Lecturas de Métodos Estadísticos Multivariantes

M.Sc. Fidel Ordoñez

Mayo-Agosto 2014

Rotaciones

Sea $x \in \mathbb{R}^p$, $\Gamma_{p\times p}$ ortogonal. Γx lo que hace es tener las nuevas coordenadas en un sistema rotado. $A_{p\times p}$ simétrica, entonces

$$\begin{aligned} A &= \Gamma\Lambda\Gamma^T \\ A^\alpha &= \Gamma\Lambda^\alpha\Gamma^T \\ A^{-1} &= \Gamma\Lambda^{-1}\Gamma^T \\ \mathrm{tra}(A) &= \sum_{i=1}^p \lambda_i \\ |A| &= \prod_{i=1}^p \lambda_i \\ \Gamma &= \mathrm{diag}(\lambda_1,\ldots,\lambda_p) \end{aligned}$$

Sea $X = (X_1,\ldots,X_p)^T$ con $X \sim (\mu,\Sigma)$ entonces

$$\begin{aligned} \mathrm{E}(X) &= (\mathrm{E}(X_1),\ldots,\mathrm{E}(X_p))^T \\ &= \mu \\ \mathrm{Var}(X) &= \mathrm{Cov}(X,X) \\ &= \Sigma \\ \Sigma &= \begin{pmatrix} \sigma_{X_1X_1} & \ldots & \sigma_{X_1X_p} \\ \vdots & \ddots & \vdots \\ \sigma_{X_pX_1} & \ldots & \sigma_{X_pX_p} \end{pmatrix}_{p\times p} \\ \rho &= \begin{pmatrix} \rho_{X_1X_1} & \ldots & \rho_{X_1X_p} \\ \vdots & \ddots & \vdots \\ \rho_{X_pX_1} & \ldots & \rho_{X_pX_p} \end{pmatrix}_{p\times p} \\ \rho_{X_iX_j} &= \frac{\mathrm{Cov}(X_iX_j)}{\sqrt{\sigma_{X_iX_i}\sigma_{X_jX_j}}} \\ \sigma_{ij} &= \sigma_{X_iX_j} \\ \sigma_i^2 &= \sigma_{X_iX_i} \\ \mathrm{Cov}(X_iX_i) &= \sigma^2_{X_i} \end{aligned}$$

Suponga que se tienen n realizaciones de X_{p+1} y se tiene la matriz de datos

$$\mathcal{X}_{n\times p} = \begin{pmatrix} x_{11} & \ldots & x_{1j} & \ldots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \ldots & x_{ij} & \ldots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \ldots & x_{nj} & \ldots & x_{np} \end{pmatrix}$$

$x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})^T \in \mathbb{R}^n$ $i = 1, 2, \ldots, n > p$ es la i-esima observacion
$x_{(j)} = (x_{1j}, \ldots, x_{ij}, \ldots, x_{nj})^T \in \mathbb{R}^n$ $j = 1, 2, \ldots, p$ es la observacion de X_j

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_j \\ \vdots \\ \bar{x}_p \end{pmatrix} \\
&= \frac{1}{n}\mathcal{X}^T \mathbf{1}_n \qquad \mathcal{X}_{n\times p} \\
\bar{x}_j &= \frac{1}{n}\sum_{i=1}^{n} x_{ij}
\end{aligned}$$

$$\begin{aligned}
\mathbf{1}_n &= \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1} \\
\mathcal{S} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \qquad \text{estimador sesgado} \\
&= \frac{1}{n}\mathcal{X}^t\mathcal{X} - \bar{x}\bar{x}^T \\
&= \frac{1}{n}\left(\mathcal{X}^T\mathcal{X} - \frac{1}{n}\mathcal{X}^T\mathbf{1}_n\mathbf{1}_n^T\mathcal{X}\right) \\
&= \frac{1}{n}\mathcal{X}^T\left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\mathcal{X} \\
&= \frac{1}{n}\mathcal{X}^T\mathcal{H}\mathcal{X} \qquad \mathcal{H} \qquad \text{simétrica e idempotente} \\
\mathcal{S}_u &= \frac{n}{n-1}\mathcal{S} \qquad \text{estimador insesgado} \\
\mathcal{S} &= \frac{1}{n}\mathcal{X}^T\mathcal{X} - \bar{x}\bar{x}^T \\
&= \frac{1}{n}\mathcal{X}^T\mathcal{H}\mathcal{X} \\
\mathcal{H} &= I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T \\
R &= D^{-1/2}\mathcal{S}D^{-1/2} \qquad D = diag(S_{X_iX_j})
\end{aligned}$$

Transformaciones Lineales $X = (X_1, \ldots, X_p)$

1. $\mathcal{X}_{n\times p} \qquad \mathcal{A}_{g\times p}$

$$\begin{aligned}
y_{n\times q} &= \mathcal{X}\mathcal{A}^T \\
&= (y_1, \ldots, y_n)^T
\end{aligned}$$

   fila $y_i = (y_{i1}, \ldots, y_{iq}) \in \mathbb{R}^q$ i esima observación de $y_{q\times 1} = \mathcal{A}X \qquad \bar{y} = \mathcal{A}\bar{x} \qquad \mathcal{S}_y = \mathcal{A}\mathcal{S}_x\mathcal{S}^T$

