Tema 2
Técnicas de Análisis Discriminante

José R. Berrendero

Departamento de Matemáticas
Universidad Autónoma de Madrid

Un problema de análisis discriminante

- Cinco medidas tomadas en cráneos recogidos en dos zonas del Tibet.
- Los datos de `craneos1.txt` corresponden a tumbas de Sikkim y alrededores. Los datos de `craneos2.txt` corresponden a cráneos encontrados en el campo de batalla de Lhasa.
- Se cree posible que los dos conjuntos pertenezcan a dos etnias diferentes.

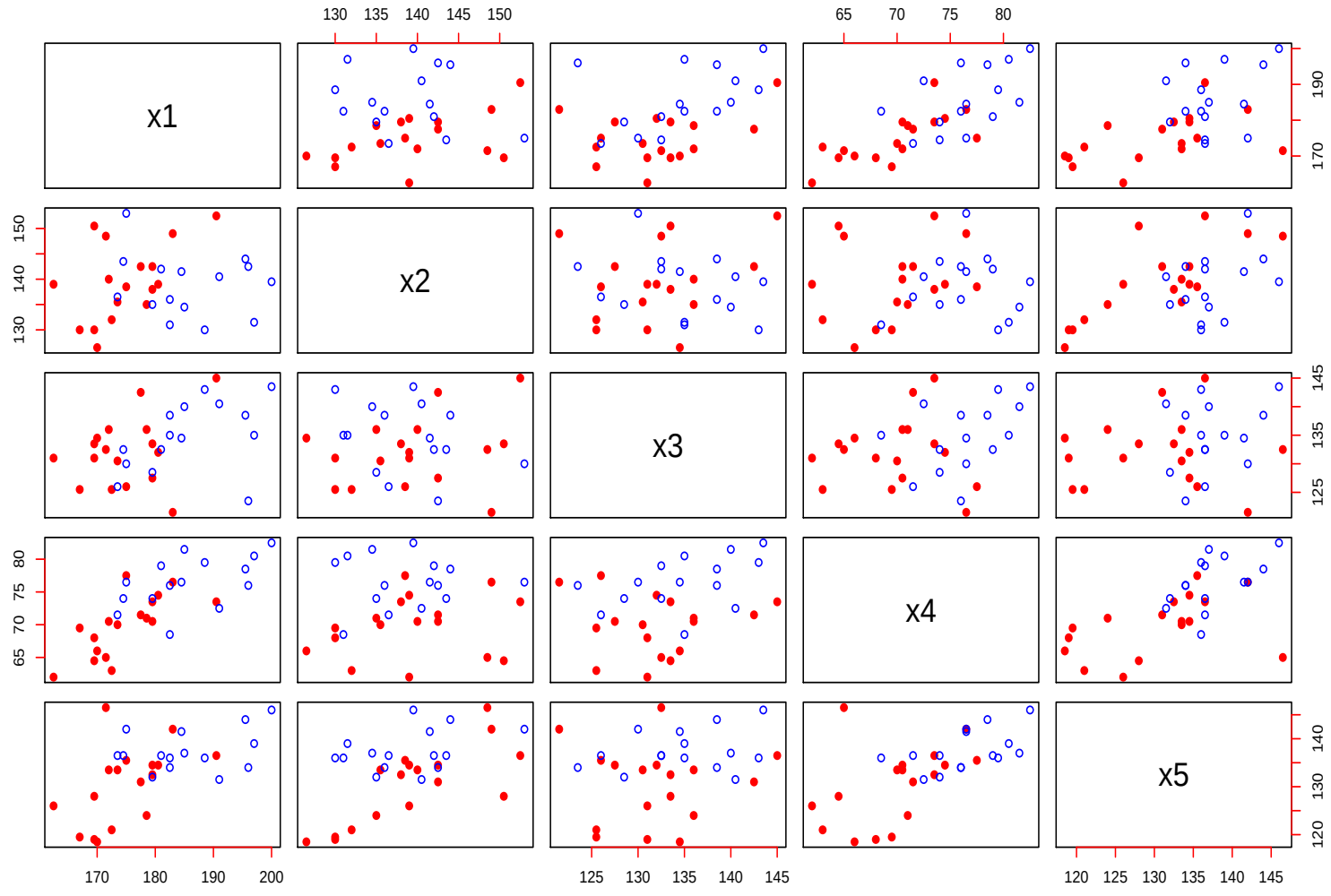
Variables

Nombre variable	Descripción
`x1`	Mayor longitud del cráneo
`x2`	Mayor anchura horizontal del cráneo
`x3`	Altura del cráneo
`x4`	Altura de la parte superior de la cara
`x5`	Anchura de la cara entre los huesos de las mejillas

Observaciones:

- Todas las variables se miden en mm.

Matriz de diagramas de dispersión



Rojo: grupo 1. Azul: grupo 2

El problema de discriminación

Sobre la base de las variables consideradas, parece que hay diferencias entre los dos grupos de cráneos.

Se recoge un nuevo cráneo, cuyo grupo es desconocido, y se miden los valores de las mismas cinco variables $(x_1, \ldots, X_5)$.

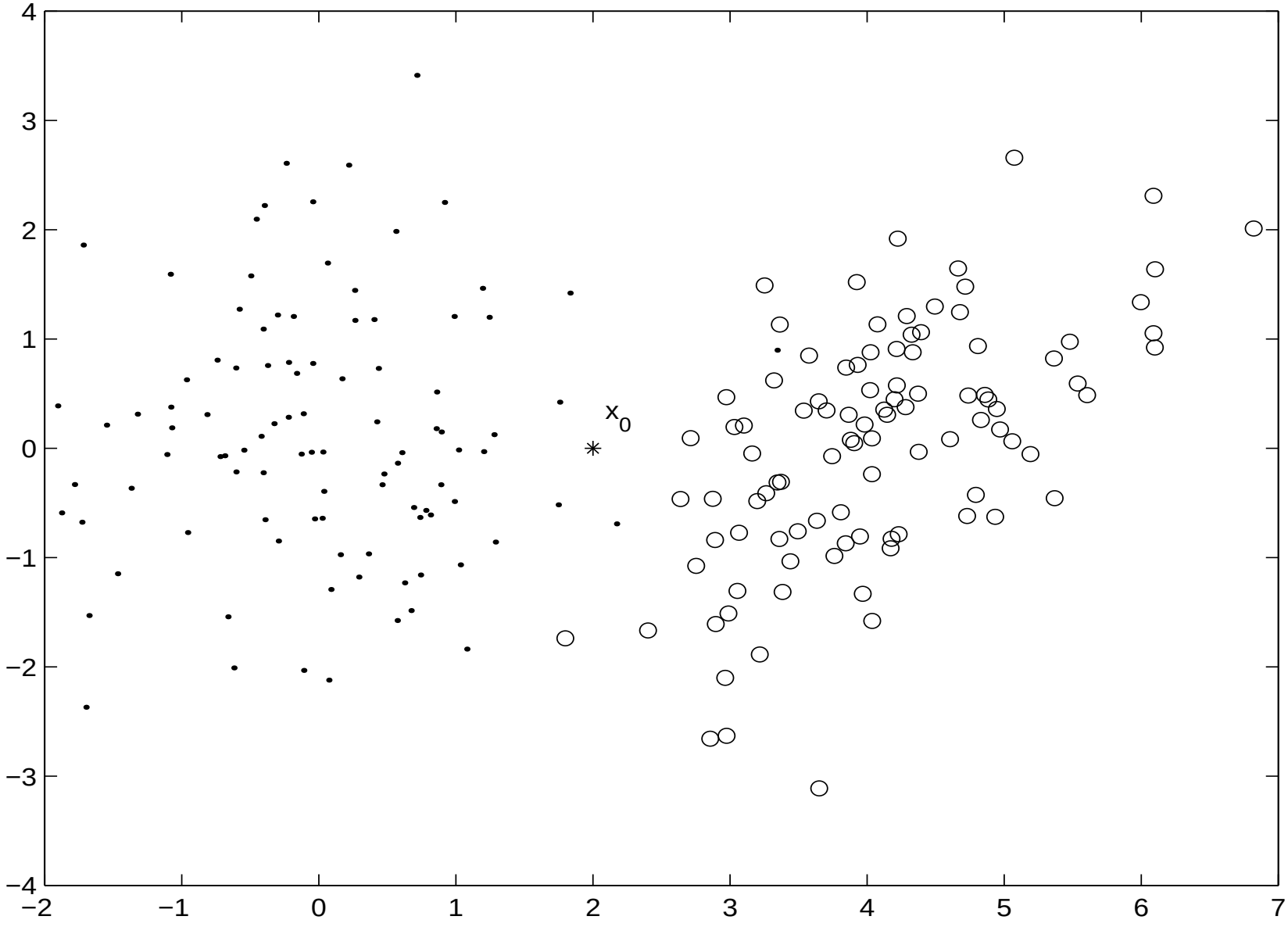
El problema de discriminación consiste en clasificar el nuevo cráneo en alguno de los dos grupos, de forma que la posibilidad de cometer un error sea lo menor posible.

