Análisis Factorial Discriminante y Análisis Discriminante

Grado en Matemáticas — Análisis de Datos

Norberto Corral Blanco
Universidad de Oviedo

Beatriz Sinova Fernández
Universidad de Oviedo

Índice

1. Preliminares: matrices de datos y centrado

1.1. La matriz de datos

Consideramos una **matriz de datos X** de dimensión $n \times p$, donde n es el número de observaciones y p el número de variables:

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

Cada fila $\boldsymbol{x}_i^t = (x_{i1}, \ldots, x_{ip})$ representa la observación i-ésima; cada columna $\boldsymbol{x}_{(j)}$ recoge los n valores de la variable j.

1.2. Matriz de centrado

Definición 1.1. La **matriz de centrado** es

$$\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t = \mathbf{I}_n - \mathbf{J}, \qquad \mathbf{J} = \frac{1}{n}\mathbf{1}\mathbf{1}^t.$$

La acción de $\mathbf{H}$ sobre un vector $\boldsymbol{x}$ es restar la media:

$$\mathbf{H}\boldsymbol{x} = \boldsymbol{x} - \mathbf{1}\bar{x} = (x_1 - \bar{x},\, \ldots,\, x_n - \bar{x})^t.$$

La **matriz de datos centrada** respecto a todas las variables es

$$\mathbf{X}_c = \mathbf{H}\mathbf{X},$$

cuya entrada (i, j) vale $x_{ij} - \bar{x}_j$.

Proposición 1.2 (Propiedades de $\mathbf{H}$). *(i)* $\mathbf{H}$ *es* ***simétrica****:* $\mathbf{H}^t = \mathbf{H}$.

(ii) $\mathbf{H}$ *es* ***idempotente****:* $\mathbf{H}^2 = \mathbf{H}$.

(iii) Los valores propios de $\mathbf{H}$ *son* 0 *(con multiplicidad* 1*) y* 1 *(con multiplicidad* $n-1$*).*

(iv) $\operatorname{rango}(\mathbf{H}) = \operatorname{tr}(\mathbf{H}) = n - 1$.

Demostración. (i) $\mathbf{H}^t = \mathbf{I}_n^t - \frac{1}{n}(\mathbf{1}\mathbf{1}^t)^t = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^t = \mathbf{H}$.

(ii) $\mathbf{J}^2 = \frac{1}{n^2}\mathbf{1}\mathbf{1}^t\mathbf{1}\mathbf{1}^t = \frac{n}{n^2}\mathbf{1}\mathbf{1}^t = \mathbf{J}$, por lo que

$$\mathbf{H}^2 = (\mathbf{I} - \mathbf{J})^2 = \mathbf{I} - 2\mathbf{J} + \mathbf{J}^2 = \mathbf{I} - 2\mathbf{J} + \mathbf{J} = \mathbf{I} - \mathbf{J} = \mathbf{H}.$$

(iii) Si $\mathbf{H}\boldsymbol{u} = \lambda\boldsymbol{u}$, aplicando $\mathbf{H}$: $\mathbf{H}^2\boldsymbol{u} = \lambda^2\boldsymbol{u}$; pero $\mathbf{H}^2 = \mathbf{H}$, así que $\lambda\boldsymbol{u} = \lambda^2\boldsymbol{u}$, es decir $\lambda(\lambda - 1) = 0$. El único vector en el núcleo de $\mathbf{H}$ es proporcional a $\mathbf{1}$ (pues $\mathbf{H}\mathbf{1} = \mathbf{1} - \mathbf{1} = \mathbf{0}$), de modo que $\lambda = 0$ tiene multiplicidad 1 y $\lambda = 1$ tiene multiplicidad $n - 1$.