2. Mahalanobis

Sea $z_i = \mathcal{S}^{-1/2}(x_i - \bar{x} \qquad i = 1, \ldots, n$, luego $Z = (z_1, \ldots, z_n) \qquad \bar{Z} = 0 \qquad \mathcal{S}_z = I_p$

$$S_{X_jX_k} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \qquad r_{X_jX_k} = \frac{S_{X_jX_k}}{\sqrt{S_{X_jX_j}S_{X_kX_k}}}$$

$$R = \begin{pmatrix} r_{X_1X_1} & \ldots & r_{X_pX_p} \\ \vdots & \ddots & \vdots \\ r_{X_pX_1} & \ldots & r_{X_pX_p} \end{pmatrix}_{p\times p} \qquad \mathcal{S} = \begin{pmatrix} S_{X_1X_1} & \ldots & S_{X_1X_p} \\ \vdots & \ddots & \vdots \\ S_{X_pX_1} & \ldots & S_{X_pX_p} \end{pmatrix}$$

$$D = \begin{pmatrix} S_{X_1X_1} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & S_{X_pX_p} \end{pmatrix} \qquad D^{-1} = \begin{pmatrix} \frac{1}{S_{X_1X_1}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{S_{X_pX_p}} \end{pmatrix}$$

$$D^{-1/2} = \begin{pmatrix} \frac{1}{\sqrt{S_{X_1X_1}}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{\sqrt{S_{X_pX_p}}} \end{pmatrix}$$

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \qquad \mathcal{S} \geq 0 \qquad \text{semidefinida positiva}$$

$$\begin{aligned} \mathcal{S} &= \frac{1}{n}\mathcal{X}^T\mathcal{H}\mathcal{X} \qquad \mathcal{S} \geq 0 \qquad \text{semidefinida positiva} \\ &= \frac{1}{n}\mathcal{X}^T\mathcal{H}\mathcal{H}\mathcal{X} \\ &= \frac{1}{n}\mathcal{X}^T\mathcal{H}^T\mathcal{H}\mathcal{X} \qquad y = \mathcal{H}\mathcal{X} \\ &= \frac{1}{n}y^Ty \geq 0 \\ \bar{y} &= \frac{1}{n} \end{aligned}$$

Distribución Normal Multivariada

Sea $X \sim N_p(\mu, \Sigma)$ si $f(x) = |2\pi\Sigma|^{-1/2}e^{\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]}$

$$\mathrm{E}(X) = \mu \qquad \mathrm{Var}(X) = \Sigma$$

Sea $X \sim N_p(\mu, \Sigma)$ si $y = \Sigma^{-1/2}(x - \mu)$, luego $Y \sim N_p(0, I)$ donde

$$\begin{aligned} I &= \begin{pmatrix} 1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1 \end{pmatrix} \\ \mathrm{E}(Y) &= \mathrm{E}\left(\Sigma^{-1}(X - \mu)\right) \\ &= \Sigma^{-1/2}[\mathrm{E} - \mu] \\ &= \Sigma^{-1/2}(\mu - \mu) \\ &= 0 \\ \mathrm{Var}(Y) &= \left(\Sigma^{-1/2}\right)^T \mathrm{Var}(X)\Sigma^{-1/2} \\ &= \Sigma^{-1/2}\Sigma\Sigma^{-1/2} \\ &= I \\ X &= \Sigma^{-1/2}Y + \mu \qquad I = \Sigma^{-1/2} \\ X - \mu &= \Sigma^{-1/2}Y \\ (x - \mu)^T\Sigma^{-1}(x - \mu) &= (\Sigma^{-1/2}y)^T\Sigma^{-1}(\Sigma^{-1/2}y) \\ &= y^Ty \end{aligned}$$

$f(y) = (2\pi)^{-p/2} e^{-\frac{1}{2} y^T y}$, por lo tanto $Y \sim N_p(0, I)$ $\diamondsuit$
Si $A_{p\times p}$ $\quad C \subset \mathbb{R}^p$ $\quad Y = AX + C$
$X \sim N_p(\mu, \Sigma)$ luego $Y \sim N_p(A\mu + C, A^T \Sigma A)$

Teorema 1

Si $X \sim N_p(\mu, \Sigma) \Longrightarrow u = (X-\mu)^T \Sigma^{-1} (X-\mu) \sim \mathcal{X}_p^2$

Distribución de Wishart

Sea $\mathcal{X}_{n\times p}$ matriz de datos de $X \sim N_p(0, \Sigma)$, luego $\mathcal{M} = \mathcal{X}^T \mathcal{X} \sim W_p(\Sigma, n)$

Nota 2

Sea $\mathcal{X}_{n\times p}$ de $X \sim N_p(0, \Sigma)$, $\mathcal{S}$ matriz de covarianza muestral, entonces:

i. $nS = \mathcal{X}^T \mathcal{H} \mathcal{X} \sim W_p(\Sigma, n-1)$

ii. $\bar{x}$, $\mathcal{S}$ son independientes

Distribución T^2 de Hotelling

Sea $Y \sim N_p(0, I)$ independientes de $\mathcal{M} \sim W_p(I, n)$, luego $ny^T \mathcal{M}^{-1} y \sim T^2(p, n)$

Teorema 3

Sea $X \sim N_p(\mu, \Sigma)$ independientes de $\mathcal{M} \sim W_p(\Sigma, n)$ entonces

$$n(X-\mu)^T \mathcal{M}^{-1} (X-\mu) \sim T^2(p, n)$$

Corolario 4

Sea $X \sim N_p(\mu, \Sigma)$ entonces $(n-1)(\bar{x}-\mu)^T \mathcal{S}^{-1} (\bar{x}-\mu) = n(\bar{x}-\mu)^T \mathcal{S}_u^{-1} (\bar{x}-\mu) \sim T^2(p, n-1)$
donde $S_u = \dfrac{n}{n-1} \mathcal{S}$