Introducción al análisis discriminante

Dos muestras de tamaño 100 de dos poblaciones normales:

- La primera con vector de medias $\mu_1 = (0, 0)'$ y matriz de covarianzas $\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
- La segunda con vector de medias $\mu_2 = (4, 0)'$ y matriz de covarianzas $\mathbf{\Sigma}_2 = \begin{pmatrix} 1 & 1/\sqrt{3} \\ 1/\sqrt{3} & 1 \end{pmatrix} \approx \begin{pmatrix} 1 & 0.58 \\ 0.58 & 1 \end{pmatrix}$

¿En cuál de las dos poblaciones tenemos que clasificar la nueva observación $\mathbf{x}_0 = (2, 0)'$?
La nueva observación dista lo mismo (si se considera la distancia euclídea) de los dos vectores de medias poblacionales.


4
3
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2
3
4
5
6
7
x_0

Regla de clasificación de Mahalanobis

Regla de clasificación: Utilizar la distancia de Mahalanobis para medir las distancias entre $\mathbf{x}_0$ y los vectores de medias $\bar{\mathbf{x}}_1$ y $\bar{\mathbf{x}}_2$. Asignar $\mathbf{x}_0$ a aquel grupo cuyo vector de medias sea el más cercano.

Se asigna $\mathbf{x}_0$ al grupo 1 si

$$(\mathbf{x}_0 - \bar{\mathbf{x}}_1)' \mathbf{S}_1^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_1) < (\mathbf{x}_0 - \bar{\mathbf{x}}_2)' \mathbf{S}_2^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_2)$$

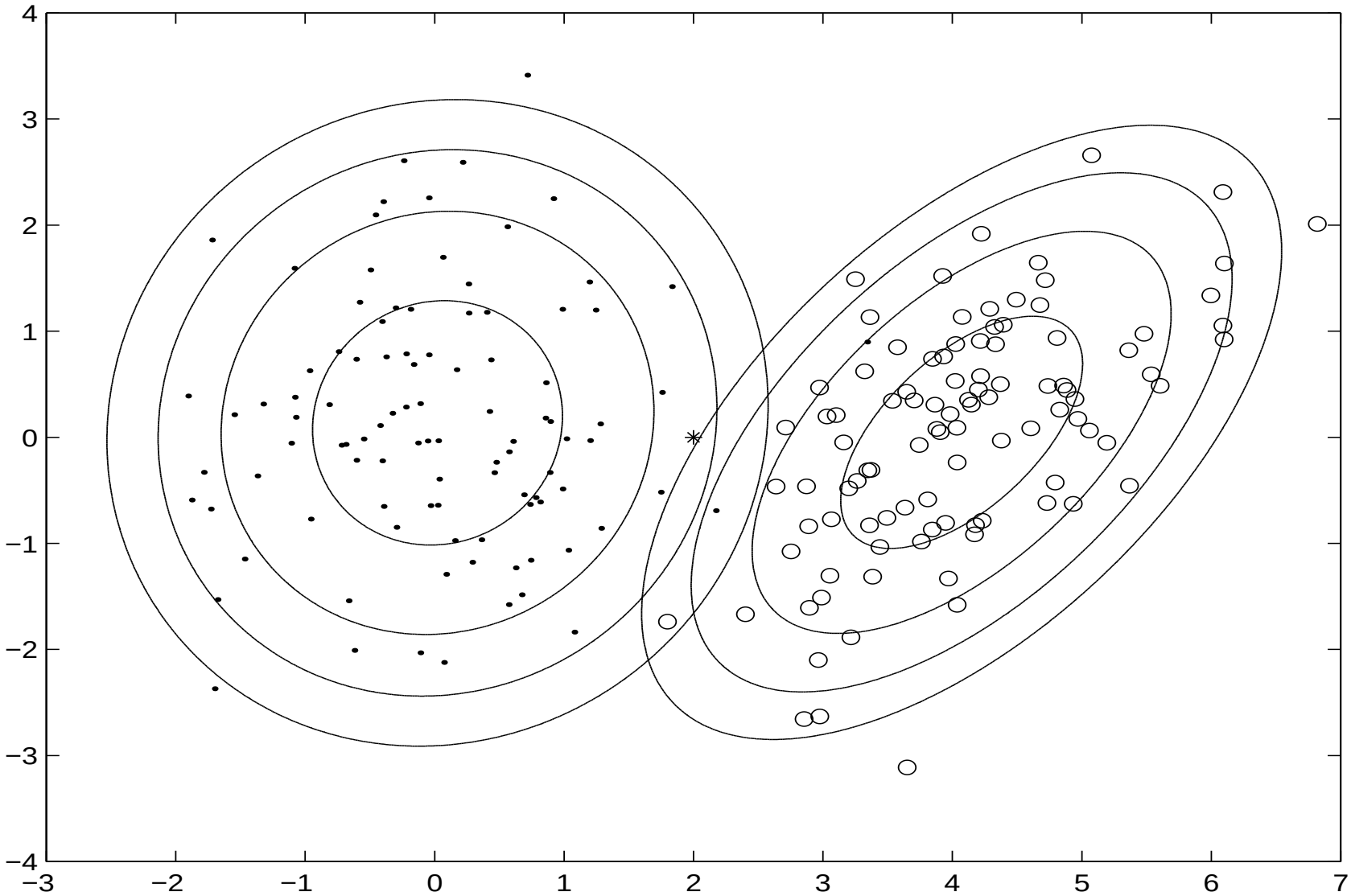
Para el ejemplo anterior se verifica:

$$\bar{\mathbf{x}}_1 = (0.02, 0.14)' \quad \text{y} \quad \mathbf{S}_1 = \begin{pmatrix} 0.93 & 0.06 \\ 0.06 & 1.33 \end{pmatrix},$$

$$\bar{\mathbf{x}}_2 = (4.07, 0.05)' \quad \text{y} \quad \mathbf{S}_2 = \begin{pmatrix} 0.87 & 0.60 \\ 0.60 & 1.20 \end{pmatrix}$$

De aquí se deduce $D_1^2 = 4.25$ y $D_2^2 = 7.36$, por lo tanto clasificamos $\mathbf{x}_0$ en el grupo 1.

Curvas de nivel



Matrices de covarianzas iguales

Para simplificar es habitual suponer $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Llamamos $\mathbf{\Sigma}$ a la matriz de covarianzas común.

¿Cómo se estima $\mathbf{\Sigma}$ a partir de las matrices de datos $\mathbf{X}_1$ $(n_1 \times p)$ y $\mathbf{X}_2$ $(n_2 \times p)$?

Tanto $\mathbf{S}_1$ como $\mathbf{S}_2$ son estimadores de $\mathbf{\Sigma}$. Parece razonable combinar los dos estimadores para usar toda la información disponible simultáneamente.

Así se obtiene lo que SPSS llama **matriz intra-grupos combinada**:

$$\mathbf{S}_w = \frac{n_1}{n_1 + n_2}\mathbf{S}_1 + \frac{n_2}{n_1 + n_2}\mathbf{S}_2$$

Algunos programas como SPSS calculan las covarianzas dividiendo por $n-1$ en lugar de n, con el fin de eliminar el sesgo.

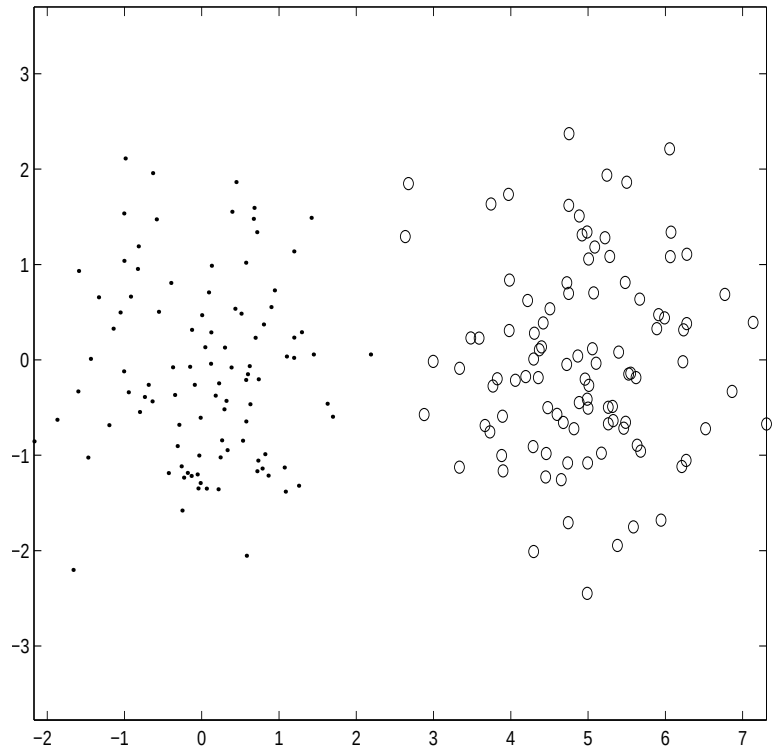
En este caso, la fórmula apropiada de la matriz intra-grupos combinada es:

$$\mathbf{S}_w = \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2$$

La diferencia entre ambas fórmulas es poco importante. Sólo cambia ligeramente la ponderación que recibe cada una de las dos matrices.

