\frac{1}{2} \widehat{\triangle}\right)$$

donde $U \equiv N_1(0, 1)$ y

$$\widehat{\triangle} = \sqrt{(\widehat{\mu}_X - \widehat{\mu}_Y)' \widehat{V}^{-1} (\widehat{\mu}_X - \widehat{\mu}_Y)}$$

es la distancia de Mahalanobis muestral.

DA16) Funciones discriminantes lineales muestrales

$$\widehat{L}_i(z) = (\widehat{\mu}^{(i)})' \widehat{V}^{-1} z - (\widehat{\mu}^{(i)})' \widehat{V}^{-1} \widehat{\mu}^{(i)} / 2,$$

clasificándose z en $G_i : \widehat{L}_i(z) \geq \widehat{L}_j(z)$ para todo j.

DA17) Proyecciones canónicas muestrales $Z^* = \widehat{V}^{-1/2} Z$.

DA18) Funciones discriminantes cuadráticas muestrales

$$\widehat{Q}_i(z) = c - 2 \log \widehat{f}_i(z) = (z - \widehat{\mu}^{(i)})' \widehat{V}_i^{-1} (z - \widehat{\mu}^{(i)}) + \log \mid \widehat{V}_i \mid,$$

clasificándose z en $G_i : \widehat{Q}_i(z) \leq \widehat{Q}_j(z)$ para todo j.

4.2 Tablas.

Tabla 4.1: Características de los modelos discretos más usuales.

Modelo	$E(X)$	$Var(X)$	γ_1	γ_2
Binomial	np	npq	$\frac{1-2p}{\sqrt{npq}}$	$\frac{1}{npq}-\frac{6}{n}$
Geométrica	q/p	q/p^2	$\frac{1+q}{\sqrt{q}}$	$6+\frac{p^2}{q}$
Bin. Neg.	nq/p	nq/p^2	$\frac{1+q}{\sqrt{nq}}$	$\frac{6q+p^2}{nq}$
Hipergeo.	$n\frac{a}{N}$	$n\frac{a}{N}\frac{N-a}{N}\frac{N-n}{N-1}$	$\frac{(N-2a)(N-2n)}{N-2}\sqrt{\frac{(N-1)}{Na(N-a)(N-n)}}$	$*$
Poisson	λ	λ	$1/\sqrt{\lambda}$	$1/\lambda$

$H(N,a,n) \approx B(n,p=a/N)$ si $N/n > 10$.
$B(n,p) \approx P(\lambda = np)$ si $np > 1$ y $p < 0.1$.
$B(n,p) \approx N(\mu = np, \sigma = \sqrt{npq})$ si $npq > 5$.
$P(\lambda) \approx N(\mu = \lambda, \sigma = \sqrt{\lambda})$ si $\lambda > 5$.

Tabla 4.2: Características de los modelos continuos más usuales.

Modelo	$E(X)$	$Var(X)$	γ_1	γ_2
Normal	μ	σ^2	0	0
Uniforme	$(a+b)/2$	$(b-a)^2/12$	0	$-6/5$
Exponencial	$1/\lambda$	$2/\lambda^2$	2	6
Gamma	a/b	a/b^2	$2/\sqrt{a}$	$6/a$
Beta	$\frac{a}{a+b}$	$\frac{ab}{(a+b+1)(a+b)^2}$	$*$	$*$
Weibull	$b\Gamma(1+1/a)$	$b^2\Gamma(1+2/a)-b^2\Gamma^2(1+1/a)$	$*$	$*$
Chi-cuadrado	n	$2n$	$\sqrt{8/n}$	$12/n$
t-Student	0	$\frac{n}{n-2}$	0	$\frac{6}{n-4}$
F-Snedecor	$\frac{n}{n-2}$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$	$*$	$*$

$\Gamma(p) = \int_0^\infty t^{p-1}\exp(-t)\,dt$.
$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Tabla 4.3: Función de distribución Normal $N(0,1)$.

z	G(z)	z	G(z)	z	G(z)
0.1	0.5398278	1.1	0.8643339	2.1	0.9821356
0.2	0.5792597	1.2	0.8849303	2.2	0.9860966
0.3	0.6179114	1.3	0.9031995	2.3	0.9892759
0.4	0.6554217	1.4	0.9192433	2.4	0.9918025
0.5	0.6914625	1.5	0.9331928	2.5	0.9937903
0.6	0.7257469	1.6	0.9452007	2.6	0.9953388
0.7	0.7580363	1.7	0.9554345	2.7	0.9965330
0.8	0.7881446	1.8	0.9640697	2.8	0.9974449
0.9	0.8159399	1.9	0.9712834	2.9	0.9981342
1	0.8413447	2	0.9772499	3	0.9986501