Corolario 5

Sea $T^2(p.n) = \dfrac{np}{n-p+1} F_{p,n-p+1}$

Análisis Factorial

Sea $X = (X_1, X_2, \ldots, X_p)^T \sim (\mu, \Sigma)$ se tienen n-observaciones formando χ.
El análisis factorial asume que hay un modelo que explica la covarianza de $X_1, X_2, \ldots, X_p$ mediante $k < p$ factores latentes.
Sea $X = QF + \mu$ con $X_{p\times 1}$, $Q_{p\times k}$, $\mu_{p\times 1}$ y $F_{k\times 1} = (F_1, F_2, \ldots, F_k)^T$

$$E(F) = 0 \qquad \mathrm{Var}(F) = I_k$$

En la práctica $X = QF + U + \mu$

Modelo Factorial Ortogonal

Sea $X = QF + U + \mu$ con $Q_{p\times k}$, $F_{k\times 1}$, $U_{p\times 1}$ y $\mu_{p\times 1}$
Q es la matríz de cargas de los factores comunes F (no aleatorio)
U matríz (aleatoria) de factores especificos
Se asume que con $i \neq j$

$$\mathrm{E}(F) = 0 \qquad \mathrm{Var}(F) = I_k \qquad \mathrm{E}(U) = 0 \qquad \mathrm{Cov}(U_i, U_j) = 0 \qquad \mathrm{Cov}(F, U) = 0$$

μ_j media de X_j con $j = 1, \ldots, p$
U_j j-esimo factor especifico
F_l l-esimo factor común $l = 1, \ldots, p$
q_{jl} carga factorial de X_j en F_l
Si $\mathrm{Var}(U) = \Psi$ donde $\Psi = diag(\psi_{11}, \ldots, \psi_{1p})$

$$\begin{aligned} X_j &= \sum_{l=1}^{k} q_{jl} F_l + U_j + \mu_j \\ \sigma_{X_j X_j} &= \mathrm{Var}(X_j) \\ &= \sum_{l=1}^{k} q_{jl}^2 + \Psi_{jj} \\ &= h_j^2 + \Psi_{jj} \end{aligned}$$

donde a h_j^2 se le llama comunalidad, y Ψ_{jj} es la varianza especifica.

Nota 6

$\mathrm{Var}(X) = Q^T \mathrm{Var}(F) Q + \mathrm{Var}(U)$
$\Sigma = Q^T Q + \Psi$ donde Σ tiene p variables y Q tiene k factores
Para interpretar factores $\Sigma_{XF} = Q$ y $\rho_{XF} = D^{-1/2}$ donde $D = \mathrm{diag}(\sigma_{x_1 x_1, \ldots, x_p x_p})$

$$\begin{aligned} \Sigma_{XF} &= E[(X - \mu)(F - 0)^T] \\ &= E[(QF + U)F^T] \\ &= QE(FF^T) + E(UF^T) \\ &= QI_k + 0 \\ &= Q \end{aligned}$$

Invarianza de Escala $X \sim (\mu, \Sigma)$

Si $Y = CX$ donde $C = \mathrm{diag}(c_1, \ldots, c_p)$ con $\Sigma = Q_X Q_X^T + \Psi_X$, luego

$$\begin{aligned} \mathrm{Var}(Y) &= C\Sigma C^T \\ &= CQ_X Q_X^T + C\Psi_X C^T \\ &= (CQ_X)(CQ_X)^T + C\Psi_X C^T \end{aligned}$$

En particular $Y = D^{-1/2}(X - \mu)$, en este caso queremos encontrar Q_X, Ψ_Y tal que

$$\begin{aligned} \rho &= Q_Y Q_Y^T + \Psi_Y \\ \rho_{XY} &= \rho_{YF} \\ &= \rho_Y \end{aligned}$$

por invarianza $Q_X = D^{-1/2} Q_Y$ y $\Psi_X = D^{-1/2} \Psi D^{-1/2}$.

La No Unicidad de las Cargas Factoriales

Si $X = QF + U + \mu$ es cierto, luego si G es ortogonal y $X = (QG)(G^T F) + U + \mu$ es cierto

$$X = Q^* F^* + U + \mu$$

Nota 7

Sea $\Sigma = QQ^T + \Psi$ donde Q tiene pk parámetros y Ψ tiene p parámetros, además Σ tiene $\dfrac{p(p+1)}{2}$ ecuaciones en el sistema.

$$\Sigma_{p\times p} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

se utilizan las siguientes restricciones:

1. $Q^T D^{-1} Q$ es diagonal
2. $Q^T \Psi^{-1} Q$ es diagonal

d:grados de libertad del sistema, con cualquiera de las restricciones:

$$\begin{aligned} d &= \frac{p(p+1)}{2} - \left[(pk+p) - \frac{k(k-1)}{2}\right] \\ &= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) \end{aligned}$$

Si $d < 0$ indeterminado (más ecuaciones que incognitas)
Si $d = 0$ solución única (excepto por rotación)
Si $d > 0$ podemos encontrar soluciones (común en la práctica)

Ejemplo 8

Si $p = 6$
$k = 1 \Longrightarrow d = 9 > 0$
$k = 2 \Longrightarrow d = 4 > 0$
$k = 3 \Longrightarrow d = 0$
$k = 4 \Longrightarrow d = -3 < 0$

$$\begin{aligned} \Sigma &= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \\ &= \begin{pmatrix} q_{11} \\ q_{21} \\ q_{31} \end{pmatrix} \begin{pmatrix} q_{11} & q_{21} & q_{31} \end{pmatrix} \begin{pmatrix} \Psi_{11} & 0 & 0 \\ 0 & \Psi_{22} & 0 \\ 0 & 0 & \Psi_{33} \end{pmatrix} \\ &= \begin{pmatrix} q_{11}^2 + \Psi_{11} & q_{11}q_{21} & q_{11}q_{31} \\ q_{21}q_{11} & q_{21}^2 + \Psi_{22} & q_{21}q_{31} \\ q_{31}q_{11} & q_{31}q_{21} & q_{31}^2 + \Psi_{33} \end{pmatrix} \end{aligned}$$

$$\begin{array}{ccc} q_{11}^2 = \dfrac{\sigma_{12}\sigma_{13}}{\sigma_{23}} & q_{21}^2 = \dfrac{\sigma_{12}\sigma_{23}}{\sigma_{13}} & q_{31}^2 = \dfrac{\sigma_{13}\sigma_{23}}{\sigma_{12}} \\ \Psi_{11} = \sigma_{11} - q_{11}^2 & \Psi_{22} = \sigma_{22} - q_{21}^2 & \Psi_{11} = \sigma_{33} - q_{31}^2 \end{array}$$