Ejemplo

Tenemos dos muestras de 100 datos normales bidimensionales con vectores de medias $(0,0)'$ y $(5,0)$. En ambos casos la matriz de covarianzas es la identidad.

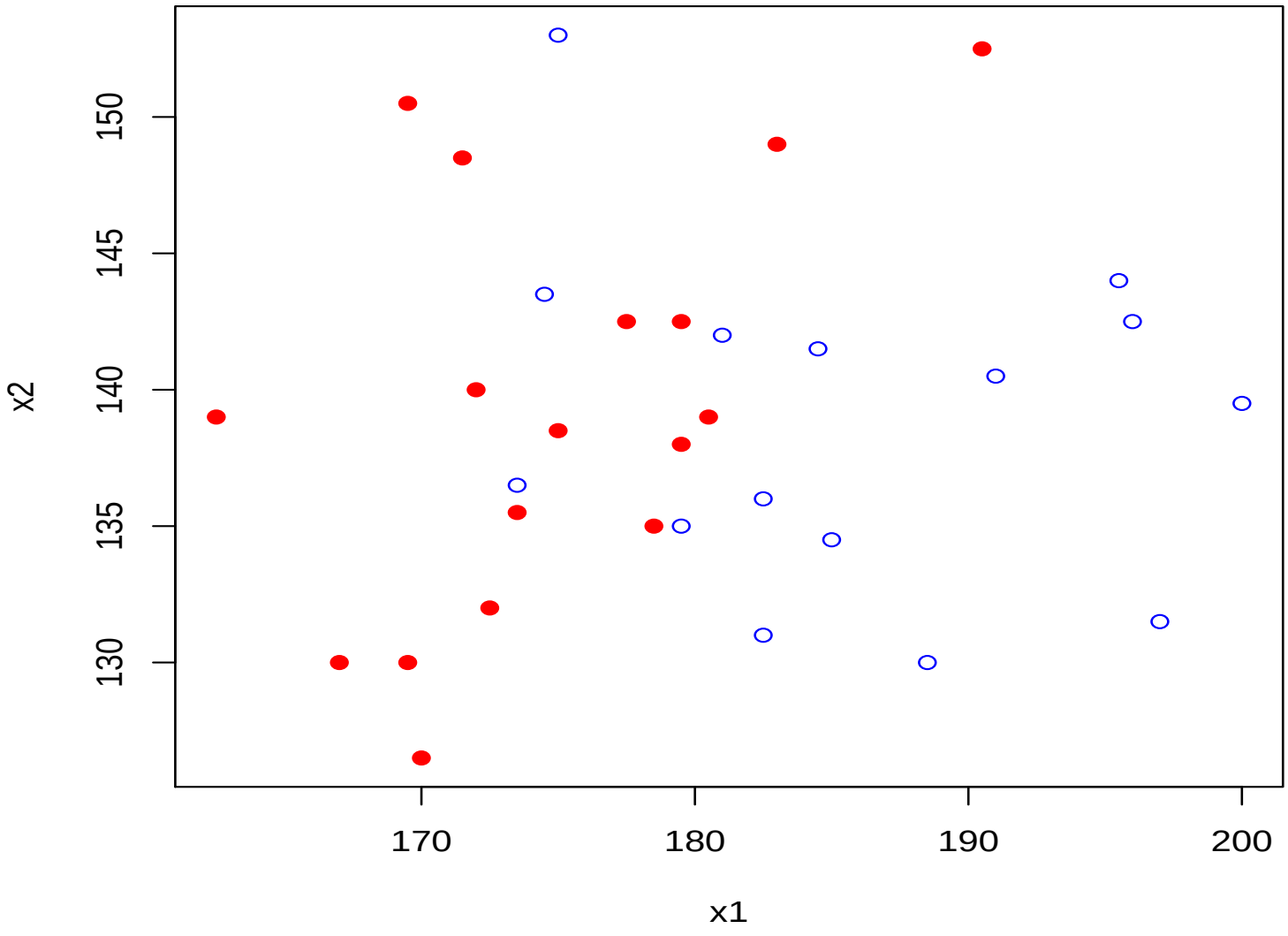


$$\mathbf{S}_1 = \begin{pmatrix} 0.7543 & -0.0146 \\ -0.0146 & 0.8924 \end{pmatrix}$$

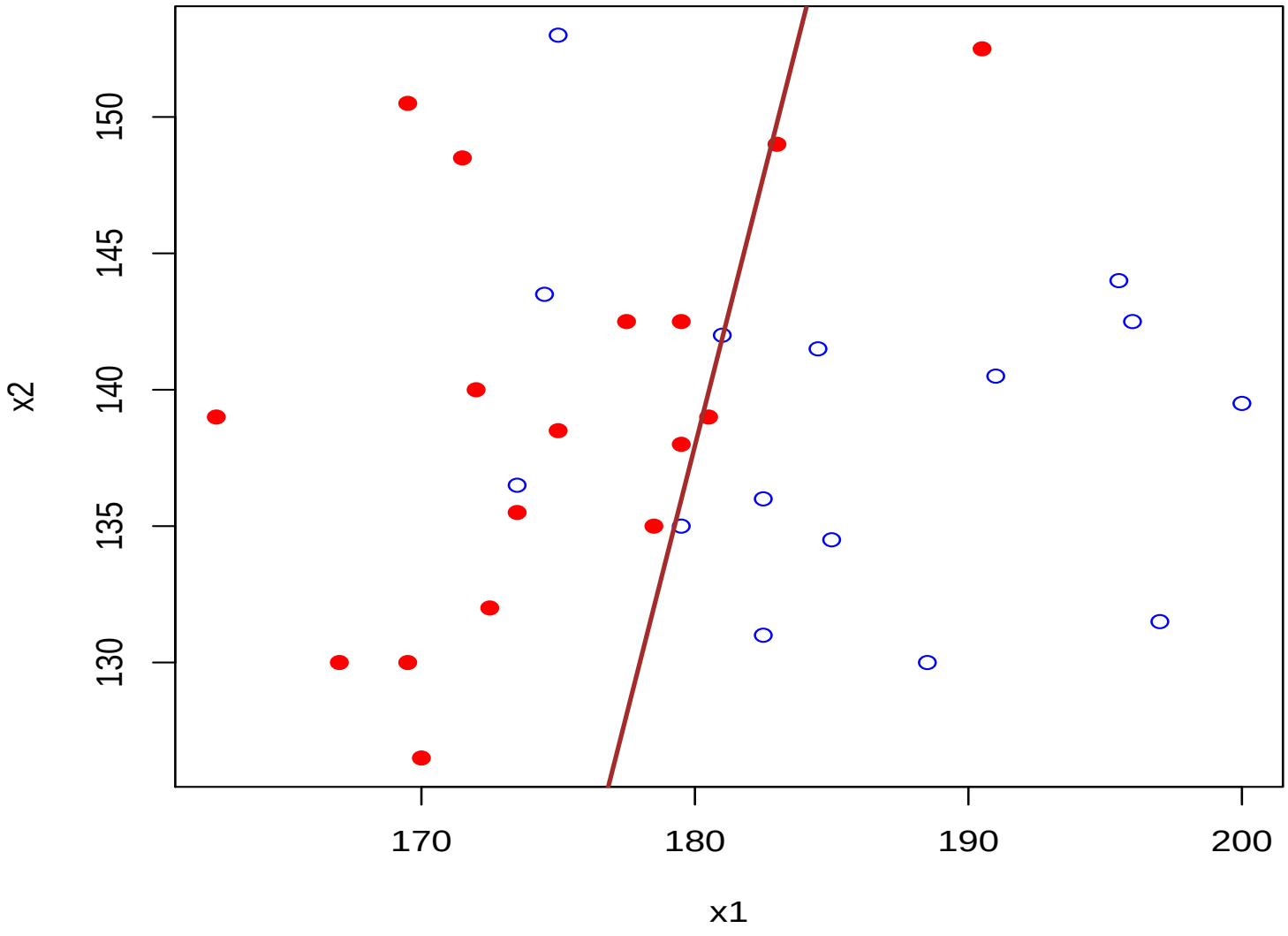
$$\mathbf{S}_2 = \begin{pmatrix} 0.9157 & -0.0041 \\ -0.0041 & 0.9955 \end{pmatrix}$$

$$\mathbf{S}_w = \begin{pmatrix} 0.8350 & -0.0093 \\ -0.0093 & 0.9440 \end{pmatrix}$$

