Coordenadas Discriminantes y Discriminación

Graciela Boente

Ejemplo

Variables medidas sobre árboles de manzana de 6 injertos. Para cada injerto hay 8 árboles. Las variables son:
x_1 =Diámetro del tronco a los 4 años en unidades de 10cm,
x_2 =Largo a los 4 años,
x_3 =Diámetro del tronco a los 15 años en unidades de 10cm,
x_4 =Peso del árbol a los 15 años, en unidades de 1000 libras.

Inj.	1	1	1	1	1	1	1	1	2	2	2	2
x_1	1.11	1.19	1.09	1.25	1.11	1.08	1.11	1.16	1.05	1.17	1.11	1.25
x_2	2.569	2.928	2.865	3.844	3.027	2.336	3.211	3.037	2.074	2.885	3.378	3.906
x_3	3.58	3.75	3.93	3.94	3.60	3.51	3.98	3.62	4.09	4.06	4.87	4.98
x_4	0.760	0.821	0.928	1.009	0.766	0.726	1.209	0.750	1.036	1.094	1.635	1.517
Inj.	2	2	2	2	3	3	3	3	3	3	3	3
x_1	1.17	1.15	1.17	1.19	1.07	0.99	1.06	1.02	1.15	1.20	1.20	1.17
x_2	2.782	3.018	3.383	3.447	2.505	2.315	2.667	2.390	3.021	3.085	3.308	3.231
x_3	4.38	4.65	4.69	4.40	3.76	4.44	4.38	4.67	4.48	4.78	4.57	4.56
x_4	1.197	1.244	1.495	1.026	0.912	1.398	1.197	1.613	1.476	1.571	1.506	1.458
Inj.	4	4	4	4	4	4	4	4	5	5	5	5
x_1	1.22	1.03	1.14	1.01	0.99	1.11	1.20	1.08	0.91	1.15	1.14	1.05
x_2	2.838	2.351	3.001	2.439	2.199	3.318	3.601	3.291	1.532	2.552	3.083	2.330
x_3	3.89	4.05	4.05	3.92	3.27	3.95	4.27	3.85	4.04	4.16	4.79	4.42
x_4	0.944	1.241	1.023	1.067	0.693	1.085	1.242	1.017	1.084	1.151	1.381	1.242
Inj.	5	5	5	5	6	6	6	6	6	6	6	6
x_1	0.99	1.22	1.05	1.13	1.11	0.75	1.05	1.02	1.05	1.07	1.13	1.11
x_2	2.079	3.366	2.416	3.100	2.813	0.840	2.199	2.132	1.949	2.251	3.064	2.469
x_3	3.47	4.41	4.64	4.57	3.76	3.14	3.75	3.99	3.34	3.21	3.63	3.95
x_4	0.673	1.137	1.455	1.325	0.800	0.606	0.790	0.853	0.610	0.562	0.707	0.952

Ejemplo

Hemos estudiado si las medias de los Injertos 1, 2 y 3 eran iguales, o sea, testeamos $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$.
Para ello supusimos que las matrices de covarianza eran iguales.

$\overline{\mathbf{x}}_1 = (1.1375, 2.9771, 3.7388, 0.8711)^{\mathrm{T}}$
$\overline{\mathbf{x}}_2 = (1.1575, 3.1091, 4.5150, 1.2805)^{\mathrm{T}}$
$\overline{\mathbf{x}}_3 = (1.1075, 2.8152, 4.4550, 1.3914)^{\mathrm{T}}$
$\overline{\mathbf{x}} = (1.1342, 2.9672, 4.2363, 1.1810)^{\mathrm{T}}$

Las matrices de covarianza estimadas son

$$\mathbf{S}_1 = \begin{pmatrix} 0.0034 & 0.0203 & 0.0037 & 0.0018 \\ 0.0203 & 0.2007 & 0.0580 & 0.0458 \\ 0.0037 & 0.0580 & 0.0352 & 0.0285 \\ 0.0018 & 0.0458 & 0.0285 & 0.0283 \end{pmatrix} \quad \mathbf{S}_2 = \begin{pmatrix} 0.0034 & 0.0258 & 0.0088 & 0.0032 \\ 0.0258 & 0.3048 & 0.1498 & 0.0832 \\ 0.0088 & 0.1498 & 0.1157 & 0.0711 \\ 0.0032 & 0.0832 & 0.0711 & 0.0565 \end{pmatrix}$$

$$\mathbf{S}_3 = \begin{pmatrix} 0.0068 & 0.0314 & 0.0087 & 0.0060 \\ 0.0314 & 0.1543 & 0.0480 & 0.0329 \\ 0.0087 & 0.0480 & 0.0951 & 0.0680 \\ 0.0060 & 0.0329 & 0.0680 & 0.0534 \end{pmatrix}$$

Ejemplo

Construímos

$$\mathbf{U} = \mathbf{Q}_1 + \mathbf{Q}_2 + \mathbf{Q}_3 = \sum_{i=1}^{3}\sum_{j=1}^{n_i}(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}}$$

$$\mathbf{H} = \sum_{i=1}^{k} n_i(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^{\mathrm{T}}$$

obteniendo

$$\mathbf{U} = \begin{pmatrix} 0.0956 & 0.5422 & 0.1490 & 0.0771 \\ 0.5422 & 4.6184 & 1.7911 & 1.1326 \\ 0.1490 & 1.7911 & 1.7221 & 1.1731 \\ 0.0771 & 1.1326 & 1.1731 & 0.9670 \end{pmatrix} \qquad \mathbf{H} = \begin{pmatrix} 0.0101 & 0.0592 & -0.0079 & -0.0346 \\ 0.0592 & 0.3466 & 0.0111 & -0.1674 \\ -0.0079 & 0.0111 & 2.9845 & 1.8233 \\ -0.0346 & -0.1674 & 1.8233 & 1.2014 \end{pmatrix}$$

El estadístico para testear $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$ era

$$V = \frac{|\mathbf{U}|}{|\mathbf{U} + \mathbf{H}|} = \Lambda(23, 4, 2)$$

y rechazamos H_0.

Ejemplo

En el caso de dos poblaciones, nosotros vimos que la dirección a la que se atribuía la responsabilidad del rechazo era

$$\widehat{\boldsymbol{\alpha}} = \mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$$

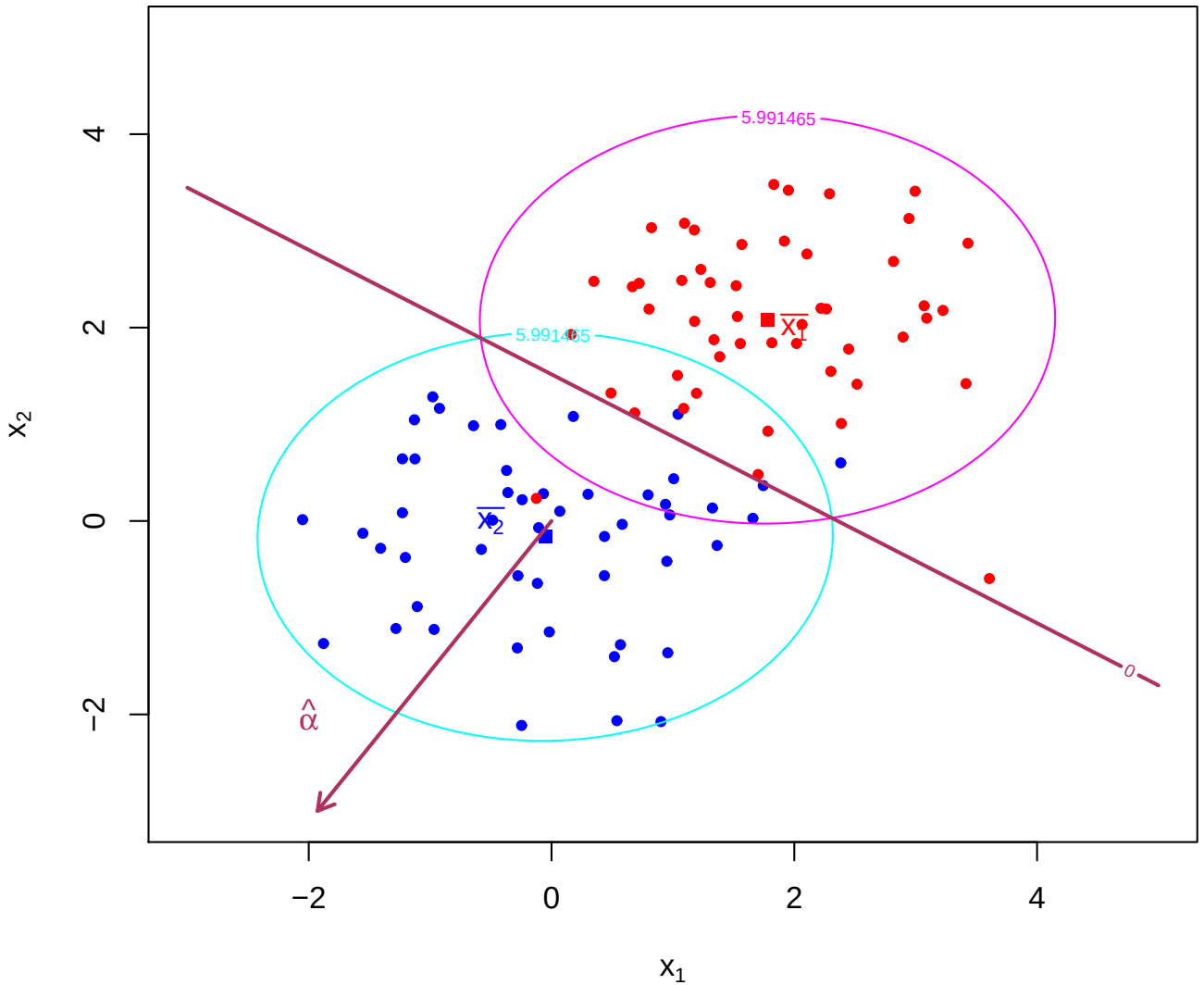
con $\mathbf{S} = ((n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2)/(n_1 + n_2 - 2)$

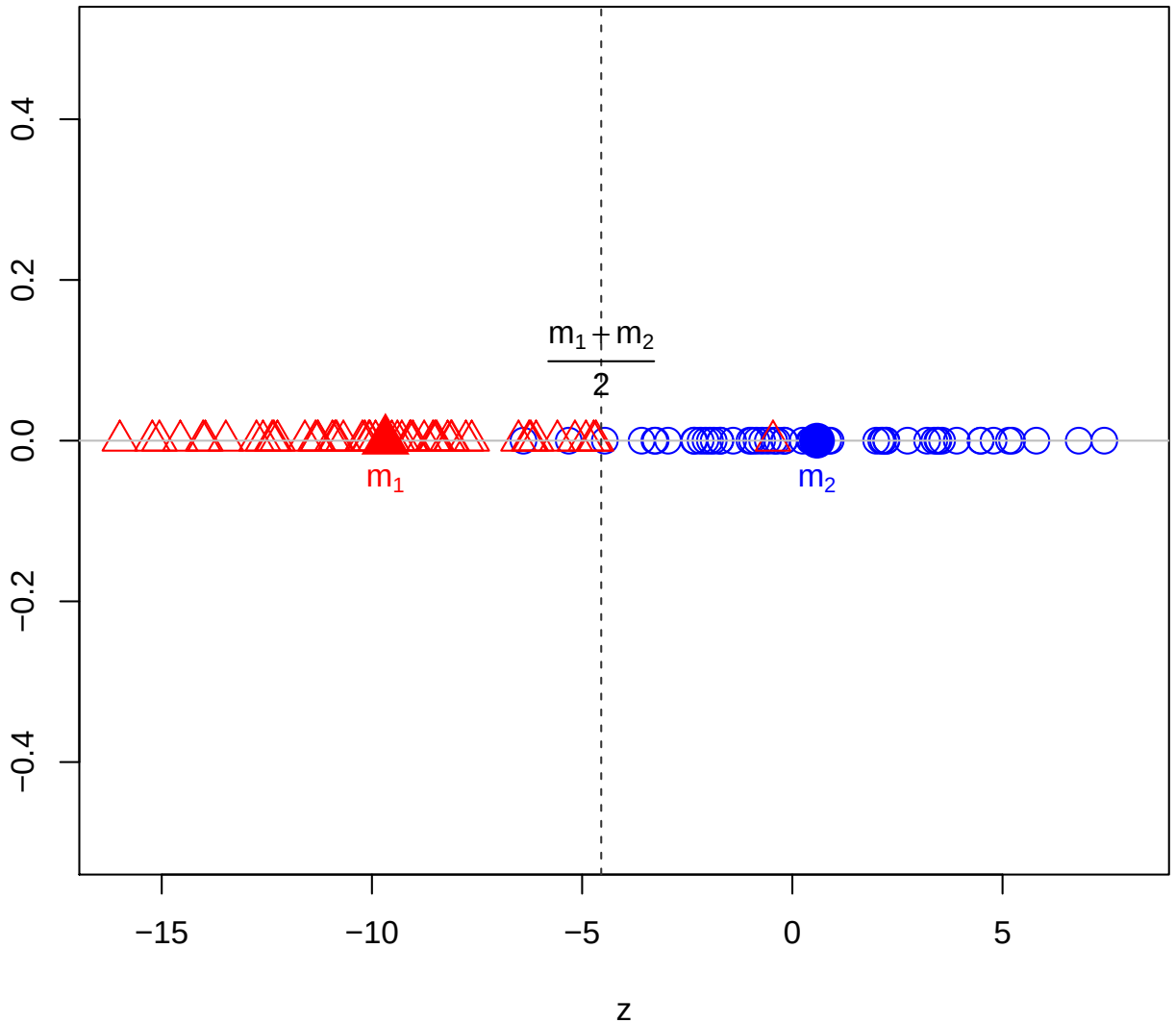
- La función $H(\mathbf{u}) = \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{u}$ es la función discriminante lineal.
- Asignábamos $\mathbf{u}$ a la población 1 si

$$z = \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{u} > \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\left(\frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) = \frac{m_1 + m_2}{2}$$

o equivalentemente, si $z - m_2 > m_1 - z$

Observemos que $m_1 - m_2 > 0$, luego, esto es análogo a clasificar $\mathbf{u}$ en aquella población donde la distancia $|z - m_j|$ sea mínima.


5.991465
5.991465
$\overline{x}_1$
$\overline{x}_2$
$\hat{\alpha}$
0
x_1
x_2
−2
0
2
4
−2
0
2
4


0.4
0.2
0.0
−0.2
−0.4
$\frac{m_1 + m_2}{2}$
m_1
m_2
−15
−10
−5
0
5
z

Propiedad

Sea $\mathbf{x} \in \mathbb{R}^p$ un vector aleatorio y G una variable aleatoria que indica la pertencia al grupo, tales que para $1 \leq i \leq k$

$$\mathbb{P}(G = i) = \pi_i \qquad \mathbb{E}(\mathbf{x}|G = i) = \boldsymbol{\mu}_i \qquad \textsc{Var}(\mathbf{x}|G = i) = \boldsymbol{\Sigma}_i$$

entonces si $\overline{\boldsymbol{\mu}} = \mathbb{E}(\mathbf{x})$ y $\overline{\boldsymbol{\Sigma}} = \textsc{Var}(\mathbf{x})$ se cumple que

$$\overline{\boldsymbol{\mu}} = \sum_{j=1}^{k} \pi_j \boldsymbol{\mu}_j \qquad \overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{\mathrm{W}} + \boldsymbol{\Sigma}_{\mathrm{B}}$$

donde

$$\boldsymbol{\Sigma}_{\mathrm{W}} = \sum_{i=1}^{k} \pi_i \boldsymbol{\Sigma}_i \qquad \boldsymbol{\Sigma}_{\mathrm{B}} = \sum_{i=1}^{k} \pi_i (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^{\mathrm{T}}$$

El Problema

Sea $z = \mathbf{a}^{\mathrm{T}}\mathbf{x}$, luego

$$\mathrm{VAR}(z) = \mathbf{a}^{\mathrm{T}}\mathrm{VAR}(\mathbf{x})\mathbf{a} = \mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{W}}\mathbf{a} + \mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{a}$$

es decir, descompusimos a la varianza de z en una componente que mide la variabilidad *dentro* de grupos y otra que mide la variablidad *entre* grupos.

Nos interesan combinaciones lineales, o sea, vectores $\mathbf{a}$ tales que la varianza de z es mucho más grande que la varianza *dentro* de grupos ya que de esto indica que la variabilidad dentro de grupos se ve aumentada por diferencias en posición.

Como $\overline{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{\mathrm{W}} + \boldsymbol{\Sigma}_{\mathrm{B}}$, nos interesarán direcciones $\mathbf{a}$ que maximizan

$$F_{\mathbf{a}} = \frac{\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{W}}\mathbf{a}}$$

Caso $k = 2$

Si $k = 2$

$$\mathbf{\Sigma}_{\mathrm{B}} = \pi_1 \pi_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}$$

luego basta maximizar

$$\frac{(\mathbf{a}^{\mathrm{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2}{\mathbf{a}^{\mathrm{T}} \mathbf{\Sigma}_{\mathrm{W}} \mathbf{a}}$$

y vimos que el máximo se alcanza en

$$\boldsymbol{\alpha} = \mathbf{\Sigma}_{\mathrm{W}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

que se estima por

$$\widehat{\boldsymbol{\alpha}} = \mathbf{S}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)$$

Caso $k > 2$

Supongamos $\mathbf{\Sigma}_{\mathrm{W}} > 0$. Sea $\mathbf{C}$ triangular tal que $\mathbf{\Sigma}_{\mathrm{W}} = \mathbf{C}^{\mathrm{T}}\mathbf{C}$ y definamos

$$\mathbf{B} = (\mathbf{C}^{-1})^{\mathrm{T}}\mathbf{\Sigma}_{\mathrm{B}}\mathbf{C}^{-1}$$

Sean $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ los autovectores de $\mathbf{B}$ asociados a los autovalores $\lambda_1 \geq \cdots \geq \lambda_p$.

$$\mathbf{B} = \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j \boldsymbol{\beta}_j^{\mathrm{T}} \qquad \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{\beta}_\ell = 0 \text{ si } j \neq \ell \quad \|\boldsymbol{\beta}_j\| = 1$$

Como rango$(\mathbf{B}) = s \leq \min(k-1, p)$, $\lambda_j = 0$ si $j > s$. Por simplicidad supondremos que los autovalores no nulos son distintos, o sea, $\lambda_1 > \cdots > \lambda_s$. Es decir, tenemos que

$$\mathbf{B} = \sum_{j=1}^{s} \lambda_j \boldsymbol{\beta}_j \boldsymbol{\beta}_j^{\mathrm{T}}\,, \quad \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{\beta}_\ell = 0 \text{ si } 1 \leq j \neq \ell \leq p \quad \|\boldsymbol{\beta}_j\| = 1\,, 1 \leq j \leq p$$

Caso $k > 2$

Luego,

$$F_{\mathbf{a}} = \frac{\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{B}}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{W}}\mathbf{a}} = \frac{(\mathbf{Ca})^{\mathrm{T}}\mathbf{B}(\mathbf{Ca})}{\|\mathbf{Ca}\|^2}$$

Por lo tanto, si $\mathbf{b} = \mathbf{Ca}$

$$\max_{\mathbf{a}\neq\mathbf{0}} F_{\mathbf{a}} = \max_{\mathbf{b}\neq\mathbf{0}} \frac{\mathbf{b}^{\mathrm{T}}\mathbf{B}\mathbf{b}}{\|\mathbf{b}\|^2}$$

El máximo de la expresión del lado derecho se alcanza en $\boldsymbol{\beta}_1$ el autovector de $\mathbf{B}$ asociado a su mayor autovalor λ_1.