Tabla 4.4: Cuantiles de la distribución Normal $N(0,1)$.

z	G(z)	z	G(z)	z	G(z)
0	0.50	0.25334710	0.60	0.52440051	0.70
0.02506891	0.51	0.27931903	0.61	0.55338472	0.71
0.05015358	0.52	0.30548079	0.62	0.58284151	0.72
0.07526986	0.53	0.33185335	0.63	0.61281299	0.73
0.10043372	0.54	0.35845879	0.64	0.64334541	0.74
0.12566135	0.55	0.38532047	0.65	0.67448975	0.75
0.15096922	0.56	0.41246313	0.66	0.70630256	0.76
0.17637416	0.57	0.43991317	0.67	0.73884685	0.77
0.20189348	0.58	0.46769880	0.68	0.77219321	0.78
0.22754498	0.59	0.49585035	0.69	0.80642125	0.79

z	G(z)	z	G(z)	z	G(z)
0.84162123	0.80	1.28155157	0.90	2.575829	0.995
0.87789630	0.81	1.34075503	0.91	3.090232	0.999
0.91536509	0.82	1.40507156	0.92	3.290527	0.9995
0.95416525	0.83	1.47579103	0.93	3.719016	0.9999
0.99445788	0.84	1.55477359	0.94		
1.03643339	0.85	1.64485363	0.95		
1.08031934	0.86	1.75068607	0.96		
1.12639113	0.87	1.88079361	0.97		
1.17498679	0.88	2.05374891	0.98		
1.22652812	0.89	2.32634787	0.99		

Tabla 4.5: Comandos en R más usuales

Comando	Significado
`m^n`	m^n
`factorial(m)`	$m!$
`choose(m,n)`	$\binom{m}{n}$
`exp(x)`	e^x
`log(x)`	$\ln(x)$
`curve(f(x),a,b,ylab='f(x)')`	Dibuja f en (a,b)
`curve(g(x),add=TRUE)`	Añade la gráfica de g
`plot(x,y)`	Dibuja los puntos (x,y)
`plot(x,y,type='l')`	Dibuja los (x,y) y los une
`plot(x)`	Dibuja la serie (i,x_i)
`text(x,y,z)`	Pone la etiqueta z en (x,y)
`barplot(x,f)`	Histograma de frecuencias (x,f_x)
`f<-function(x) 2x+3`	Define la función $f(x)=2x+3$
`gamma(p)`	$\Gamma(p)=\int_0^\infty t^{p-1}\exp(-t)\,dt$
`beta(a,b)`	$\beta(a,b)=\int_0^1 x^{a-1}(1-x)^{b-1}$
`mean(X)`	$\overline{X}=\frac{1}{n}=\sum_{i=1}^n X_i$
`var(X)`	$S^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\overline{X})^2$
`v<-vector(length=3)`	Define el vector (columna) v
`M<-matrix(nrow=2,ncol=2)`	Define la matriz M
`M[1,2]`	Elemento de la fila 1 y columna 2 de M
`t(M)`	Matriz (o vector) transpuesto
`A%*%B`	Producto de matrices
`solve(M)`	Matriz inversa
`eigen(M)`	Vectores y valores propios de M
`mahalanobis(x,y,V)`	Distancia de Mahalanobis

Tabla 4.6: Nombres en R de los modelos discretos más usuales.

Modelo	$p(x)$	x	Nombre en R
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	0,. . . ,n	$binom(x,n,p)$
Geométrica	$p(1-p)^x$	0,1,. . .	$geom(x,p)$
Bin. Neg.	$\binom{n+x-1}{x}p^n(1-p)^x$	$0,1,\ldots$	$nbinom(x,n,p)$
Hipergeo.	$\binom{a}{x}\binom{b}{n-x}/\binom{N}{n}$	$0,\ldots,\min(n,a)$	$hyper(x,a,b,n)$
Poisson	$\exp(-\lambda)\lambda^x/x!$	$0,1,\ldots$	$pois(x,\lambda)$

La función de distribución $F(x)$ se obtiene haciendo *pnombre*, la función puntual de probabilidad $p(x)$ con *dnombre*, los cuantiles $F^{-1}(x)$ con *qnombre* y podemos generar m datos con $rnombre(m,a,b)$.