Cráneos: x1 frente a x2



Función de discriminación lineal de Fisher



Datos de esclerosis

En un estudio sobre esclerosis múltiple se registran las respuestas a dos estímulos visuales diferentes (R1 y R2) del ojo izquierdo (I) y del ojo derecho (D) . Hay dos grupos formados por 29 individuos enfermos y 69 individuos sanos. Se registran las variables:

Edad	
R1suma	R1I + R1D
R1dif	|R1I - R1D|
R2suma	R2I + R2D
R2dif	|R2I - R2D|

Estadísticos descriptivos

PACIENTE		Media	Desv. típ.	N válido (según lista)	
				No ponderados	Ponderados
0	EDAD	37,9855	16,66230	69	69,000
	R1SUMA	147,2899	10,59692	69	69,000
	R1DIF	1,5623	1,34351	69	69,000
	R2SUMA	195,6029	13,60988	69	69,000
	R2DIF	1,6203	1,53475	69	69,000
1	EDAD	42,0690	11,00627	29	29,000
	R1SUMA	178,2690	29,06339	29	29,000
	R1DIF	12,2759	17,81191	29	29,000
	R2SUMA	236,9310	34,35160	29	29,000
	R2DIF	13,0828	18,73625	29	29,000
Total	EDAD	39,1939	15,26782	98	98,000
	R1SUMA	156,4571	22,90336	98	98,000
	R1DIF	4,7327	10,81701	98	98,000
	R2SUMA	207,8327	28,80996	98	98,000
	R2DIF	5,0122	11,42985	98	98,000

Matrices de covarianzas

Matrices de covarianza

PACIENTE		EDAD	R1SUMA	R1DIF	R2SUMA	R2DIF
0	EDAD	277,632	95,398	5,361	103,724	3,241
	R1SUMA	95,398	112,295	1,766	106,785	2,042
	R1DIF	5,361	1,766	1,805	2,235	,501
	R2SUMA	103,724	106,785	2,235	185,229	2,351
	R2DIF	3,241	2,042	,501	2,351	2,355
1	EDAD	121,138	52,795	-20,220	68,133	-29,820
	R1SUMA	52,795	844,681	244,463	912,415	106,764
	R1DIF	-20,220	244,463	317,264	232,365	297,319
	R2SUMA	68,133	912,415	232,365	1180,032	81,097
	R2DIF	-29,820	106,764	297,319	81,097	351,047

Matrices intra-grupo combinadas[a]

		EDAD	R1SUMA	R1DIF	R2SUMA	R2DIF
Covarianza	EDAD	231,988	82,972	-2,100	93,343	-6,402
	R1SUMA	82,972	325,907	72,553	341,760	32,586
	R1DIF	-2,100	72,553	93,814	69,356	87,073
	R2SUMA	93,343	341,760	69,356	475,380	25,319
	R2DIF	-6,402	32,586	87,073	25,319	104,057

a. La matriz de covarianza tiene 96 grados de libertad

Función discriminante lineal

$$\mathbf{w} = \mathbf{S}_w^{-1}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1) = (-0.023, 0.034, -0.210, 0.084, 0.253)'$$

Término independiente: $\mathbf{w}' \left(\frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2}\right) = 23.23$

Coeficientes de las funciones canónicas discriminantes

	Función
	1
EDAD	-,010
R1SUMA	,015
R1DIF	-,093
R2SUMA	,037
R2DIF	,112
(Constante)	-9,834

Coeficientes no tipificados

La relación entre la salida de SPSS ($\tilde{\mathbf{w}}$) y el valor del vector $\mathbf{w}$ es

$$\mathbf{w} = [\tilde{\mathbf{w}}'(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)]\tilde{\mathbf{w}} \approx 2.24\tilde{\mathbf{w}},$$

es decir, son proporcionales.

Regla de clasificación

Clasificamos a la observación $\mathbf{x}$ como individuo sano si $\mathbf{w}'\mathbf{x} < \mathbf{w}'\left(\frac{\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2}{2}\right) = 23.23$

X_1	X_2	X_3	X_4	X_5	Paciente/Control	$\mathbf{w}'\mathbf{x}$	Clasificado
18	152.0	1.6	198.4	0.0	0	21.13	0
19	138.0	0.4	180.8	1.6	0	19.80	0
20	144.0	0.0	186.4	0.8	0	20.34	0
20	143.6	3.2	194.8	0.0	0	20.15	0
20	148.8	0.0	217.6	0.0	0	22.92	0
23	148.0	0.8	205.4	0.6	1	23.50	1
25	195.2	3.2	262.8	0.4	1	21.78	0
25	158.0	8.0	209.8	12.2	1	27.62	1
28	134.4	0.0	198.4	3.2	1	23.88	1
29	190.2	14.2	243.8	10.6	1	21.44	0

Tasas de error

Para evaluar la calidad de la función discriminante se utiliza la información sobre las tasas de error:

Resultados de la clasificación[b,c]

			Grupo de pertenencia pronosticado		
		PACIENTE	0	1	Total
Original	Recuento	0	66	3	69
		1	7	22	29
	%	0	95,7	4,3	100,0
		1	24,1	75,9	100,0
Validación cruzada[a]	Recuento	0	64	5	69
		1	8	21	29
	%	0	92,8	7,2	100,0
		1	27,6	72,4	100,0

a. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

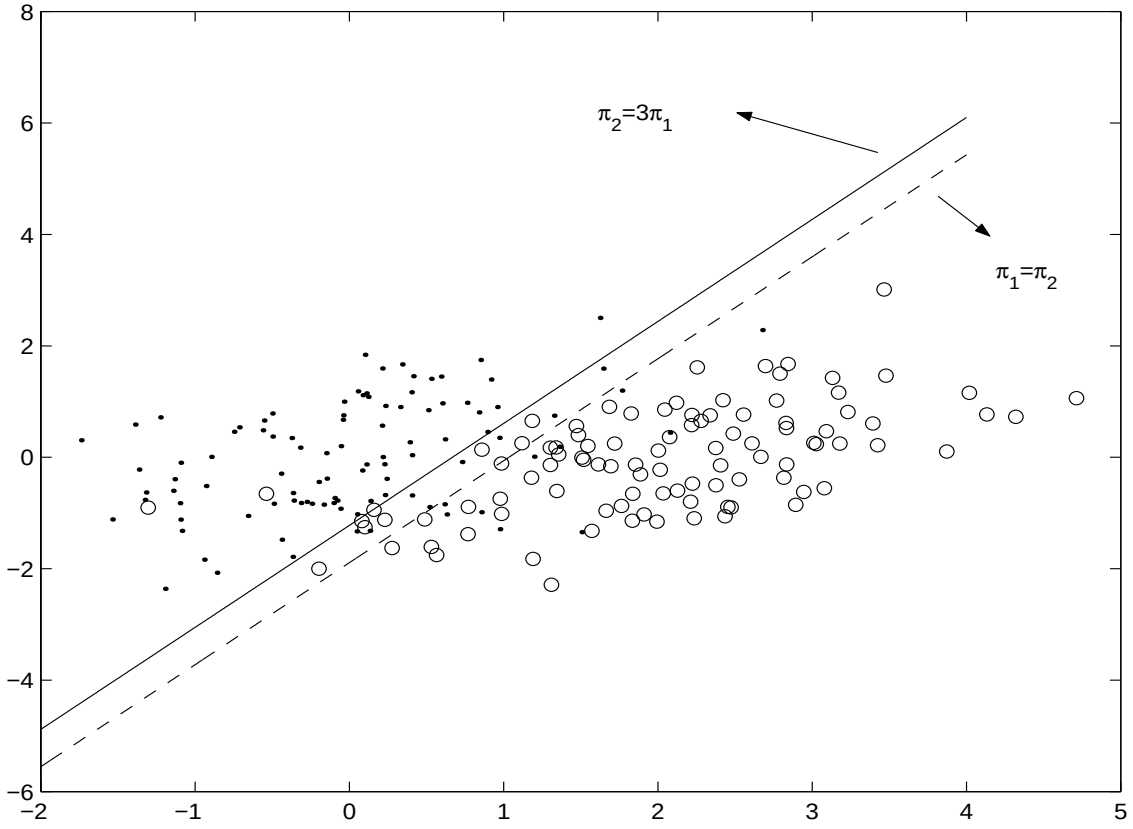
b. Clasificados correctamente el 89,8% de los casos agrupados originales.

c. Clasificados correctamente el 86,7% de los casos agrupados validados mediante validación cruzada.

