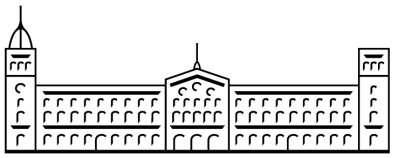


Treball final de grau

GRAU DE MATEMÀTIQUES

Facultat de Matemàtiques
Universitat de Barcelona

ANÀLISI DISCRIMINANT

Autor: María José Cañas Porcuna

Director:	**Dra. Carme Florit**
Co-Director:	**Dr. Josep Fortiana**
Realitzat a:	**Departament de Probabilitat, Lògica i Estadística**
Barcelona,	**17 de gener de 2016**

Abstract

It is common to find the necessity of identifying the characteristics that allow to distinguish two or more groups of individuals. The discriminant analysis consists in studying and analyzing these characteristics that you can use at the time of classifying in two or more groups. To know how to distinguish the groups you need to get the information, evaluated in variables in how it is suposed to distinguish. With the discriminant analysis you can find these variables and which of these are necessary to achieve the best classification. You use the name of class to identify the groups as an answer, for example, a categoric variable with as many discreet values as groups they have. The variables that are used to distinguish the groups are used as predictors or discriminant variables. There are several ways to deal with these analysis. In particular, this work is focused on the classical discriminant analysis: Fisher's linear discriminant and quadratic discriminant. Another technique is the canonical analysis of populations that it is applied in the case of more than two populations with the objective of representing them in orthogonal axes that allow to explain better the relation between the different groups. It will also be treated the logistic's regression model and the discriminant based on distances.

Resum

És freqüent trobar-se amb la necessitat d'identificar les característiques que permetin diferenciar a dos o més grups d'individus. L'anàlisi discriminant consisteix a estudiar i analitzar aquestes característiques que es poden utilitzar a l'hora de fer una classificació en dos o més grups. Per conèixer en què es diferencien els grups es necessita disposar de la informació, quantificada en unes variables, en la que se suposa que es diferencien. Amb l'anàlisi discriminant es troben aquestes variables i quines d'aquestes són necessàries per aconseguir la millor classificació. L'etiqueta de classe que identifica els grups s'utilitza com a resposta, és a dir, una variable categòrica amb tants valors discrets com grups hi hagin. Les variables que es fan servir per diferenciar els grups s'utilitzen com a predictors o variables discriminants. Existeixen diverses formes de tractar aquesta anàlisi. Concretament, aquest treball es concentra en l'anàlisi discriminant clàssica: discriminador lineal de Fisher i discriminador quadràtic. Una altre técnica és l'anàlisi canònica de poblacions que s'aplica en el cas de més de dues poblacions amb l'objectiu de representar aquestes en uns eixos ortogonals que permeten explicar millor la relació entre els diferents grups. També, es tractarà el model de regressió logística i el discriminador basat en distàncies.

Continguts

1 Introducció **4**

2 Distribucions **5**

2.1 Distribució normal multivariant 5

2.1.1 Definició 5

2.1.2 Propietats 6

2.2 Distribució de Wishart 6

2.2.1 Definició 6

2.2.2 Propietats 7

2.3 Distribució lambda de Wilks 7

3 Anàlisi discriminant **8**

3.1 Classificació en dues poblacions 9

3.1.1 Discriminador lineal 9

3.1.2 Regla de la màxima versemblança 10

3.1.3 Regla de Bayes 10

3.2 Classificació en poblacions normals 11

3.2.1 Discriminador lineal 11

3.2.2 Regla de Bayes 12

3.2.3 Probabilitat de classificació errònia 12

3.2.4 Discriminador quadràtic 13

3.2.5 Classificació quan els paràmetres són estimats 13

3.3 Exemple amb dues poblacions 13

3.4 Discriminació en el cas de k poblacions 15

3.4.1 Discriminadors lineals 15

3.4.2 Regla de la màxima versemblança 16

3.4.3 Regla de Bayes 16

3.5 Exemple amb tres poblacions 16

4 Anàlisi canònica de poblacions **20**

4.1 Distància de Mahalanobis i transformació canònica 21
4.2 Aspectes inferencials 22

5 Discriminador de Fisher per $k > 2$ poblacions **24**

6 Discriminació logística **26**

6.1 Anàlisi discriminant logístic 26
6.1.1 Model de regresió logística 26
6.1.2 Distribució asimptòtica i test de Wald 27
6.1.3 Ajust del model 28
6.1.4 Corba ROC 29
6.2 Anàlisi discriminant basat en distàncies 30
6.2.1 La funció de proximitat 30
6.2.2 La regla discriminant DB 31
6.2.3 La regla DB comparada amb altres 32
6.2.4 La regla DB en el cas de mostres 33

7 Exemples d'aplicacions en R **34**

7.1 Classificació en dues classes 34
7.1.1 Package *ElemStatLearn*: SAheart 34
7.1.2 Package *ElemStatLearn*: Spam 40
7.2 Classificació en més de dues classes 41
7.2.1 Package *car*: Skulls 41

8 Conclusions **46**

1 Introducció

Què és l'anàlisi discriminant?

L'anàlisi discriminant és una tècnica estadística multivariant amb la finalitat d'analitzar si existeixen diferències significatives entre grups d'objectes respecte a un conjunt de variables mesurades sobre aquests. Es tracta d'identificar les característiques que permetin diferenciar a dos o més grups d'individus a l'hora de fer la seva classificació. És a dir, donada una població, dividida en grups, l'anàlisi discriminant troba una funció que permet explicar aquesta divisió. Una vegada obtinguda, pot utilitzar-se per classificar nous individus en algun dels grups en que està dividida la població.

Motivació

La motivació per aquest Treball Final de Grau va sorgir pel meu interés per l'Estadística i, en concret, per l'anàlisi de dades, dins de l'anàlisi discriminant multivariant.

L'Estadística sempre ha sigut un tema pel qual he mostrat molta atenció i em semblava interessant aprofundir més en aquest àmbit.

Objectius

L'objectiu d'aquest treball és fer un estudi de l'anàlisi discriminant clàssica: el discriminador lineal de Fisher, tant en el cas de la classificació en dos grups com en el cas de més de dos grups, el discriminador quadràtic o la regla de Bayes, entre d'altres.

Una vegada estudiats els discriminadors, l'objectiu del treball és aplicar la teoria estudiada a un problema real donades unes dades on s'implementaran diferents funcions discriminants amb el programa R per identificar les variables que millor discriminin dos o més grups i fer la corresponent classificació.

Per mesurar la qualitat dels resultats obtinguts s'analitzará la matriu de confusió que s'explica en el capítol corresponent als exemples d'aplicacions en R.

2 Distribucions

Abans de començar amb la teoria del tema, es defineixen alguns conceptes previs referents a les diferents distribucions que s'utilitzaran més endavant.

2.1 Distribució normal multivariant

2.1.1 Definició

Sigui X una variable aleatòria amb distribució $N(\mu, \sigma^2)$, és a dir, amb mitjana μ i variància σ^2. La funció de densitat de X és:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} = \frac{(\sigma^2)^{-1/2}}{\sqrt{2\pi}} e^{\frac{1}{2}(x-\mu)\frac{1}{\sigma^2}(x-\mu)}. \tag{2.1}$$

Efectivament, es verifica:

$$X = \mu + \sigma Y \text{ on } Y \sim N(0,1), \tag{2.2}$$

on el símbol $\sim$ significa "distribuït com".