(iv) Como $\mathbf{H}$ es simétrica e idempotente, admite diagonalización ortogonal $\mathbf{H} = \mathbf{U}\Lambda\mathbf{U}^t$ con $\mathbf{U}$ ortogonal. Por tanto $\operatorname{rango}(\mathbf{H}) = \operatorname{rango}(\Lambda) = \operatorname{tr}(\Lambda) = \operatorname{tr}(\mathbf{U}\Lambda\mathbf{U}^t) = \operatorname{tr}(\mathbf{H}) = n - 1$. □

2. Sumas de cuadrados y descomposición de la variabilidad

2.1. Suma de cuadrados total

Definición 2.1. La **suma de cuadrados total** es la matriz $p \times p$

$$\mathbf{T} = \mathbf{X}^t \mathbf{H} \mathbf{X}.$$

La matriz de varianzas-covarianzas muestral verifica $\mathbf{S} = \frac{1}{n}\mathbf{T}$. Cuando $n > p$, el estimador insesgado de la matriz de covarianzas poblacional es $\hat{\mathbf{S}} = \frac{1}{n-p}\mathbf{T}$.

2.2. Estructura por grupos

Supongamos que las n observaciones provienen de g subpoblaciones: el grupo k aporta n_k observaciones ($\sum_{k=1}^{g} n_k = n$). La matriz de datos se particiona como $\mathbf{X}^t = (\mathbf{X}_1^t \mid \cdots \mid \mathbf{X}_g^t)$, donde $\mathbf{X}_k$ tiene dimensión $n_k \times p$.

Definición 2.2. Definimos la **matriz diagonal por bloques**

$$\mathbf{D} = \operatorname{diag}(\mathbf{J}_1, \ldots, \mathbf{J}_g), \qquad \mathbf{J}_k = \frac{1}{n_k} \mathbf{1}_{n_k} \mathbf{1}_{n_k}^t.$$

Proposición 2.3 (Propiedades de $\mathbf{D}$)**.** $\mathbf{D}$ *es simétrica e idempotente con* $\operatorname{rango}(\mathbf{D}) = \operatorname{tr}(\mathbf{D}) = g$.

Demostración. Simetría: cada $\mathbf{J}_k$ es simétrica, luego $\mathbf{D}^t = \mathbf{D}$. Idempotencia: $\mathbf{D}^2 = \operatorname{diag}(\mathbf{J}_1^2, \ldots, \mathbf{J}_g^2) = \operatorname{diag}(\mathbf{J}_1, \ldots, \mathbf{J}_g) = \mathbf{D}$ porque $\mathbf{J}_k^2 = \mathbf{J}_k$. Rango: $\operatorname{tr}(\mathbf{J}_k) = n_k \cdot \frac{1}{n_k} = 1$, así que $\operatorname{rango}(\mathbf{D}) = \operatorname{tr}(\mathbf{D}) = \sum_{k=1}^{g} 1 = g$. □

2.3. Descomposición $\mathbf{T} = \mathbf{W} + \mathbf{B}$

Teorema 2.4 (Descomposición de la variabilidad total)**.**

$$\mathbf{T} = \mathbf{X}^t \mathbf{H} \mathbf{X} = \underbrace{\mathbf{X}^t (\mathbf{I} - \mathbf{D}) \mathbf{X}}_{\mathbf{W}} + \underbrace{\mathbf{X}^t (\mathbf{D} - \mathbf{J}) \mathbf{X}}_{\mathbf{B}}.$$

Demostración. Basta escribir $\mathbf{H} = \mathbf{I} - \mathbf{J} = (\mathbf{I} - \mathbf{D}) + (\mathbf{D} - \mathbf{J})$. □

La **variabilidad intragrupos** ($\mathbf{W}$) y la **variabilidad entre grupos** ($\mathbf{B}$) admiten las siguientes expresiones explícitas.

2.3.1. Variabilidad entre grupos: B

Observemos que $\mathbf{DJ} = \mathbf{J}$ (cada bloque $\mathbf{J}_k$ aplicado a $\mathbf{J}$ da $\mathbf{J}$ por ser todas las columnas de $\mathbf{J}$ iguales a $\frac{1}{n}$). Por simetría, $\mathbf{JD} = \mathbf{J}$ también.

Proposición 2.5 (Propiedades de $\mathbf{D}-\mathbf{J}$)**.** $\mathbf{D}-\mathbf{J}$ *es simétrica e idempotente con* $\operatorname{rango}(\mathbf{D}-\mathbf{J}) = g - 1$.