Diferentes probabilidades a priori

Dos muestras de tamaño 100 de dos poblaciones normales:

- Vectores de medias $\mu_1 = (0, 0)'$, $\mu_2 = (2, 0)'$
- Las matrices de covarianzas valen $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \begin{pmatrix} 1 & 1/\sqrt{3} \\ 1/\sqrt{3} & 1 \end{pmatrix}$

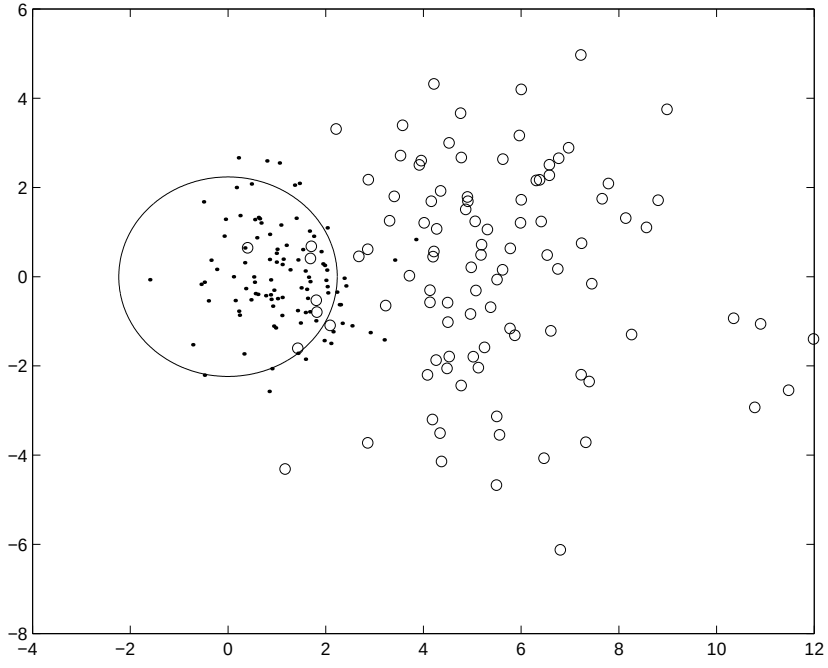


Matrices de covarianzas distintas

Dos muestras de tamaño 100 de dos poblaciones normales:

- Vectores de medias $\mu_1 = (1, 0)'$, $\mu_2 = (5, 0)'$
- Las matrices de covarianzas valen $\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ y $\mathbf{\Sigma}_2 = 5\mathbf{\Sigma}_1$

Si los valores verdaderos de los parámetros son conocidos la función discriminante es $x_1^2 + x_2^2 = 5$. En la práctica tendremos que estimarlos.



Más de dos grupos

Tres muestras de tamaño 100 de tres poblaciones normales:

- Vectores de medias $\mu_1 = (0,0)'$, $\mu_2 = (4,0)'$ y $\mu_1 = (4,4)'$
- Las tres matrices de covarianzas valen $\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$

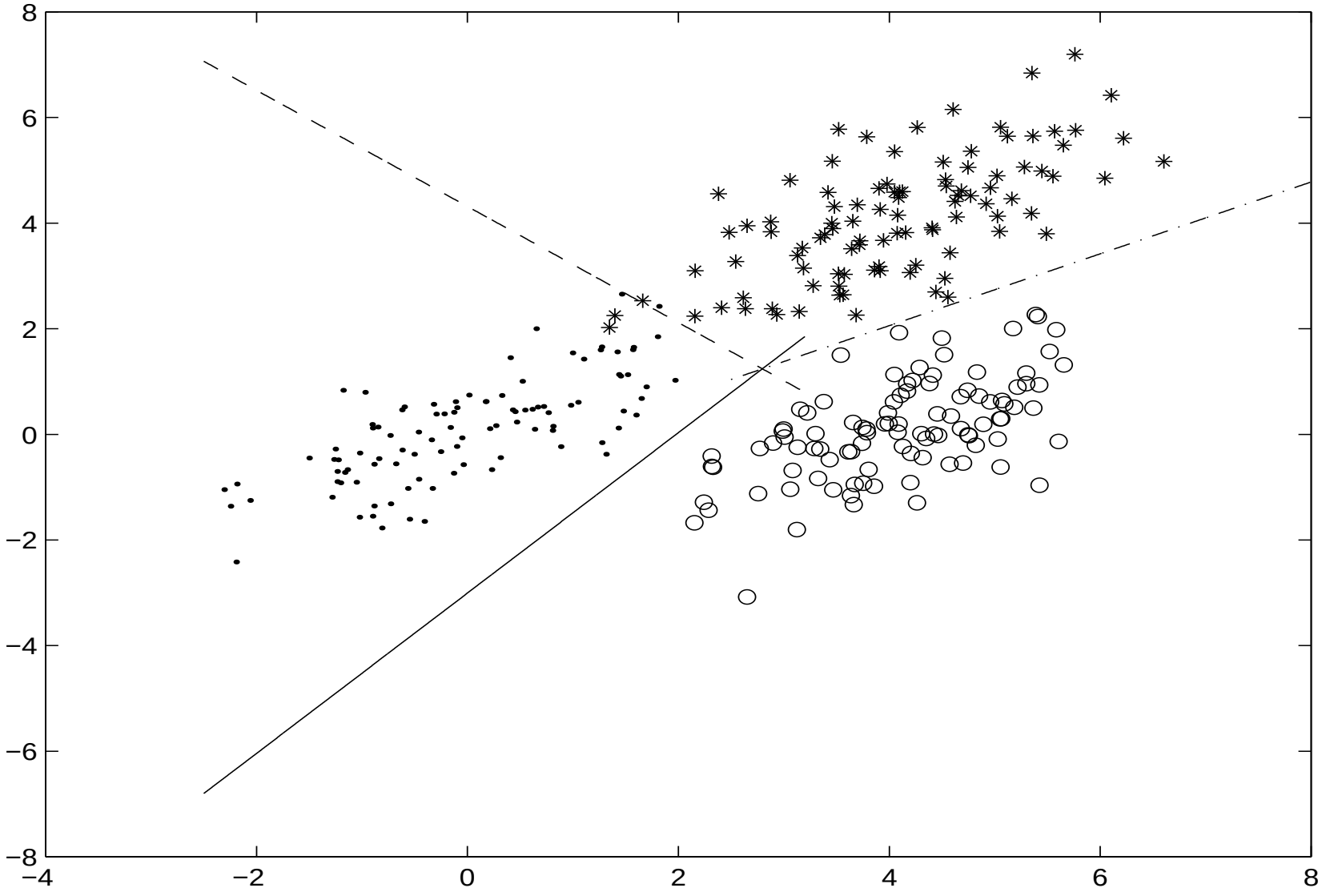
Ahora hay tres funciones discriminantes:

$$\mathbf{w}_{12}'\mathbf{x} = \mathbf{w}_{12}' \left(\frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2} \right), \quad \text{donde } \mathbf{w}_{12} = \mathbf{S}_w^{-1}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1).$$

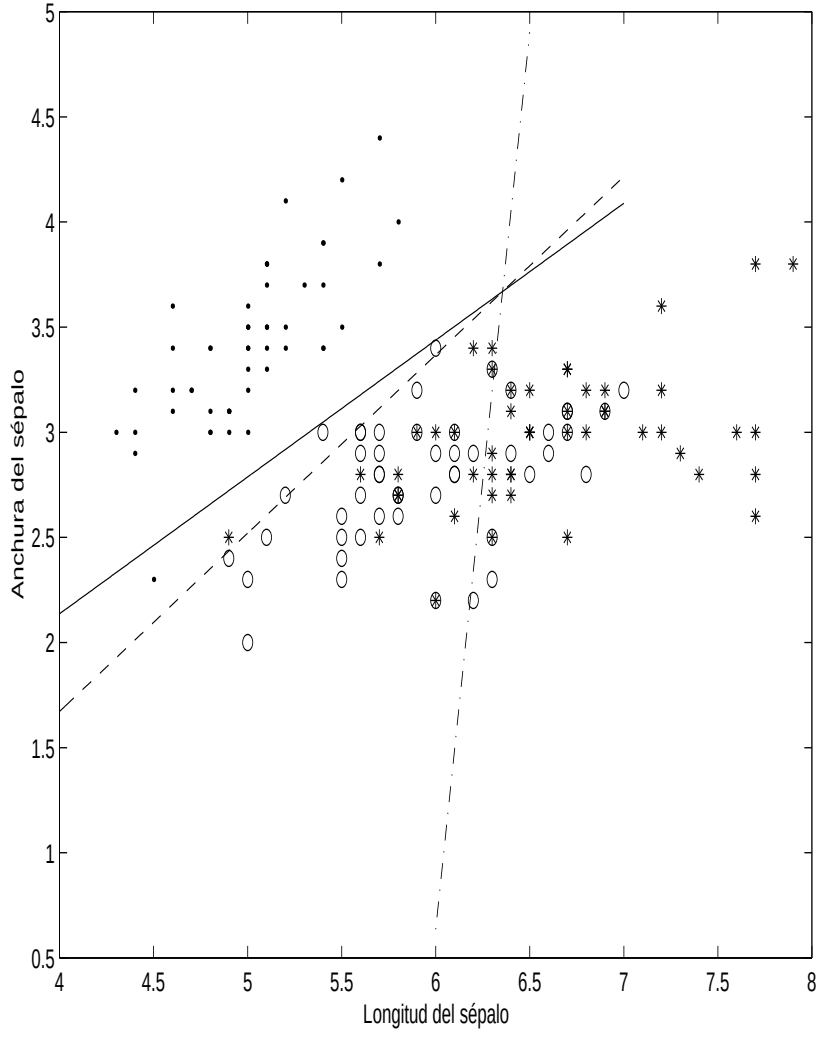
$$\mathbf{w}_{13}'\mathbf{x} = \mathbf{w}_{13}' \left(\frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_3}{2} \right), \quad \text{donde } \mathbf{w}_{13} = \mathbf{S}_w^{-1}(\bar{\mathbf{x}}_3 - \bar{\mathbf{x}}_1).$$