S'introdueix la distribució normal multivariant $N_p(\mu, \Sigma)$ com una generalització de la normal univariant. D'una banda, (2.1) suggereix definir la densitat de X $= (X_1, ..., X_p)' \sim N_p(\mu, \Sigma)$ segons:

$$f(\mathrm{x}; \mu, \Sigma) = \frac{|\Sigma|^{-1/2}}{(\sqrt{2\pi})^p} e^{-\frac{1}{2}(\mathrm{X}-\mu)'\Sigma^{-1}(\mathrm{X}-\mu)}, \tag{2.3}$$

on x $= (x_1, ..., x_p)'$, $\mu = (\mu_1, ..., \mu_p)'$ i $\Sigma = (\sigma_{ij})$ una matriu definida positiva, que és la matriu de covariàncies. D'altra banda, (2.2) suggereix definir la distribució X $= (X_1, ..., X_p)' \sim N_p(\mu, \Sigma)$ com una combinació lineal de p variables $Y_1, ..., Y_p$ independents amb distribució $N(0,1)$

$$\begin{aligned} X_1 &= \mu_1 + a_{11}Y_1 + ... + a_{1p}Y_p, \\ &\vdots \\ X_p &= \mu_p + a_{p1}Y_1 + ... + a_{pp}Y_p \end{aligned} \tag{2.4}$$

que es pot escriure com:

$$\mathrm{X} = \mu + A\mathrm{Y}, \tag{2.5}$$

sent Y $= (Y_1, ..., Y_p)'$ i A $= (a_{ij})$ una matriu $p \times p$ que verifica AA' $= \Sigma$.

2.1.2 Propietats

1. De (2.5) és immediat que $E(\mathrm{X}) = \mu$ i que la matriu de covariàncies és

$$E[(\mathrm{X} - \mu)(\mathrm{X} - \mu)'] = E(\mathrm{AYY'A'}) = \mathrm{AI}_p\mathrm{A}' = \Sigma.$$

2. La distribució de cada variable marginal X_i és normal univariant:

$$X_i \sim N(\mu_i, \sigma_{ii}),\ i = 1, ..., p$$

   És conseqüència de la definció (2.4).

3. Tota combinació lineal de les variables $X_1, ..., X_p$

$$Z = b_0 + b_1X_1 + ... + b_pX_p$$

   és també normal univariant. En efecte, de (2.4) resulta que Z és combinació lineal de $N(0, 1)$ independent.

4. Si $\Sigma = \mathrm{diag}(\sigma_{11}, ..., \sigma_{pp})$ és matriu diagonal, és a dir, $\sigma_{ij} = 0$, $i \neq j$, llavors les variables $(X_1, ..., X_p)$ són estocàsticament independents. Efectivament, la funció de densitat conjunta resulta igual al producte de les funcions de densitat marginals:

$$f(x_1, ..., x_p; \mu, \Sigma) = f(x_1; \mu_1, \sigma_{11}) \times ... \times f(x_p; \mu_p, \sigma_{pp})$$

5. La distribució de la forma quadràtica

$$U = (\mathrm{x} - \mu)'\Sigma^{-1}(\mathrm{x} - \mu)$$

   és khi-quadrat amb p graus de llibertat. En efecte, de (2.5) $U = \mathrm{Y'Y} = \Sigma_{i=1}^{p} Y_i^2$ és suma dels quadrats de p variables $N(0, 1)$ independents.

2.2 Distribució de Wishart

2.2.1 Definició

Si les files de la matriu $\mathrm{Z}_{n\times p}$ són independents $N_p(0, \Sigma)$ llavors la matriu $\mathrm{Q} = \mathrm{Z'Z}$ és Wishart $W_p(\Sigma, n)$, amb paràmetres Σ i n graus de llibertat.

Quan Σ és definida positiva i $n \geq p$, la densitat de Q és

$$f(\mathrm{Q}) = c|\mathrm{Q}|^{(n-p-1)}\exp(-\tfrac{1}{2}\mathrm{tr}(\Sigma^{-1}\mathrm{Q}))$$

on $c^{-1} = 2^{np/2}\pi^{p(p-1)/4}|\Sigma|^{n/2}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+1-i)\right]$.

2.2.2 Propietats

1. Si Q_1, Q_2 són independents Wishart $W_p(\Sigma, m)$, $W_p(\Sigma, n)$, llavors la suma $\mathrm{Q}_1 + \mathrm{Q}_2$ és també Wishart $W_p(\Sigma, m+n)$.

2. Si Q és $W_p(\Sigma, n)$, es separen les p variables en dos conjunts de p_1 i p_2 variables i es considera el següent

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \qquad \mathrm{Q} = \begin{pmatrix} \mathrm{Q}_{11} & \mathrm{Q}_{12} \\ \mathrm{Q}_{21} & \mathrm{Q}_{22} \end{pmatrix}$$

   llavors Q_{11} és $W_{p_1}(\Sigma_{11}, n)$ i Q_{22} és $W_{p_2}(\Sigma_{22}, n)$.

3. Si Q és $W_p(\Sigma, n)$ i T és una matriu $p \times q$ de constants, aleshores T'QT és $W_q(\mathrm{T}'\Sigma\mathrm{T}, n)$.

2.3 Distribució lambda de Wilks

Si les matrius A, B d'ordre $p \times q$ són independents Wishart $W_p(\Sigma, m)$, $W_p(\Sigma, n)$, respectivament, amb $m \geq p$, la distribució del quocient de determinants

$$\Lambda = \frac{|\mathrm{A}|}{|\mathrm{A}+\mathrm{B}|}$$

és la distribució lambda de Wilks, que s'indica com $\Lambda(p, m, n)$.

3 Anàlisi discriminant

Siguin Ω_1, Ω_2 dues poblacions, $X_1, ..., X_p$ variables observables. S'indica com x $= (x_1, ..., x_p)$ les observacions de les variables sobre un individu ω. Una regla discriminant és un criteri que permet assignar ω on es coneix $(x_1, ..., x_p)$ i que es planteja mitjançant una funció discriminant $D(x_1, ..., x_p)$. Llavors, la regla de classificació és:

$$\begin{cases} \text{Si } D(x_1, ..., x_p) \geq 0 & \text{s'assigna } \omega \text{ a } \Omega_1 \\ \text{En cas contrari} & \text{s'assigna } \omega \text{ a } \Omega_2 \end{cases}$$

Amb aquesta regla resulta la divisió de R^p en aquestes dues regions:

$$R_1 = \{\text{x } | D(\text{x}) > 0\} \text{ i } R_2 = \{\text{x } | D(\text{x}) < 0\}$$

A l'hora d'identficar ω, pot haver errors si s'assigna ω a una població a la que no pertany. La probabilitat de classificació errònia (pce) és:

$$pce = P(R_1/\Omega_1)P(\Omega_1) + P(R_2/\Omega_2)P(\Omega_2)$$

La Figura 1 mostra un exemple hipotètic on es consideren les variables X_1 i X_2 que es mesuren en un conjunt de 13 individus. Es tracta de trobar la recta que millor separi en dues regions els 13 individus analitzats.

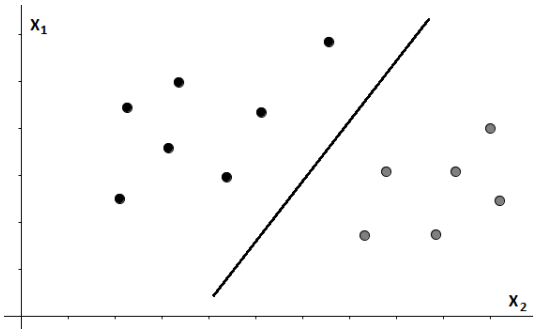


Figura 1: Núvol de punts resultants on es veu la separació en dues regions.