Demostración. Simetría: inmediata por serlo $\mathbf{D}$ y $\mathbf{J}$. *Idempotencia*:

$$(\mathbf{D}-\mathbf{J})^2 = \mathbf{D}^2 - \mathbf{DJ} - \mathbf{JD} + \mathbf{J}^2 = \mathbf{D} - \mathbf{J} - \mathbf{J} + \mathbf{J} = \mathbf{D} - \mathbf{J}.$$

Rango: $\text{rango}(\mathbf{D}-\mathbf{J}) = \text{tr}(\mathbf{D}-\mathbf{J}) = \text{tr}(\mathbf{D}) - \text{tr}(\mathbf{J}) = g - 1$. □

Dado que $(\mathbf{D}-\mathbf{J})$ es simétrica e idempotente, usando que $\mathbf{DX} = (\mathbf{J}_1\mathbf{X}_1^t \mid \cdots \mid \mathbf{J}_g\mathbf{X}_g^t)^t$ con $\mathbf{J}_k\mathbf{X}_k$ igual a la matriz cuyas filas son $\bar{\boldsymbol{x}}_k^t$, se obtiene:

$$\mathbf{B} = \mathbf{X}^t(\mathbf{D}-\mathbf{J})\mathbf{X} = \big[(\mathbf{D}-\mathbf{J})\mathbf{X}\big]^t\big[(\mathbf{D}-\mathbf{J})\mathbf{X}\big] = \sum_{k=1}^{g} n_k(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})^t,$$

donde $\bar{\boldsymbol{x}}_k$ es el vector de medias del grupo k y $\bar{\boldsymbol{x}}$ la media global.

2.3.2. Variabilidad intragrupos: W

Proposición 2.6 (Propiedades de $\mathbf{I}-\mathbf{D}$)**.** $\mathbf{I}-\mathbf{D}$ *es simétrica e idempotente con* $\text{rango}(\mathbf{I}-\mathbf{D}) = n - g$.

Demostración. Análogo al caso de $\mathbf{H}$. Idempotencia: $(\mathbf{I}-\mathbf{D})^2 = \mathbf{I} - 2\mathbf{D} + \mathbf{D}^2 = \mathbf{I} - \mathbf{D}$. Rango: $\text{tr}(\mathbf{I}-\mathbf{D}) = n - g$. □

La variabilidad intragrupos satisface:

$$\mathbf{W} = \mathbf{X}^t(\mathbf{I}-\mathbf{D})\mathbf{X} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(\boldsymbol{x}_{ki} - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_{ki} - \bar{\boldsymbol{x}}_k)^t = \sum_{k=1}^{g} n_k\mathbf{S}_k,$$

donde $\boldsymbol{x}_{ki}$ es la i-ésima observación del grupo k y $\mathbf{S}_k$ es la matriz de covarianzas muestral del grupo k.

3. Análisis Factorial Discriminante: factores canónicos

3.1. Motivación

Clásicamente, el análisis discriminante distingue dos vertientes:

- **Aspecto descriptivo (geométrico)**: buscar combinaciones lineales de las variables originales que separen lo mejor posible los g grupos.
- **Aspecto decisional (probabilístico)**: asignar un nuevo individuo a uno de los grupos a partir de sus valores observados.

A menudo, cada variable marginal x_j se comporta de modo similar en todos los grupos, pero las diferencias resultan muy claras al considerar el comportamiento conjunto. Los **factores discriminantes canónicos** responden al aspecto descriptivo.

3.2. Planteamiento del problema

Dada la combinación lineal centrada $\boldsymbol{y} = \mathbf{H}\mathbf{X}\boldsymbol{a}$, su variabilidad total se descompone como:

$$n\ \mathrm{Var}(\boldsymbol{y}) = \boldsymbol{y}^t\boldsymbol{y} = \boldsymbol{a}^t\mathbf{X}^t\mathbf{H}\mathbf{X}\boldsymbol{a} = \boldsymbol{a}^t\mathbf{T}\boldsymbol{a} = \underbrace{\boldsymbol{a}^t\mathbf{W}\boldsymbol{a}}_{\text{intragrupos}} + \underbrace{\boldsymbol{a}^t\mathbf{B}\boldsymbol{a}}_{\text{entre grupos}}.$$