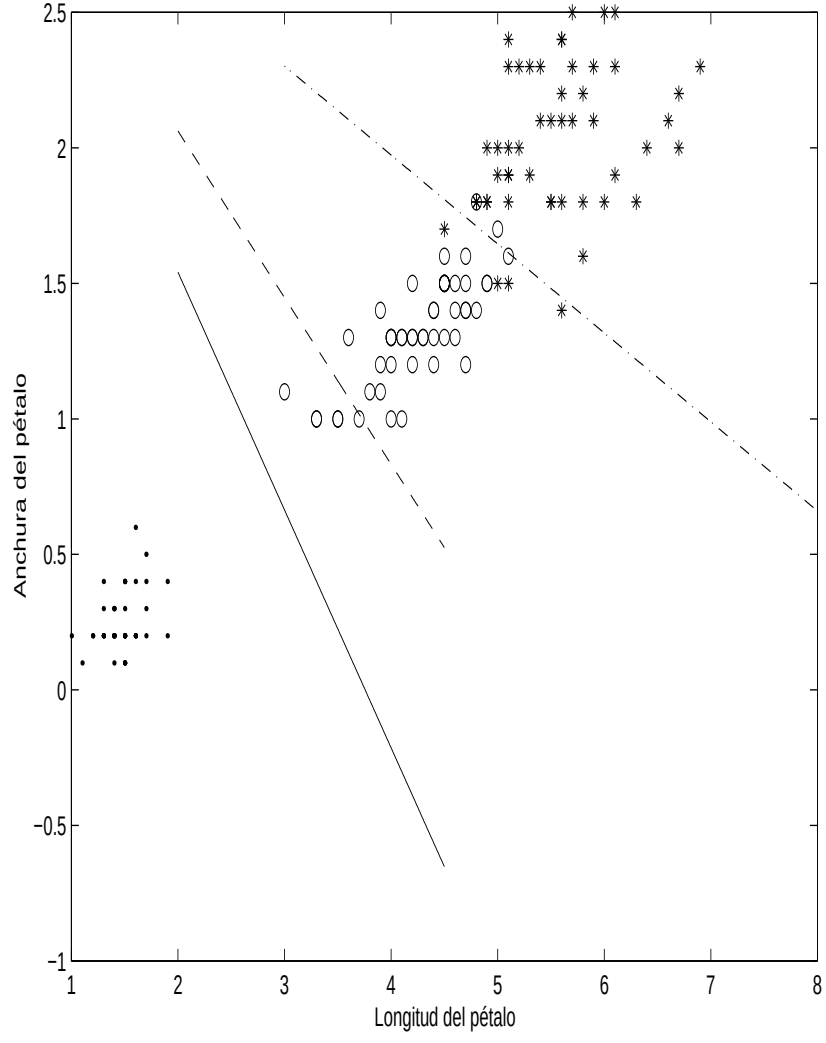
$$\mathbf{w}_{23}'\mathbf{x} = \mathbf{w}_{23}' \left(\frac{\bar{\mathbf{x}}_2 + \bar{\mathbf{x}}_3}{2} \right), \quad \text{donde } \mathbf{w}_{23} = \mathbf{S}_w^{-1}(\bar{\mathbf{x}}_3 - \bar{\mathbf{x}}_2).$$

Representación gráfica



Método de Mahalanobis (variables 2 a 2)





Método de Fisher (las 4 variables)

Coeficientes de las funciones canónicas discriminantes

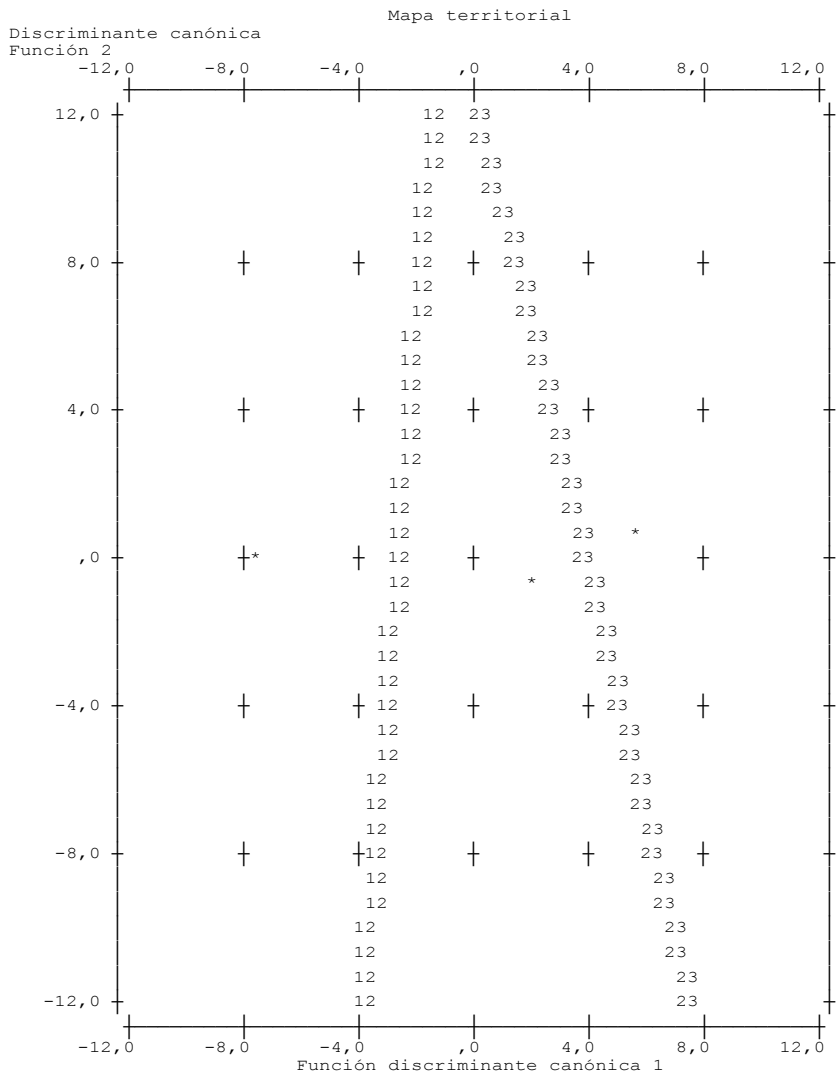
	Función	
	1	2
SL	-,819	,033
SW	-1,548	2,155
PL	2,185	-,930
PW	2,854	2,806
(Constante)	-2,119	-6,639

Coeficientes no tipificados

Funciones en los centroides de los grupos

	Función	
ESPECIE	1	2
1,00	-7,616	,213
2,00	1,822	-,718
3,00	5,793	,505

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos



Resultados

Estadísticos por casos

		Puntuaciones discriminantes	
	Número de casos	Función 1	Función 2
Original	1	-9,869	1,615
	2	-9,496	1,850
	3	-9,180	2,756
	4	-9,162	1,251
	5	-8,704	,900
Validación cruzada[a]	1		
	2		
	3		
	4		
	5		

Para los datos originales, la distancia de Mahalanobis al cuadrado se basa en las funciones canónicas.
Para los datos validados mediante validación cruzada, la distancia de Mahalanobis al cuadrado se basa en las observaciones.

a. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

Resultados de la clasificación[b,c]