3.1 Classificació en dues poblacions

3.1.1 Discriminador lineal

Siguin μ_1, μ_2 els vectors de mitjanes de les variables en Ω_1, Ω_2, respectivament, i es suposa que la matriu de covariàncies Σ és comuna. Les distàncies de Mahalanobis de les observacions $\mathrm{x} = (x_1, ..., x_p)'$ d'un individu ω a les poblacions són

$$M^2(\mathrm{x},\, \mu_i) = (\mathrm{x} - \mu_i)'\Sigma^{-1}(\mathrm{x} - \mu_i),\ i = 1, 2.$$

Un primer criteri de classificació consisteix a assignar ω a la població més próxima:

$$\begin{cases} \text{Si } M^2(\mathrm{x},\, \mu_1) < M^2(\mathrm{x},\, \mu_2) & \text{s'assigna } \omega \text{ a } \Omega_1 \\ \text{En cas contrari} & \text{s'assigna } \omega \text{ a } \Omega_2 \end{cases} \tag{3.1}$$

Expressant aquesta regla com una funció discriminant, es té:

$$\begin{aligned} M^2(\mathrm{x},\, \mu_1) - M^2(\mathrm{x},\, \mu_2) &= \mathrm{x}'\Sigma^{-1}\mathrm{x} + \mu_2\Sigma^{-1}\mu_2 - 2\mathrm{x}'\Sigma^{-1}\mu_2 - \\ &\quad - \mathrm{x}'\Sigma^{-1}\mathrm{x} - \mu_1\Sigma^{-1}\mu_1 + 2\mathrm{x}'\Sigma^{-1}\mu_1 = \\ &= (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1) + 2\mathrm{x}'\Sigma^{-1}(\mu_1 - \mu_2) \end{aligned}$$

Es defineix la funció discriminant com

$$L(\mathrm{x}) = [\mathrm{x} - \frac{1}{2}(\mu_1 + \mu_2)]'\Sigma^{-1}(\mu_1 - \mu_2) \tag{3.2}$$

Aleshores, $M^2(\mathrm{x},\, \mu_2) - M^2(\mathrm{x},\, \mu_1) = 2L(\mathrm{x}) - L((\mu_1 - \mu_2)/2)$ i el criteri (3.1) és

$$\begin{cases} \text{Si } L(\mathrm{x}) > 0 & \text{s'assigna } \omega \text{ a } \Omega_1 \\ \text{En cas contrari} & \text{s'assigna } \omega \text{ a } \Omega_2 \end{cases}$$

La funció (3.2) és el discriminador lineal de Fisher (basat en l'article de Fisher, 1936: *The use of multiple measuraments in taxonomic problems*).

A la Figura 2 es mostra la idea de Fisher. Hi han dues classes que estan ben separades a l'espai original (x_1, x_2) però hi ha un solapament quan es projecta sobre la línea unint les seves mitjanes. La idea proposada per Fisher és maximitzar una funció que donarà una llarga separació entre les mitjanes de les classes projectades a la vegada que es dona una petita variància dins de cada classe, minimitzant així el solapament de classes.

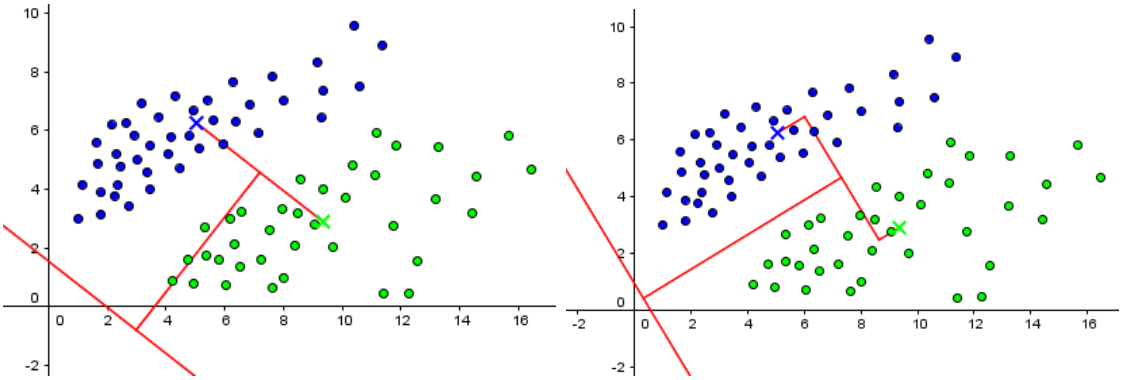


Figura 2: El dibuix de l'esquerra mostra les dades de dos classes, representades en blau i verd, amb la projecció sobre la recta que uneix les mitjanes de les classes. S'observa que hi ha un solapament de classes a l'espai projectat. El dibuix de la dreta mostra la projecció corresponent basada en el discriminant lineal de Fisher on es veu la millora de la separació de classes.

3.1.2 Regla de la màxima versemblança

Es suposa que $f_1(\mathrm{x})$, $f_2(\mathrm{x})$ són les densitats de x en Ω_1, Ω_2, respectivament. Un criteri de classificació consisteix a assignar ω a la població on la versemblança de les observacions x sigui més gran:

$$\begin{cases} \text{Si } f_1(\mathrm{x}) > f_2(\mathrm{x}) & \text{s'assigna } \omega \text{ a } \Omega_1 \\ \text{En cas contrari} & \text{s'assigna } \omega \text{ a } \Omega_2 \end{cases}$$

La funció discriminant és $V(\mathrm{x}) = \log\, f_1(\mathrm{x}) - \log\, f_2(\mathrm{x})$.

3.1.3 Regla de Bayes

Hi ha casos en que es coneixen les probabilitats *a priori* de que ω pertanyi a cada una de les poblacions, és a dir, $q_1 = P(\Omega_1)$, $q_2 = P(\Omega_2)$, $q_1 + q_2 = 1$. Una vegada que es disposa de les observacions $\mathrm{x} = (x_1, ..., x_p)$, les probabilitats *a posteriori* de que ω pertanyi a les poblacions (teorema de Bayes) són:

$$P(\Omega_i|\mathrm{x}) = \frac{q_i f_i(\mathrm{x})}{q_1 f_1(\mathrm{x}) + q_2 f_2(\mathrm{x})},\ i = 1, 2$$

La regla de classificació de Bayes és

$$\begin{cases} \text{Si } P(\Omega_1|\mathrm{x}) > P(\Omega_2|\mathrm{x}) & \text{s'assigna } \omega \text{ a } \Omega_1 \\ \text{En cas contrari} & \text{s'assigna } \omega \text{ a } \Omega_2 \end{cases}$$

El discriminador de Bayes és $B(\mathrm{x}) = \log\, f_1(\mathrm{x}) - \log\, f_2(\mathrm{x}) + \log\, (q_1/q_2)$. Quan $q_1 = q_2 = 1/2$, llavors $B(\mathrm{x}) = V(\mathrm{x})$ i aquest discriminador és òptim.

Teorema 3.1.3.1. La regla de Bayes minimitza la probabilitat de classificació errònia.

Demostració. Es suposa que es té una altre regla que classifica a Ω_1 si $\mathrm{x} \in R_1^*$, i a Ω_2 si $\mathrm{x} \in R_2^*$, on R_1^* i R_2^* són regions complementàries de l'espai mostral. S'indica que $d\mathrm{x} = dx_1, ..., dx_p$, llavors la probabilitat de classificació errònia és:

$$\begin{aligned} pce^* &= q_1 \int_{R_2^*} f_1(\mathrm{x})d\mathrm{x} + q_2 \int_{R_1^*} f_2(\mathrm{x})d\mathrm{x} = \\ &= \int_{R_2^*} (q_1 f_1(\mathrm{x}) - q_2 f_2(\mathrm{x}))d\mathrm{x} + q_2(\int_{R_1^*} f_2(\mathrm{x})d\mathrm{x} + \int_{R_2^*} f_2(\mathrm{x})d\mathrm{x}) = \\ &= \int_{R_2^*} (q_1 f_1(\mathrm{x}) - q_2 f_2(\mathrm{x}))d\mathrm{x} + q_2 \end{aligned}$$

Sigui $z = q_1 f_1(\mathrm{x}) - q_2 f_2(\mathrm{x})$, es té que l'última integral és mínima si R_2^* inclou totes les x tals que $z < 0$ i exclou totes les x tals que $z > 0$, és a dir, pce^* és mínima si $R_2^* = R_2$, on $R_2 = \{\mathrm{x}|B(\mathrm{x}) < 0\}$. □

3.2 Classificació en poblacions normals

Es suposa ara que les poblacions són normals, és a dir, que la distribució de $X_1, ..., X_p$ en Ω_1 és $N_p(\mu_1, \Sigma_1)$ i en Ω_2 és $N_p(\mu_2, \Sigma_2)$ i, per tant, que

$$f_i(\mathrm{x}) = (2\pi)^{-p/2}|\Sigma_i^{-1}|^{1/2}\, \exp\{-\frac{1}{2}(\mathrm{x} - \mu_i)'\Sigma_i^{-1}(\mathrm{x} - \mu_i)\}$$

3.2.1 Discriminador lineal

Si es suposa que $\mu_1 \neq \mu_2$, $\Sigma_1 = \Sigma_2 = \Sigma$, es té que

$$V(\mathrm{x}) = -\frac{1}{2}(\mathrm{x} - \mu_1)'\Sigma^{-1}(\mathrm{x} - \mu_1) + \frac{1}{2}(\mathrm{x} - \mu_2)'\Sigma^{-1}(\mathrm{x} - \mu_2) = L(\mathrm{x})$$

i, per tant, els discriminadors màxima versemblança i lineal coincideixen.