Estimación del Modelo

Con los datos $\mathcal{X}$ encontrar $\hat{Q}$ y $\hat{\Psi}$ tal que $S = \hat{Q}\hat{Q}^T + \hat{\Psi}$
Más facil cuando usamos $Y = \mathcal{H}\mathcal{X}D^{1/2}$
$S_Y = R$ matriz de correlaciones de $\mathcal{X}$
Queremos $R = \hat{Q}_Y\hat{Q}_Y^T + \hat{\Psi}_Y$

Método de Componentes Principales

Se comienza con aproximación de Q, sea $\hat{Q}$
$S = \mathcal{G}\mathcal{L}\mathcal{G}^T$ S es simétrica, donde $\mathcal{L} = \text{diag}(l_1, \ldots, l_p)$ y $l_1 \geq \ldots \geq l_p$ autovalores de S con autovectores $g_1, \ldots, g_p$ que forman $\mathcal{G}$.
Considerando solamente los primeros k autovalores más grandes y que sean positivos, se aproxima

$$S = \mathcal{G}_1\mathcal{L}_1\mathcal{G}_1^T$$

$\mathcal{L}_1 = \text{diag}(l_1, \ldots, l_k)$ y $\mathcal{G}_1$ tiene los respectivos autovectores, luego

$$\begin{aligned}
\hat{Q} &= \mathcal{G}_1\mathcal{L}_1^{1/2} \\
&= (\sqrt{l_1}g_1, \ldots, \sqrt{l_k}g_k) \\
\hat{\Psi} &= \text{diag}(S - \hat{Q}\hat{Q}^T) \\
\hat{\Psi}_{jj} &= s_{jj} - \sum_{l=1}^{k} \hat{q}_{jl}^2
\end{aligned}$$

Para evaluar la estimación ver la matriz residual $S = \hat{Q}\hat{Q}^T + \hat{\Psi}$. Observemos que es diagonal

$$\begin{aligned}
\hat{Q}\hat{Q}^T &= (\mathcal{G}_1\mathcal{L}_1^{1/2})(\mathcal{G}_1\mathcal{L}_1^{1/2})^T \\
&= \mathcal{G}_1\mathcal{L}_1^{1/2}\mathcal{L}_1^{1/2}\mathcal{G}_1^T \\
&= \mathcal{G}_1\mathcal{L}_1\mathcal{G}_1^T
\end{aligned}$$

Método del Factor Principal

Se puede utilizar S(observado) o R(estimado).

1. Como estimar Ψ en el método del factor principal:
   i. h_j^2=cuadrado de coeficiente de correlación múltiple en la regresión de X_j sobre el resto de X'
   ii. Con $\hat{\Psi}_{jj} = 1 - h_j^2$ se tiene que $h_j^2 = \max_{l \neq j}\{r_{jl}\}$
2. De $R = \hat{Q}\hat{Q}^T + \hat{\Psi}$ luego $R - \hat{\Psi} = \hat{Q}\hat{Q}^T$
3. $R - \hat{\Psi}$ es simétrica, $R - \hat{\Psi} = \mathcal{G}\mathcal{L}\mathcal{G}^T$ descomposición espectral.
4. De $\mathcal{L}$ tomar los k autovalores mayores, digamos $l_1 \geq \ldots \geq l_k > 0$ y formamos $\mathcal{L}_1$ y $\mathcal{G}_1$ con los respectivos autovectores.
   Luego $\hat{Q} = \mathcal{G}_1\mathcal{L}_1^{1/2}$ i.e. $\hat{q}_l = \sqrt{l_l}g_l$ con $l = 1, \ldots, k$
5. Construir $\hat{\Psi}$(nuevo) $\hat{\Psi}_{jj} = 1 - \sum_{i=1}^{k} q_{jl}^2 \longrightarrow \hat{\Psi}$
6. Se itera, comenzando en el paso 3 hasta que $\|\hat{Q}_{n+1} - \hat{Q}_n\| < \epsilon$ o $\hat{\Psi}_{jj}$ son estables

Método de Máxima Verosimilitud

Sea $\mathcal{X}_{n\times p}$ de $X \sim N_p(\mu, \Sigma)$. Recordar que

$$\begin{aligned}
(\mathcal{X}; \mu, \Sigma) &= -\frac{n}{2}\ln|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i - \Sigma)\Sigma^{-1}(x_i - \mu)^T \\
&= -\frac{n}{2}\ln|2\pi\Sigma| - \frac{n}{2}\,\text{tra}\,|\Sigma^{-1}S| - \frac{n}{2}(\bar{\mathcal{X}} - \mu)\Sigma^{-1}(\bar{\mathcal{X}} - \mu)^T
\end{aligned}$$

EMV de μ en $\bar{\mathcal{X}}$: $l(\mathcal{X};\mu,\Sigma) = -\frac{n}{2}\ln|2\pi\Sigma| - \frac{n}{2}\operatorname{tra}|\Sigma^{-1}S|$ sustituyendo $\Sigma = QQ^T + \Psi$ tenemos que

$$l(\mathcal{X};\hat{\mu},Q,\Psi) = -\frac{n}{2}\{\ln|2\pi(QQ^T+\Psi)| + \operatorname{tra}[(QQ^T+\Psi)^{-1}]S\} \tag{1}$$