Con lo cual,

$$\max_{\mathbf{a}\neq\mathbf{0}} F_{\mathbf{a}} = F_{\boldsymbol{\alpha}_1} = \lambda_1$$

donde

$$\boldsymbol{\alpha}_1 = \mathbf{C}^{-1}\boldsymbol{\beta}_1$$

La combinación lineal $z_1 = \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{x}$ se llama la primer **coordenada discriminante** y da la mejor separación entre grupos.

Caso $k > 2$

Tenemos que si rechazamos H_3

$$0 < \text{rango}(\mathbf{B}) = \text{rango}(\mathbf{\Sigma}_{\text{B}}) = s \leq \min(k-1, p)\,.$$

Luego, para elegir las siguientes direcciones y no repetir información, buscaremos maximizar $F_{\mathbf{a}}$ sujeto a la condición de que las nuevas coordenadas sean no–correlacionadas con z_1, es decir, tales que

$$\text{Cov}(\mathbf{a}^{\text{T}}\mathbf{x}, \boldsymbol{\alpha}_1^{\text{T}}\mathbf{x}) = \mathbf{a}^{\text{T}}\mathbf{\Sigma}_{\text{W}}\boldsymbol{\alpha}_1 = 0\,.$$

Este problema se puede escribir como

$$\max_{\substack{\mathbf{a}\neq\mathbf{0} \\ \mathbf{a}^{\text{T}}\mathbf{\Sigma}_{\text{W}}\boldsymbol{\alpha}_1=0}} F_{\mathbf{a}} = \max_{\substack{\mathbf{b}\neq\mathbf{0} \\ \mathbf{b}^{\text{T}}\boldsymbol{\beta}_1=0}} \frac{\mathbf{b}^{\text{T}}\mathbf{B}\mathbf{b}}{\|\mathbf{b}\|^2}$$

El lado derecho de la expresión se alcanza en $\boldsymbol{\beta}_2$ el autovector de $\mathbf{B}$ asociado al segundo mayor autovalor λ_2.

Caso $k > 2$

$$\max_{\substack{\mathbf{a}\neq\mathbf{0} \\ \mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{W}}\boldsymbol{\alpha}_1=0}} F_{\mathbf{a}} = F_{\boldsymbol{\alpha}_2} = \lambda_2 \qquad \boldsymbol{\alpha}_2 = \mathbf{C}^{-1}\boldsymbol{\beta}_2$$

donde $\boldsymbol{\beta}_2$ el autovector de $\mathbf{B}$ asociado a λ_2.

Definición:

Definimos la $j-$ésima variable canónica o variable discriminante z_j como

$$z_j = \boldsymbol{\alpha}_j^{\mathrm{T}}\mathbf{x} \qquad \textit{donde} \qquad \boldsymbol{\alpha}_j = \mathbf{C}^{-1}\boldsymbol{\beta}_j$$

El vector $\mathbf{z} = (z_1, \ldots, z_p)^{\mathrm{T}}$ es el vector de variables canónicas o variables discriminantes.

$$\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x} \qquad \textit{con} \qquad \mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p)$$

Observemos que $\boldsymbol{\alpha}_j$ es un autovector de $\boldsymbol{\Sigma}_{\mathrm{W}}^{-1}\boldsymbol{\Sigma}_{\mathrm{B}}$.

Caso $k > 2$ y $s < p$

Como rango$(\mathbf{B}) = s \leq \min(k-1, p)$ si $s < p$, para todo $s+1 \leq m \leq p$ tendremos que

$$\max_{\substack{\mathbf{a} \neq \mathbf{0} \\ \mathbf{a}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{W}} \boldsymbol{\alpha}_{\ell} = 0, \ell < m}} \qquad F_{\mathbf{a}} = F_{\boldsymbol{\alpha}_m} = \lambda_m = 0$$

Es decir, una combinación lineal no correlacionada con $z_1, \ldots, z_s$ no dará información sobre las diferencias en posición.

En particular, $z_{s+1}, \ldots, z_p$ no dan información sobre las diferencias de posición.

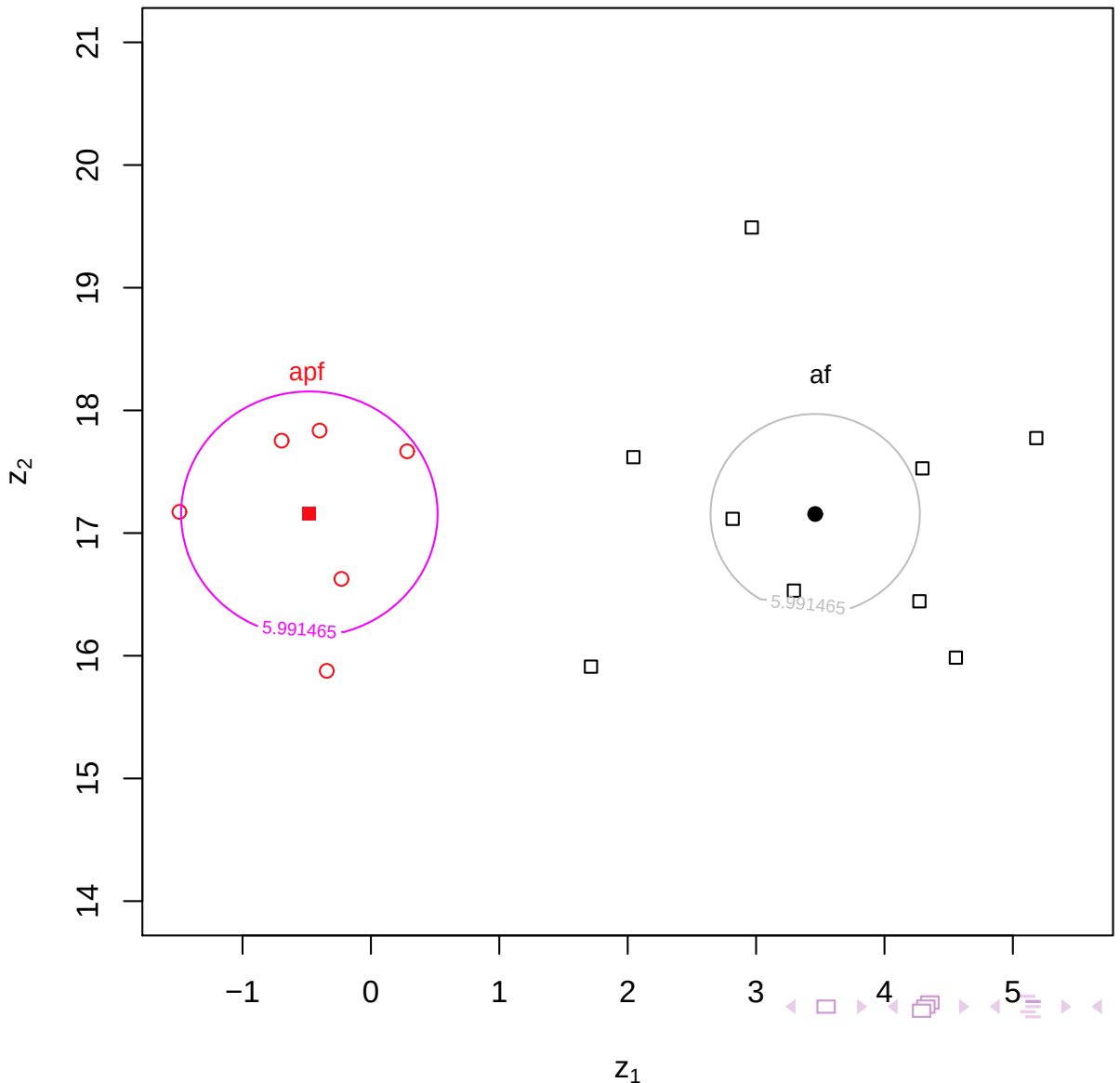
O sea, si $\overline{\boldsymbol{\nu}} = \mathbf{A}^{\mathrm{T}} \overline{\boldsymbol{\mu}}$ y si $\boldsymbol{\nu}_i = \mathbf{A}^{\mathrm{T}} \boldsymbol{\mu}_i$, entonces

$$\boldsymbol{\nu}_{im} = \overline{\boldsymbol{\nu}}_m \qquad \text{para} \quad s+1 \leq m \leq p$$

Más aún, si $\boldsymbol{\Lambda}_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_s)$

$$\mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{B}} \mathbf{A} = \begin{pmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Ejemplo



Caso $k > 2$ y $s < p$

Por lo tanto, sólo basta considerar el vector de variables discriminantes

$$\mathbf{z}^{(1)} = (z_1, \ldots, z_s) = \mathbf{A}_1^{\mathrm{T}}\mathbf{x}$$

donde

$$\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2) \quad \text{con} \quad \mathbf{A}_1 = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_s)$$

Sean además, $\boldsymbol{\nu}_i^{(1)} = \mathbf{A}_1^{\mathrm{T}}\boldsymbol{\mu}_i$ $\mathbf{z}^{(2)} = \mathbf{A}_2^{\mathrm{T}}\mathbf{x}$, $\mathbf{z} = \begin{pmatrix} \mathbf{z}^{(1)} \\ \mathbf{z}^{(2)} \end{pmatrix}$

Si $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$, $1 \leq i \leq k$ entonces $\boldsymbol{\Sigma}_{\mathrm{W}} = \boldsymbol{\Sigma}$ por lo tanto,

$$\textsc{Var}(\mathbf{z}|G = i) = \mathbf{I}_p \qquad \mathbb{E}(\mathbf{z}|G = i) = \boldsymbol{\nu}_i$$

Si $\mathbf{x}|G = i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ entonces

$$\mathbf{z}|G = i \sim N(\boldsymbol{\nu}_i, \mathbf{I}_p)$$

Caso $k > 2$

En particular

$$\mathbf{z}^{(1)} | G = i \sim N(\boldsymbol{\nu}_i^{(1)}, \mathbf{I}_s)$$

Luego, una manera heurística de definir una región para clasificar una nueva observación es asignar el vector $\mathbf{x}_0$ al grupo i si la media $\boldsymbol{\nu}_i^{(1)}$ de las variables transformadas es la más cercana a $\mathbf{v}_0 = \mathbf{A}_1^{\mathrm{T}}\mathbf{x}_0$.

O sea, asigno $\mathbf{x}_0$ al grupo i si $\mathbf{v}_0 = \mathbf{A}_1^{\mathrm{T}}\mathbf{x}_0 \in \mathcal{G}_i$ donde

$$\begin{aligned}
\mathcal{G}_i &= \{\mathbf{v} \in \mathbb{R}^s : \quad \|\mathbf{v} - \boldsymbol{\nu}_i^{(1)}\| < \|\mathbf{v} - \boldsymbol{\nu}_\ell^{(1)}\| \quad \forall \ell \neq i\} \\
&= \{\mathbf{v} \in \mathbb{R}^s : \quad (\boldsymbol{\nu}_i^{(1)})^{\mathrm{T}}(\mathbf{v} - \frac{1}{2}\boldsymbol{\nu}_i^{(1)}) > (\boldsymbol{\nu}_\ell^{(1)})^{\mathrm{T}}(\mathbf{v} - \frac{1}{2}\boldsymbol{\nu}_\ell^{(1)}) \quad \forall \ell \neq j\}
\end{aligned}$$

Si $s = 1$ partimos la recta de las observaciones transformadas en dos semi–rectas.

Si $s = 2$ la regla de clasificación genera una partición de $\mathbb{R}^2$ en regiones limitadas por semi–rectas.

Inferencia

Supongamos tener ahora observaciones $\mathbf{x}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $1 \leq j \leq n_i$, $1 \leq i \leq k$, entonces

- $\overline{\mathbf{x}}_i$ es el EMV de $\boldsymbol{\mu}_i$.
- El EMV de $\boldsymbol{\Sigma}_{\mathrm{W}}$ es
$$\frac{\mathbf{U}}{n}$$
- El EMV de $\boldsymbol{\Sigma}_{\mathrm{B}}$ es
$$\frac{\mathbf{H}}{n}$$
y $\mathbb{P}(\text{rango}(\mathbf{H}) = \min(p, k-1)) = 1$
- Un estimador insesgado de $\boldsymbol{\Sigma}_{\mathrm{W}}$ es
$$\frac{\mathbf{U}}{n-k}$$

Inferencia

- $\mathbf{U} = \mathbf{T}^{\mathrm{T}}\mathbf{T}$

- $\mathbf{b}_j$ los autovectores de $\widehat{\mathbf{B}} = (\mathbf{T}^{-1})^{\mathrm{T}}\mathbf{H}\mathbf{T}^{-1}$ tales que $\|\mathbf{b}_j\| = 1$, $\mathbf{b}_j^{\mathrm{T}}\mathbf{b}_\ell = 0$ si $\ell \neq j$

  $\mathbf{b}_j$ es el autovector asociado al $j-$ésimo autovalor $\widehat{\lambda}_j$ de $\widehat{\mathbf{B}}$, donde

$$\mathbb{P}(\widehat{\lambda}_1 > \widehat{\lambda}_2 > \cdots > \widehat{\lambda}_s > 0 \quad \text{y} \quad s = \min(p, k-1)) = 1$$

- $\mathbf{a}_j = (\mathbf{T}^{-1})^{\mathrm{T}}\mathbf{b}_j\sqrt{n-k}$, $\widehat{\mathbf{A}}_1 = (\mathbf{a}_1, \ldots, \mathbf{a}_s)$

- $\overline{\mathbf{z}}_i^{(1)} = \widehat{\mathbf{A}}_1^{\mathrm{T}}\overline{\mathbf{x}}_i$. Entonces, si n_i es grande $\overline{\mathbf{z}}_i^{(1)} \approx N(\boldsymbol{\nu}_i^{(1)}, \mathbf{I}_s/n_i)$.

Inferencia

Una región de confianza asintótica de nivel $1-\alpha$ para $\boldsymbol{\nu}_i^{(1)}$ está dada por

$$\{\mathbf{v}\in\mathbb{R}^s: \quad n_i\|\overline{\mathbf{z}}_i^{(1)}-\mathbf{v}\|^2\leq\chi^2_{s,\alpha}\}=\{\mathbf{v}\in\mathbb{R}^s: \quad \|\overline{\mathbf{z}}_i^{(1)}-\mathbf{v}\|\leq\sqrt{\frac{\chi^2_{s,\alpha}}{n_i}}\}$$

Además, asignamos $\mathbf{x}_0$ al grupo i si $\widehat{\mathbf{v}}_0=\widehat{\mathbf{A}}_1^{\mathrm{T}}\mathbf{x}_0\in\widehat{\mathcal{G}}_i$ donde

$$\begin{aligned}
\widehat{\mathcal{G}}_i &= \{\mathbf{v}\in\mathbb{R}^s: \quad \|\mathbf{v}-\overline{\mathbf{z}}_i^{(1)}\|<\|\mathbf{v}-\overline{\mathbf{z}}_\ell^{(1)}\| \quad \forall\ell\neq i\}\\
&= \{\mathbf{v}\in\mathbb{R}^s: \quad (\overline{\mathbf{z}}_i^{(1)})^{\mathrm{T}}(\mathbf{v}-\frac{1}{2}\overline{\mathbf{z}}_i^{(1)})>(\overline{\mathbf{z}}_\ell^{(1)})^{\mathrm{T}}(\mathbf{v}-\frac{1}{2}\overline{\mathbf{z}}_\ell^{(1)}) \quad \forall\ell\neq j\}\\
&= \{\mathbf{v}\in\mathbb{R}^s: \quad (\overline{\mathbf{z}}_i^{(1)}-\overline{\mathbf{z}}_\ell^{(1)})^{\mathrm{T}}\left(\mathbf{v}-\frac{1}{2}\left(\overline{\mathbf{z}}_i^{(1)}+\overline{\mathbf{z}}_\ell^{(1)}\right)\right)>0 \quad \forall\ell\neq j\}
\end{aligned}$$

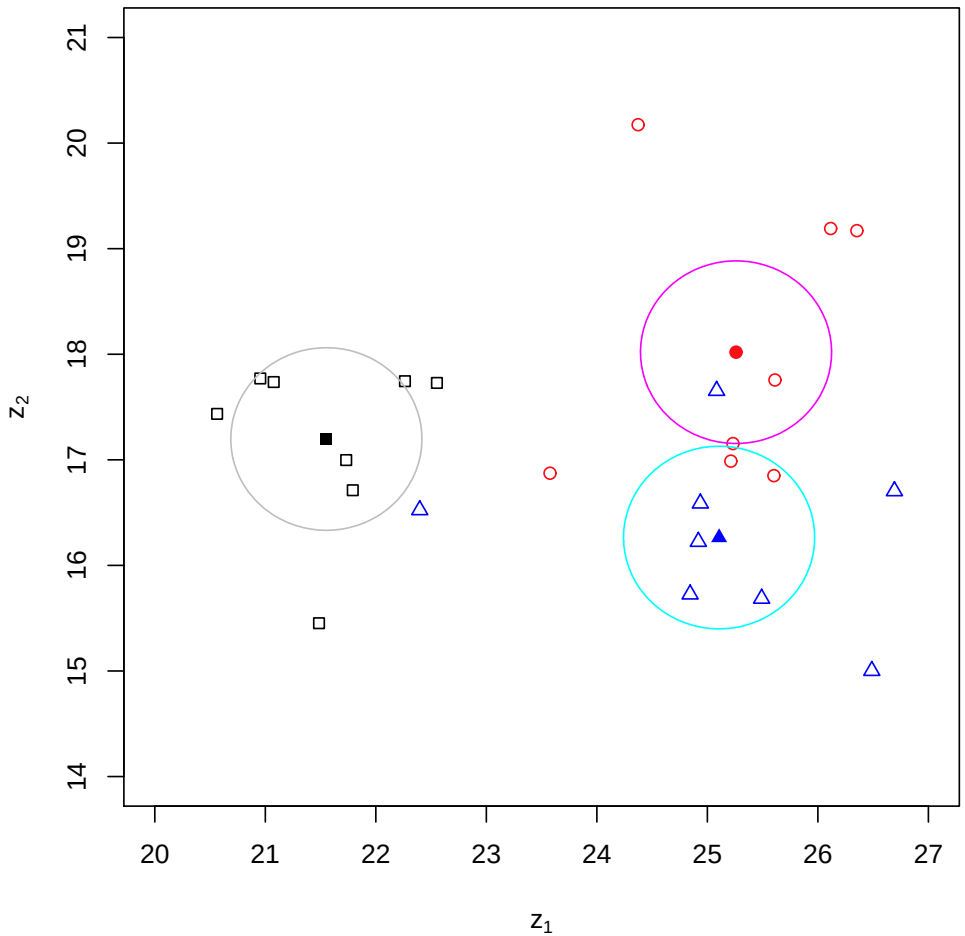
Ejemplo

En el ejemplo de los árboles tenemos $k = 3$, $p = 4$ luego $s = 2$, los autovalores no nulos de $\widehat{\mathbf{B}} = (\mathbf{T}^{-1})^{\mathrm{T}}\mathbf{H}\mathbf{T}^{-1}$ son 3.3522, 0.5879 y