Sigui α la distància de Mahalanobis entre les dues poblacions

$$\alpha = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$$

Es considera que $U = (\mathrm{x} - \mu_1)'\Sigma^{-1}(\mu_1 - \mu_2)$. Si x prové de $N_p(\mu_1, \Sigma)$, llavors $E(U) = 0$, $\mathrm{var}(U) = E[(\mu_1 - \mu_2)'\Sigma^{-1}(\mathrm{x} - \mu_1)(\mathrm{x} - \mu_1)'\Sigma^{-1}(\mu_1 - \mu_2)] = \alpha$, per ser $E[(\mathrm{x} - \mu_1)(\mathrm{x} - \mu_1)'] = \Sigma$.

D'altra banda, de $\mathrm{x} - \frac{1}{2}(\mu_1 + \mu_2) = \mathrm{x} - \mu_1 + \frac{1}{2}(\mu_1 - \mu_2)$, es veu que $L(\mathrm{x}) = U + \frac{1}{2}\alpha$. Aleshores, $E(L(\mathrm{x})) = \alpha/2$, $\mathrm{var}(L(\mathrm{x})) = \alpha$.

De $\mathrm{x} - \mu_1 = \mathrm{x} - \mu_2 + \mu_2 - \mu_1$, es té que $U = (\mathrm{x} - \mu_2)'\Sigma^{-1}(\mathrm{x} - \mu_2) + \alpha$. Llavors, si x prové de $N_p(\mu_2, \Sigma)$, es veu que $E(U) = -\alpha$, $\mathrm{var}(U) = \alpha$. Al ser $L(\mathrm{x}) = U + \frac{1}{2}\alpha$, es dedueix que $E(L(\mathrm{x})) - \alpha/2$, $\mathrm{var}(L(\mathrm{x})) = \alpha$.

Amb això s'ha trobat la distribució de la funció discriminant $L(\mathrm{x})$:

$$\begin{cases} L(\mathrm{x}) \text{ és } N(+\frac{1}{2}\alpha, \alpha) \text{ si x prové de } N_p(\mu_1, \Sigma) \\ L(\mathrm{x}) \text{ és } N(-\frac{1}{2}\alpha, \alpha) \text{ si x prové de } N_p(\mu_2, \Sigma) \end{cases}$$

3.2.2 Regla de Bayes

Es suposa que $\mu_1 \neq \mu_2$, $\Sigma_1 = \Sigma_2 = \Sigma$, i que ja es coneixen les probabilitats $q_1 = P(\Omega_1)$, $q_2 = P(\Omega_2)$, $q_1 + q_2 = 1$, aleshores la funció discriminant de Bayes és el discriminador lineal més la constant $\log(q_1/q_2)$, és a dir, la funció és la següent $B(\mathrm{x}) = L(\mathrm{x}) + \log(q_1/q_2)$.

3.2.3 Probabilitat de classificació errònia

La probabilitat d'assignar x a Ω_2 quan prové de $N_p(\mu_1, \Sigma)$ és

$$P(L(\mathrm{x}) < 0|\Omega_1) = P((L(\mathrm{x}) - \tfrac{1}{2}\alpha)/\sqrt{\alpha}) = \Phi(-\tfrac{1}{2}\sqrt{\alpha})$$

on $\Phi(z)$ és la funció de distribució $N(0, 1)$. La probabilitat de classificació errònia és $pce = q_1 P(L(\mathrm{x}) < 0|\Omega_1) + q_2 P(L(\mathrm{x}) > 0|\Omega_2) = \Phi(-\frac{1}{2}\sqrt{\alpha})$.

3.2.4 Discriminador quadràtic

(Mardia et al, 1979). Es suposa que $\mu_1 \neq \mu_2$, $\Sigma_1 \neq \Sigma_2$. Aleshores, el criteri de màxima versemblança proporciona el discriminador quadràtic $Q(\mathrm{x})$:

$$\begin{aligned} Q(\mathrm{x}) =& \frac{1}{2}\mathrm{x}'(\Sigma_2^{-1} - \Sigma_1^{-1})\mathrm{x} + \mathrm{x}'(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) + \\ &+ \frac{1}{2}\mu_2'\Sigma_2^{-1}\mu_2 - \frac{1}{2}\mu_1'\Sigma_1^{-1}\mu_1 + \frac{1}{2}\log|\Sigma_2| - \frac{1}{2}\log|\Sigma_1| \end{aligned}$$

3.2.5 Classificació quan els paràmetres són estimats

A l'hora de fer un cas pràctic es tindrà que μ_1, μ_2, Σ_1, Σ_2 seran desconeguts i s'hauran d'estimar a partir de mostres de mides n_1, n_2 de les poblacions. Això es fa substituint μ_1, μ_2 pels vectors de mitjanes $\overline{x}_1$, $\overline{x}_2$ i Σ_1, Σ_2 per les matrius de covariàncies S_1, S_2. Si s'agafa l'estimador lineal, llavors l'estimació de Σ serà $\mathrm{S} = = (n_1\mathrm{S}_1 + n_2\mathrm{S}_2)/(n_1 + n_2)$ i la versió mostral del discriminador lineal quedarà així

$$\widehat{L}(\mathrm{x}) = [\mathrm{x} - \tfrac{1}{2}(\overline{x}_1 + \overline{x}_2)]'\mathrm{S}^{-1}(\overline{x}_1 - \overline{x}_2)$$

La distribució asimptòtica de $\widehat{L}(\mathrm{x})$ és normal, on $\alpha = (\overline{x}_1 - \overline{x}_2)'\mathrm{S}^{-1}(\overline{x}_1 - \overline{x}_2)$:

$$\begin{cases} \widehat{L}(\mathrm{x}) \text{ és } N(+\frac{1}{2}\alpha, \alpha) \text{ si x prové de } N_p(\mu_1, \Sigma) \\ \widehat{L}(\mathrm{x}) \text{ és } N(-\frac{1}{2}\alpha, \alpha) \text{ si x prové de } N_p(\mu_2, \Sigma) \end{cases}$$

3.3 Exemple amb dues poblacions

(Peña 2002, pp. 401-402). Es desitja classificar un retrat entre dos possibles pintors. Per fer això, es mesuren dues variables: la profunditat del traç i la proporció que ocupa el retrat sobre la superfície del llenç. Les mitjanes d'aquestes variables pel primer pintor, A, són 2 i 0.8, i pel segon pintor, B, 2.3 i 0.7. Les desviacions típiques d'aquestes variables en tots dos pintors són 0.5 i 0.1, i la correlació entre aquestes mesures és 0.5. L'obra a classificar té mesures d'aquestes variables 2.1 i 0.75. Calcular les probabilitats d'error.