Maximizando al derivar con respecto a Q y Ψ, además con el supuesto de que $Q^T\Psi^{-1}Q = D$ es diagonal, se obtienen las siguientes ecuaciones:

$$\begin{cases} \hat{\Psi} = \operatorname{diag}(S - \hat{Q}\hat{Q}^T) \\ (\hat{\Psi}^{-1/2}(S-I)\hat{\Psi}^{-1/2})(\hat{\Psi}^{-1/2}\hat{Q}) = (\hat{\Psi}^{-1/2}\hat{Q})D \\ \hat{Q}^T\Psi^{-1}\hat{Q} = D \end{cases}$$

Algortimo 9

1. Partir de $\hat{Q}$ (puede usar factor principal), luego $\hat{\Psi} = diag(S - QQ^T)$
2. A (simétrica) donde:
$$\begin{aligned} A &= \hat{\Psi}^{-1/2}(S-\hat{\Psi})\hat{\Psi}^{-1/2} \\ &= \hat{\Psi}^{-1/2}S\hat{\Psi})\hat{\Psi}^{-1/2} - I \end{aligned}$$
3. Encontrar la descomposición espectral de A, $A = \mathcal{G}\mathcal{L}\mathcal{G}^T$
donde $\mathcal{L} = diag(l_1,\ldots,l_p)$ donde $l_1 \geq l_2 \geq \ldots \geq l_p$ con autovectores $g_1,\ldots,g_p$ de $\mathcal{G}$. Tomar los k autovalores más grandes y positivos, i.e. $l_1 \geq l_2 \geq \ldots \geq l_k > 0$ y se tiene $\mathcal{L}_1 = diag(l_1,\ldots,l_k)$ y sus respectivos autovectores en $\mathcal{G}_1$
4. Tomar $\hat{Q} = \hat{\Psi}^{\frac{1}{2}}\mathcal{G}_1\mathcal{L}^{-\frac{1}{2}}$ y sustituir $\hat{Q}$ en (1), maximizar para Ψ, iterar a partir de 2 hasta la convergencia.

Prueba de Razón de Verosimilitud para el Número de Factores Comunes

$H_0 : \Sigma = QQ^T + \Psi$
H_1:no modelo factorial
Sean $\hat{\Psi}$ y $\hat{Q}$ estimador de máxima verosimilitud(EMV) con $S \doteq \hat{Q}\hat{Q}^T + \hat{\Psi}$, luego

$$\begin{aligned} -2\ln\left(\frac{MVH_0}{MVSR}\right) &= n\ln\left(\frac{|\hat{Q}\hat{Q}^T+\hat{\Psi}|}{|S|}\right) \\ &\sim \chi^2_{\frac{1}{2}(p-k)^2+\frac{1}{2}(p+k)} \end{aligned}$$

donde $MVSR$ es la máxima verosimilitud sin restricción.
La corrección de *Bartletts* reemplaza n por $\dfrac{(n-1)-(2p+4k+5)}{6}$, además rechaza H_0 si

$$\left[n-1-\left(\frac{2p+4k+5}{6}\right)\right] n\ln\left(\frac{|\hat{Q}\hat{Q}^T+\hat{\Psi}|}{|S|}\right) > \chi^2_{1-\alpha;\frac{1}{2}(p-k)^2-\frac{1}{2}(p+k)}$$

Método Varimax

Estandarizar las cargas factoriales $\tilde{q}$: $\tilde{q}_{jl} = \dfrac{\hat{q}^v_{jl}}{\hat{h}^v_j}$, queremos que V sea máxima

$$V = \sum_{j=1}^{k}\left\{\frac{1}{p}\sum_{j=1}(q^v_{jl})^4 - \left(\frac{1}{4}\sum_{j=1}^{p}\hat{q}^v_{jl}\right)^2\right\}$$

Estimación de los Factores(Puntajes Factoriales)

1. Sea $X-\mu = QF + U$ donde $U \sim N(0,\Psi)$ y $X-\mu \sim N(QF,\Psi)$

$$\hat{F} = (Q^T\Psi^{-1}Q)^{-1}Q^T\Psi^{-1}(X-\mu)$$
$$\hat{f}_i = (\hat{Q}^T\Psi^{-1}\hat{Q})^{-1}\hat{Q}^T\hat{\Psi}^{-1}(x_i-\mu)$$

2. Sea $X-\mu = QF + U$ con F variable aletoria

$$\mathrm{E}(F|X=x) = Q^T\Sigma^{-1}(X-\mu)$$
$$\hat{f}_i = Q^T S^{-1}(x_i-\mu)$$

Análisis de Conglomerados

El objetivo es formar grupos que sean entre ellos (heterogéneos) y dentro de ellos (homogéneos).
Pasos para realizar el análisis de conglomerados.

1. Seleccionar una medida de proximidad (similaridad), así se conoce que tan cercanos son dos unidades si sus valores están cerca.
   $i \to x_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$
   $j \to x_j^T = (x_{j1}, x_{j2}, \ldots, x_{jp})$
2. Seleccionar algoritmo de agrupación. Tal que las unidades dentro de los conglomerados sean lo más homogéneas posibles, y entre los grupos lo más heterogéneos posibles (basados en la medida de la proximidad seleccionada).

Proximidad Entre Objetos

$$D = \begin{pmatrix} d_{11} & d_{12} & \ldots & d_{1n} \\ d_{21} & d_{22} & \ldots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \ldots & d_{nn} \end{pmatrix}$$

Similaridad Entre Objetos en Estructura Binaria

Sea (x_i, x_j) donde $x_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$ y $x_j^T = (x_{j1}, x_{j2}, \ldots, x_{jp})$ con $x_{ik}, x_{jk} \in (0,1)$

$$d_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)}$$

donde

$$a_1 = \sum_{k=1}^{p} I_{x_{ik}=x_{jk}=1} \qquad a_2 = \sum_{k=1}^{p} I_{x_{ik}=0, x_{jk}=1} \qquad a_3 = \sum_{k=1}^{p} I_{x_{ik}=1, x_{jk}=0} \qquad a_4 = \sum_{k=1}^{p} I_{x_{ik}=x_{jk}=0}$$

La naturaleza de las variables determinar la medida de similaridad.