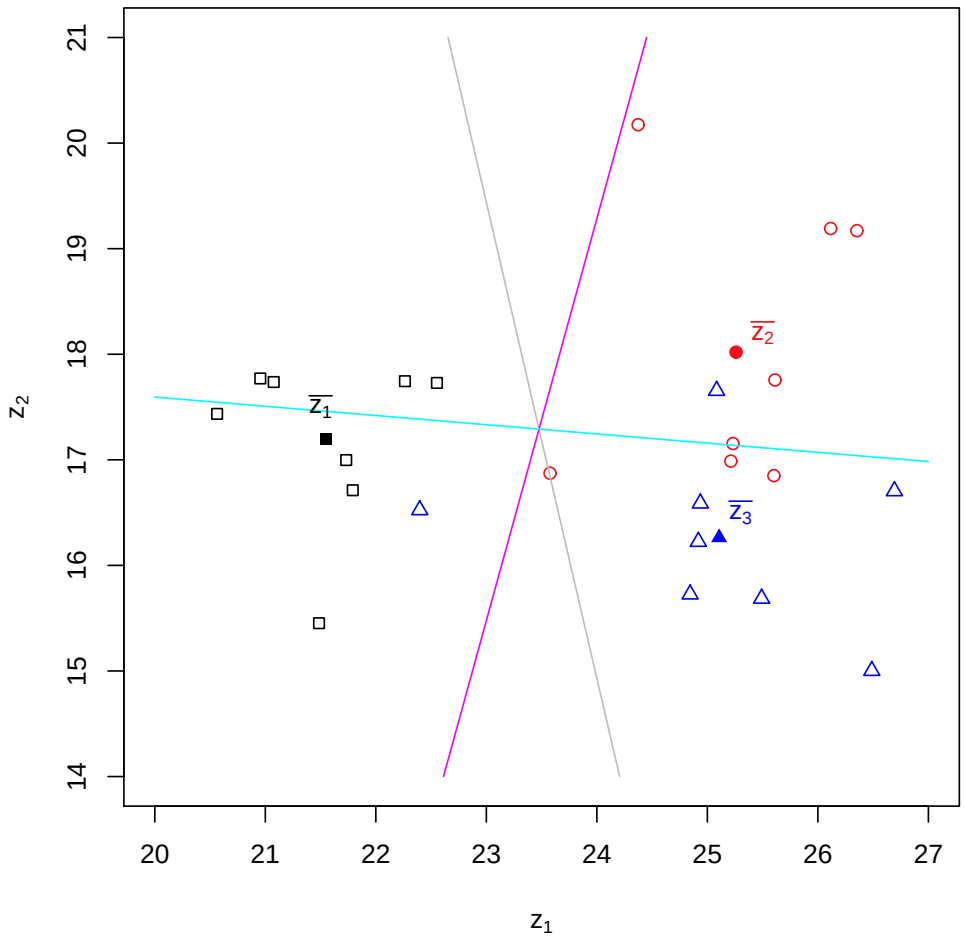
$$\widehat{\mathbf{B}} = \begin{pmatrix} 0.0139 & -0.4232 & 0.0000 & 0.9059 \\ -0.0139 & -0.0207 & 0.9996 & -0.0094 \\ -0.9845 & 0.1523 & -0.0098 & 0.0863 \\ 0.1742 & 0.8929 & 0.0248 & 0.4145 \end{pmatrix}$$

$$\widehat{\mathbf{A}}_1 = \begin{pmatrix} 7.7167 & 0.0711 \\ -2.7909 & 0.8135 \\ 6.1042 & 6.3137 \\ -1.9958 & -10.2292 \end{pmatrix}$$

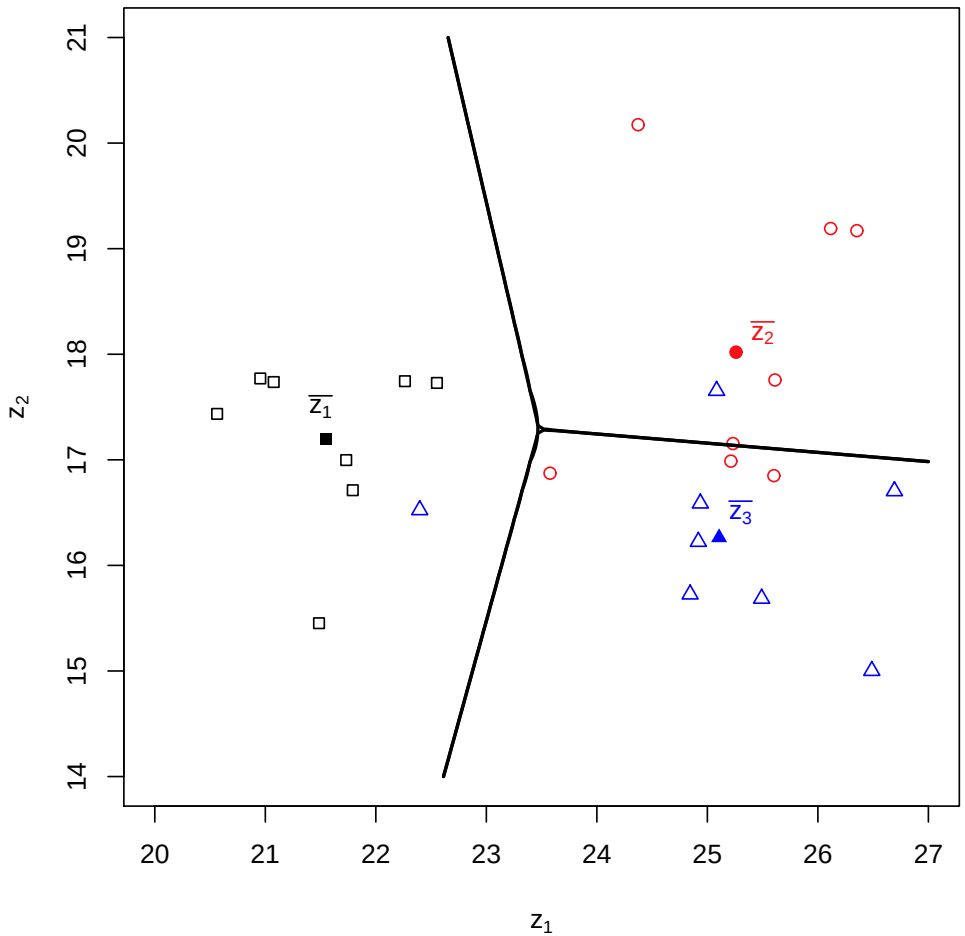
Ejemplo: Se grafican los círculos de radio $\{\chi^2_{2,0.05}/n_i\}^{1/2} = 2.4477/\sqrt{n_i}$



Ejemplo


21
20
19
18
17
16
15
14
z_2
20
21
22
23
24
25
26
27
z_1
$\overline{z_1}$
$\overline{z_2}$
$\overline{z_3}$

Ejemplo


21
20
19
18
17
16
15
14
z_2
20
21
22
23
24
25
26
27
z_1
$\overline{z_1}$
$\overline{z_2}$
$\overline{z_3}$

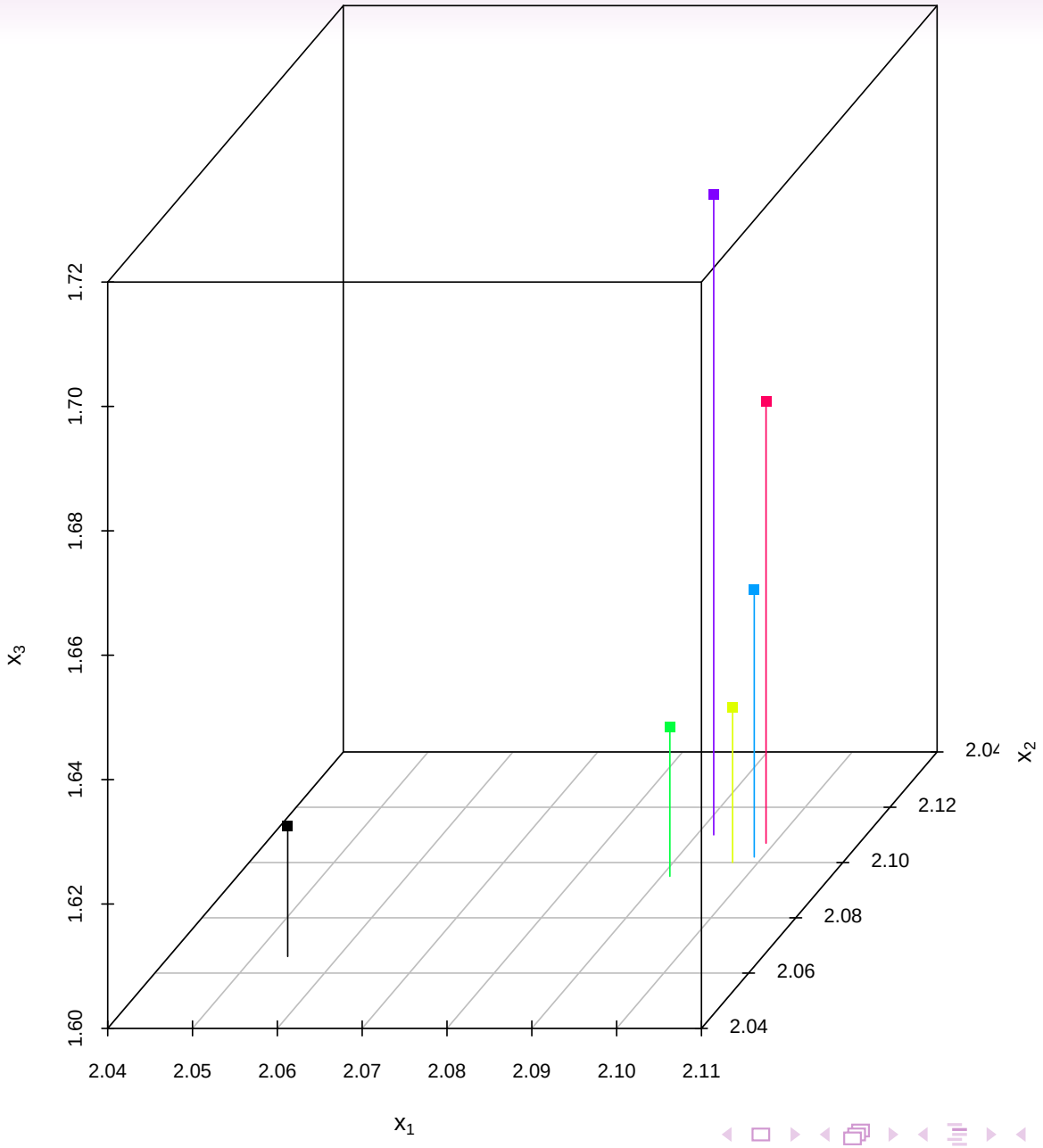
Ejemplo 2

Medidas tridiensionales en cráneos de 4 subespecies de oso hormiguero. Las mediciones hechas fueron

- $y_1 =$ Largo de base, excluyendo premaxilar
- $y_2 =$ Largo oxipitonasal
- $y_3 =$ Largo de los nasales.

Los datos consisten en $x_j = \log(y_j)$ para las subespecies

- *Instabilis*, (Colombia) $n_1 = 21$, $\overline{\mathbf{x}}_1 = (2.054, 2.066, 1.621)^{\mathrm{T}}$.
- *Chapadensis*, en tres localidades
  - Minas Gerais: $n_2 = 6$, $\overline{\mathbf{x}}_2 = (2.097, 2.1, 1.625)^{\mathrm{T}}$,
  - Matto Grosso: $n_3 = 9$, $\overline{\mathbf{x}}_3 = (2.091, 2.095, 1.624)^{\mathrm{T}}$,
  - Santa Cruz: $n_3 = 3$ $\overline{\mathbf{x}}_4 = (2.099, 2.102, 1.643)^{\mathrm{T}}$
- *Chiriquensis*, (Panamá) $n_5 = 4$, $\overline{\mathbf{x}}_5 = (2.092, 2.11, 1.703)^{\mathrm{T}}$
- *Mexicana* $n_6 = 5$, $\overline{\mathbf{x}}_5 = (2.099, 2.107, 1.671)^{\mathrm{T}}$


1.72
1.70
1.68
1.66
1.64
1.62
1.60
x_3
2.04
2.05
2.06
2.07
2.08
2.09
2.10
2.11
x_1
2.04
2.06
2.08
2.10
2.12
x_2

Ejemplo 2

La matrices **U** y **H** son

$$\mathbf{U} = \begin{pmatrix} 0.01363 & 0.01277 & 0.01644 \\ & 0.01293 & 0.01714 \\ & & 0.03615 \end{pmatrix} \qquad \mathbf{H} = \begin{pmatrix} 0.02002 & 0.01744 & 0.01308 \\ & 0.01585 & 0.01507 \\ & & 0.03068 \end{pmatrix}$$

Luego, $\nu_1 = pk + \frac{p(p+1)}{2} = 24$ y $\nu_2 = p + \frac{p(p+1)}{2} = 9$ de donde

$$-48 \log\left(\frac{|\mathbf{U}|}{|\mathbf{U}+\mathbf{H}|}\right) \approx \chi^2_{15}$$

$$\frac{|\mathbf{U}|}{|\mathbf{U}+\mathbf{H}|} = 0.1468 \quad -48 \log\left(\frac{|\mathbf{U}|}{|\mathbf{U}+\mathbf{H}|}\right) = 92.08 \quad \chi^2_{15,0.01} = 30.57791$$

y rechazo la igualdad de medias.

Ejemplo 2

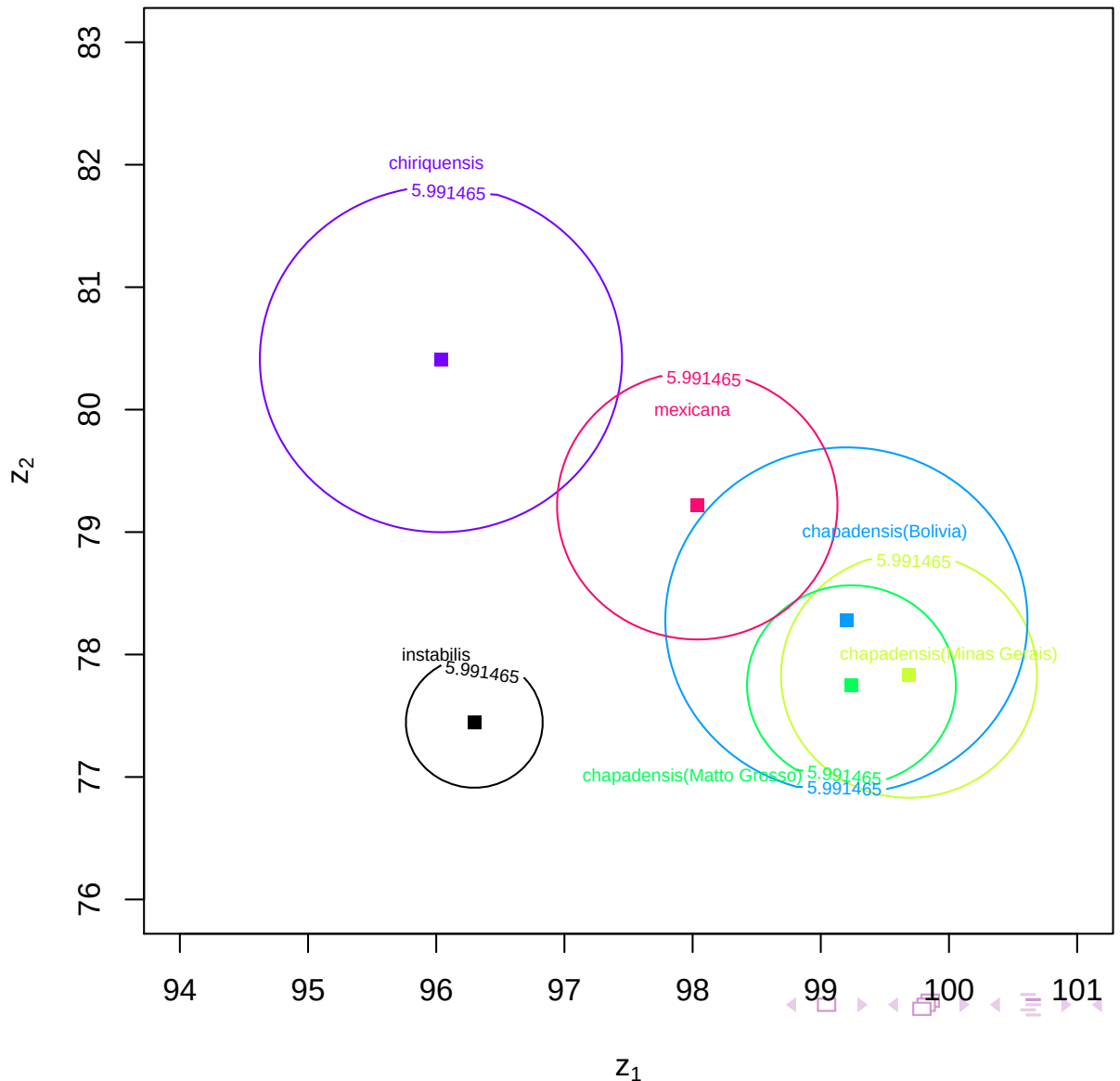
En este ejemplo $s = 3$. Los autovalores de $\widehat{B}$ son