Objetivo: Encontrar $\boldsymbol{a} \in \mathbb{R}^p$ que maximice la proporción de variabilidad explicada entre grupos respecto a la variabilidad intragrupos:

$$\max_{\boldsymbol{a}} \frac{\boldsymbol{a}^t\mathbf{B}\boldsymbol{a}}{\boldsymbol{a}^t\mathbf{W}\boldsymbol{a}}.$$

Maximizar únicamente $\boldsymbol{a}^t\mathbf{B}\boldsymbol{a}$ no tiene sentido: una dirección con mayor variabilidad entre grupos puede tener aún mayor dispersión intragrupos, lo que empeoraría la separación real.

3.3. Resolución: reducción a un problema de valores propios

Suponemos que $\mathbf{W}$ es **definida positiva** (lo cual requiere $n - g \geq p$ y que ninguna variable sea combinación lineal de las restantes). En ese caso, $\mathbf{W} = \mathbf{U}\Lambda\mathbf{U}^t$ con Λ diagonal de entradas positivas, y podemos definir

$$\mathbf{W}^{1/2} = \mathbf{U}\Lambda^{1/2}\mathbf{U}^t, \quad \mathbf{W}^{-1/2} = \mathbf{U}\Lambda^{-1/2}\mathbf{U}^t.$$

Ambas son **simétricas** e inversas la una de la otra. El cambio de variable $\boldsymbol{b} = \mathbf{W}^{1/2}\boldsymbol{a}$ transforma el cociente de Rayleigh generalizado en uno estándar:

$$\frac{\boldsymbol{a}^t\mathbf{B}\boldsymbol{a}}{\boldsymbol{a}^t\mathbf{W}\boldsymbol{a}} = \frac{\boldsymbol{b}^t\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\boldsymbol{b}}{\boldsymbol{b}^t\boldsymbol{b}}.$$

Teorema 3.1 (Primer factor discriminante canónico)**.** *El máximo de* $\dfrac{\boldsymbol{a}^t\mathbf{B}\boldsymbol{a}}{\boldsymbol{a}^t\mathbf{W}\boldsymbol{a}}$ *se alcanza con*

$$\boldsymbol{a}_1 = \mathbf{W}^{-1/2}\boldsymbol{b}_1,$$

donde $\boldsymbol{b}_1$ *es el vector propio unitario asociado al mayor valor propio* λ_1 *de la matriz simétrica semidefinida positiva* $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$*. El primer factor discriminante canónico es*

$$\boldsymbol{y}_1 = \mathbf{H}\mathbf{X}\boldsymbol{a}_1.$$

Demostración. El cociente $\dfrac{\boldsymbol{b}^t\mathbf{M}\boldsymbol{b}}{\boldsymbol{b}^t\boldsymbol{b}}$, con $\mathbf{M} = \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ simétrica, alcanza su máximo λ_1 en el vector propio $\boldsymbol{b}_1$ asociado a dicho valor propio (cociente de Rayleigh clásico). Deshaciendo el cambio, $\boldsymbol{a}_1 = \mathbf{W}^{-1/2}\boldsymbol{b}_1$. □

3.3.1. Conexión con los valores propios de $\mathbf{W}^{-1}\mathbf{B}$

Proposición 3.2. $\boldsymbol{a}_1 = \mathbf{W}^{-1/2}\boldsymbol{b}_1$ *es vector propio de* $\mathbf{W}^{-1}\mathbf{B}$ *asociado al mismo valor propio* λ_1*, es decir,* $\mathbf{W}^{-1}\mathbf{B}\boldsymbol{a}_1 = \lambda_1\boldsymbol{a}_1$*.*

Demostración. De $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\boldsymbol{b}_1 = \lambda_1\boldsymbol{b}_1$ se tiene, premultiplicando por $\mathbf{W}^{-1/2}$:

$$\mathbf{W}^{-1}\mathbf{B}\underbrace{\mathbf{W}^{-1/2}\boldsymbol{b}_1}_{\boldsymbol{a}_1} = \lambda_1\underbrace{\mathbf{W}^{-1/2}\boldsymbol{b}_1}_{\boldsymbol{a}_1}.$$ □

En la práctica es más eficiente trabajar con la matriz simétrica $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$, pero conceptualmente el resultado se enuncia equivalentemente en términos de $\mathbf{W}^{-1}\mathbf{B}$.