Solució:

Sigui $\mathrm{x} = (2.1, 0.75)$ les observacions de les variables, $\mu_A = (2, 0.8)$ el vector de mitjanes de les variables respecte del pintor A, $\mu_B = (2.3, 0.7)$ el vector de mitjanes de les variables respecte del pintor B i la matriu de covariàncies calculada com el producte de la correlació per les desviacions típiques és la següent:

$$\Sigma = \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}$$

Les distàncies de Mahalanobis són:

$$\begin{aligned} M_A^2(\mathrm{x}, \mu_A) =& (\mathrm{x} - \mu_A)' \Sigma^{-1} (\mathrm{x} - \mu_A) = \\ =& (2.1 - 2, 0.75 - 0.8)' \Sigma^{-1} (2.1 - 2, 0.75 - 0.8) = \\ =& (0.1, -0.05)' \Sigma^{-1} (0.1, -0.05) = 0.52 \\ M_B^2(\mathrm{x}, \mu_B) =& (\mathrm{x} - \mu_B)' \Sigma^{-1} (\mathrm{x} - \mu_B) = \\ =& (2.1 - 2.3, 0.75 - 0.7)' \Sigma^{-1} (2.1 - 2.3, 0.75 - 0.7) = \\ =& (-0.2, 0.05)' \Sigma^{-1} (-0.2, 0.05) = 0.8133 \end{aligned}$$

Si s'aplica el criteri de classificació es té que:

$$M_A^2(\mathrm{x}, \mu_A) = 0.52 < 0.8133 = M_B^2(\mathrm{x}, \mu_B) \Rightarrow \text{s'assigna l'obra al primer pintor, A.}$$

Ara, es calcula la probabilitat de classificació errònia. La distància de Mahalanobis, α, entre les dues poblacions és:

$$\begin{aligned} \alpha =& (\mu_A - \mu_B)' \Sigma^{-1} (\mu_A - \mu_B) = \\ =& (2 - 2.3, 0.8 - 0.7)' \Sigma^{-1} (2 - 2.3, 0.8 - 0.7) = \\ =& (-0.3, 0.1)' \Sigma^{-1} (-0.3, 0.1) = 2.6133 \end{aligned}$$

Llavors, resulta que la *pce* és:

$$pce = \Phi(-\tfrac{1}{2}\sqrt{\alpha}) = \Phi(-\tfrac{1}{2}\sqrt{2.6133}) = \Phi(-0.808) = 0.1894$$

De manera que la classificació mitjançant aquestes variables no és molt precisa, ja que es pot tenir un 18.94 % de probabilitat d'error.

Per últim, es calcula la probabilitat *a posteriori* de que el quadre pertanyi al pintor A suposant que, *a priori*, els dos pintors són igualment probables.

$$\begin{aligned} P(A|\mathrm{x}) =& \frac{P(A) f_A(\mathrm{x})}{P(A) f_A(\mathrm{x}) + P(B) f_B(\mathrm{x})} = \frac{1}{1 + \frac{P(B)}{P(A)} \exp\{-\frac{1}{2}(M_B^2 - M_A^2)\}} = \\ =& \frac{1}{1 + \frac{P(B)}{P(A)} \exp\{-\frac{1}{2}(0.8133 - 0.52)\}} = \frac{1}{1.863} = 0.5376 \end{aligned}$$

Aquesta probabilitat indica que al classificar l'obra com pertanyent al pintor A existeix molta incertesa en la decisió, ja que les probabilitats de que pertanyi a cada pintor són semblants (0.5376 i 0.4624).

3.4 Discriminació en el cas de k poblacions

En aquest cas es suposarà que l'individu ω pot provenir de k poblacions $\Omega_1, ..., \Omega_k$, on $k \geqslant 3$. Es vol establir una regla que permeti assignar ω a una de les k poblacions a partir de les observacions $\mathrm{x} = (x_1, ..., x_p)'$ de p variables.

3.4.1 Discriminadors lineals

Sigui μ_i la mitjana de les variables en Ω_i i es suposa que la matriu de covariàncies Σ és comuna. Considerant les distàncies de Mahalanobis de ω a les poblacions

$$M^2(\mathrm{x},\, \mu_i) = (\mathrm{x} - \mu_i)'\Sigma^{-1}(\mathrm{x} - \mu_i),\ i = 1, ..., k$$

un criteri de classificació consisteix en assignar ω a la població més próxima:

$$\text{Si } M^2(\mathrm{x}, \mu_i) = \text{mín}\{M^2(\mathrm{x}, \mu_1), ..., M^2(\mathrm{x}, \mu_k)\}, \text{ s'assigna } \omega \text{ a } \Omega_i \tag{3.3}$$

Introduint les funcions discriminants lineals (Figura 3)

$$L_{ij}(\mathrm{x}) = (\mu_i - \mu_j)'\Sigma^{-1}\mathrm{x} - \tfrac{1}{2}(\mu_i - \mu_j)'\Sigma^{-1}(\mu_i + \mu_j)$$

resulta que (3.3) equival a:

$$\text{Si } L_{ij}(\mathrm{x}) > 0,\ \forall j \neq i, \text{ s'assigna } \omega \text{ a } \Omega_i.$$

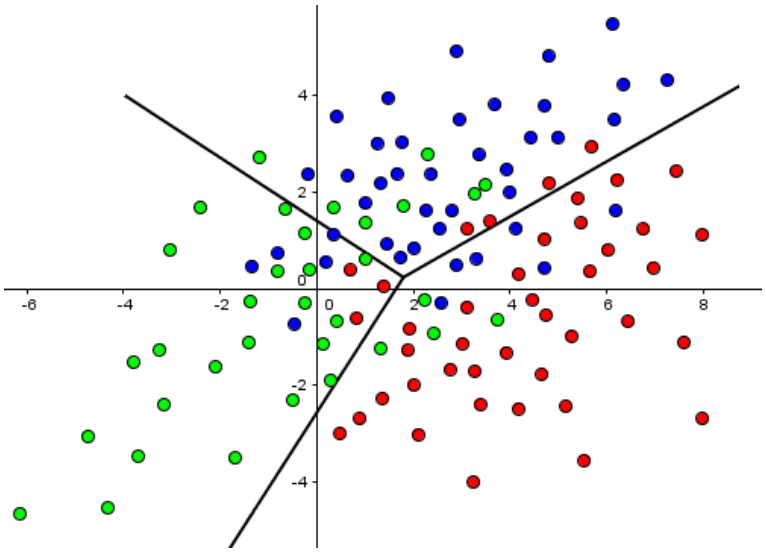


Figura 3: Discriminació entre més de dues poblacions normals.

3.4.2 Regla de la màxima versemblança

Sigui $f_i(\mathrm{x})$ la funció de x en la població Ω_i, la regla de classificació s'obté assignant ω a la població on la versemblança és més gran:

$$\text{Si } f_i(\mathrm{x}) = \text{màx}\{f_1(\mathrm{x}), ..., f_k(\mathrm{x})\}, \text{ s'assigna } \omega \text{ a } \Omega_i$$

I les funcions discriminants $V_{ij}(\mathrm{x}) = \log\, f_i(\mathrm{x}) - \log\, f_j(\mathrm{x})$.

En el cas de normalitat multivariant i matriu de covariàncies comuna, es verifica que $V_{ij}(\mathrm{x}) = L_{ij}(\mathrm{x})$, és a dir, els discriminants versemblants coincideixen amb els lineals. Però si les matrius de covariàncies són diferents $\Sigma_1, ..., \Sigma_k$, aleshores aquest criteri resultarà als discriminants quadràtics

$$\begin{aligned} Q_{ij}(\mathrm{x}) =& \frac{1}{2}\mathrm{x}'(\Sigma_j^{-1} - \Sigma_i^{-1})\mathrm{x} + \mathrm{x}'(\Sigma_i^{-1}\mu_1 - \Sigma_j^{-1}\mu_2) + \\ &+ \frac{1}{2}\mu_j'\Sigma_j^{-1}\mu_j - \frac{1}{2}\mu_i'\Sigma_i^{-1}\mu_i + \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}\log|\Sigma_i| \end{aligned}$$

3.4.3 Regla de Bayes

Sigui $f_i(\mathrm{x})$ la funció de densitat i es suposa que es coneixen les probabilitats *a priori* $q_1 = P(\Omega_1), ..., q_k = P(\Omega_k)$. Aleshores, la regla de Bayes que assigna ω a la població tal que la probabilitat *a posteriori* és màxima és la següent:

$$\text{Si } q_i f_i(\mathrm{x}) = \text{màx}\{q_1 f_1(\mathrm{x}), ..., q_k f_k(\mathrm{x})\}, \text{ s'assigna } \omega \text{ a } \Omega_i$$

que està associada a les funcions discriminants

$$B_{ij}(\mathrm{x}) = \log\, f_i(\mathrm{x}) - \log\, f_j(\mathrm{x}) + \log\, (q_i/q_j)$$

Sigui $P(j|i)$ la probabilitat d'assignar ω a Ω_j quan en realitat és de Ω_i, llavors, la probabilitat de classificació errònia és

$$pce = \sum_{i=1}^{k} q_i \left(\sum_{j \neq i}^{k} P(j|i) \right)$$

3.5 Exemple amb tres poblacions

(Peña 2002, pp. 404-406). Una màquina que admet monedes realitza tres mesures de cada moneda per determinar el seu valor: pes (x_1), gruix (x_2) i la densitat

d'estries en el seu cant (x_3). Els instruments de mesurament d'aquestes variables no són molt precises i s'ha comprovat en una amplia experimentació amb tres tipus de monedes utilitzades, M_1, M_2, M_3, on les mesures s'atribueixen normalment, amb mitjanes per cada tipus de moneda donades per:

$$\mu_1 = (20, 8, 8)'$$
$$\mu_2 = (19.5, 7.8, 10)'$$
$$\mu_3 = (20.5, 8.3, 5)'$$

i matriu de covariàncies:

$$V = \begin{pmatrix} 4 & 0.8 & -5 \\ 0.8 & 0.25 & -0.9 \\ -5 & -0.9 & 9 \end{pmatrix}$$

El que planteja el problema és indicar com es classificaria una moneda amb mesures $(22, 8.5, 7)'$ i analitzar la regla de classificació. Calcular les probabilitats d'error.