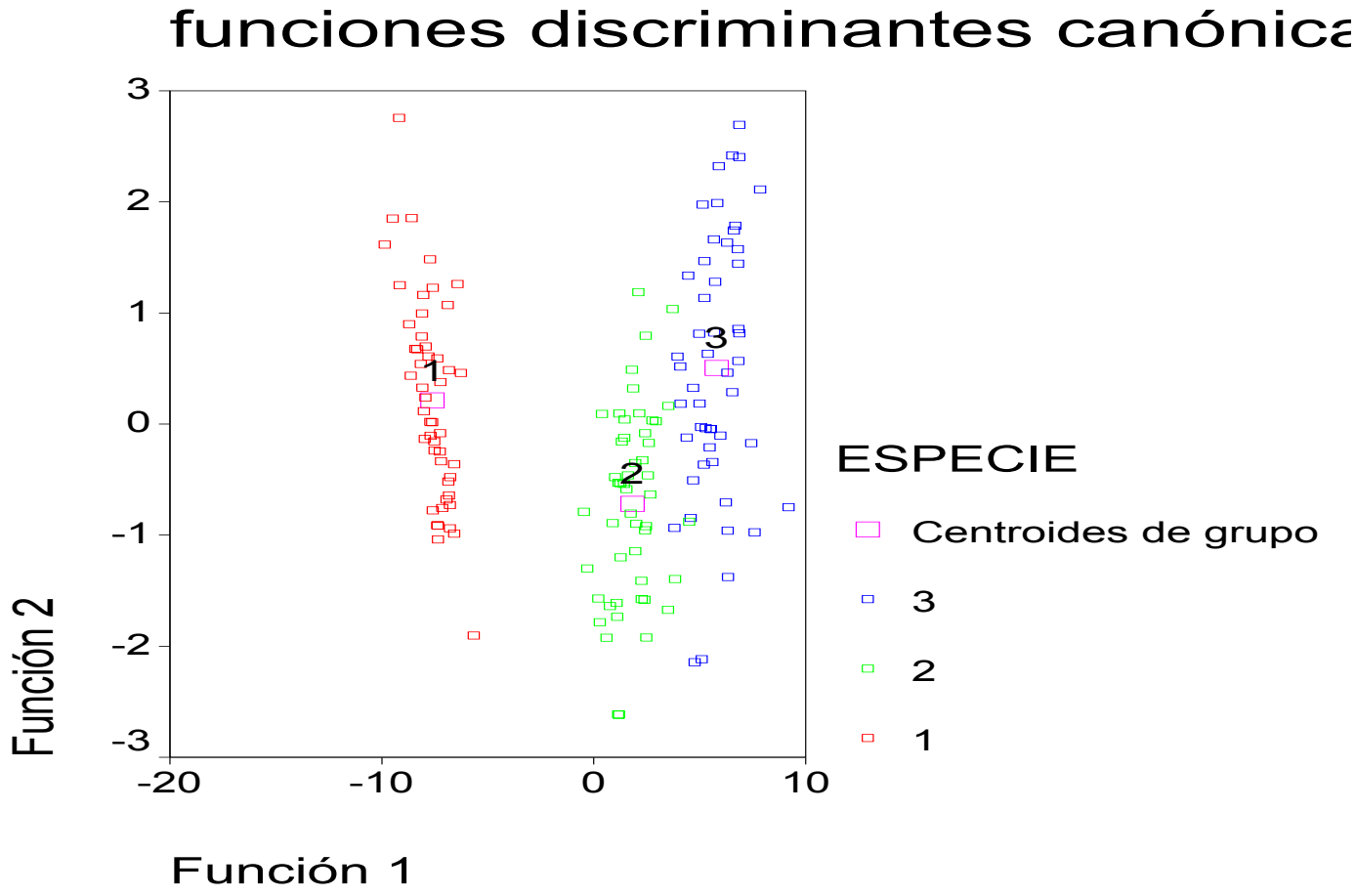
			Grupo de pertenencia pronosticado			
		ESPECIE	1,00	2,00	3,00	Total
Original	Recuento	1,00	50	0	0	50
		2,00	0	48	2	50
		3,00	0	1	49	50
	%	1,00	100,0	,0	,0	100,0
		2,00	,0	96,0	4,0	100,0
		3,00	,0	2,0	98,0	100,0
Validación cruzada[a]	Recuento	1,00	50	0	0	50
		2,00	0	48	2	50
		3,00	0	1	49	50
	%	1,00	100,0	,0	,0	100,0
		2,00	,0	96,0	4,0	100,0
		3,00	,0	2,0	98,0	100,0

a. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

b. Clasificados correctamente el 98,0% de los casos agrupados originales.

c. Clasificados correctamente el 98,0% de los casos agrupados validados mediante validación cruzada.

Puntuaciones discriminantes



Ejemplo: gorriones

- Tras una fuerte tormenta en febrero de 1898, un grupo de gorriones moribundos fueron llevados a la Universidad Brown (Rhode Island).
- Alrededor de la mitad de los gorriones murieron. Se consideró la situación como una oportunidad de estudiar el efecto de la selección natural sobre los pájaros.
- Se tomaron diversas medidas morfológicas, de las que se incluyen 5 en el fichero.

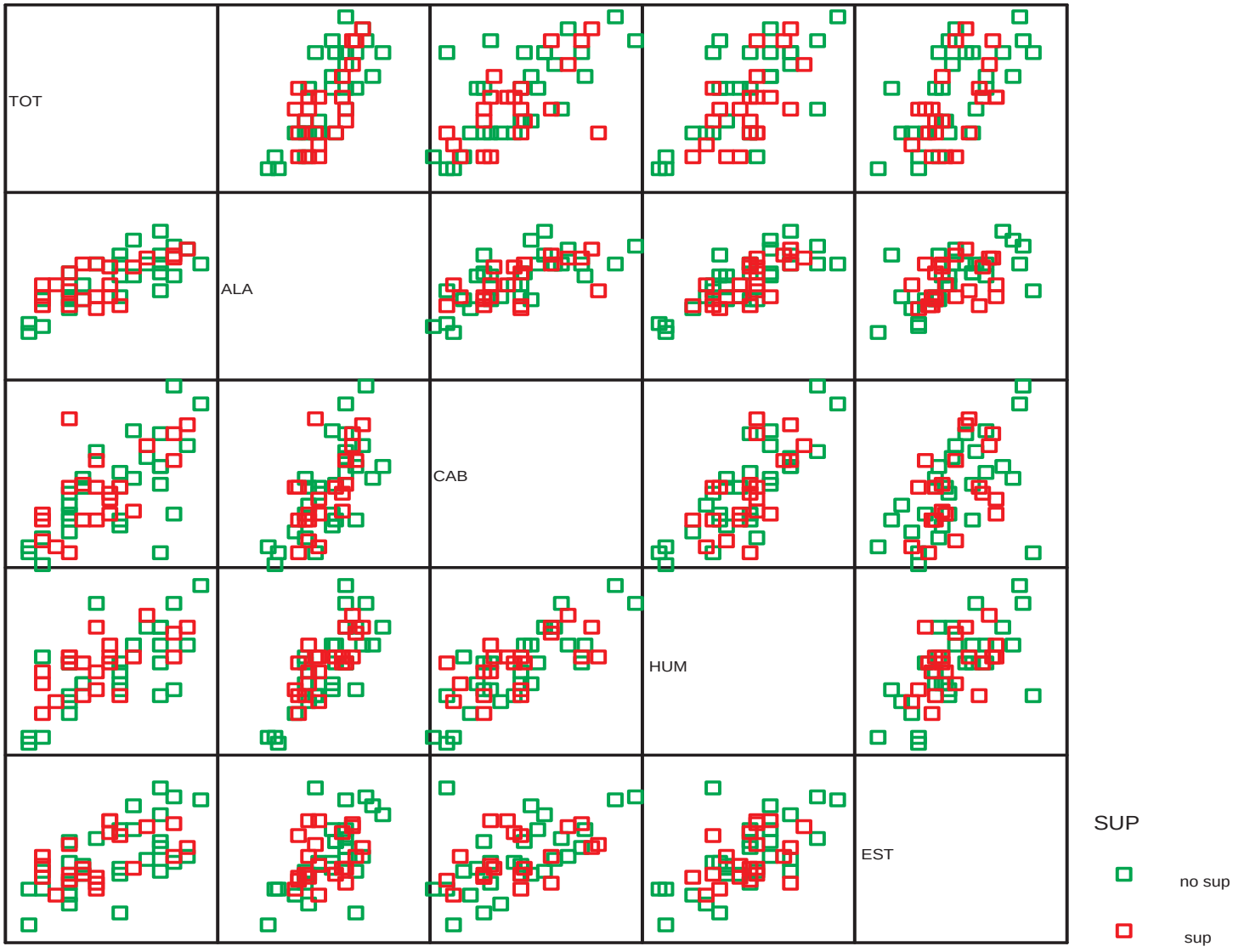
Variables

Nombre variable	Descripción
`TOT`	Longitud total
`ALA`	Extensión de las alas
`CAB`	Longitud del pico y la cabeza
`HUM`	Longitud del húmero
`EST`	Longitud del esternón

Observaciones:

- Todas las variables se miden en mm.
- El fichero contiene datos de 49 gorriones.
- Los 21 primeros gorriones fueron los supervivientes.

Matriz de diagramas de dispersión



Análisis discriminante lineal de Fisher

Veamos hasta qué punto es posible discriminar entre supervivientes y no supervivientes mediante un análisis discriminante lineal de Fisher.
La salida de SPSS que se muestra se ha obtenido utilizando

- Marcando en `Estadísticos` las opciones: Medias, M de Box, coeficientes no tipificados, covarianza intra-grupos, covarianzas de grupos separados.
- Marcando en `Clasificar...` las opciones: Probabilidades previas todos los grupos iguales, usar matriz de covarianza intra-grupos, mostrar tabla de resumen, mostrar clasificación dejando uno fuera.