3.4. Número de factores discriminantes canónicos

Teorema 3.3. *Si* $\text{rango}(\mathbf{X}) = p$*, el número máximo de factores discriminantes canónicos es*

$$s = \text{mín}(p,\, g-1).$$

Demostración. Como $\mathbf{W}$ es invertible, $\text{rango}(\mathbf{W}^{-1}\mathbf{B}) = \text{rango}(\mathbf{B})$. Ahora bien,

$$\text{rango}(\mathbf{B}) = \text{rango}\big(\mathbf{X}^t(\mathbf{D}-\mathbf{J})\mathbf{X}\big) \le \text{mín}\big(p,\, \text{rango}(\mathbf{D}-\mathbf{J})\big) = \text{mín}(p,\, g-1).$$

La igualdad se alcanza cuando $\text{rango}(\mathbf{X}) = p$. La cota $g-1$ refleja que conocer las medias de $g-1$ grupos determina la del último: $\bar{\boldsymbol{x}}_g = \big(n\bar{\boldsymbol{x}} - \sum_{k=1}^{g-1} n_k \bar{\boldsymbol{x}}_k\big)/n_g$. □

3.5. Los s factores discriminantes son no correlacionados

Definición 3.4. Los **factores discriminantes canónicos** son las variables $\boldsymbol{y}_k = \mathbf{H}\mathbf{X}\boldsymbol{a}_k$, $k = 1, \ldots, s$, con $\boldsymbol{a}_k = \mathbf{W}^{-1/2}\boldsymbol{b}_k$ y $\boldsymbol{b}_k$ el k-ésimo vector propio de $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ (ordenados por valor propio decreciente).

Teorema 3.5. $\text{Cov}(\boldsymbol{y}_1, \boldsymbol{y}_2) = 0$*; más en general, los s factores discriminantes canónicos tienen covarianzas dos a dos nulas.*

Demostración. Como los factores están centrados ($\mathbf{H}\mathbf{X}\boldsymbol{a}_k$ tiene media cero),

$$\begin{aligned} n\ \text{Cov}(\boldsymbol{y}_1, \boldsymbol{y}_2) &= \boldsymbol{a}_1^t \mathbf{X}^t \mathbf{H}\mathbf{X}\boldsymbol{a}_2 = \boldsymbol{a}_1^t \mathbf{T}\boldsymbol{a}_2 = \boldsymbol{a}_1^t(\mathbf{W}+\mathbf{B})\boldsymbol{a}_2 \\ &= \boldsymbol{a}_1^t \mathbf{W}\boldsymbol{a}_2 + \boldsymbol{a}_1^t \mathbf{B}\boldsymbol{a}_2. \end{aligned}$$

Para el primer sumando, con $\boldsymbol{a}_k = \mathbf{W}^{-1/2}\boldsymbol{b}_k$:

$$\boldsymbol{a}_1^t \mathbf{W}\boldsymbol{a}_2 = \boldsymbol{b}_1^t \underbrace{\mathbf{W}^{-1/2}\mathbf{W}\mathbf{W}^{-1/2}}_{\mathbf{I}} \boldsymbol{b}_2 = \boldsymbol{b}_1^t \boldsymbol{b}_2 = 0,$$

pues $\boldsymbol{b}_1, \boldsymbol{b}_2$ son vectores propios de una matriz simétrica para valores propios distintos, luego ortogonales. Para el segundo sumando, usando la simetría de $\mathbf{W}^{-1/2}$:

$$\boldsymbol{a}_1^t \mathbf{B}\boldsymbol{a}_2 = \boldsymbol{b}_1^t \underbrace{\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}}_{\mathbf{M}} \boldsymbol{b}_2 = \boldsymbol{b}_1^t(\lambda_2 \boldsymbol{b}_2) = \lambda_2\, \boldsymbol{b}_1^t \boldsymbol{b}_2 = 0.$$ □