Solució:

Observant les dades, la moneda a classificar està més pròxima a M_3 en les dues primeres coordenades, però més pròxima a M_3 per x_3, la densitat d'estries. La variable indicador per classificar entre M_1 i M_3 és:

$$z_1 = (\mu_1 - \mu_3)'\mathrm{V}^{-1}\mathrm{x} = 1.77x_1 - 3.31x_2 + 0.98x_3$$

la mitjana d'aquesta variable per la primera moneda, M_1, és $1.77 \times 20 - 3.31 \times 8 + 0.98 \times 8 = 16.71$ i per la tercera, M_3, $1.77 \times 20.5 - 3.31 \times 8.3 + 0.98 \times 5 = 13.65$. El punt de tall és la mitjana d'aquestes dues quantitats, 15.17. Com per la moneda a classificar es té:

$$1.77 \times 22 - 3.31 \times 8.5 + 0.98 \times 7 = 17.61$$

es classificarà com M_1. Aquest anàlisi és equivalent a calcular les distàncies de Mahalanobis a cada població que resulten ser $M_1^2 = 1.84$, $M_2^2 = 2.01$ i $M_3^2 = 6.69$. Per tant, es classifica primer en M_1, després en M_2, i per últim, com M_3. La regla per classificar entre la primera i la segona és

$$z_2 = (\mu_1 - \mu_2)'\mathrm{V}^{-1}\mathrm{x} = -0.93x_1 + 1.74x_2 - 0.56x_3$$

de les dues regles per obtenir z_1 i z_2 es dedueix la regla per classificar entre la segona i la tercera, ja que

$$z_3 = (\mu_2 - \mu_3)'\mathrm{V}^{-1}\mathrm{x} = (\mu_1 - \mu_3)'\mathrm{V}^{-1}\mathrm{x} - (\mu_1 - \mu_2)'\mathrm{V}^{-1}\mathrm{x} = z_1 - z_2$$

S'analitzen les regles de classificació obtingudes. S'expressa la regla inicial per classificar M_1 i M_3 per les variables estandaritzades, i així s'evita el problema de les unitats. S'anomena $\tilde{x}_i$ a les variables dividides per les seves desviacions típiques $\tilde{x}_1 = x_1/2$, $\tilde{x}_2 = x_2/2$ i $\tilde{x}_3 = x_3/2$, i la regla en variables estandaritzades és:

$$\tilde{z}_1 = 3.54\tilde{x}_1 - 1.65\tilde{x}_2 + 2.94\tilde{x}_3$$

que indica que les variables amb més pes per decidir la classificació són la primera i la tercera, que són les que tenen majors coeficients. Es veu que amb variables estandaritzades, la matriu de covariàncies és la de correlació

$$R = \begin{pmatrix} 1 & 0.8 & -0.83 \\ 0.8 & 1 & -0.6 \\ -0.83 & -0.6 & 1 \end{pmatrix}$$

L'origen d'aquestes correlacions entre els errors de mesura és que si la moneda adquireix brutícia i augmenta lleugerament el seu pes, també augmenta el seu gruix i fa més difícil determinar la seva densitat d'estries. Per tant, hi ha correlacions positives entre pes i gruix, si augmenta el pes augmenta el gruix, però negatives amb les estries. La moneda que es vol classificar té bastant pes i gruix, i això indicaria que pertany a la classe 3, però llavors la densitat d'estries hauria de mesurar-se com baixa, ja que hi ha correlacions negatives entre les dues mesures i, però, es mesura relativament alta en la moneda. Les tres mesures són coherents amb una moneda bruta del tipus 1, i per tant es classifica amb facilitat en aquest grup.

Ara, es calcula la probabilitat *a posteriori* de que l'observació sigui de la classe M_1. Suposant que les probabilitats *a priori* són iguals, aquesta probabilitat serà:

$$\begin{aligned} P(1|\mathrm{x}_0) =& \frac{\exp(-M_1^2/2)}{\exp(-M_1^2/2) + \exp(-M_2^2/2) + \exp(-M_3^2/2)} = \\ =& \frac{\exp(-1.84/2)}{\exp(-1.84/2) + \exp(-2.01/2) + \exp(-6.69/2)} = 0.50 \end{aligned}$$

i, anàlogament, $P(2|\mathrm{x}_0) = 0.46$ i $P(3|\mathrm{x}_0) = 0.04$.

Es pot calcular les probabilitats d'error de classificar una moneda de qualsevol tipus en una altre classe. Per exemple, la probabilitat de classificar una moneda M_3 amb aquesta regla com tipus M_1 és:

$$P(z_1 > 15.73|N(13.64, \sqrt{3.07})) = P\left(y > \frac{15.17 - 13.64}{1.75}\right) = P(y > 0.87) = 0.192$$

com es veu, aquesta probabilitat és bastant alta. Si es vol reduir, s'ha d'augmentar la distància de Mahalanobis entre les mesures dels grups, i això suposa "augmentar" la matriu V^{-1} o "reduir" V. Per exemple, si es redueix a la meitat l'error en la mesura de les estries introduint mesuradors més precisos, i es manté les correlacions amb les altres mesures, s'obté la matriu de covariàncies següent:

$$V_2 = \begin{pmatrix} 4 & 0.8 & -2.5 \\ 0.8 & 0.25 & -0.45 \\ -1 & -0.2 & 2.25 \end{pmatrix}$$

Aleshores, ara la regla de classificació entre la primera i la tercera és:

$$z_1 = (\mu_1 - \mu_3)'\mathrm{V}^{-1}\mathrm{x} = 3.44x_1 - 4.57x_2 + 4.24x_3$$

i la distància de Mahalanobis entre les poblacions 1 i 3 (monedes M_1 i M_3) han canviat de 3.01 a 12.38, la qual cosa implica que la probabilitat d'error entre aquestes dues poblacions ha disminuït a $\Phi(-\frac{1}{2}\sqrt{12.38}) = \Phi(-1.76) = 0.04$ i s'observa que la probabilitat d'error ha disminuït considerablament. Es calcularà així la precisió en les mesures necessària per aconseguir unes probabilitats d'error determinades.

Comentari respecte el capítol:

El métode per mínims quadrats ordinaris és un altre model clàssic de l'anàlisi discriminant, encara que a efectes de comparació, és un métode sense interés ja que hi ha métodes més interessants i millors a l'hora de fer una classificació. Com es veurà més endavant, aquest métode s'utilitza al capítol d'exemples d'aplicacions amb R i es compara amb els altres discriminadors quan es fa la classificació.

4 Anàlisi canònica de poblacions

L'anàlisi canònica de poblacions és una tècnica que s'aplica si es tenen diverses matrius de dades, com a resultat d'observar les variables en diverses poblacions, i el que es vol és representar les poblacions.