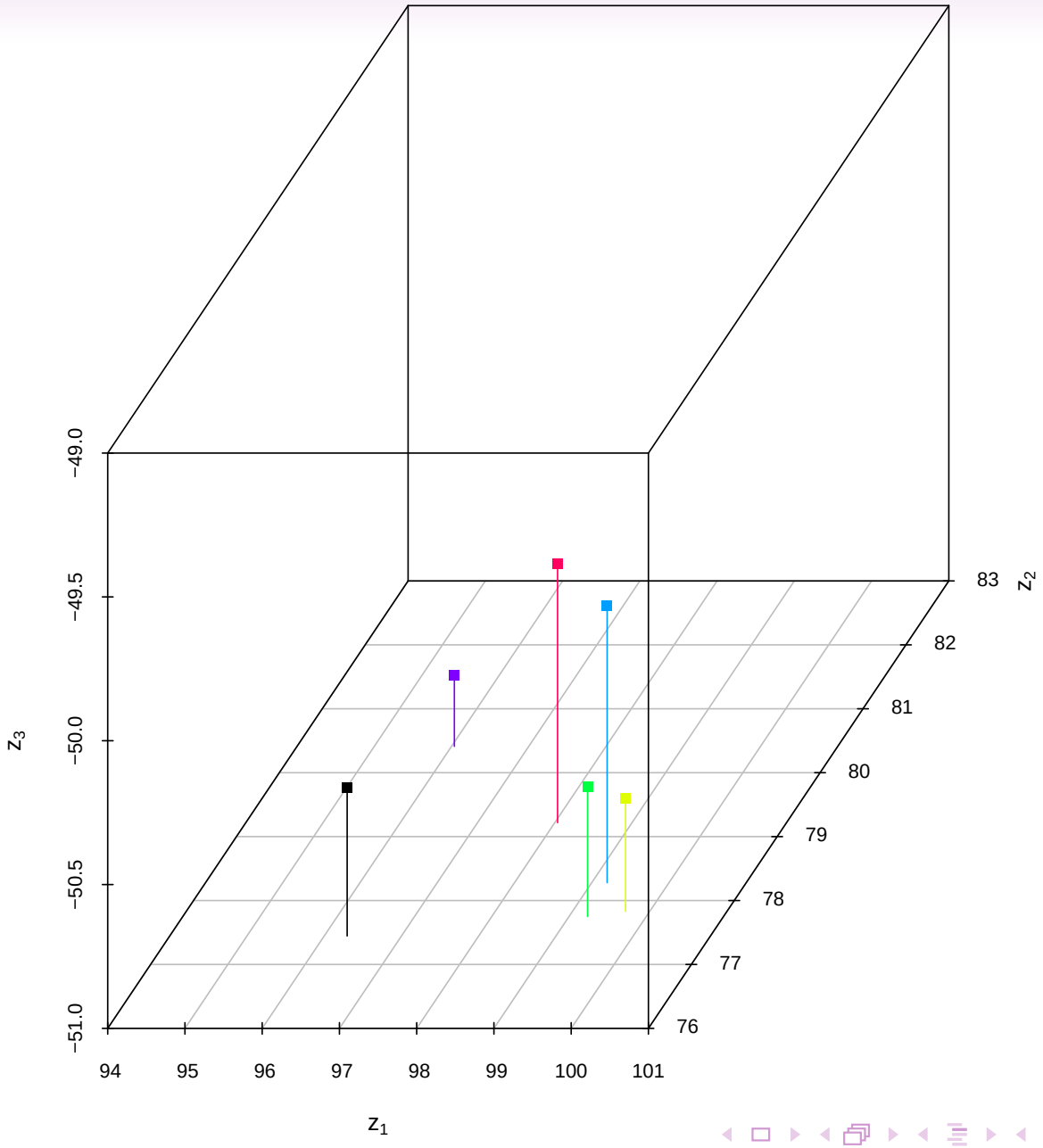
$$\widehat{\lambda}_1 = 2.4001 \quad \widehat{\lambda}_2 = 0.9050 \quad \widehat{\lambda}_3 = 0.0515$$

$$\widehat{\mathbf{A}}_1 = \begin{pmatrix} 108.9220 & -40.6120 & 169.3528 \\ -33.8285 & 59.9577 & -222.1998 \\ -35.4962 & 22.8195 & 37.4651 \end{pmatrix}$$

Graficaremos las dos primeras coordenadas discriminantes.

Ejemplo 2: Se grafican los círculos de radio $\{\chi^2_{2,0.05}/n_i\}^{1/2} = 2.4477/\sqrt{n_i}$




−49.0
−49.5
−50.0
−50.5
−51.0
z_3
94
95
96
97
98
99
100
101
z_1
76
77
78
79
80
81
82
83
z_2

Tenemos k poblaciones o grupos diferentes $\mathcal{P}_1, \ldots, \mathcal{P}_k$ y un vector $\mathbf{x} \in \mathbb{R}^p$ que puede pertenecer a cualquiera de esas poblaciones.

a) G una variable aleatoria que indica la pertenencia al grupo,

$$\mathbb{P}(G = j) = \pi_j$$

Enfoque bayesiano,

- $\Theta = \{\mathcal{P}_1, \ldots, \mathcal{P}_k\}$ o en forma simplificada tenemos una variable aleatoria G que toma valores $\{1, \ldots, k\}$.
- Sobre Θ definimos una probabilidad a priori τ que es discreta y es tal que $\mathbb{P}(G = j) = \pi_j$

b) La distribución de $\mathbf{x}$ varía según el grupo de pertenencia.
Si $\mathbf{x}$ pertenece a la población $\mathcal{P}_j$ entonces $\mathbf{x}$ tiene densidad f_j,

$$\mathbf{x}|G = j \sim f_j$$

c) Para dar una regla de clasificación daremos una partición de $\mathbb{R}^p$ en k conjuntos disjuntos

$$\mathbb{R}^p = \cup_{i=1}^k \mathcal{G}_j \qquad \mathcal{G}_j \cap \mathcal{G}_s = \emptyset$$

Definiciones

La densidad marginal de $\mathbf{x}$ está dada por

$$f_{\mathbf{x}}(\mathbf{x}) = \sum_{j=1}^{k} \pi_j f_j(\mathbf{x})$$

y la probabilidad condicional de que una observación pertenezca a $\mathcal{P}_j$ dado que $\mathbf{x} = \mathbf{x}_0$, está dada por

$$q_j(\mathbf{x}_0) = \mathbb{P}(G = j | \mathbf{x} = \mathbf{x}_0) = \frac{\pi_j f_j(\mathbf{x}_0)}{\sum_{\ell=1}^{k} \pi_\ell f_\ell(\mathbf{x}_0)}$$

La cantidad $q_j(\mathbf{x}_0)$ es la probabilidad a posteriori.

Definición 1.
Una regla de clasificación es una variable aleatoria $G^\star(\mathbf{x})$ tal que

$$G^\star(\mathbf{x}) = j \qquad \text{si} \quad \mathbf{x} \in \mathcal{G}_j$$

donde $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$ es una partición de $\mathbb{R}^p$.

Definiciones

Podemos ver a $G^\star$ como la pertenencia predicha mientras que G es la pertenencia real.

La teoría de clasificación trata de encontrar reglas de clasificación óptimas en algún sentido. Lo ideal sería que $\mathbb{P}(G^\star = G) = 1$, pero esto no es posible.

Definición 2. Para una regla de clasificación $G^\star$ con regiones de clasificación $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$, la probabilidad de asignar la observación $\mathbf{x}$ a $\mathcal{P}_i$ cuando en realidad, $\mathbf{x} \in \mathcal{P}_j$ es

$$p_{i|j} = \mathbb{P}(G^\star = i | G = j) = \mathbb{P}(\mathbf{x} \in \mathcal{G}_i | G = j) = \int_{\mathcal{G}_i} f_j(\mathbf{x}) d\mathbf{x}$$

Observemos que $\sum_{i=1}^{k} p_{i|j} = 1$.

Definiciones

A veces es posible asignar un costo $c_{i|j} \geq 0$ a la clasificación de una observación del grupo j en el grupo i. En muchos casos se elige $c_{i|j} = 1$ si $i \neq j$.

Definimos

a) La función de pérdida como

$$L(\mathcal{P}_j, i) = \begin{cases} c_{i|j} & \text{si} \quad i \neq j \\ 0 & \text{si} \quad i = j \end{cases}$$

$$L(\mathcal{P}_j, G^\star) = L(j, G^\star) = \sum_{i=1}^{k} c_{i|j} \mathbb{I}(\mathbf{x} \in \mathcal{G}_i | \mathbf{x} \in \mathcal{P}_j)$$

donde $c_{j|j} = 0$.

b) El riesgo de $G^\star$ es

$$R(\mathcal{P}_j, G^\star) = \mathbb{E}\, L(\mathcal{P}_j, G^\star) = \sum_{i=1}^{k} c_{i|j} p_{i|j} = \sum_{i=1}^{k} c_{i|j} \int_{\mathcal{G}_i} f_j(\mathbf{x}) d\mathbf{x}$$

Definiciones

El Riesgo de Bayes de una regla de clasificación $G^\star$ será

$$r(\tau, G^\star) = \mathbb{E}R(\Theta, G^\star) = \sum_{j=1}^{k} \sum_{i \neq j} \pi_j c_{i|j} p_{i|j} = \sum_{j=1}^{k} \sum_{i \neq j} \pi_j c_{i|j} \int_{\mathcal{G}_i} f_j(\mathbf{x}) d\mathbf{x}$$

En particular, si $c_{i|j} = 1$ si $i \neq j$ tenemos que

$$r(\tau, G^\star) = \sum_{j=1}^{k} \sum_{i \neq j} \pi_j p_{i|j} = 1 - \sum_{j=1}^{k} \pi_j p_{j|j} = 1 - \sum_{j=1}^{k} \pi_j \int_{\mathcal{G}_j} f_j(\mathbf{x}) d\mathbf{x}$$

que se llama la probabilidad total de mala clasificación ya que coincide con $\mathbb{P}(G^\star \neq G)$.

Si $k = 2$ y $c_{i|j} = 1$ si $i \neq j$ tenemos que

$$r(\tau, G^\star) = \pi_1 + \int_{\mathcal{G}_1} \left[\pi_2 f_2(\mathbf{x}) - \pi_1 f_1(\mathbf{x})\right] d\mathbf{x}$$

Definiciones

1. Diremos que una regla de clasificación $G_0^\star$ es Bayes respecto de la distribución a priori τ si

$$r(\tau, G_0^\star) = \min_{G^\star} r(\tau, G^\star)$$

2. Diremos que una regla de clasificación $G_0^\star$ es minimax si

$$\max_{1 \leq j \leq k} R(\mathcal{P}_j, G_0^\star) = \min_{G^\star} \max_{1 \leq j \leq k} R(\mathcal{P}_j, G^\star)$$

Propiedad

La regla Bayes respecto de τ clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \mathcal{G}_{i,0}$ donde

$$\mathcal{G}_{i,0} = \{\mathbf{x} \in \mathbb{R}^p : \quad \sum_{j=1}^{k} \pi_j c_{i|j} f_j(\mathbf{x}) < \sum_{j=1}^{k} \pi_j c_{\ell|j} f_j(\mathbf{x}) \quad \forall \ell \neq i\}$$

o sea, clasifico $\mathbf{x} \in \mathcal{P}_i$ si

$$\sum_{j=1}^{k} \pi_j c_{i|j} f_j(\mathbf{x}) = \min_{\ell} \sum_{j=1}^{k} \pi_j c_{\ell|j} f_j(\mathbf{x})$$

siendo la asignación en la frontera de $\mathcal{G}_{i,0}$ arbitraria.

Casos particulares

a) Supongamos que $c_{i|j} = 1$ si $i \neq j$ entonces la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \mathcal{G}_{i,0}$ donde

$$\begin{aligned} \mathcal{G}_{i,0} &= \{\mathbf{x} \in \mathbb{R}^p : \quad \pi_\ell \, f_\ell(\mathbf{x}) < \pi_i \, f_i(\mathbf{x}) \quad \forall \ell \neq i\} \\ &= \{\mathbf{x} \in \mathbb{R}^p : \quad q_\ell(\mathbf{x}) < q_i(\mathbf{x}) \quad \forall \ell \neq i\} \end{aligned}$$

es decir, clasifico $\mathbf{x} \in \mathcal{P}_i$ si $q_i(\mathbf{x}) = \max_{1 \leq \ell \leq k} q_\ell(\mathbf{x})$ siendo la asignación en la frontera de $\mathcal{G}_{i,0}$ arbitraria. Por lo tanto,

i) la regla Bayes coincide con el criterio de minimizar la probabilidad total de mala clasificación.

ii) la regla Bayes coincide con el criterio de maximizar la probabilidad a posteriori.

iii) si además, $\pi_j = 1/k$, $1 \leq j \leq k$, la regla Bayes coincide con el criterio de máxima verosimilitud, que asigna $\mathbf{x}$ a la población que máximiza la verosimilitud de $\mathbf{x}$.

Casos particulares

b) Supongamos que $k = 2$ entonces la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_1$ si $\mathbf{x} \in \mathcal{G}_{1,0}$ donde

$$\mathcal{G}_{1,0} = \left\{ \mathbf{x} \in \mathbb{R}^p : \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2 c_{1|2}}{\pi_1 c_{2|1}} \right\}$$

i) Si $\pi_2 c_{1|2} = \pi_1 c_{2|1}$ la regla Bayes da el criterio de máxima verosimilitud. En particular, si $c_{i|j} = 1$ si $i \neq j$, el criterio de máxima verosimilitud es la regla Bayes asociada a $\pi_1 = \pi_2 = \frac{1}{2}$

ii) si $c_{i|j} = 1$ si $i \neq j$ y $\pi_1 = 1 - \alpha$, $\pi_2 = \alpha$, $0 < \alpha < 1$, entonces la regla Bayes clasifica

- $\mathbf{x} \in \mathcal{P}_1$ si

$$\mathbf{x} \in \mathcal{G}_{1,0} = \{ \mathbf{x} \in \mathbb{R}^p : \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > a = \frac{\alpha}{1-\alpha} \}$$

- $\mathbf{x} \in \mathcal{P}_2$ si

$$\mathbf{x} \in \mathcal{G}_{2,0} = \{ \mathbf{x} \in \mathbb{R}^p : \quad \frac{f_2(\mathbf{x})}{f_1(\mathbf{x})} > \frac{1}{a} = \frac{1-\alpha}{\alpha} \}$$

Casos particulares

b) ii) Sea a_0 tal que

$$\int_{\mathcal{G}_{1,0}} f_2(\mathbf{x})d\mathbf{x} = \int_{\mathcal{G}_{2,0}} f_1(\mathbf{x})d\mathbf{x}$$

Luego, la regla Bayes respecto de $\tau = (\alpha_0, 1-\alpha_0)$ con $\alpha_0 = 1/(1+a_0)$ iguala riesgos y es la regla minimax.

c) Supongamos que $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)^{\mathrm{T}}$ con $\mathbf{x}_1 \in \mathbb{R}^q$ y que $\mathbf{x}_1$ y $\mathbf{x}_2$ son independientes en todas las poblaciones, o sea, $f_j(\mathbf{x}) = h_j(\mathbf{x}_1)\ell_j(\mathbf{x}_2)$. Más aún, supongamos que $\ell_j(\mathbf{x}_2) = \ell(\mathbf{x}_2)$ para todo j. Entonces, la regla de clasificación se basa solamente en $\mathbf{x}_1$, es decir, la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \mathcal{G}_{i,0}$ donde

$$\mathcal{G}_{i,0} = \{\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^p : \sum_{j=1}^{k} \pi_j c_{i|j} h_j(\mathbf{x}_1) < \sum_{j=1}^{k} \pi_j c_{\ell|j} h_j(\mathbf{x}_1) \quad \forall \ell \neq i\}$$

o sea, clasifico $\mathbf{x} \in \mathcal{P}_i$ si $\sum_{j=1}^{k} \pi_j c_{i|j} h_j(\mathbf{x}_1) = \min_\ell \sum_{j=1}^{k} \pi_j c_{\ell|j} h_j(\mathbf{x}_1)$.

Ejemplo

Supongamos que $k = 2$, $p = 2$

- $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ con x_i Bernouilli independientes
- $\mathbb{P}(x_1 = 1|G = 1) = p_{11}$, $\mathbb{P}(x_1 = 1|G = 2) = p_{21}$,
- $\mathbb{P}(x_2 = 1|G = 1) = p_{12}$, $\mathbb{P}(x_2 = 1|G = 2) = p_{22}$

Luego, la regla de clasificación óptima (Bayes) con costos iguales lleva a una función discriminate lineal, o sea, decido que $\mathbf{x} \in \mathcal{P}_1$ si

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 > \log\left(\frac{\pi_2}{\pi_1}\right)$$

donde

$$\beta_0 = \log\left(\frac{(1-p_{11})(1-p_{12})}{(1-p_{21})(1-p_{22})}\right)$$

$$\beta_1 = \log\left(\frac{p_{11}(1-p_{21})}{(1-p_{11})p_{21}}\right) \qquad \beta_2 = \log\left(\frac{p_{12}(1-p_{22})}{(1-p_{12})p_{22}}\right)$$

Problema

Hasta ahora supusimos que la distribución de $\mathbf{x}$ en cada población es conocida. En la mayoría de los casos esto no ocurre y tenemos alguna de las siguientes situaciones

a) la distribución es conocida salvo por algunos parámetros que deberemos estimar, $f_j = f_j(\cdot, \boldsymbol{\theta}_j)$
b) la distribución es parcialmente desconocida, o sea, sabemos por ejemplo que

$$\log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$$

c) la distribución es desconocida

En a) y b) estimamos los parámetros. La regla en este caso se estima por $\widehat{G}_0^\star$ reemplazando los parámetros desconocidos por sus estimadores.

Entonces, necesitamos conocer las probabilidades de error cometido, o sea, aproximar $r(\tau, G_0^\star)$ y $R(\mathcal{P}_j, G_0^\star)$.