3.6. Interpretación de los factores

Cada factor $\boldsymbol{y}_k$ proyecta los datos en la dirección $\boldsymbol{a}_k$ que, dentro de la restricción $\boldsymbol{a}^t\mathbf{W}\boldsymbol{a} = 1$, maximiza $\boldsymbol{a}^t\mathbf{B}\boldsymbol{a}$ sujeto a ser no correlacionado con los factores anteriores. El valor propio λ_k mide la separación entre grupos en esa dirección relativa a la dispersión intragrupos:

$$\lambda_k = \frac{\boldsymbol{a}_k^t \mathbf{B}\boldsymbol{a}_k}{\boldsymbol{a}_k^t \mathbf{W}\boldsymbol{a}_k}, \quad \text{con } \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_s \ge 0.$$

La proporción de discriminación explicada por el k-ésimo factor es

$$\rho_k = \frac{\lambda_k}{\sum_{j=1}^{s} \lambda_j}.$$

4. Análisis Discriminante: clasificación

4.1. El problema de clasificación

Sea $\boldsymbol{x} = (x_1, \ldots, x_p)^t$ el vector de características observadas en un individuo cuyo grupo de pertenencia se desconoce. Hay g subpoblaciones y se dispone de muestras de entrenamiento bien clasificadas. El objetivo es asignar $\boldsymbol{x}$ a uno de los g grupos.

Las situaciones habituales son:

1. La distribución de $\boldsymbol{x}$ está *completamente especificada* en cada subpoblación.
2. La distribución es *conocida salvo por el valor de algunos parámetros*.
3. No se conoce la distribución de $\boldsymbol{x}$ (métodos no paramétricos).

4.2. El método de máxima verosimilitud

Definición 4.1. El **criterio de máxima verosimilitud** asigna el individuo $\boldsymbol{x}$ al grupo k si

$$\mathcal{L}(\boldsymbol{x} \mid k) = \max_{j=1,\ldots,g} \mathcal{L}(\boldsymbol{x} \mid j),$$

donde $\mathcal{L}(\boldsymbol{x} \mid j)$ denota la función de verosimilitud de $\boldsymbol{x}$ en la subpoblación j.

4.2.1. Caso univariante con igual varianza

Para $g = 2$ con $X \mid 1 \sim \mathcal{N}(\mu_1, \sigma)$ y $X \mid 2 \sim \mathcal{N}(\mu_2, \sigma)$, el cociente de verosimilitudes es:

$$\frac{\mathcal{L}(x \mid 1)}{\mathcal{L}(x \mid 2)} = \exp\left(-\frac{(x-\mu_1)^2 - (x-\mu_2)^2}{2\sigma^2}\right).$$

Proposición 4.2. *El individuo x se clasifica en la subpoblación 1 si y solo si x está más próximo a μ_1 que a μ_2, es decir,*

$$|x-\mu_1| < |x-\mu_2| \iff x > \bar{\mu} := \frac{\mu_1+\mu_2}{2} \text{ si } \mu_1 > \mu_2 \quad (\text{y al revés si } \mu_1 < \mu_2).$$

Demostración. $\mathcal{L}(x \mid 1) > \mathcal{L}(x \mid 2)$ iff $-(x-\mu_1)^2 + (x-\mu_2)^2 > 0$ iff $2x(\mu_1-\mu_2) > \mu_1^2 - \mu_2^2 = (\mu_1-\mu_2)(\mu_1+\mu_2)$. Dividiendo por $\mu_1-\mu_2$ (y cambiando la desigualdad si es negativo) se obtiene el resultado. □

4.2.2. Caso univariante con desiguales varianzas

Para $X \mid j \sim \mathcal{N}(\mu_j, \sigma_j)$, $j = 1, 2$, la frontera de decisión es *cuadrática*:

$$z_1^2 + 2\ln\sigma_1 < z_2^2 + 2\ln\sigma_2, \quad z_j = \frac{x-\mu_j}{\sigma_j}.$$

4.3. Caso multivariante: discriminante lineal (LDA)

4.3.1. Dos subpoblaciones

Suponemos $\boldsymbol{x} \mid j \sim \mathcal{N}_p(\boldsymbol{\mu}_j, \Sigma)$, $j = 1, 2$, con la *misma* matriz de covarianzas Σ.