Ejemplo 10

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
i	1	0	1	1	0	0	0	0	0	0
j	0	1	1	1	1	0	1	0	0	0

$a_1 = 2$, $a_2 = 3$, $a_3 = 1$, $a_4 = 4$
usando Jaccard $d_{ij} = \frac{1}{3}$

Nombre	δ	λ	Definición
Jaccard	0	1	$\frac{a_1}{a_1+a_2+a_3}$
Tanimoto	1	2	
Pareo Simple	1	1	

Medidas de Distancia para Variables Continuas

Norma L_r con $r \geq 1$

$$\begin{aligned} d_{ij} &= \|x_i - x_j\| \\ &= \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r\right)^{1/r} \\ &= \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \end{aligned}$$

Al utilizar la norma L_r es conveniente que las mediciones esten en la misma escala, si no entonces estandarizamos

$$\begin{aligned} d_{ij}^2 &= (x_i - x - j)^T A(x_i - x_j) \\ &= \|x_i - x_j\|_A \end{aligned}$$

En particular si $A = \operatorname{diag}\left(\frac{1}{S_{x_1x_1}}, \dots, \frac{1}{S_{x_px_p}}\right)$ entonces $d_i^2 j = \sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{S_{x_kx_k}}$ que no depende de la escala de medida.

Ejemplo 11

$r = 2$, norma L_2

Métrica χ^2 para Comparar Filas o Columnas de una Tabla de Contingencia

	1	...	j	...	p	
1	x_{11}	...	x_{1j}	...	x_{1p}	$x_{1.}$
⋮	⋮		⋮		⋮	⋮
i	x_{i1}	...	x_{ij}	...	x_{ip}	$x_{i.}$
⋮	⋮		⋮		⋮	⋮
n	x_{n1}	...	x_{nj}	...	x_{np}	$x_{n.}$
	$x_{.1}$	...	$x_{.j}$	...	$x_{.p}$	$x_{..}$

Distribución marginal de fila i, $\frac{x_{i.}}{x_{..}}$ donde $x_{i.} = \sum_{j=1}^{p} x_{ij} \qquad x_{..} = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}$

Para columna j: $\frac{x_{.j}}{x_{..}}$ con $x_{.j} = \sum_{i=1}^{n} x_{ij}$

Distribución condicional de fila i: $\frac{x_{.j}}{x_{..}} \longrightarrow \left(\frac{x_{i1}}{x_{i.}}, \dots, \frac{x_{ij}}{x_{i.}}, \dots, \frac{x_{ip}}{x_{i.}}\right)$

Para columna j: $\frac{x_{ij}}{x_{.j}} \longrightarrow \left(\frac{x_{1j}}{x_{.j}}, \dots, \frac{x_{ij}}{x_{.j}}, \dots, \frac{x_{nj}}{x_{.j}}\right)$

Distancia entre la fila i_1 y la fila i_2: $d^2(i_1, i_2) = \sum_{j=1}^{p} \frac{1}{\frac{x_{.j}}{x_{..}}} \left(\frac{x_{i_1j}}{x_{i_1.}} - \frac{x_{i_2j}}{x_{i_2.}}\right)^2$

Distancia entre la columna j_1 y la columna j_2: $d^2(j_1, j_2) = \sum_{i=1}^{p} \frac{1}{\frac{x_{i.}}{x_{..}}} \left(\frac{x_{ij_1}}{x_{.j_1}} - \frac{x_{ij_2}}{x_{.j_2}}\right)^2$

Coeficiente de Correlación Q como Medida de Similaridad

Sean $x_i^T = (x_{i1}, \dots, x_{ip})$ y $x_j^T = (x_{j1}, \dots, x_{jp})$

$$d_{ij} = \frac{\sum_{k=1}^{p} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)^2}{\sqrt{[\sum_{k=1}^{p} (x_{ik} - \bar{x}_i)^2] - [\sum_{k=1}^{p} (x_{jk} - \bar{x}_j)^2]}}$$

Clasificación Automática

1. Seleccionar medida de proximidad o distancia
2. Seleccionar algoritmo de conglomeración

Algoritmos de Conglomeración

De los algoritmos más usuales están los

1. Jerárquicos
   a. Jerárquicos conglomerativos (asociativos)
   b. Jerárquicos divisivos (disasociativos)
2. De partición

Jerárquicos Conglomerativos

Parten con n conglomerados (cada observación es un conglomerado).
Se unen los dos más cercanos para formar $(n-1)$ conglomerados, se une hasta formar un sólo conglomerado conformado por $\mathcal{X}$

Jerárquicos Divisivos

Parte de un sólo conglomerado que es $\mathcal{X}$, se va dividiendo hasta tener n conglomerados (conformados por cada observación)

De Partición

Parte de un número preestablecido de conglomerados y se van intercambiando las observaciones hasta optimizar algún puntaje.