Error óptimo de clasificación

El error de mala clasificación de la población j

$$R(\mathcal{P}_j, G_0^\star) = \sum_{i \neq j}^{k} p_{i|j} = \sum_{i \neq j}^{k} \int_{\mathcal{G}_i} f_j(\mathbf{x}) d\mathbf{x}$$

Si $k = 2$, llamaremos

$$\begin{aligned} e_{1,opt} = R(\mathcal{P}_1, G_0^\star) &= \int_{\mathcal{G}_{2,0}} f_1(\mathbf{x}) d\mathbf{x} = 1 - \int_{\mathcal{G}_{1,0}} f_1(\mathbf{x}) d\mathbf{x} \\ e_{2,opt} = R(\mathcal{P}_2, G_0^\star) &= \int_{\mathcal{G}_{1,0}} f_2(\mathbf{x}) d\mathbf{x} = 1 - \int_{\mathcal{G}_{2,0}} f_2(\mathbf{x}) d\mathbf{x} \\ e_{opt} &= \pi_1 e_{1,opt} + \pi_2 e_{2,opt} \end{aligned}$$

Error óptimo de clasificación si $f_j = f_j(\cdot, \boldsymbol{\theta}_j)$

El error de mala clasificación de la población j

$$R(\mathcal{P}_j, G_0^\star) = \sum_{i \neq j}^{k} p_{i|j} = \sum_{i \neq j}^{k} \int_{\mathcal{G}_i} f_j(\mathbf{x}, \boldsymbol{\theta}_j) d\mathbf{x}$$

Si $k = 2$, llamaremos

$$\begin{aligned} e_{1,opt} = R(\mathcal{P}_1, G_0^\star) &= \int_{\mathcal{G}_{2,0}} f_1(\mathbf{x}, \boldsymbol{\theta}_1) d\mathbf{x} = 1 - \int_{\mathcal{G}_{1,0}} f_1(\mathbf{x}, \boldsymbol{\theta}_1) d\mathbf{x} \\ e_{2,opt} = R(\mathcal{P}_2, G_0^\star) &= \int_{\mathcal{G}_{1,0}} f_2(\mathbf{x}, \boldsymbol{\theta}_2) d\mathbf{x} = 1 - \int_{\mathcal{G}_{2,0}} f_2(\mathbf{x}, \boldsymbol{\theta}_2) d\mathbf{x} \\ e_{opt} &= \pi_1 e_{1,opt} + \pi_2 e_{2,opt} \end{aligned}$$

$f_j = f_j(\cdot, \boldsymbol{\theta}_j)$, $\boldsymbol{\theta}_j$ desconocido

Hasta ahora supusimos que la distribución de $\mathbf{x}$ en cada población es conocida. Supongamos que la distribución es conocida salvo por algunos parámetros que deberemos estimar, $f_j = f_j(\cdot, \boldsymbol{\theta}_j)$ y sea

- $\widehat{\boldsymbol{\theta}}_j$ un estimador de $\boldsymbol{\theta}_j$ basado en la muestra $\mathbf{x}_{j,1}, \ldots, \mathbf{x}_{j,n_j}$.
- $\widehat{f}_j(\mathbf{x}) = f_j(\cdot, \widehat{\boldsymbol{\theta}}_j)$

entonces la regla Bayes, con $c_{i|j} = 1$ si $i \neq j$, se estima por la regla $\widehat{G}_0^\star$ que clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \widehat{\mathcal{G}}_{i,0}$ donde

$$\widehat{\mathcal{G}}_{i,0} \quad = \quad \{\mathbf{x} \in \mathbb{R}^p : \quad \pi_\ell \, \widehat{f}_\ell(\mathbf{x}) < \pi_i \, \widehat{f}_i(\mathbf{x}) \quad \forall \ell \neq i\}$$

Se sugiere que n_j sea tres veces por lo menos la cantidad de parámetros $\boldsymbol{\theta}_j$ a estimar y el número puede ser mayor si los grupos no están bien separados.

Cálculo errores de clasificación f_j, $j = 1, 2$

Se definen varios tipos de errores

a) El error actual

$$\begin{aligned} e_{1,act} = R(\mathcal{P}_1, \widehat{G}_0^\star) &= \int_{\widehat{\mathcal{G}}_{2,0}} f_1(\mathbf{x})d\mathbf{x} = 1 - \int_{\widehat{\mathcal{G}}_{1,0}} f_1(\mathbf{x})d\mathbf{x} \\ e_{2,act} = R(\mathcal{P}_2, \widehat{G}_0^\star) &= \int_{\widehat{\mathcal{G}}_{1,0}} f_2(\mathbf{x})d\mathbf{x} = 1 - \int_{\widehat{\mathcal{G}}_{2,0}} f_2(\mathbf{x})d\mathbf{x} \\ e_{act} &= \pi_1 e_{1,act} + \pi_2 e_{2,act} \end{aligned}$$

Claramente, $e_{opt} \leq e_{act}$

b) La tasa de error actual esperada

$$\mathbb{E}e_{act} = \pi_1 \mathbb{E}e_{1,act} + \pi_2 \mathbb{E}e_{2,act}$$

Estimación de los errores de clasificación

1) El estimador *plug–in*

$$\widehat{e}_{j,act} = \sum_{\ell \neq j} \int_{\widehat{\mathcal{G}}_{\ell,0}} \widehat{f}_j(\mathbf{x}) d\mathbf{x} = 1 - \int_{\widehat{\mathcal{G}}_{j,0}} \widehat{f}_j(\mathbf{x}) d\mathbf{x}$$

$$\widehat{e}_{act} = \sum_{j=1}^{k} \pi_j \widehat{e}_{k,act}$$

Este error se basa en la corrrecta especificación del modelo pero además en muchos casos, como veremos, subestima el error real e_{opt}.

Si $k = 2$

$$\widehat{e}_{1,act} = \int_{\widehat{\mathcal{G}}_{2,0}} \widehat{f}_1(\mathbf{x}) d\mathbf{x} = 1 - \int_{\widehat{\mathcal{G}}_{1,0}} \widehat{f}_1(\mathbf{x}) d\mathbf{x}$$

$$\widehat{e}_{2,act} = \int_{\widehat{\mathcal{G}}_{1,0}} \widehat{f}_2(\mathbf{x}) d\mathbf{x} = 1 - \int_{\widehat{\mathcal{G}}_{2,0}} \widehat{f}_2(\mathbf{x}) d\mathbf{x} \qquad \widehat{e}_{act} = \pi_1 e_{1,act} + \pi_2 e_{2,act}$$

Estimación de los errores de clasificación

2) La tasa de error aparente.

Consideremos la regla basada en las regiones $\widehat{\mathcal{G}}_{j,0}$, $1 \leq j \leq k$ y sean

$$n_{i,j} = \#\{\mathbf{x}_{i\ell} \text{ clasificadas en la población } \mathcal{P}_j\} = \#\{\mathbf{x}_{i\ell} \in \widehat{\mathcal{G}}_{j,0}\}$$

$$n_i = \sum_{j=1}^{k} n_{ij} \text{ el total de observaciones de la población } i-\text{ésima},$$

$$\widehat{\pi}_i = \frac{n_i}{n} \quad \text{con} \quad n = \sum_{i=1}^{k} n_i$$

$$m_i = \#\{\mathbf{x}_{i\ell} \text{ mal clasificadas }\} = \sum_{j \neq i} n_{ij}$$

Estimación de los errores de clasificación

2) La tasa de error aparente es

$$e_{i,app} = \frac{m_i}{n_i} \qquad e_{app} = \sum_{j=1}^{k} \pi_i e_{i,app}$$

$$\widehat{e}_{app} = \sum_{j=1}^{k} \widehat{\pi}_i e_{i,app} = \frac{\sum_{i=1}^{k} m_i}{n}$$

El método basado en $\widehat{e}_{app}$ se llama también de resustitución.

Este estimador del error es muy optimista ya que tiende a subestimar la probabilidad real de error, pues los mismos datos se usan para armar la regla (estimar los parámetros) y para evaluar la regla resultante. Los estimadores de los parámetros obtenidos son los que mejor ajustan a los datos y por ello tiendo a clasificar mejor.

Estimación de los errores de clasificación

2) Veamos un ejemplo de como la tasa de error aparente subestima.

Sea $X \in \mathbb{R}$ y supongamos que $n_1 = n_2 = 1$, o sea, tenemos las observaciones x_1 y x_2 de la población 1 y 2, respectivamente. Supongamos $x_1 > x_2$.

Consideremos la regla de clasificación que asigna x a $\mathcal{P}_1$ si $x \leq (x_1 + x_2)/2$ y al grupo 2 en otro caso.

Entonces, $\widehat{e}_{app} = 0$, lo cual es demasiado optimista.

Estimación de los errores de clasificación

3) El estimador de convalidación cruzada.
En este método se sacan las observaciones de a una. Con los $n-1$ datos restantes se arma la regla y se clasifica la observación extraída. Sea

$$a_i = \#\{\mathbf{x}_{i\ell} \text{ mal clasificadas }, 1 \leq \ell \leq n_i\}$$

$$e_{i,cv} = \frac{a_i}{n_i} \qquad e_{cv} = \sum_{i=1}^{k} \pi_i e_{i,cv}$$

$$\widehat{e}_{cv} = \sum_{i=1}^{k} \widehat{\pi}_i e_{i,cv} = \frac{\sum_{i=1}^{k} a_i}{n}$$

Estimación de los errores de clasificación

3) El estimador de convalidación cruzada.

   Este método da estimadores consistentes del error pero con varianza grande.
   Obviamente, es más costoso computacionalmente pero da resultados más honestos y debería ser usado si es posible.
   En el caso normal, las fórmulas para los estimadores de los parámetros evitan efectuar el cálculo de la regla en cada paso.

4) Otra opción es usar el método de $M-$fold que divide la muestra total en M grupos y construye la regla con $M-1$ grupos mientras clasifica el grupo restante, sucesivamente.

Estimación de los errores de clasificación

5) El estimador bootstrap.

Como $e_{i,app}$ es sesgado, Efron (1979) sugiere estimar su sesgo usando boostrap.

a) Para cada $1 \leq i \leq k$, tomamos una muestra $\mathbf{x}^*_{i\ell}$ con reemplazo de la muestra original de la población $\mathcal{P}_i$, de tamaño n_i.

b) Construyamos la regla de clasificación basada en esta muestras que llamaremos $\widehat{G}^{\star,*}_0$ con regiones $\widehat{\mathcal{G}}^*_{i,0}$. Sean

$$m_i^* = \#\{\mathbf{x}^*_{i\ell} \text{ mal clasificadas } 1 \leq \ell \leq n_i\} = \#\{\mathbf{x}^*_{i\ell} \notin \widehat{\mathcal{G}}^*_{j,0}\ 1 \leq \ell \leq n_i\}$$

$$m_i^{**} = \#\{\mathbf{x}_{i\ell} \text{ mal clasificadas } 1 \leq \ell \leq n_i\} = \#\{\mathbf{x}_{i\ell} \notin \widehat{\mathcal{G}}^*_{j,0}\ 1 \leq \ell \leq n_i\}$$

$$d_i = \frac{m_i^{**} - m_i^*}{n_i}$$

c) Repitase a) y b) un número B grande de veces. Sea $d_{i,s}$ el valor de d_i en la replicación s y defina $\overline{d}_i = \sum_{s=1}^{B} d_{i,s}$

Estimación de los errores de clasificación

5) El estimador bootstrap.
El estimador bootstrap se define como

$$e_{i,boot} = \frac{m_i}{n_i} + \overline{d}_i \qquad e_{boot} = \sum_{i=1}^{k} \pi_i e_{i,boot}$$

$$\widehat{e}_{boot} = \sum_{i=1}^{k} \widehat{\pi}_i e_{i,boot}$$

$$f_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

Supongamos que $c_{i|j} = 1$ si $i \neq j$ entonces la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \mathcal{G}_{i,0}$ donde

$$\begin{aligned} \mathcal{G}_{i,0} &= \{\mathbf{x} \in \mathbb{R}^p : \quad \frac{\pi_\ell}{\pi_i} < \frac{f_i(\mathbf{x})}{f_\ell(\mathbf{x})} \quad \forall \ell \neq i\} \\ &= \{\mathbf{x} \in \mathbb{R}^p : \quad \log\left(\frac{f_i(\mathbf{x})}{f_\ell(\mathbf{x})}\right) > \log\left(\frac{\pi_\ell}{\pi_i}\right) \quad \forall \ell \neq i\} \end{aligned}$$

donde

$$\begin{aligned} \log\left(\frac{f_i(\mathbf{x})}{f_\ell(\mathbf{x})}\right) &= \frac{1}{2}\log\frac{\det(\boldsymbol{\Sigma}_\ell)}{\det(\boldsymbol{\Sigma}_j)} - \frac{1}{2}\left(\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^{\mathrm{T}}\boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\mu}_\ell\right) \\ &\quad -\frac{1}{2}\left\{\mathbf{x}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_i^{-1} - \boldsymbol{\Sigma}_\ell^{-1}\right)\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \boldsymbol{\Sigma}_\ell^{-1}\boldsymbol{\mu}_\ell\right)\right\} \end{aligned}$$

o sea, obtenemos una forma cuadrática en $\mathbf{x}$.

Caso $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$

En este caso,

$$\log\left(\frac{f_i(\mathbf{x})}{f_\ell(\mathbf{x})}\right) = \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i) - \boldsymbol{\mu}_\ell^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_\ell)$$

Por lo tanto, si llamamos

$$L_i(\mathbf{x}) = \log \pi_i + \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i)$$

asigno $\mathbf{x}$ al grupo con mayor $L_i(\mathbf{x})$, o sea, clasifico $\mathbf{x} \in \mathcal{P}_i$ si

$$L_i(\mathbf{x}) = \max_{1 \le \ell \le k} L_\ell(\mathbf{x})$$

Si $\pi_j = \frac{1}{k}$ para todo j, esta regla de clasificación es la obtenida antes con las coordenadas discriminantes.

Caso $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$

En este caso, asigno $\mathbf{x}$ al grupo con mayor $L_i(\mathbf{x})$ donde llamamos

$$L_i(\mathbf{x}) = \log \pi_i + \boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i)$$

Las funciones

$$d_{i\ell}(\mathbf{x}) = L_i(\mathbf{x}) - L_\ell(\mathbf{x}) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \frac{(\boldsymbol{\mu}_i + \boldsymbol{\mu}_\ell)}{2} \right) + \log \pi_i - \log \pi_\ell$$

se llaman funciones discriminantes y $d_{i\ell}(\mathbf{x}) = -d_{\ell i}(\mathbf{x})$. Sea $\boldsymbol{\alpha}_{i,\ell} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)$, Luego

$$d_{i\ell}(\mathbf{x}) = \boldsymbol{\alpha}_{i,\ell}^{\mathrm{T}} \left(\mathbf{x} - \frac{(\boldsymbol{\mu}_i + \boldsymbol{\mu}_\ell)}{2} \right) + \log \pi_i - \log \pi_\ell$$

Caso $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$

Si tenemos k poblaciones, sólo necesitamos encontrar $r = \min(p, k-1)$ direcciones de proyección en lugar de $\binom{k}{2}$.

Efectivamente, basta conocer $\boldsymbol{\alpha}_{i,i+1}$, $1 \leq i \leq k-1$ ya que

$$\boldsymbol{\alpha}_{i,i+2} = \boldsymbol{\alpha}_{i,i+1} - \boldsymbol{\alpha}_{i+1,i+2}$$

Ejemplo: Si $k = 3$ y obtenemos que $L_1(\mathbf{x}) > L_2(\mathbf{x})$ y $L_2(\mathbf{x}) > L_3(\mathbf{x})$ entonces $L_1(\mathbf{x}) > L_3(\mathbf{x})$.

Si además $p = 2$ cada ecuación $d_{i\ell}(\mathbf{x}) = 0$ es una recta y las tres rectas se cortan en el mismo punto ya que

$$d_{13}(\mathbf{x}) = L_1(\mathbf{x}) - L_3(\mathbf{x}) = L_1(\mathbf{x}) - L_2(\mathbf{x}) + (L_2(\mathbf{x}) - L_3(\mathbf{x})) = d_{12}(\mathbf{x}) + d_{23}(\mathbf{x})$$

Caso $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$

Si $k = 2$, como $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

$$d_{12}(\mathbf{x}) = \boldsymbol{\alpha}^{\mathrm{T}} \left(\mathbf{x} - \frac{(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}{2} \right) + \log \pi_1 - \log \pi_2$$

es la regla discriminante lineal de Fisher que clasifica en el grupo 1 si $d_{12}(\mathbf{x}) > 0$.

El hiperplano $d_{12}(\mathbf{x}) = 0$ determina un hiperplano que separa los dos grupos.

Caso $\boldsymbol{\Sigma}_\ell = \boldsymbol{\Sigma}$

Veamos que si $\pi_j = \frac{1}{k}$ para todo j, la regla de clasificación Bayes es la obtenida antes con las coordenadas discriminantes. Sea

- $\boldsymbol{\Sigma}_{\mathrm{B}} = \sum_{i=1}^{k} \pi_i (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^{\mathrm{T}}$, $s = \text{rango}(\boldsymbol{\Sigma}_{\mathrm{B}})$
- $\mathbf{z} = \mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{x} = (\mathbf{z}_1^{\mathrm{T}}, \mathbf{z}_2^{\mathrm{T}})^{\mathrm{T}}$ el vector de variables discriminantes con

$$\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p) = (\mathbf{A}_1, \mathbf{A}_2) \quad \text{donde} \quad \mathbf{A}_1 = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_s)$$

- $\mathbf{z}^{(1)} = \mathbf{A}_1^{\mathrm{T}}\mathbf{x}$, $\mathbf{z}^{(2)} = \mathbf{A}_2^{\mathrm{T}}\mathbf{x}$
- $\boldsymbol{\nu}_i = \mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}_i$, entonces $\boldsymbol{\nu}_i^{(2)} = \boldsymbol{\nu}^{(2)}$

Si $\mathbf{x}|G = i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ vimos que

$$\mathbf{z}|G = i \sim N(\boldsymbol{\nu}_i, \mathbf{I}_p)$$

o sea,

$$\mathbf{z}^{(1)}|G = i \sim N(\boldsymbol{\nu}_i^{(1)}, \mathbf{I}_s) \qquad \mathbf{z}^{(2)} \sim N(\boldsymbol{\nu}^{(2)}, \mathbf{I}_{p-s})$$

Caso $\mathbf{\Sigma}_\ell = \mathbf{\Sigma}$

La regla que vimos asignaba $\mathbf{x}_0$ al grupo i si $\mathbf{v}_0 = \mathbf{A}_1^{\mathrm{T}}\mathbf{x}_0 \in \mathcal{G}_i$ donde

$$\begin{aligned}\mathcal{G}_i &= \{\mathbf{v} \in \mathbb{R}^s : \quad \|\mathbf{v} - \boldsymbol{\nu}_i^{(1)}\| < \|\mathbf{v} - \boldsymbol{\nu}_\ell^{(1)}\| \quad \forall \ell \neq i\} \\ &= \{\mathbf{v} \in \mathbb{R}^s : \quad (\boldsymbol{\nu}_i^{(1)})^{\mathrm{T}}(\mathbf{v} - \frac{1}{2}\boldsymbol{\nu}_i^{(1)}) > (\boldsymbol{\nu}_\ell^{(1)})^{\mathrm{T}}(\mathbf{v} - \frac{1}{2}\boldsymbol{\nu}_\ell^{(1)}) \quad \forall \ell \neq i\}\end{aligned}$$

Esta regla es Bayes cuando $\pi_j = \frac{1}{k}$ para todo j.