Teorema 4.3 (Regla lineal de clasificación)**.** *El criterio de máxima verosimilitud clasifica $\boldsymbol{x}$ en la subpoblación* 1 *sii*

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} \left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right) > 0.$$

Demostración. El cociente de verosimilitudes es

$$\frac{\mathcal{L}(\boldsymbol{x} \mid 1)}{\mathcal{L}(\boldsymbol{x} \mid 2)} = \exp\left\{ -\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) + \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \right\}.$$

Expandiendo las formas cuadráticas:

$$\begin{aligned} & -(\boldsymbol{x} - \boldsymbol{\mu}_1)^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) + (\boldsymbol{x} - \boldsymbol{\mu}_2)^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \\ &= 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} \boldsymbol{x} - \boldsymbol{\mu}_1^t \Sigma^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2^t \Sigma^{-1} \boldsymbol{\mu}_2 \\ &= 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \Sigma^{-1} \left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right). \end{aligned}$$

Esta cantidad es positiva iff se cumple la condición del enunciado. □

4.3.2. g subpoblaciones con igual Σ

Teorema 4.4. *Con g subpoblaciones gaussianas de igual covarianza Σ, el criterio de máxima verosimilitud equivale a la distancia de Mahalanobis:*

$$\textit{Clasificar en } k = \arg \min_{j=1,\ldots,g} \; D^2(\boldsymbol{x}, \boldsymbol{\mu}_j) \;:=\; (\boldsymbol{x} - \boldsymbol{\mu}_j)^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j).$$

Demostración. Maximizar $\mathcal{L}(\boldsymbol{x} \mid j) = |2\pi\Sigma|^{-1/2} \exp\{-\frac{1}{2} D^2(\boldsymbol{x}, \boldsymbol{\mu}_j)\}$ sobre j equivale a minimizar $D^2(\boldsymbol{x}, \boldsymbol{\mu}_j)$ sobre j, dado que el factor $|2\pi\Sigma|^{-1/2}$ es constante en j. □

Expandiendo $D^2(\boldsymbol{x}, \boldsymbol{\mu}_j)$ y omitiendo el término $\boldsymbol{x}^t \Sigma^{-1} \boldsymbol{x}$ (constante en j), la regla equivale a maximizar las **funciones discriminantes lineales**:

$$\delta_j(\boldsymbol{x}) = \boldsymbol{\mu}_j^t \Sigma^{-1} \boldsymbol{x} - \tfrac{1}{2} \boldsymbol{\mu}_j^t \Sigma^{-1} \boldsymbol{\mu}_j, \quad j = 1, \ldots, g.$$

4.4. Discriminante cuadrático (QDA)

Cuando las matrices de covarianzas Σ_j son distintas entre grupos, el criterio de máxima verosimilitud ya no produce fronteras lineales.

Teorema 4.5. *Con $\boldsymbol{x} \mid j \sim \mathcal{N}_p(\boldsymbol{\mu}_j, \Sigma_j)$ y matrices Σ_j posiblemente distintas, se clasifica $\boldsymbol{x}$ en el grupo k donde se maximiza la* ***función discriminante cuadrática****:*

$$q_j(\boldsymbol{x}) = -\tfrac{1}{2} \boldsymbol{x}^t \Sigma_j^{-1} \boldsymbol{x} + \boldsymbol{\mu}_j^t \Sigma_j^{-1} \boldsymbol{x} - \tfrac{1}{2} \ln |\Sigma_j| - \tfrac{1}{2} \boldsymbol{\mu}_j^t \Sigma_j^{-1} \boldsymbol{\mu}_j.$$

Demostración. Tomando logaritmos en la verosimilitud y descartando la constante $-\frac{p}{2} \ln(2\pi)$:

$$\ln \mathcal{L}(\boldsymbol{x} \mid j) = -\tfrac{1}{2} \ln |\Sigma_j| - \tfrac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_j)^t \Sigma_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j).$$

Expandiendo la forma cuadrática se obtiene directamente $q_j(\boldsymbol{x})$. □

Las fronteras de decisión son ahora superficies cuadráticas (hiperboloides, elipsoides o paraboloides según los Σ_j), en contraste con los hiperplanos del caso lineal.