Es suposa que de l'observació de p variables quantitatives $X_1, ..., X_p$ en g poblacions s'obtenen g matrius de dades $\mathrm{X} = (X_1, ..., X_g)'$, on X_i és la matriu $n_i \times p$ de la població i. Siguin $\overline{\mathrm{x}}_1', ..., \overline{\mathrm{x}}_g'$ els vectors fila de les mitjanes de cada població. X és d'ordre $n \times p$, on $n = \sum_{i=1}^{g} n_i$. La matriu $g \times p$ amb les mitjanes de les g poblacions és $\overline{\mathrm{X}} = (\overline{\mathrm{x}}_1' - \overline{\mathrm{x}}', ..., \overline{\mathrm{x}}_g' - \overline{\mathrm{x}}')'$. Les dues formes de quantificar matricialment la dispersió entre les poblacions són:

- La matriu de dispersió no ponderada entre grups:

$$\mathrm{A} = \overline{\mathrm{X}}'\overline{\mathrm{X}} = \sum_{i=1}^{g} (\overline{\mathrm{x}}_i - \overline{\mathrm{x}})(\overline{\mathrm{x}}_i - \overline{\mathrm{x}})'$$

- La matriu de dispersió ponderada entre grups:

$$\mathrm{B} = \sum_{i=1}^{g} n_i (\overline{\mathrm{x}}_i - \overline{\mathrm{x}})(\overline{\mathrm{x}}_i - \overline{\mathrm{x}})'$$

La matriu A és proporcional a una matriu de covariàncies prenent com a dades només les mitjanes de les poblacions. Aleshores, aquesta matriu serà la matriu de covariàncies entre les poblacions. La matriu B participa, juntament amb W que és la matriu de dispersió dins de grups, en el test de comparació de mitjanes de g poblacions. D'altra banda, es té la matriu $\mathrm{S} = \dfrac{1}{n-g} \sum_{i=1}^{g} n_i \mathrm{S}_i$ que serà la matriu de covariàncies dins de les poblacions.

Continuem definint les variables canòniques. Siguin $\mathrm{V} = [\mathrm{v}_1, ..., \mathrm{v}_p]$ els vectors propis d'A respecte de S amb valors propis $\lambda_1 > ... > \lambda_p$, és a dir, $\mathrm{A}\mathrm{v}_i = \lambda_i \mathrm{S}_i \mathrm{v}_i$, normalitzats segons $\mathrm{v}_i' \mathrm{S}_i \mathrm{v}_i = 1$.
Els vectors $\mathrm{v}_1, ..., \mathrm{v}_p$ són els vectors canònics i les variables canòniques són les variables compostes $Y_i = \mathrm{X}\mathrm{v}_i$.

Si $\mathrm{v}_i = (\upsilon_{1i}, ..., \upsilon_{pi})'$ i $\mathrm{X} = [X_1, ..., X_p]$, la variable canònica Y_i és la variable composta

$$Y_i = \mathrm{X}\mathrm{v}_i = \upsilon_{1i} X_1 + ... + \upsilon_{pi} X_p$$

que té S-variància 1 i A-variància λ_i, és a dir:

$$\mathrm{var}_A(Y_i) = \mathrm{v}_i'\mathrm{A}\mathrm{v}_i = \lambda_i, \quad \mathrm{var}_S(Y_i) = \mathrm{v}_i'\mathrm{S}_i\mathrm{v}_i = 1$$

4.1 Distància de Mahalanobis i transformació canònica

En aquest apartat es definirà la distància entre dues poblacions quan hi ha més de dues poblacions. Es consideren mostres multivariants de g poblacions amb vectors de mitjanes $\overline{\mathrm{x}}_1, ..., \overline{\mathrm{x}}_g$ i matriu de covariàncies comuna S. La distància al quadrat de Mahalanobis entre les poblacions i, j és

$$M^2(i,j) = (\overline{\mathrm{x}}_i - \overline{\mathrm{x}}_j)'\mathrm{S}^{-1}(\overline{\mathrm{x}}_i - \overline{\mathrm{x}}_j)$$

Si $\overline{\mathrm{X}}$ és la matriu centrada amb els vector de mitjanes i $\mathrm{V} = [\mathrm{v}_1, ..., \mathrm{v}_p]$ és la matriu amb els vectors canònics, la transformació canònica és $\mathrm{Y} = \overline{\mathrm{X}}\mathrm{V}$, que és una matriu d'ordre $g \times p$ que conté les coordenades canòniques de les g poblacions.

Teorema 4.1.1. La distància de Mahalanobis entre cada par de poblacions i, j coincideix amb la distància euclídea entre les files i, j de la matriu de coordenades canòniques Y. Si $\mathrm{y}_i = \overline{\mathrm{x}}_i\mathrm{V}$ aleshores

$$d_E^2(i,j) = (\mathrm{y}_i - \mathrm{y}_j)'(\mathrm{y}_i - \mathrm{y}_j) = (\overline{\mathrm{x}}_i - \overline{\mathrm{x}}_j)'\mathrm{S}^{-1}(\overline{\mathrm{x}}_i - \overline{\mathrm{x}}_j)$$

Pel que fa a la representació mitjançant les coordenades canòniques Y amb la mètrica euclidiana, es realitza al llarg d'eixos ortogonals. Si, a més, es prenen les m primeres coordenades canòniques, la representació és totalment factible i és òptima en dimensió reduïda, en el sentit que maximitza la variabilitat geomètrica.

Teorema 4.1.2. La variabilitat geomètrica de les distàncies de Mahalanobis entre les poblacions és proporcional a la suma dels valors propis:

$$V_M(\overline{\mathrm{X}}) = \frac{1}{2g^2}\sum_{i,j=1}^{g} M(i,j)^2 = \frac{1}{g}\sum_{i=1}^{p}\lambda_i$$

Si $\mathrm{Y} = \overline{\mathrm{X}}\mathrm{V}$, on V és la matriu de la transformació canònica d'ordre $p \times m$ en dimensió m i

$$\delta_{ij}^2(m) = (\mathrm{y}_i - \mathrm{y}_j)(\mathrm{y}_i - \mathrm{y}_j)' = \sum_{h=1}^{m}(y_{ih} - y_{jh})^2$$

és la distància euclídea al quadrat entre dos files de Y, la variabilitat geomètrica en dimensió $m \leq p$ és

$$V_\delta(\mathrm{X})_m = \frac{1}{2g^2} \sum_{i,j=1}^{g} \delta_{ij}^2(m) = \frac{1}{g} \sum_{i=1}^{p} \lambda_i$$

i aquesta quantitat és màxima entre totes les transformacions lineals possibles en dimensió m.

4.2 Aspectes inferencials

Ara es suposarà que les matrius de dades $\mathrm{X}_1, ..., \mathrm{X}_g$ provenen de g poblacions normals $N_p(\mu_1, \Sigma_1), ..., N_p(\mu_g, \Sigma_g)$.

- **Comparació de mitjanes.** Es construeix el test següent per decidir si es pot acceptar la hipòtesi d'igualtat de mitjanes

  $$H_0 : \mu_1 = \mu_2 = ... = \mu_g$$

  Siguin

  $$\mathrm{B} = \sum_{i=1}^{g} n_i(\overline{\mathrm{x}}_i - \overline{\mathrm{x}})(\overline{\mathrm{x}}_i - \overline{\mathrm{x}})' \qquad \text{(dispersió entre grups)}$$

  $$\mathrm{W} = \sum_{i=1}^{g} \sum_{\alpha=1}^{n_i} (\mathrm{x}_{i\alpha} - \overline{\mathrm{x}}_i)(\mathrm{x}_{i\alpha} - \overline{\mathrm{x}}_i)' \qquad \text{(dispersió dins de grups)}$$

  $$\mathrm{T} = \sum_{i=1}^{g} \sum_{\alpha=1}^{n_i} (\mathrm{x}_{i\alpha} - \overline{\mathrm{x}})(\mathrm{x}_{i\alpha} - \overline{\mathrm{x}})' \qquad \text{(dispersió total)}$$

  el test es decideix calculant l'estadístic $\Lambda = |\mathrm{W}|/|\mathrm{B} + \mathrm{W}|$ amb distribució lambda de Wilks, és a dir, $\Lambda(p, n - g, g - 1)$.
  Si s'accepta H_0, les mitjanes de les poblacions són teòricament iguals i l'anàlisi canònic no tindria sentit. Per tant, és convenient rebutjar H_0.