Para una probabilidad a priori τ general tenemos que modificar $\mathcal{G}_i$ por $\mathcal{G}_{i,\tau}$, o sea, asigno $\mathbf{x}_0$ al grupo i si $\mathbf{v}_0 = \mathbf{A}_1^{\mathrm{T}}\mathbf{x}_0 \in \mathcal{G}_{i,\tau}$ donde

$$\begin{aligned}\mathcal{G}_{i,\tau} &= \{\mathbf{v} \in \mathbb{R}^s : \|\mathbf{v} - \boldsymbol{\nu}_i^{(1)}\|^2 - 2\log(\pi_i) < \|\mathbf{v} - \boldsymbol{\nu}_\ell^{(1)}\|^2 - 2\log(\pi_\ell)\ \forall \ell \neq i\} \\ &= \{\mathbf{v} \in \mathbb{R}^s : (\boldsymbol{\nu}_i^{(1)})^{\mathrm{T}}(\mathbf{v} - \frac{1}{2}\boldsymbol{\nu}_i^{(1)}) > (\boldsymbol{\nu}_\ell^{(1)})^{\mathrm{T}}(\mathbf{v} - \frac{1}{2}\boldsymbol{\nu}_\ell^{(1)}) \\ &\quad + \log(\pi_\ell) - \log(\pi_i) \quad \forall \ell \neq i\}\end{aligned}$$

$$f_j \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j),\ j = 1, 2$$

Definición. Supongamos que $\mathbf{x} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, 2$. Se dice que $\mathbf{x}$ está dada en forma canónica si $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_p$ y $\boldsymbol{\Sigma}_2$ es diagonal. Llamaremos $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_2 = \text{diag}(\lambda_1, \ldots, \lambda_p)$ y $\boldsymbol{\mu}_2 = \boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^{\mathrm{T}}$.

Propiedad. Supongamos que $\mathbf{x} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, 2$ donde $\boldsymbol{\Sigma}_1 > 0$, $\boldsymbol{\Sigma}_2 > 0$. Entonces existe $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times p}$ no singular tal que $\mathbf{z} = \boldsymbol{\Gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_1)$ está en forma canónica.

Por lo tanto, si $c_{i|j} = 1$ si $i \neq j$ entonces la regla Bayes clasifica $\mathbf{x} \in \mathcal{P}_1$ si $\mathbf{z} = \mathbf{\Gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_1) \in \mathcal{G}_{1,0}$ donde

$$\mathcal{G}_{1,0} = \{\mathbf{z} \in \mathbb{R}^p : \quad \log\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) > \log\left(\frac{\pi_2}{\pi_1}\right)\} = \{\mathbf{z} \in \mathbb{R}^p : \ Q(\mathbf{z}) > 0\}$$

con

$$\begin{aligned} Q(\mathbf{z}) &= \sum_{i=1}^{p} a_{ii} z_i^2 + \sum_{i=1}^{p} b_i z_i + c \\ a_{ii} &= \frac{1}{2}\left(\frac{1}{\lambda_i} - 1\right) \qquad\qquad b_i = -\frac{\delta_i}{\lambda_i} \\ c &= \log\left(\frac{\pi_1}{\pi_2}\right) + \frac{1}{2}\sum_{s=1}^{p} \log \lambda_s + \frac{1}{2}\sum_{s=1}^{p} \frac{\delta_s^2}{\lambda_s} \end{aligned}$$

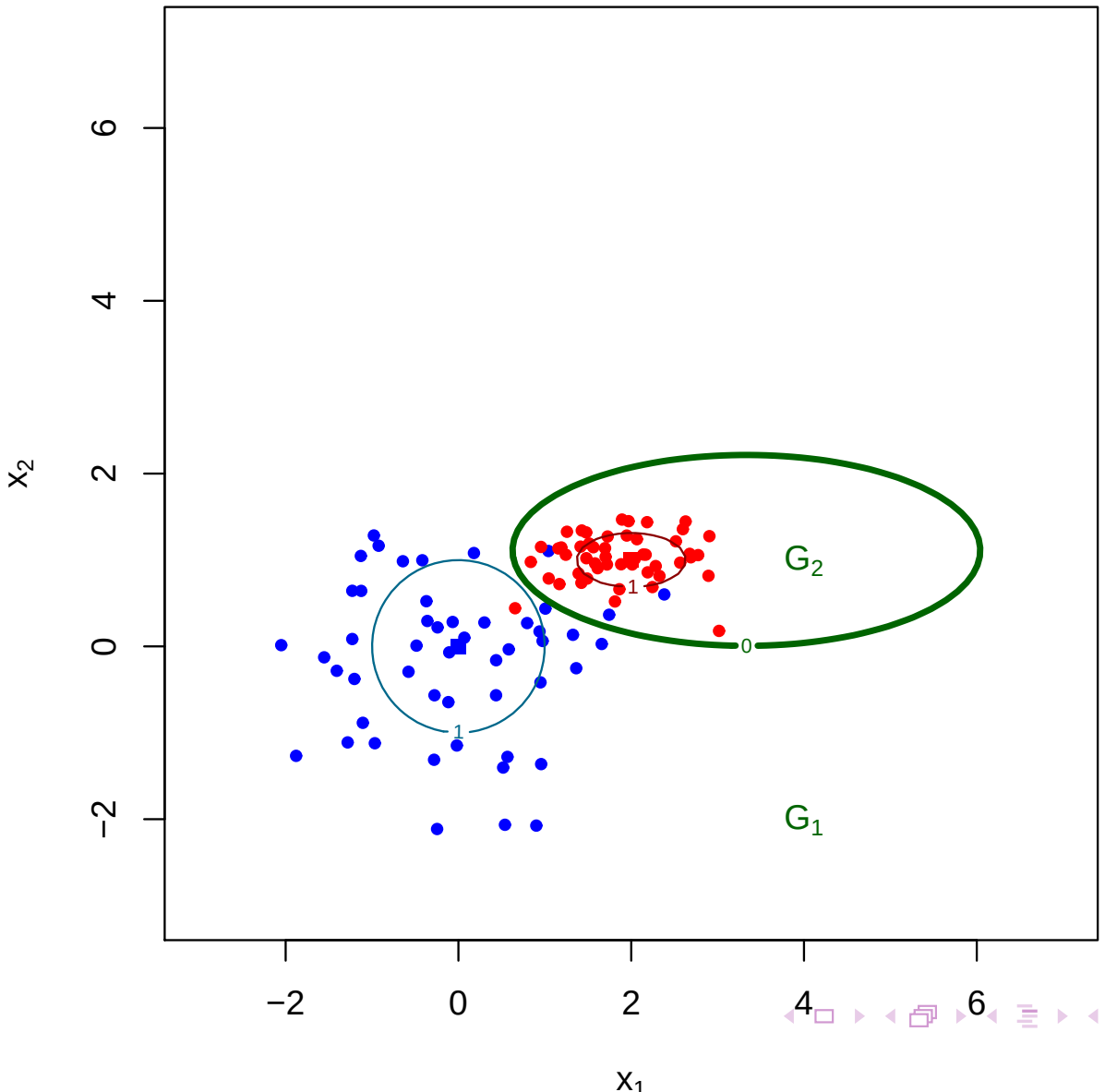
La ventaja de la forma canónica es que los términos de la forma $a_{ij}x_i x_j$ desaparecen lo que hace más fácil de entender la regla de clasificación.

Ejemplo $p = 2$

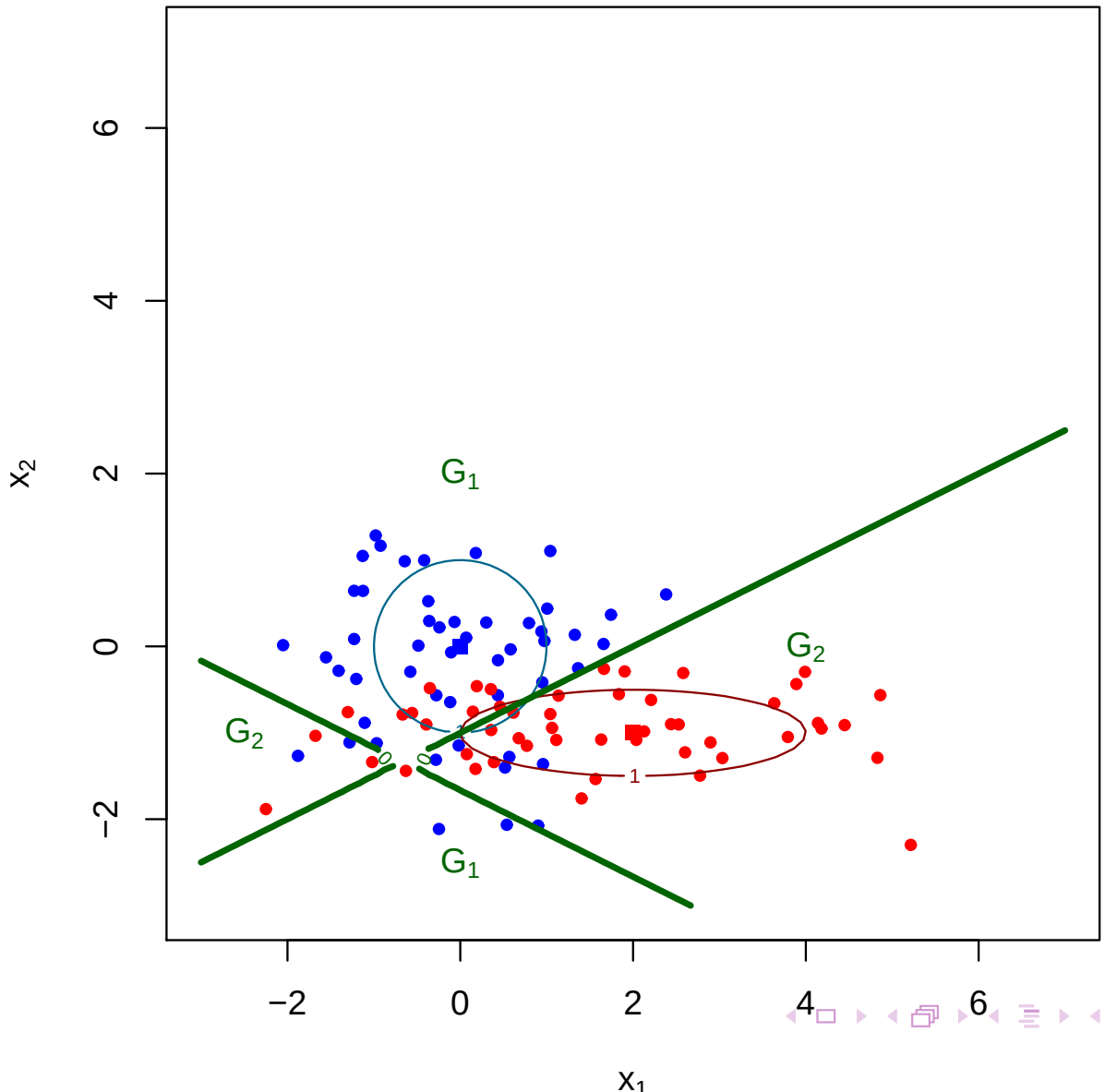
Tomemos $\pi_1 = \pi_2 = 1/2$ y los siguientes valores para $\boldsymbol{\delta}$ y $\boldsymbol{\Lambda}$

Caso	δ_1	δ_2	λ_1	λ_2	a_{11}	a_{22}	b_1	b_2	c
A	2	1	0.4	0.1	0.75	4.5	-5	-10	8.391
B	2	-1	4	0.25	-0.375	1.5	-0.5	4	2.5
C	3	1	4	1	-0.375	0	-0.75	-1	2.318
D	2	0	10	1	-0.45	0	-0.2	0	1.351