Algoritmo Aglomerativo

1. Construir n grupos cada con uno con una observación
2. Calcular la matriz de distancia D
3. Encontrar los conglomerados con la distancia más cercana
4. Unir en un sólo conglomerado los encontrados en 3
5. Calcular D restringida entre los grupos nuevos

Repetir $3, 4, 5$ hasta tener un sólo conglomerado formado por $\mathcal{X}$

Distancia Utilizada Entre Dos Grupos

Sea $P+Q$ que resulta de unir P y Q. R otro grupo

$$d(R, P+Q = \delta_1 d(R,P) + \delta_2 d(R,Q) + \delta_3 d(P,Q) + \delta_4 |d(R,P) - d(R,Q)|$$

Nombre	δ_1	δ_2	δ_3	δ_4
Encadenamiento simple	1/2	1/2	0	-1/2
Encadenamiento completo	1/2	1/2	0	1/2
Encadenamiento promedio	1/2	1/2	0	0
Encadenamiento promedio ponderado	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	0	0
Centroide	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	$-\frac{n_P n_Q}{(n_P+n_Q)^2}$	0
Mediana	1/2	1/2	1/4	0
Ward	$\frac{n_R+n_P}{n_R+n_P+n_Q}$	$\frac{n_R+n_Q}{n_R+n_P+n_Q}$	$-\frac{n_R}{n_R+n_P+n_Q}$	0

Ward

$$n_P = \sum_{i=1}^{n} I(x_i \in P)$$

Encadanamiento simple modificado $d(R, P+Q) = \min\{d(P,R), d(Q,R)\}$
Encadanamiento completo modificado $d(R, P+Q) = \max\{d(P,R), d(Q,R)\}$

Ejemplo 12

Sea

$$\begin{aligned} D &= \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & & & \\ 3 & 0 & & \\ 7 & 9 & 0 & \\ 8 & 5 & 6 & 0 \end{pmatrix} \end{aligned}$$

Análisis de Discriminante

- Descriptivo
- Predictivo (el objetivo es clasificar observaciones en grupos ya conocidos)

Reglas de Clasificación para Distribuciones Conocidas

Suponga que tenemos las poblaciones Π_j; $j = 1, \ldots, J$ y se tiene que clasificar una observación con $x^T = (x_1, \ldots, x_p)$ a una de estas poblaciones. Regla discriminante es una separación del espacio muestral $\mathbf{R}^p$ en conjuntos R_j tal que si $x \in R_j$ identificamos la observación como de la población Π_j

Regla Discriminante de Máxima Verosimilitud (RDML)

Sea $f_i(x)$ la densidad de la población Π_i. La $RDML$ clasificara a x en Π_j si $f_j(x)$ es el máximo de la verosimilitud, i.e.

$$\begin{aligned} L_j &= f_j(x) \\ &= \max_i f_i(x) \end{aligned}$$

En caso de que hayan varias se clasifican en cualquiera $R_j = \{x : L_j(x) > L_i(x); i = 1, \ldots, J; i \neq j\}$

Regla que Minimiza el Costo Esperado de la Mala Clasificacion (ECM)

Suponga $J = 2$

$$\begin{aligned} p_{21} &= \Pr(x \in R_2|\Pi_1) \\ &= \int_{R_2} f_1(x)dx \\ p_{12} &= \Pr(x \in R_1|\Pi_2) \\ &= \int_{R_1} f_2(x)dx \end{aligned}$$

Las observaciones mal clasificadas crean un costo $C(i|j)$: costo de asignarlos a R_i dado que es de Π_j, tenemos

	Π_1	Π_2
Π_1	0	$C(2\|1)$
Π_2	$C(1\|2)$	0

Suponga Π_j la probabilidad a priori de que pertenece a Π_j

$$ECM = C(2|1)p_{21}\Pi_1 + C(1|2)p_{12}\Pi_2$$

La regla que minimiza el ECM viene dado por

$$\begin{aligned} R_1 &= \left\{ x : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{C(1|2)}{C(2|1)} \right) \left(\frac{\Pi_2}{\Pi_1} \right) \right\} \\ R_2 &= \left\{ x : \frac{f_1(x)}{f_2(x)} < \left(\frac{C(1|2)}{C(2|1)} \right) \left(\frac{\Pi_2}{\Pi_1} \right) \right\} \end{aligned}$$

Ejemplo 13

Sea $\Pi_1 = N(\mu_1, \sigma_1^2)$ y $\Pi_2 = N(\mu_2, \sigma_2^2)$

$$f_i(x) = (2\pi\sigma_i^2)^{-1/2} e^{-1/2\left(\frac{x-\mu_i}{\sigma_i}\right)^2}$$

$x \in R_1$, luego

$$\begin{aligned} f_1(x) &\geq f_2(x) \\ \frac{f_1(x)}{f_2(x)} &\geq 1 \end{aligned}$$

$$\begin{aligned} \frac{\sigma_2}{\sigma_1} e^{-1/2\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]} &\geq 1 \\ -\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right] &\geq \ln\left(\frac{\sigma_2}{\sigma_1}\right) \end{aligned}$$

$\therefore x$ se clasifica en Π_1 si

$$x^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \left(\frac{\mu_1^1}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) \leq \ln\left(\frac{\sigma_2}{\sigma_1}\right)$$

Si $\mu_1 < \mu_2$ y $\sigma_1 = \sigma_2 = \sigma$. x se clasifica en Π_1 si

$$\begin{aligned}
\frac{3}{4}x^2 + \frac{1}{2}x - \frac{1}{4} &\leq 2\ln 2 = \ln 4 \\
\frac{3}{4}x^2 + \frac{1}{2}x - \left(\ln 4 + \frac{1}{4}\right) &\leq 0 \\
3x^2 + 2x - (\ln 256 + 1) &\leq 0 \\
s &= 4 + 4(3)(\ln 256 + 1) \\
&= 4 + 12(\ln 256 + 1) \\
x &= \frac{-2 \pm \sqrt{4 + 12(\ln 256 + 1)}}{6} \\
&= \begin{cases} -1.85 \\ 1.18 \end{cases}
\end{aligned}$$