Medias por grupos

Estadísticos de grupo

SUP		Media	Desv. típ.	N válido (según lista)	
				No ponderados	Ponderados
,00	TOT	157,3810	3,32380	21	21,000
	ALA	241,0000	4,18330	21	21,000
	CAB	31,4333	,72893	21	21,000
	HUM	18,5000	,41952	21	21,000
	EST	20,8095	,75822	21	21,000
1,00	TOT	158,4286	3,88185	28	28,000
	ALA	241,5714	5,70528	28	28,000
	CAB	31,4786	,85347	28	28,000
	HUM	18,4464	,65911	28	28,000
	EST	20,8393	1,14934	28	28,000
Total	TOT	157,9796	3,65428	49	49,000
	ALA	241,3265	5,06782	49	49,000
	CAB	31,4592	,79475	49	49,000
	HUM	18,4694	,56429	49	49,000
	EST	20,8265	,99137	49	49,000

Estimación de la matriz de covarianzas

Matrices de covarianza

SUP		TOT	ALA	CAB	HUM	EST
,00	TOT	11,048	9,100	1,557	,870	1,286
	ALA	9,100	17,500	1,910	1,310	,880
	CAB	1,557	1,910	,531	,189	,240
	HUM	,870	1,310	,189	,176	,133
	EST	1,286	,880	,240	,133	,575
1,00	TOT	15,069	17,190	2,243	1,746	2,931
	ALA	17,190	32,550	3,398	2,950	4,066
	CAB	2,243	3,398	,728	,470	,559
	HUM	1,746	2,950	,470	,434	,506
	EST	2,931	4,066	,559	,506	1,321

Matrices intra-grupo combinadas[a]

		TOT	ALA	CAB	HUM	EST
Covarianza	TOT	13,358	13,748	1,951	1,373	2,231
	ALA	13,748	26,146	2,765	2,252	2,710
	CAB	1,951	2,765	,645	,350	,423
	HUM	1,373	2,252	,350	,324	,347
	EST	2,231	2,710	,423	,347	1,004

a. La matriz de covarianza tiene 47 grados de libertad

¿Es aceptable la hipótesis $\Sigma_1 = \Sigma_2$?

Para contrastar $H_0 : \Sigma_1 = \Sigma_2$ se utiliza el contraste M de Box. En este ejemplo,

Resultados de la prueba

M de Box		11,786
F	Aprox.	,692
	gl1	15
	gl2	7429,395
	Sig.	,795

Contrasta la hipótesis nula de que las matrices de covarianza poblacionales son iguales.

El p-valor es bastante grande, 0.795, por lo que no existe evidencia en los datos para rechazar la hipótesis nula. Podemos suponer que las matrices de covarianzas de los dos grupos son iguales.

Función discriminante lineal

Coeficientes de las funciones canónicas discriminantes

	Función
	1
TOT	,320
ALA	,055
CAB	,191
HUM	-2,128
EST	-,143
(Constante)	-27,499

Coeficientes no tipificados

Funciones en los centroides de los grupos

	Función
SUP	1
,00	-,277
1,00	,208

Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos

Cómo clasificar una nueva observación

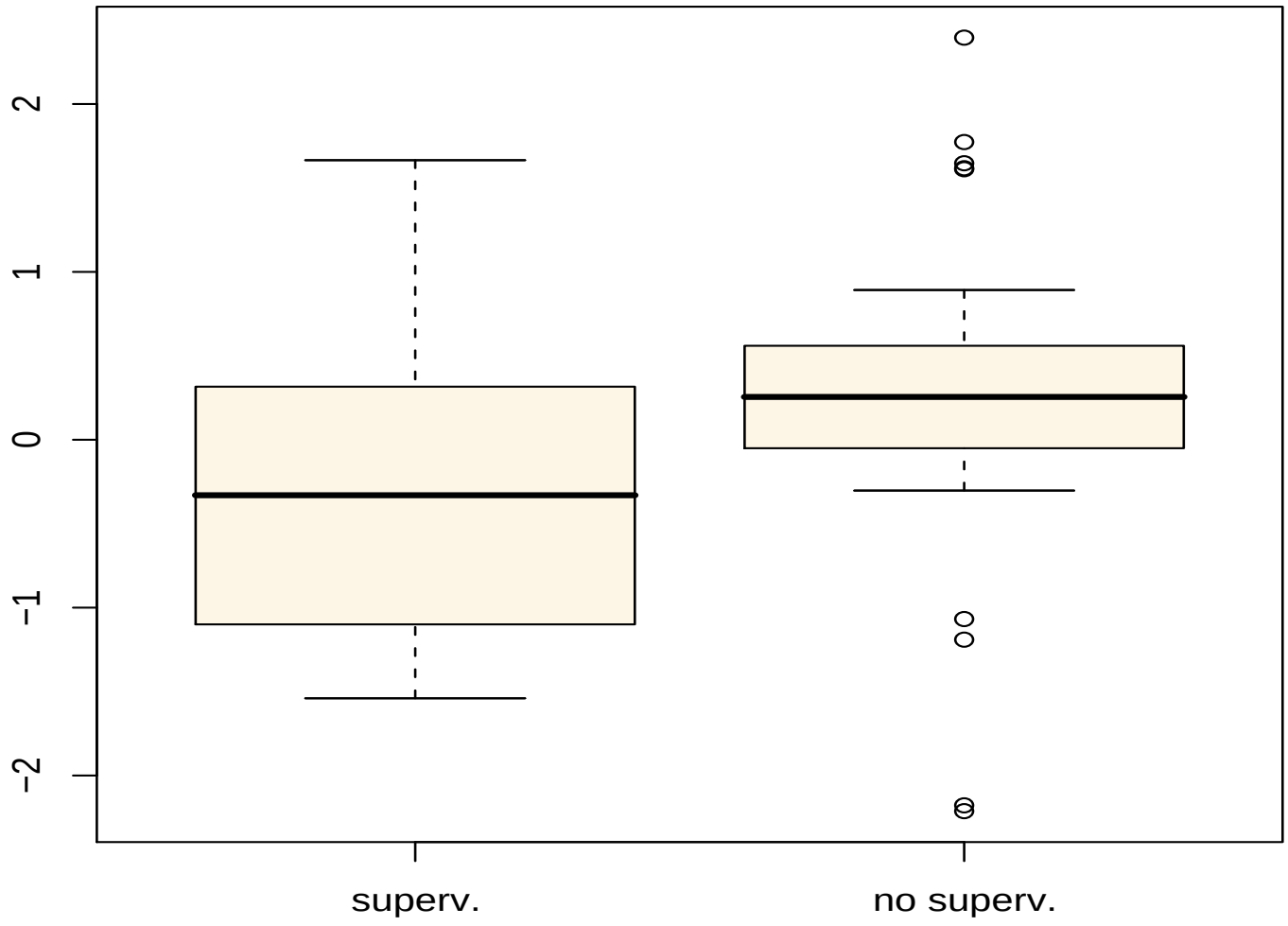
Supongamos que, a partir de la salida anterior, queremos clasificar una nueva observación x_0 tal que $x_0 = (156, 245, 31.6, 18.5, 20.5)$
La forma más sencilla consiste en calcular la puntuación discriminante de x_0 (teniendo en cuenta el término independiente):

$$\tilde{w}' x_0 + \tilde{w}_0 = 0.32\texttt{TOT} + 0.055\texttt{ALA} + \cdots - 0.143\texttt{EST} - 27.499$$

El número obtenido se compara con las `funciones en los centroides de los grupos` y se asigna al grupo más cercano.

En el ejemplo, $\tilde{w}' x_0 + \tilde{w}_0 \approx -0.42$, que está más cerca de -0.277 que de 0.208, por lo que clasificamos x_0 entre los supervivientes.

Puntuaciones discriminantes



Tasas de error

Resultados de la clasificación[b,c]

			Grupo de pertenencia pronosticado		
		SUP	,00	1,00	Total
Original	Recuento	,00	13	8	21
		1,00	9	19	28
	%	,00	61,9	38,1	100,0
		1,00	32,1	67,9	100,0
Validación cruzada[a]	Recuento	,00	10	11	21
		1,00	16	12	28
	%	,00	47,6	52,4	100,0
		1,00	57,1	42,9	100,0

a. La validación cruzada sólo se aplica a los casos del análisis. En la validación cruzada, cada caso se clasifica mediante las funciones derivadas a partir del resto de los casos.

b. Clasificados correctamente el 65,3% de los casos agrupados originales.

c. Clasificados correctamente el 44,9% de los casos agrupados validados mediante validación cruzada.

Tasas de error

Las tasas de error no son buenas.

Por validación cruzada el porcentaje de errores supera el 50%

Se debe a que los dos grupos están muy mezclados. Las variables que se han medido no tienen mucha capacidad para discriminar entre supervivientes y no supervivientes, al menos mediante funciones lineales.