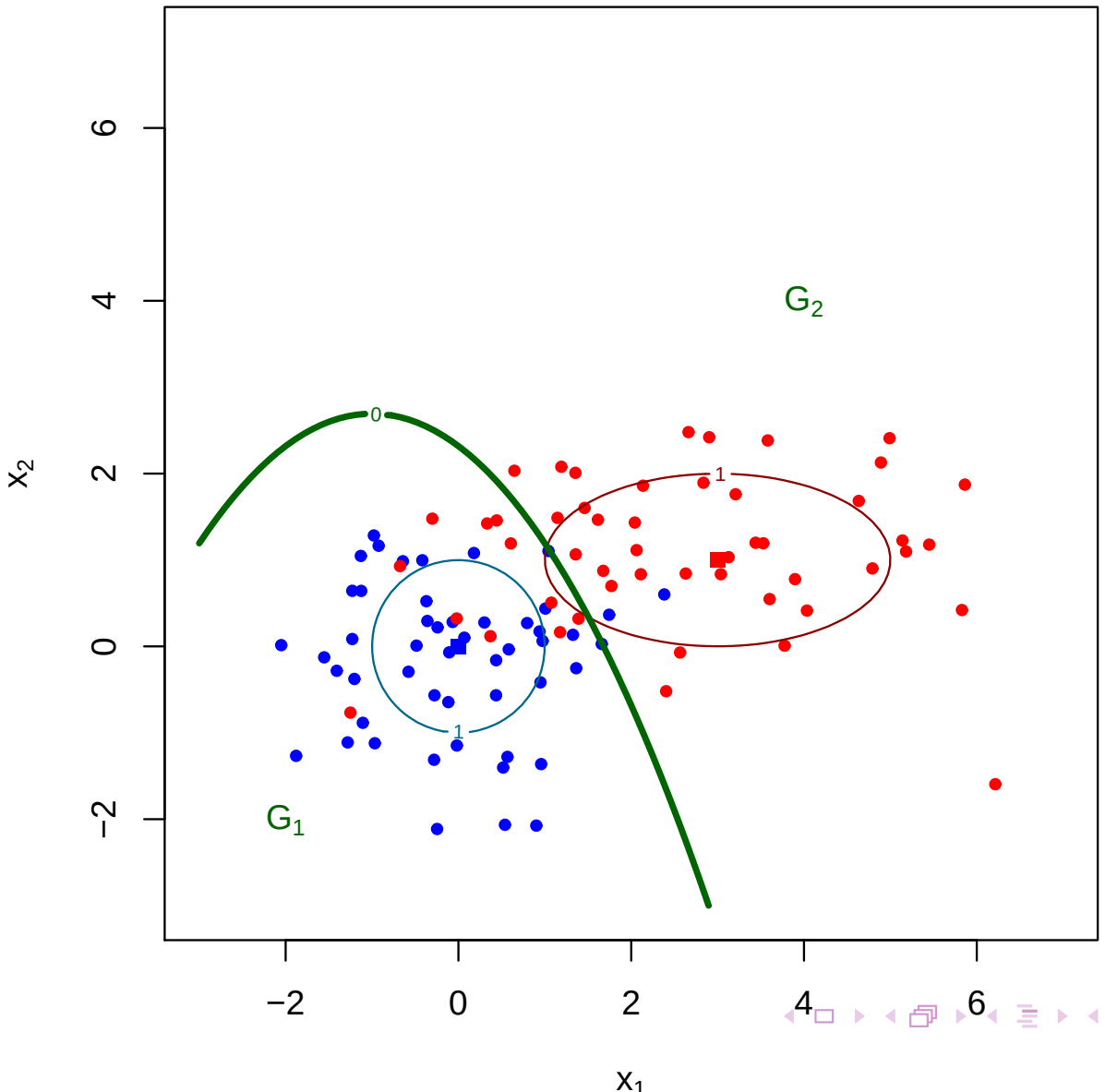
Ejemplo $p = 2$, Caso A



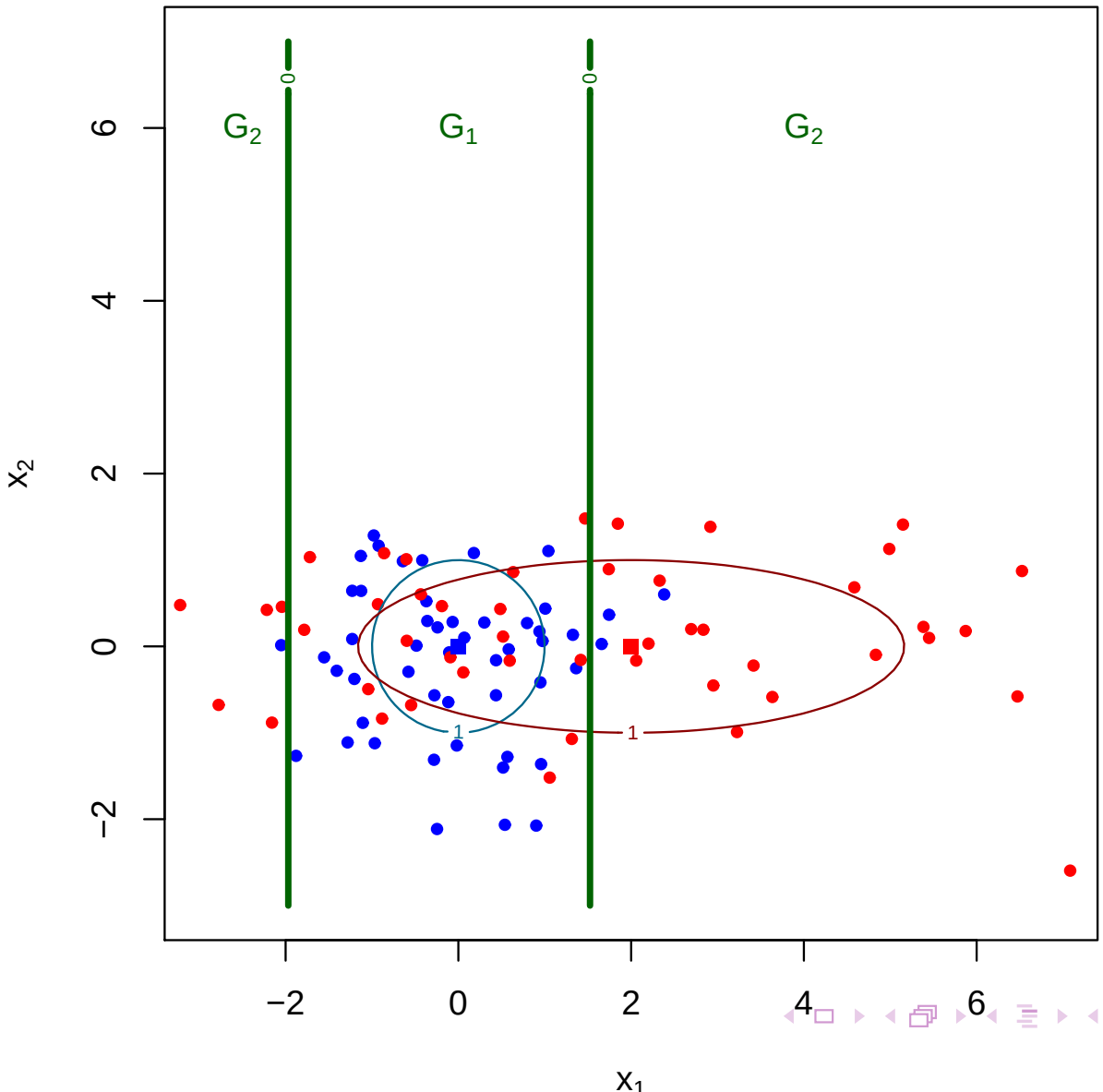
Ejemplo $p = 2$, Caso B



Ejemplo $p = 2$, Caso C, $a_{22} = 0$

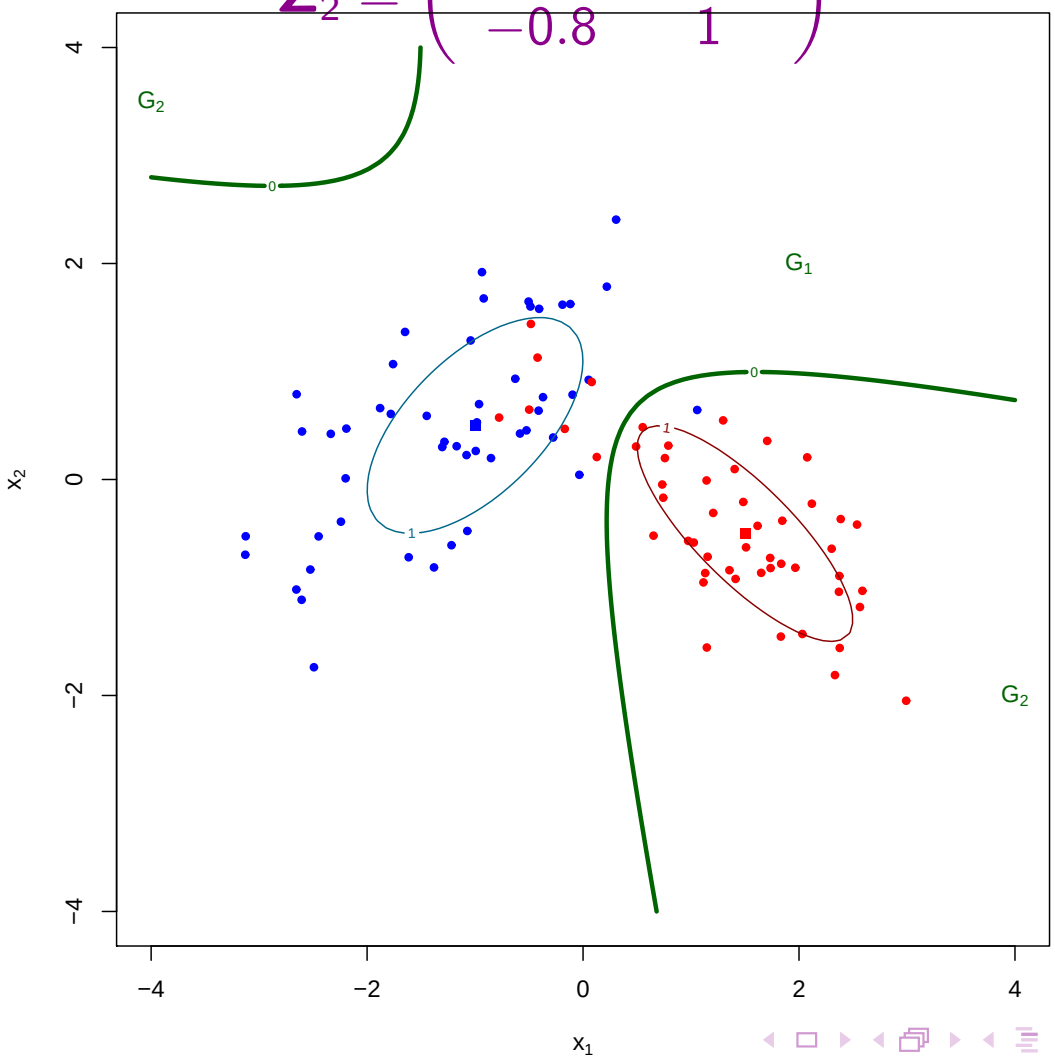


Ejemplo $p = 2$, Caso D, $a_{22} = 0$, $b_2 = 0$



Ejemplo $\boldsymbol{\mu}_1 = (-1, 0.5)$, $\boldsymbol{\mu}_2 = (1.5, -0.5)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$,

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$$



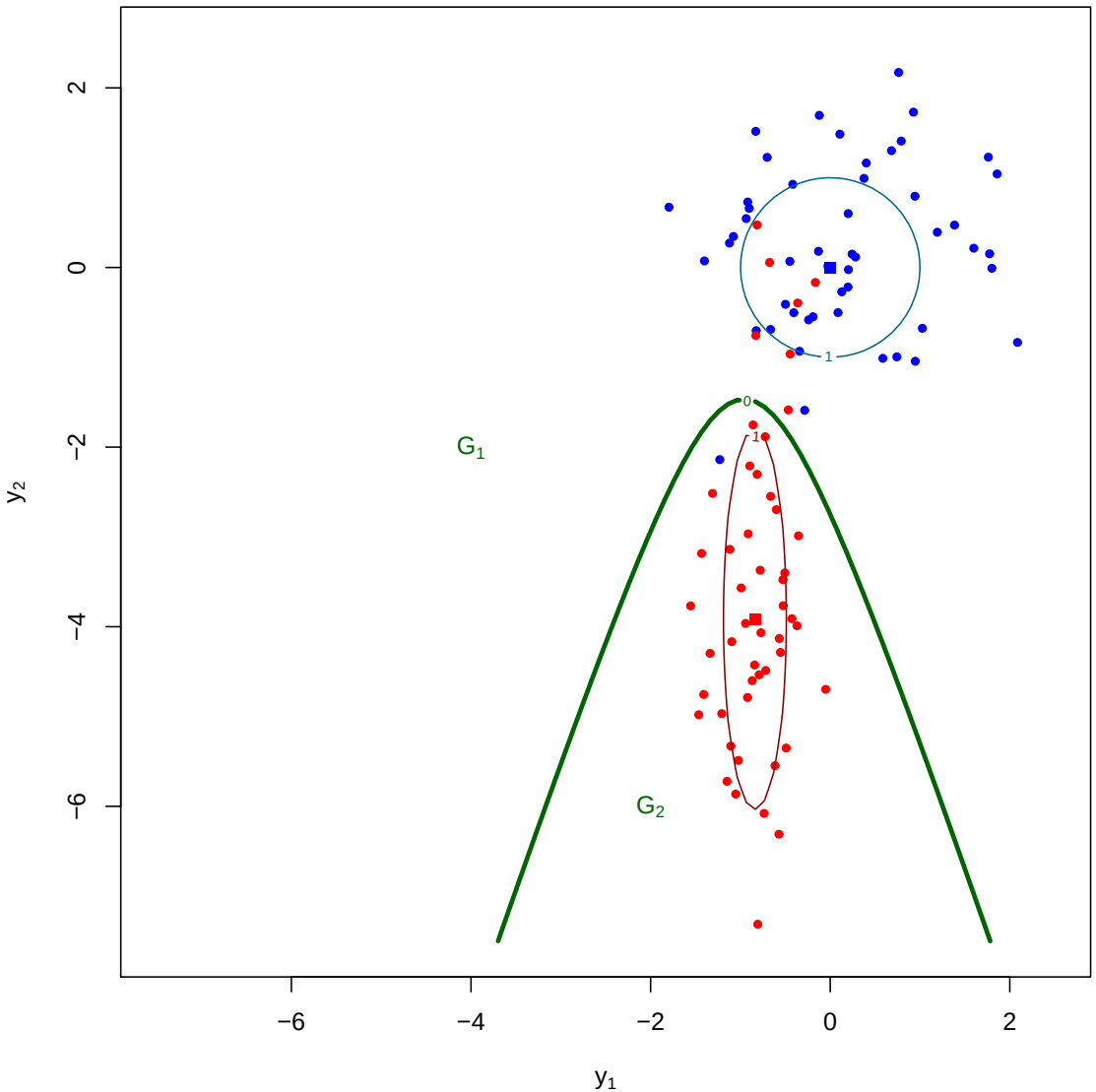
Ejemplo $\boldsymbol{\mu}_1 = (-1, 0.5)$, $\boldsymbol{\mu}_2 = (1.5, -0.5)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$,

$$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$$

Datos transformados $\mathbf{z} = \boldsymbol{\Gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_1)$.

$$\begin{aligned} \boldsymbol{\Gamma} &= \begin{pmatrix} -0.5590 & -1.1180 \\ -0.5590 & 1.1180 \end{pmatrix} \\ \lambda_1 &= 0.125 \qquad \lambda_2 = 4.5 \\ a_{11} &= 3.500 \qquad a_{22} = -0.3889 \\ b_1 &= 6.708 \qquad b_2 = 0.8696 \qquad c = 5.325 \end{aligned}$$

Datos transformados $\mathbf{z} = \mathbf{\Gamma}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}_1)$



Cálculo errores de clasificación $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, 2$

- $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$,
- la regla $G_0^\star$ clasifica en el grupo 1 si $\mathbf{x} \in \mathcal{G}_{1,0}$, con $\mathcal{G}_{1,0} = \{\mathbf{x} : d_{12}(\mathbf{x}) > 0\}$

$$d_{12}(\mathbf{x}) = \boldsymbol{\alpha}^{\mathrm{T}} \left(\mathbf{x} - \frac{(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)}{2} \right) + \log \pi_1 - \log \pi_2$$

$$\Delta_p^2 = \boldsymbol{\alpha}^{\mathrm{T}} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\alpha}$$

Luego

$$R(\mathcal{P}_1, G_0^\star) = \int_{\mathcal{G}_{2,0}} f_1(\mathbf{x}) d\mathbf{x} = 1 - \int_{\mathcal{G}_{1,0}} f_1(\mathbf{x}) d\mathbf{x}$$

Cálculo errores de clasificación $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, 2$

En $\mathcal{P}_1$, $\mathbf{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ luego

$$\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} \sim N(\boldsymbol{\alpha}^{\mathrm{T}}\mu_1, \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}) = N(\boldsymbol{\alpha}^{\mathrm{T}}\mu_1, \Delta_p^2)$$

de donde

$$R(\mathcal{P}_1, G_0^\star) = \Phi\left(\frac{\log\left(\frac{\pi_2}{\pi_1}\right) - \frac{1}{2}\Delta_p^2}{\Delta_p}\right)$$

$$R(\mathcal{P}_2, G_0^\star) = \Phi\left(-\frac{\log\left(\frac{\pi_2}{\pi_1}\right) + \frac{1}{2}\Delta_p^2}{\Delta_p}\right)$$

$$e_{opt} = r(\tau, G_0^\star) = \pi_1\Phi\left(\frac{\log\left(\frac{\pi_2}{\pi_1}\right) - \frac{1}{2}\Delta_p^2}{\Delta_p}\right) + \pi_2\Phi\left(-\frac{\log\left(\frac{\pi_2}{\pi_1}\right) + \frac{1}{2}\Delta_p^2}{\Delta_p}\right)$$

Cálculo errores de clasificación $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, 2$

Si $\pi_1 = \pi_2 = 1/2$ entonces

$$R(\mathcal{P}_1, G_0^\star) = R(\mathcal{P}_2, G_0^\star) = \Phi\left(-\frac{1}{2}\Delta_p\right)$$

y $G_0^\star$ es minimax.

En general, la regla minimax asigna al grupo 1 si $D(\mathbf{x}) = \boldsymbol{\alpha}^{\mathrm{T}}\left(\mathbf{x} - \frac{(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)}{2}\right) > \log(c)$ donde c se elige de modo que

$$\Phi\left(\frac{\log(c) - \frac{1}{2}\Delta_p^2}{\Delta_p}\right) = \Phi\left(\frac{-\log(c) - \frac{1}{2}\Delta_p^2}{\Delta_p}\right)$$

que tiene como solution $c = 1$ coincidiendo con el método de cociente de verosimilitud.

Cálculo errores de clasificación $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, \ldots, k$

- la regla $G_0^\star$ clasifica en el grupo i si $\mathbf{x} \in \mathcal{G}_{i,0}$, con

$$\mathcal{G}_{i,0} = \{\mathbf{x} : d_{i\ell}(\mathbf{x}) > 0 \forall \ell \neq i\} = \{\mathbf{x} : L_i(\mathbf{x}) = \max_\ell L_\ell(\mathbf{x})\}$$

$$d_{i\ell}(\mathbf{x}) = L_i(\mathbf{x}) - L_\ell(\mathbf{x}) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \frac{(\boldsymbol{\mu}_i + \boldsymbol{\mu}_\ell)}{2} \right) + \log \pi_i - \log \pi_\ell$$

- $\Delta^2_{i\ell} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)$

Luego

$$R(\mathcal{P}_i, G_0^\star) = \sum_{\ell \neq i} \int_{\mathcal{G}_{\ell,0}} f_i(\mathbf{x}) d\mathbf{x} = 1 - \int_{\mathcal{G}_{i,0}} f_i(\mathbf{x}) d\mathbf{x}$$

Cálculo errores de clasificación $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, $j = 1, \ldots, k$

En $\mathcal{P}_i$, $\mathbf{x} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ luego

$$d_{i\ell}(\mathbf{x}) \sim N\left(\frac{1}{2}\Delta^2_{i\ell} + \log \pi_i - \log \pi_\ell, \Delta^2_{i\ell}\right)$$

Más aún, el vector $\mathbf{d}_i(\mathbf{x}) = (d_{i\ell}(\mathbf{x}))_{\ell \neq i}$ tiene distribución normal $(k-1)-$variada y

$$\operatorname{Cov}(d_{i\ell}(\mathbf{x}), d_{ij}(\mathbf{x})) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

Si $k = 3$ se pueden calcular, fácilmente.

$f_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ con $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ desconocidos

Supongamos tener $\mathbf{x}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $1 \leq j \leq n_i$, $1 \leq i \leq k$ entonces estimamos $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}_i$ por

$$\widehat{\boldsymbol{\mu}}_i = \overline{\mathbf{x}}_i \qquad \mathbf{S}_i = \frac{\mathbf{Q}_i}{n_i - 1} = \frac{1}{n_i} \sum_{j=1}^{n_i} (\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{i,j} - \overline{\mathbf{x}}_i)^{\mathrm{T}}$$

Clasificamos $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \widehat{\mathcal{G}}_{i,0}$ con

$$\widehat{\mathcal{G}}_{i,0} = \{\mathbf{x} \in \mathbb{R}^p : \quad \log\left(\frac{\widehat{f}_i(\mathbf{x})}{\widehat{f}_\ell(\mathbf{x})}\right) > \log\left(\frac{\pi_\ell}{\pi_i}\right) \quad \forall \ell \neq i\}$$

donde $\widehat{f}_i(\mathbf{x}) = f(\mathbf{x}, \widehat{\boldsymbol{\mu}}_i, \mathbf{S}_i)$ con $f(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\begin{aligned}\log\left(\frac{\widehat{f}_i(\mathbf{x})}{\widehat{f}_\ell(\mathbf{x})}\right) &= \frac{1}{2}\log\frac{\det(\mathbf{S}_\ell)}{\det(\mathbf{S}_j)} - \frac{1}{2}\left(\widehat{\boldsymbol{\mu}}_i^{\mathrm{T}}\mathbf{S}_i^{-1}\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_\ell^{\mathrm{T}}\mathbf{S}_\ell^{-1}\widehat{\boldsymbol{\mu}}_\ell\right) \\ &\quad -\frac{1}{2}\left\{\mathbf{x}^{\mathrm{T}}\left(\mathbf{S}_i^{-1} - \mathbf{S}_\ell^{-1}\right)\mathbf{x} - 2\mathbf{x}^{\mathrm{T}}\left(\mathbf{S}_i^{-1}\widehat{\boldsymbol{\mu}}_i - \mathbf{S}_\ell^{-1}\widehat{\boldsymbol{\mu}}_\ell\right)\right\}\end{aligned}$$

$f_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ con $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}$ desconocidos

Supongamos tener $\mathbf{x}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $1 \leq j \leq n_i$, $1 \leq i \leq k$ entonces estimamos $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}$ por

$$\widehat{\boldsymbol{\mu}}_i = \overline{\mathbf{x}}_i \qquad \mathbf{S} = \frac{1}{n-k}\sum_{i=1}^{k}\mathbf{Q}_i = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{i,j}-\overline{\mathbf{x}}_i)(\mathbf{x}_{i,j}-\overline{\mathbf{x}}_i)^{\mathrm{T}}$$

Clasificamos $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \widehat{\mathcal{G}}_{i,0}$ con

$$\widehat{\mathcal{G}}_{i,0} = \{\mathbf{x} \in \mathbb{R}^p : \quad \widehat{L}_i(\mathbf{x}) > \widehat{L}_\ell(\mathbf{x}) \quad \forall \ell \neq i\} = \{\mathbf{x} \in \mathbb{R}^p : \quad \widehat{d}_{i\ell}(\mathbf{x}) > 0\}$$

donde

$$\widehat{L}_i(\mathbf{x}) = \log \pi_i + \widehat{\boldsymbol{\mu}}_i^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x} - \frac{1}{2}\widehat{\boldsymbol{\mu}}_i)$$

$$\begin{aligned}\widehat{d}_{i\ell}(\mathbf{x}) &= \widehat{L}_i(\mathbf{x}) - \widehat{L}_\ell(\mathbf{x}) \\ &= \log\left(\frac{\pi_i}{\pi_\ell}\right) + (\widehat{\boldsymbol{\mu}}_i - \widehat{\boldsymbol{\mu}}_\ell)^{\mathrm{T}}\mathbf{S}^{-1}\left(\mathbf{x} - \frac{\widehat{\boldsymbol{\mu}}_i + \widehat{\boldsymbol{\mu}}_\ell}{2}\right)\end{aligned}$$

Estimación de los errores de clasificación $k = 2$, $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$