Tabla 4.7: Nombres en R de los modelos continuos más usuales.

Modelo	$f(x)$	x	Nombre en R
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp(-(x-\mu)^2/(2\sigma^2))$	$\mathbb{R}$	$norm(x,\mu,\sigma)$
Uniforme	$1/(b-a)$	(a,b)	$unif(x,a,b)$
Exponencial	$\lambda\exp(-\lambda x)$	$(0,\infty)$	$exp(x,\lambda)$
Gamma	$\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$	$(0,\infty)$	$gamma(x,a,b)$
Beta	$\frac{1}{\beta(a,b)}x^{a-1}(1-x)^{b-1}$	$(0,1)$	$beta(x,a,b)$
Weibull	$b^{-a}x^{a-1}\exp(-(x/b)^a)$	$(0,\infty)$	$weibull(x,a,b)$
Chi-cuadrado	$\frac{2^{-n/2}}{\Gamma(n/2)}x^{n/2-1}e^{-x/2}$	$(0,\infty)$	$chisq(x,n)$
t-Student	$\frac{1}{\beta(1/2,n/2)}\sqrt{\frac{n^n}{(n+x^2)^{n+1}}}$	$\mathbb{R}$	$t(x,n)$
F-Snedecor	$\frac{n^{n/2}m^{m/2}}{\beta(n/2,m/2)}\frac{x^{m/2-1}}{(n+mx)^{(n+m)/2}}$	$(0,\infty)$	$f(x,m,n)$
$N_k(\mu,V)$	$\frac{1}{\sqrt{\lvert V\rvert(2\pi)^k}}\exp(-\Delta^2(x,\mu)/2)$	$\mathbb{R}^k$	$mvnorm(x,\mu,V)$

La función de distribución $F(x)$ se obtiene haciendo *pnombre*, la función de densidad de probabilidad $f(x)$ con *dnombre*, los cuantiles $F^{-1}(x)$ con *qnombre* y podemos generar m datos con $rnombre(m,a,b)$. Para la normal multivariante hay que cargar el paquete **mvtnorm**.

Tabla 4.8: Comandos en R para Análisis de Componentes Principales.

Comando	Significado
`data()`	Datos en R
`d<-LifeCycleSavings`	Guardar datos
`summary(d)`	Resumen de los datos
`View(d)`	Ver los datos
`plot(d)`	Gráficos bidimensionales
`cov(d)`	Matriz de covarianzas
`cor(d)`	Matriz de correlaciones
`boxplot(d)`	Diagramas caja-bigote
`boxplot(d[,1])`	Diagrama caja-bigote
`which.max(d[,1])`	Índice que da el máximo
`sort(d[,1])`	Valores ordenados
`hist(d[,1])`	Histograma
`PCA<-princomp(d)`	PCA de covarianzas
`PCA<-princomp(d,cor=TRUE)`	PCA de correlaciones
`princomp(covmat=cor(d))`	PCA basado en una matriz
`PCAbis<-prcomp(d,scale=TRUE)`	PCA de correlaciones
`summary(PCA,loadings=TRUE)`	Resumen PCA
`PCA$loadings->T`	Matriz de cargas
`PCA$scores->S`	Matriz de puntuaciones
`z<-scale(d)`	Variables estandarizadas
`biplot(PCA,pc.biplot=TRUE)`	Gráfico $Y_1 - Y_2$
`biplot(PCA,pc.biplot=TRUE,choices=c(3,4))`	Gráfico $Y_3 - Y_4$
`biplot(PCA,pc.biplot=TRUE,xlabs=1:50)`	Gráfico con etiquetas
`plot(S[,1],S[,2],xlab='Y1',ylab='Y2')`	Gráfico $Y_1 - Y_2$
`text(S[38,1],S[38,2],labels='Esp')`	Etiqueta del dato 38
`pairs(PCA$scores[,1:3])`	Gráficos $Y_1 - Y_2 - Y_3$
`SAT<-cor(d,S)`	Matriz de saturaciones
`SAT[,1] ^ 2`	Informaciones en Y_1
`SAT[,1] ^ 2+ SAT[,2] ^ 2`	Comunalidades en $Y_1 - Y_2$
`screeplot(PCA)`	Gráfico de sedimentación
`plot(eigen(cor(d))$values,type='l')`	Gráfico de sedimentación