Teorema 14

a. La RDML clasifica x a Π_j con $j = 1, \ldots, J$ cuando se minimiza la distancia al cuadrado de Mahalanoubis entre x y μ_j con $i \neq j$ si $\delta^2(x, \mu_j) \leq \delta^2(x, \mu_i)$

$$\delta^2(x, \mu_j) = (x - \mu)^T \Sigma^{-1} (x - \mu_j)$$

b. En el caso de $J = 2$, $x \in R_1 \iff \alpha^T(x - \mu) \geq 0$ donde $\alpha^T = (\mu_1 - \mu_2)^T \Sigma^{-1}$ y $\mu = \frac{1}{2}(\mu_1 + \mu_2)$

Demostración

b. $x \in R_1$ si $(x - \mu_1)^T \Sigma^{-1} (x - \mu_1) - (x - \mu_2)^T \Sigma^{-1} (x - \mu_2) \leq 0$

$$\begin{aligned}
x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_2 + \mu_2^T\Sigma^{-1}x - \mu_2^T\Sigma^{-1}\mu_2 &\leq 0 \\
-2\mu_1^T\Sigma^{-1}x + 2\mu_2^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2 &\leq 0 \\
-2(\mu_1 - \mu_2)^T\Sigma^{-1}x + (\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 + \mu_2) &\leq 0 \\
(\mu_1 - \mu_2)^T\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 + \mu_2) &\geq 0 \\
(\mu_1 - \mu_2)^T\Sigma^{-1}\left(\frac{1}{2}(\mu_1 - \mu_2)\right) &\geq 0
\end{aligned}$$

$\therefore \alpha^T(x - \mu) \geq 0 \qquad \diamondsuit$

Regla Discriminante de Bayes

Sea Π_i la probabilidad a priori de que x pertenece a Π_i, $i = 1, \ldots, J$.
Clasificamos x como de Π_j si

$$\Pi_j f_j(x) = \max_i \{\Pi_i f_i(x)\}$$

Nota 15

$\Pi_i = 1/J$ luego la regla discriminante de Bayes es la RDML

Probabilidades de Mala Clasificación RML

Sea $J=2$, recordar $\alpha^T=(\mu_1-\mu_2)^T\Sigma^{-1}$, $\mu=frac12(\mu_1+\mu_2)$, $\Pi_1=N(\mu_1,\Sigma)$, $\Pi_2=N(\mu_2,\Sigma)$

$$\begin{aligned}
p_{12} &= Pr(x\in R_1|\Pi_2)\\
&= \Pr\left(\alpha^T(x-\mu)>0|\Pi_2\right)\\
R_1 &: (\mu_1-\mu_2)^T\Sigma^{-1}\left(x-\frac{1}{2}(\mu_1+\mu_2)\right)>0\\
&: y>\frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2)\\
R_2 &: (\mu_1-\mu_2)^T\Sigma^{-1}\left(x-\frac{1}{2}(\mu_1+\mu_2)\right)\le 0\\
&: y\le\frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2)\\
y &= (\mu_1-\mu_2)^T\Sigma^{-1}x\\
&= \alpha^Tx
\end{aligned}$$

Como y es combinación lineal de x, entonces $y\sim N$

$$\begin{aligned}
\mu_{1y} &= \alpha^T\mu_1\\
&= (\mu_1-\mu_2)^T\Sigma^{-1}\mu_1\\
\mu_{2y} &= \alpha^T\mu_2\\
&= (\mu_1-\mu_2)^T\Sigma^{-1}\mu_2\\
\sigma_y^2 &= \alpha^T\Sigma\alpha\\
&= (\mu_1-\mu_2)^T\Sigma^{-1}\Sigma\Sigma^{-1}(\mu_1-\mu_2)\\
&= (\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)\\
&= (\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2)
\end{aligned}$$

Donde δ^2 es la distancia de Mahalanobis al cuadrado entre Π_1 y Π_2

$$\begin{aligned}
p_{12} &= \Pr\left(y>\frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2)|\Pi_2\right)\\
&= \left(z>\frac{\frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2)-(\mu_1-\mu_2)^T\Sigma^{-1}\mu_2}{\delta}\right)\\
&= \Pr\left(z>\frac{\frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)}{\delta}\right)\\
&= \Pr\left(z>\frac{\frac{1}{2}\delta^2}{\delta}\right)\\
&= \Pr\left(z\le-\frac{\delta}{2}\right)\\
&= \Phi\left(-\frac{\delta}{2}\right)
\end{aligned}$$

Clasificación con Matrices de Covarianza Distintas

Suponga $J=2$, $\Pi_1=N(\mu_1,\Sigma_1)$, $\Pi_2=N(\mu_2,\Sigma_2)$, las regiones de clasificación son definidas por funciones cuadráticas

Reglas Discriminantes en la Práctica

Suponga los datos que provienen de $\Pi_j=N(\mu_j,\Sigma)$ y tenemos J grupos con n_j observaciones cada uno.

$$\hat{\mu_j}=\bar{x_j}\qquad\hat{\Sigma}=S_j$$

Estimación de la Matriz de Covarianzas Común

$$S_u = \sum_{j=1}^{J} n_j \frac{S_j}{n-J} \qquad n = \sum_{j=1}^{J} n_j$$

La regla empírica ML clasifica x a Π_j si j minimiza $(x - \bar{x}_i)^T S_u^{-1}(x - \bar{x}_i)$

Estimación de la Probabilidad de Mala Clasificación

Sea $\hat{p}_{12} = \hat{p}_{21} = \Phi\left(-\frac{\hat{\delta}}{2}\right)$ entonces $\hat{\delta}^2 = (\bar{x}_1 - \bar{x}_2)^T S^{-1}(\bar{x}_1 - \bar{x}_2)$

Se puede utilizar el método de resustitución para tener una aproximación de la calidad de la regla discriminante, estimando p_{ij} con $\hat{p}_{ij} = \frac{n_{ij}}{n_j}$

n_j: numero de observaciones en Π_j

n_{ij}: número de observaciones de Π_j clasificado como de Π_i, la matriz $(\hat{p}_{ij})$ es llamada matriz de confusión.