En el caso normal clasificamos $\mathbf{x} \in \mathcal{P}_1$ si $\widehat{d}_{12}(\mathbf{x}) > 0$,

$$\begin{aligned}\widehat{d}_{12}(\mathbf{x}) &= \log\left(\frac{\pi_1}{\pi_2}\right) + (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}}\mathbf{S}^{-1}\left(\mathbf{x} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) \\ &= \log\left(\frac{\pi_1}{\pi_2}\right) + \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\left(\mathbf{x} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right)\end{aligned}$$

Por lo tanto, como si $\mathbf{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ y si llamamos $\sigma^2 = \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\Sigma}\widehat{\boldsymbol{\alpha}}$, $\mathbf{X}_1 = (\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,n_1})$ y $\mathbf{X}_2 = (\mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,n_2})$, entonces

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\left(\mathbf{x} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right)|(\mathbf{X}_1, \mathbf{X}_2) &\sim N\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\left(\boldsymbol{\mu}_1 - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right), \sigma^2\right) \\ \widehat{d}_{12}(\mathbf{x}) &\sim N\left(\widehat{d}_{12}(\boldsymbol{\mu}_1), \sigma^2\right)\end{aligned}$$

Estimación de los errores de clasificación $k = 2$, $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$

$$
\begin{aligned}
e_{1,act} &= \int_{\widehat{\mathcal{G}}_{2,0}} f_1(\mathbf{x})d\mathbf{x} = \int_{\widehat{d}_{12}(\mathbf{x})<0} f_1(\mathbf{x})d\mathbf{x} \\
&= \mathbb{P}\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\left(\mathbf{x} - \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2}\right) + \log\left(\frac{\pi_1}{\pi_2}\right) < 0 | \mathbf{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\right) \\
&= \Phi\left(-\frac{\widehat{d}_{12}(\boldsymbol{\mu}_1)}{\sigma}\right) = \Phi\left(\frac{\widehat{d}_{21}(\boldsymbol{\mu}_1)}{\sigma}\right)
\end{aligned}
$$

Por lo tanto,

$$
e_{act} = \pi_1 \Phi\left(\frac{\widehat{d}_{21}(\boldsymbol{\mu}_1)}{\sigma}\right) + \pi_2 \Phi\left(\frac{\widehat{d}_{12}(\boldsymbol{\mu}_2)}{\sigma}\right)
$$

Estimación de los errores de clasificación $k = 2$, $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$

$$\widehat{e}_{act} = \pi_1 \Phi \left(\frac{\log\left(\frac{\pi_2}{\pi_1}\right) - \frac{1}{2} D_p^2}{D_p} \right) + \pi_2 \Phi \left(- \frac{\log\left(\frac{\pi_2}{\pi_1}\right) + \frac{1}{2} D_p^2}{D_p} \right)$$

donde

$$D_p^2 = \widehat{\boldsymbol{\alpha}}^{\mathrm{T}} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2) = (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^{\mathrm{T}} \mathbf{S}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)$$

Si π_1 y π_2 son desconocidos se estiman por n_1/n y n_2/n, con $n = n_1 + n_2$. Si $\pi_1 = \pi_2 = 0.5$, entonces

$$\widehat{e}_{act} = \Phi \left(-\frac{1}{2} D_p \right)$$

Estimación de los errores de clasificación $k = 2$, $f_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$

Como $\mathbb{E}\mathcal{F}_{\nu_1,\nu_2}(\lambda) = \nu_2(\nu_1 + \lambda)/[\nu_1(\nu_2 - 2)]$ si $\nu_2 > 2$ y

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T_0^2 \sim \mathcal{F}_{p,n_1+n_2-p-1}(\lambda^2)$$

con $\lambda^2 = \frac{n_1\, n_2}{n_1+n_2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ y

$$T_0^2 = \frac{n_1\, n_2}{n_1 + n_2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^{\mathrm{T}}\mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$$

tenemos que

$$\mathbb{E}D_p^2 = \frac{n-2}{n-p-3}\left(\Delta_p^2 + \frac{pn}{n_1 n_2}\right)$$

D_p^2 sobreestima a Δ_p^2 y por lo tanto, $\widehat{e}_{act}$ subestima el error real e_{act}.

Estimación de los errores de clasificación $k = 2$, $\mathbf{x}|\mathbf{x} \in \mathcal{P}_j \sim f_j = f_j(\mathbf{x}, \boldsymbol{\theta}_j)$

Supongamos que $c_{i|j} = 1$ si $i \neq j$ entonces la regla Bayes clasificaba $\mathbf{x} \in \mathcal{P}_i$ si $\mathbf{x} \in \mathcal{G}_{i,0}$ donde

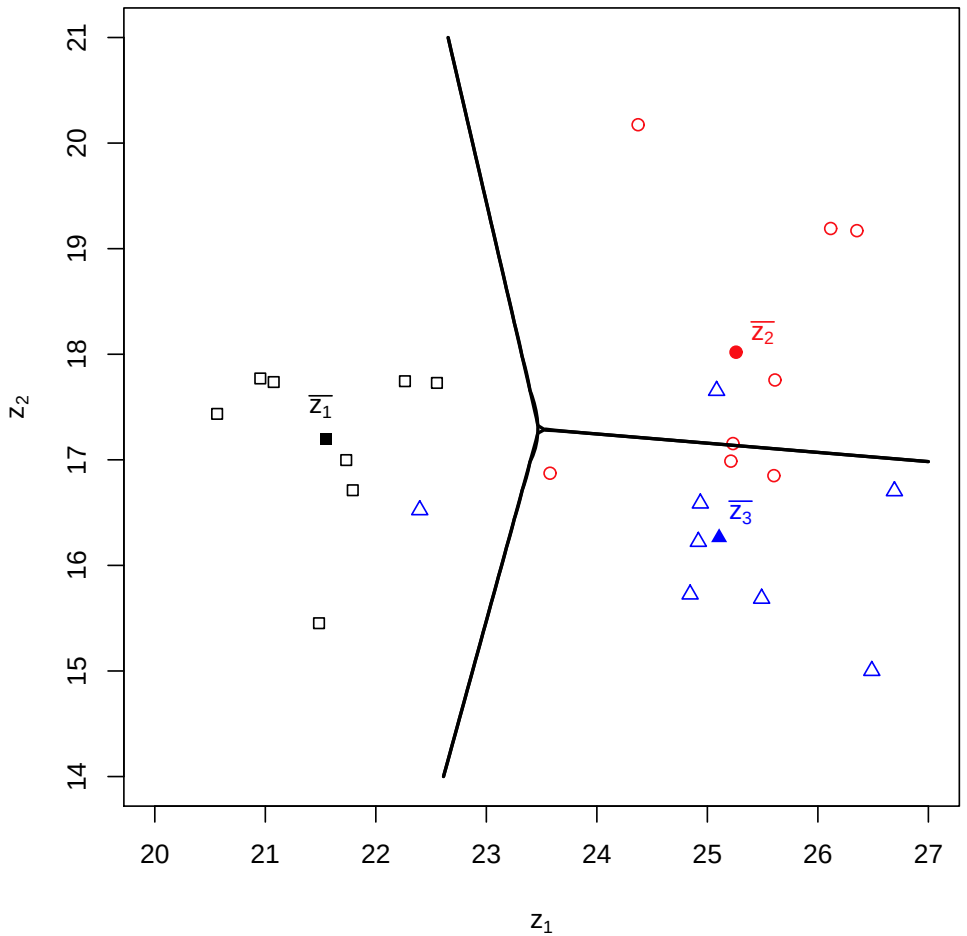
$$\begin{aligned} \mathcal{G}_{i,0} &= \{\mathbf{x} \in \mathbb{R}^p : \quad \pi_\ell \, f_\ell(\mathbf{x}) < \pi_i \, f_i(\mathbf{x}) \quad \forall \ell \neq i\} \\ &= \{\mathbf{x} \in \mathbb{R}^p : \quad r_\ell(\mathbf{x}) < r_i(\mathbf{x}) \quad \forall \ell \neq i\} \end{aligned}$$

es decir, clasifico $\mathbf{x} \in \mathcal{P}_i$ si $r_i(\mathbf{x}) = \max_{1 \leq \ell \leq k} r_\ell(\mathbf{x})$, donde $r_i(\mathbf{x}) = f_i(\mathbf{x}, \boldsymbol{\theta}_i)\pi_i$.

Lema. Sea $\widehat{r}_i(\mathbf{x}) = f_i(\mathbf{x}, \widehat{\boldsymbol{\theta}}_i)\pi_i$ un estimador insesgado de $r_i(\mathbf{x})$, o sea, $\mathbb{E}_{\widehat{\theta}} \widehat{r}_i(\mathbf{x})|\mathbf{x} = r_i(\mathbf{x})$ para casi todo $\mathbf{x}$. Luego

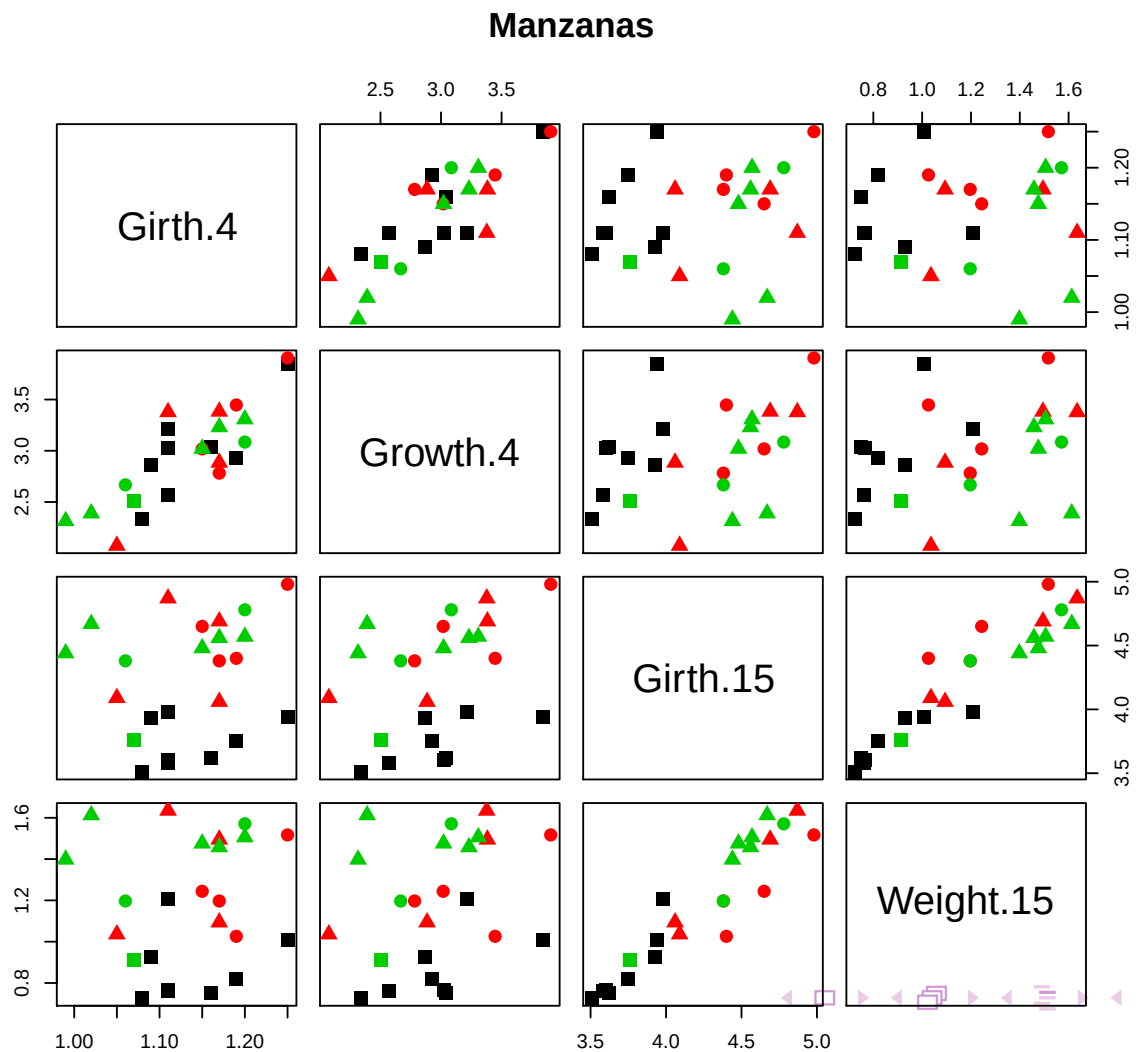
$$\mathbb{E}\widehat{e}_{act} \leq e_{opt} < \mathbb{E}e_{act}$$

Ejemplo Manzanos

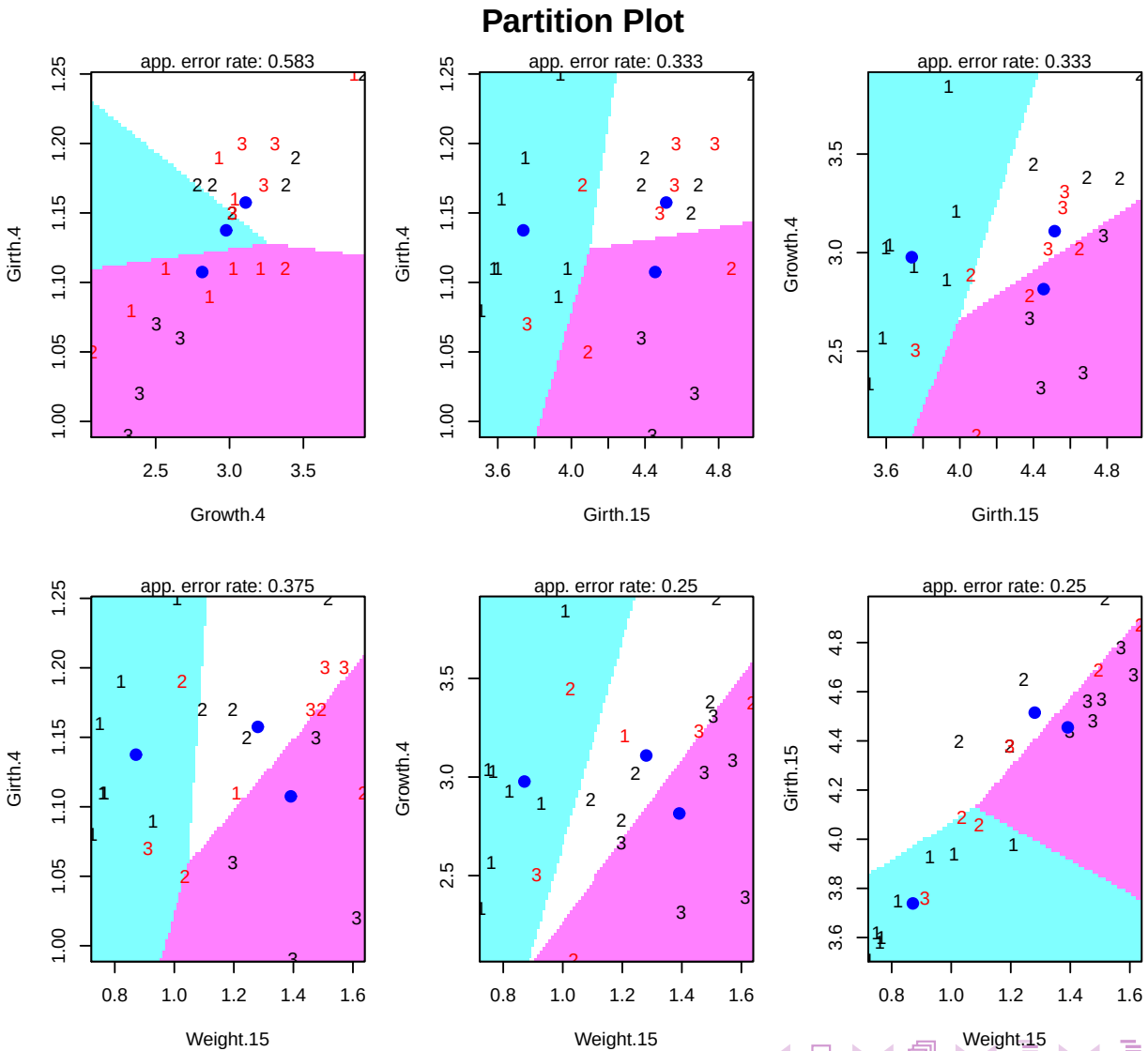


Ejemplo Manzanos

■: Asignado al grupo 1, • asignado al grupo 2, ▲ asignado al grupo 3



Ejemplo Manzanos: En negro bien clasificados, en rojo mal clasificados



Comparación entre LDF y QDF, $k = 2$

- En general la decisión de elegir entre la regla lineal (LDF) y cuadrática (QDF) se hace en base al resultado del test para $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Si el test rechaza se usa QDF.

- A pesar de que esta decisión es razonable ya que LDF es óptima si H_0 es cierta, hay un número importante de trabajos que muestran que aunque no lo sea, LDF es tan buena como QDF.

- Uno podría basar su decisión en elegir el método que da menor error aparente, $\widehat{e}_{app}$, lo cual es peligroso ya que este estimador del error subestima el error real, e_{act}. El cálculo del error actual esperado sólo puede hacerse por simulación.

- Una opción es utilizar el e_{cv} para elegir entre ambas reglas.

Comparación entre LDF y QDF, $k = 2$

En general, LDF es buena para pequeños alejamientos de $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. El mejor comportamiento de QDF depende del tamaño de las muestras y de la dimensión.

- Para n_1 y n_2 pequeñas y $p \leq 6$ hay poca pérdida al elegir LDF.
- Para $n_1, n_2 \leq 25$ y p grande y/o diferencias entre $\mathbf{\Sigma}_1$ y $\mathbf{\Sigma}_2$, LDF es preferible.
- Sin embargo, cuando p grande y $\mathbf{\Sigma}_1$ y $\mathbf{\Sigma}_2$ son muy distintas, la probabilidades de mala clasificación $e_{1,act}$ y $e_{2,act}$ pueden ser muy grandes para un uso práctico

Comparación entre LDF y QDF, $k = 2$

- Si $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ y la diferencia es grande y $p > 6$, QDF es mucho mejor que LDF si el tamaño de muestra es grande.
  - Se recomienda para $p = 4$, $n_1 = n_2 = 25$
  - 25 observaciones adicionales cada dos dimensiones, o sea, para $p = 6, 8, 10$ se necesitan $n_1 = n_2 = 50, 75, 100$
- Para $n_i \geq 100$ y p moderado los resultados asintóticos que favorecen QDF se alcanzan bastante rápido.
- QDF se deteriora rápidamente si p crece porque $\mathbf{S}_i$ no provee una estimación confiable de $\mathbf{\Sigma}_i$ si p es una fracción moderada de n_i.