4.5. El método bayesiano

En muchas aplicaciones existe información a priori sobre la frecuencia relativa de cada grupo en la población.

Definición 4.6. Sean $\pi_j > 0$, $\sum_j \pi_j = 1$, las **probabilidades a priori** de pertenencia al grupo j. El **método bayesiano** asigna $\boldsymbol{x}$ al grupo k donde se maximiza la probabilidad a posteriori:

$$\max_{j=1,\dots,g} \mathcal{L}(\boldsymbol{x} \mid j)\, \pi_j.$$

Observación 4.7. La maximización de $\mathcal{L}(\boldsymbol{x} \mid j)\pi_j$ es equivalente a la de $\mathcal{L}(j \mid \boldsymbol{x}) = \mathcal{L}(\boldsymbol{x} \mid j)\pi_j \,/\, \mathcal{L}(\boldsymbol{x})$, pues el denominador no depende de j. Si $\pi_j = 1/g$ para todo j, el método bayesiano se reduce al de máxima verosimilitud.

4.5.1. LDA bayesiano

Para subpoblaciones gaussianas con igual Σ, el método bayesiano maximiza sobre j:

$$\delta_j^{\mathrm{B}}(\boldsymbol{x}) = \boldsymbol{\mu}_j^t \Sigma^{-1} \boldsymbol{x} - \tfrac{1}{2} \boldsymbol{\mu}_j^t \Sigma^{-1} \boldsymbol{\mu}_j + \ln \pi_j.$$

4.5.2. QDA bayesiano

Con matrices Σ_j distintas, se maximiza:

$$q_j^{\mathrm{B}}(\boldsymbol{x}) = q_j(\boldsymbol{x}) + \ln \pi_j.$$

5. Estimación de parámetros desconocidos

En la práctica, los parámetros $\boldsymbol{\mu}_j$ y Σ (o Σ_j) son desconocidos y se sustituyen por sus estimadores de máxima verosimilitud a partir de las muestras de entrenamiento:

$$\hat{\boldsymbol{\mu}}_j = \bar{\boldsymbol{x}}_j, \qquad \hat{\Sigma} = \frac{1}{n}\mathbf{W} = \frac{1}{n}\sum_{k=1}^{g} n_k \mathbf{S}_k \quad \text{(caso LDA)},$$

$$\hat{\Sigma}_j = \mathbf{S}_j \quad \text{(caso QDA)}.$$

Las reglas de clasificación resultantes se denominan, respectivamente, **LDA muestral** y **QDA muestral**.

Observación 5.1. Cuando p es grande respecto a n, la estimación de Σ es inestable. En esos casos es habitual usar versiones regularizadas: $\hat{\Sigma}_\alpha = (1-\alpha)\hat{\Sigma} + \alpha \mathbf{I}$, con $\alpha \in [0,1]$ escogido por validación cruzada.

6. Resumen y relación entre métodos

Método	**Supuesto**	**Frontera**	**Criterio**
LDA	$\Sigma_j = \Sigma$	Lineal	$\max_j \delta_j(\boldsymbol{x})$
LDA bayesiano	$\Sigma_j = \Sigma$, π_j conocidas	Lineal	$\max_j \delta_j^{\mathrm{B}}(\boldsymbol{x})$
QDA	Σ_j distintas	Cuadrática	$\max_j q_j(\boldsymbol{x})$
QDA bayesiano	Σ_j distintas, π_j conocidas	Cuadrática	$\max_j q_j^{\mathrm{B}}(\boldsymbol{x})$
Análisis factorial	Descriptivo	—	Vectores propios de $\mathbf{W}^{-1}\mathbf{B}$

La conexión entre los dos grandes enfoques es la siguiente: el **análisis factorial discriminante** (aspecto descriptivo) produce las direcciones $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_s$ que maximizan progresivamente la separación entre grupos; el **análisis discriminante clásico** (aspecto decisional) usa reglas probabilísticas para asignar nuevos individuos. Ambos comparten la descomposición $\mathbf{T} = \mathbf{W}+\mathbf{B}$ como herramienta algebraica fundamental.