Tabla 4.9: Comandos en R para Análisis Discriminante

Comando	Significado
`load('e:/tal/escarabajos.rda')`	Leer escarabajos.rda
`dump('d','g:/nombre/datos1.R')`	Guardar el objeto d
`source('g:/nombre/datos1.R'`	Leer el objeto d
`tapply(d$surco,d$especie,summary)`	Resumen por grupos
`plot(d$surco,d$codigo)`	Gráfica por grupos
`text(d$surco[40],1.5,labels='40')`	Etiqueta para el dato 40
`boxplot(d$surco~d$especie)`	Gráficos caja-bigote por grupos
`plot(d$surco,d$long,pch=as.integer(d$especie))`	Gráficos $x-y$ por grupos
`legend('topright',legend=c('40','2','1'),pch=1:3)`	Explicación símbolos
`plot(d$surco,d$long)`	Gráfico bidimensional
`text(d$surco,d$long,d$especie)`	Etiquetas
`pca<-princomp(d[,1:4],cor=TRUE)`	Cálculo PCA
`biplot(pca,pc.biplot=TRUE,xlabs=d$especie)`	Gráfico PCA por grupos
`library('MASS')`	Carga paquete MASS
`LDA<-lda(d[1:39,1:4],d[1:39,6],prior=c(1/2,1/2))`	Cálculo LDA
`a<-LDA$scaling`	Vector a
`L<-function(z) sum(a*z)`	Función de Fisher
`z<-d[,1:4]`	Función discriminante
`D<-1:40`	Función discriminante
`for (i in 1:40) D[i]<-L(z[i,])`	Función discriminante FD
`plot(D,d$codigo)`	Gráfico FD
`text(D,d$codigo,pos=3,col='red')`	Etiquetas D
`text(D[40],labels='*')`	Etiqueta dato 40
`text(D[40],labels='40')`	Etiqueta dato 40
`P<-predict(LDA,d[,1:4])`	Predicciones con LDA
`P$x`	Puntuaciones
`ldahist(P$x,g=d$especie)`	Histograma por grupos
`P$class`	Predicciones
`P$class==d[,6]`	Aciertos y fallos
`table(P$class,d[,6])`	Resumen aciertos
`P$posterior`	Probabilidades a posteriori
`predict(LDA,c(185,280,150,200))`	Predicción
`CV<-lda(...,CV=TRUE)`	Validación cruzada
`table(CV$class,d[1:39,6])`	Aciertos CV
`QDA<-qda(d[1:39,1:4],d[1:39,6],prior=c(0.5,0.5))`	QDA
`predict(QDA,d[,1:4])->P`	Predicciones QDA
`table(P$class,d$codigo)`	Aciertos QDA
`qda(...,CV=TRUE)`	Validación cruzada
`table(QDACV$class,d[1:39,6])`	Aciertos CV
`S1<-cov(d[1:19,1:4])`	Matriz covarianza MC1
`S2<-cov(d[20:39,1:4])`	Matriz covarianza MC2
`(18*S1+19*S2)/37->S`	MC ponderada

5

Índice alfabético

6

Bibliografía

Anderson, T.W. (1974). A Introduction to Multivariate Statistical Analysis. Wiley.

Burgos, J. (1994). Curso de Álgebra y Geometría. Alhambra Longman.

Cuadras, C.M. (1991). Métodos de análisis multivariante. PPU.

Guillamón, A.; Navarro, J. (2002). Probabilidad y Estadística. Fundamentos (2^a ed.). DM.

Kotz, S.; Balakrishnan, N.; Jonhson, N.L. (2000). Continuous multivariate distributions.

Mardia, K.V.; Kent, J.T; Bibby, J.M. (1997). Multivariate Analysis. Academic Press.

Navarro, J.; Franco, M.; Guillamón, A. (1999). Probabilidad y Estadística. Problemas. DM.

Peña, D. (2002). Análisis de datos multivariantes. McGraw Hill.

Rencher, A.C. (1995). Methods of Multivariate Analysis. Wiley.

Srivastava, M.S.; Carter, E.M. (1983). A Introduction to Applied Multivariate Statistics. North-Holland.

Zoroa, P.; Zoroa, N. (2008). Elementos de Probabilidades. Diego Marín.