- **Comparació de covariàncies.** El test

  $$H_0' : \Sigma_1 = \Sigma_2 = ... = \Sigma_g$$

  es resol mitjançant el test de la raó de versemblança

$$\lambda_R = \frac{|\mathrm{S}_1|^{n_1/2} \times ... \times |\mathrm{S}_g|^{n_g/2}}{|\mathrm{S}|^{n/2}}$$

on S_i és la matriu de covariàncies de les dades de la població i, estimació màxim versemblant de Σ_i i $\mathrm{S} = (n_1\mathrm{S}_1 + ... + n_g\mathrm{S}_g)/n = \mathrm{W}/n$ és l'estimació màxim versemblant de Σ, matriu de covariàncies comuna sota H_0'. Es rebutjarà H_0' si l'estadístic

$$-2\log\lambda_R \sim \chi^2_q$$

és significatiu, on $q = gp(p+1)/2 - p(p+1)/2 = (g-1)p(p+1)/2$ són els graus de llibertat de la khi-quadrat. Si es rebutja H_0', resulta que no es disposa d'uns eixos comuns per representar totes les poblacions i l'anàlisi canònic és teòricament incorrecte. Aleshores, és convenient acceptar H_0'. Aquest test s'anomena test de Bartlett.

5 Discriminador de Fisher per $k > 2$ poblacions

En aquest capítol, es considerarà $K > 2$ poblacions. Es suposa que la dimensió D de l'entrada és major que el nombre K de les classes. Sigui $D' > 1$ i que els elements lineals $y_i = \mathbf{a}_i'\mathbf{x}$, on $i = 1, ..., D'$. Aquests valors es poden agrupar junts formant el vector $\mathbf{y}$. De la mateixa manera, els vectors de ponderació $\{\mathbf{a}_i\}$ poden ser considerats com les columnes d'una matriu $\mathbf{A}$, de manera que $\mathbf{y} = \ = \mathbf{A}'\mathbf{x}$.

Es considera una mostra amb n_i casos en cada grup, i la mida total de la mostra és $n = \sum_{i=1}^{K} n_i$. Es denotarà per $\mathbf{x}_{ij}$ l'observació i-èsima en el grup j-èsim. Aleshores, definim:

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\mathbf{x}_{ij} = \frac{1}{n}\sum_{i=1}^{K} n_i\overline{\mathbf{x}}_i, \qquad \overline{\mathbf{x}}_i = \frac{1}{n}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$$

Ara, es té que la generalització de la matriu de covariàncies dins de grups és

$$\mathrm{W} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)'$$

i la matriu total de covariàncies és

$$\mathrm{T} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}})'$$

La matriu total de covariàncies es pot descomposar en la suma de la matriu de covariàncies dins de grups i de la matriu de covariàncies entre grups, B, és a dir:

$$\mathrm{T} = \mathrm{W} + \mathrm{B}$$

Aleshores, resulta que:

$$\mathrm{B} = \sum_{i=1}^{K} n_i(\overline{\mathbf{x}_i} - \overline{\mathbf{x}})(\overline{\mathbf{x}_i} - \overline{\mathbf{x}})'$$

Per tant, el criteri pot ser escrit com una funció explícita de la matriu de projecció $\mathbf{A}$ de la forma:

$$J(\mathbf{a}) = \mathrm{Tr}\{(\mathbf{A}'\mathrm{B}\mathbf{A})^{-1}(\mathbf{A}'\mathrm{W}\mathbf{A})\}$$

és a dir, el procediment de Fisher determina el vector $\mathbf{a}$ que maximitza el quocient

$$\frac{\sum_{i=1}^{K}\left[\mathbf{a}'\sqrt{n_i}(\overline{\mathbf{x}}_i-\overline{\mathbf{x}})\right]^2}{\sum_{i=1}^{K}\left[\mathbf{a}'\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\overline{\mathbf{x}}_i)\right]^2}=\frac{\mathbf{a}'\mathrm{W}\mathbf{a}}{\mathbf{a}'\mathrm{B}\mathbf{a}}=\lambda$$

Derivant aquesta expressió respecte a $\mathbf{a}$ s'obté

$$(\mathrm{B}-\lambda\mathrm{W})\,\mathbf{a}=0$$

Donat que W té inversa, la igualtat anterior és equivalent a

$$(\mathrm{W}^{-1}\mathrm{B}-\lambda Id)\mathbf{a}=0$$

Aquesta igualtat té solució per valors λ i vectors $\mathbf{a}$ que són respectivament valors i vectors propis de la matriu $\mathrm{W}^{-1}\mathrm{B}$.

La matriu B té rang com a màxim o igual a $(K-1)$ i, per tant, hi ha $(K-1)$ valors propis no nuls. Això demostra que la projecció sobre el subespai dimensional $(K-1)$ travessat pels vectors propis de B no altera el valor de $J(\mathbf{a})$.

Per últim, s'observa que hi ha una relació entre la solució anterior i els resultats obtinguts al capítol 4. És a dir, l'anàlisi canònica de poblacions i el discriminador de Fisher són el mateix. Són métodes per determinar aquell vector unitari $\mathbf{v}$ que separa millor els grups que formen el vectors $\mathbf{x}_i$, és a dir, es tracta de trobar aquella direcció al llarg de la qual els grups quedin més separats.

6 Discriminació logística

6.1 Anàlisi discriminant logístic

(McCullagh i Nelder, 1989). El model de regressió logística permet estimar la probabilitat d'un succés que depèn dels valors de certes covariables.

Es suposa que un succés d'interés A pot presentar-se o no en cada un dels individus d'una certa població. Considerem una variable binària y que pren els valors:

$$y = 1 \text{ si } A \text{ es presenta, } y = 0 \text{ si } A \text{ no es presenta}$$

Si la probabilitat d'A no depèn d'altres variables, indicant $P(A) = p$, la versemblança d'una única observació y és

$$L = p^y(1-p)^{1-y},$$

doncs, $L = p$ si $y = 1$, $L = 1 - p$ si $y = 0$.

6.1.1 Model de regresió logística

Es suposa ara que la probabilitat p depèn dels valors de certes variables $X_1, ..., X_p$. És a dir, si $\mathrm{x} = (x_1, ..., x_p)'$ són les observacions d'un cert individu ω sobre les variables, llavors la probabilitat per esdevenir A donat x és $p(y = 1|\ \mathrm{x}) = p(\mathrm{x})$. La probabilitat contrària de que no succeeixi A donat x serà $p(y = 0|\ \mathrm{x}) = 1 - p(\mathrm{x})$.

Es suposa un model lineal per la transformació logística de la probablitat

$$\ln\left[\frac{p(\mathrm{x})}{1-p(\mathrm{x})}\right] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = \beta_0 + \beta'\mathrm{x} \tag{6.1}$$

on $\beta = (\beta_1, ..., \beta_p)'$ són els paràmetres de regressió. El model (6.1) equival a suposar les següents probabilitats per A i el seu contrari, ambdues en funció de x

$$p(\mathrm{x}) = \frac{e^{\beta_0+\beta'\mathrm{x}}}{1+e^{\beta_0+\beta'\mathrm{x}}},\ 1-p(\mathrm{x}) = \frac{1}{1+e^{\beta_0+\beta'\mathrm{x}}}$$

Suposant que y és una variable resposta quantitativa i que e és un error amb mitjana 0 i variància σ^2, el model de regressió lineal és $y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + e$.

Si es considera que els paràmetres són coneguts o estimats i sigui ω un individu, la regla de discriminació logísitica només decideix que ω posseeix la característica A si $p(\mathrm{x}) > 0.5$, y no la posseeix si $p(\mathrm{x}) \leqslant 0.5$.

S'introdueix la funció discriminant

$$L_g(\mathrm{x}) = \ln\left[\frac{p(\mathrm{x})}{1-p(\mathrm{x})}\right],$$

llavors, la regla de decisió logística queda així

$$\text{Si } L_g(\mathrm{x}) > 0, \text{ aleshores } y = 1; \text{ si } L_g(\mathrm{x}) \leqslant 0, \text{ llavors } y = 0.$$

A la Figura 4 tenim un possible exemple corresponent al model de regressió logística.

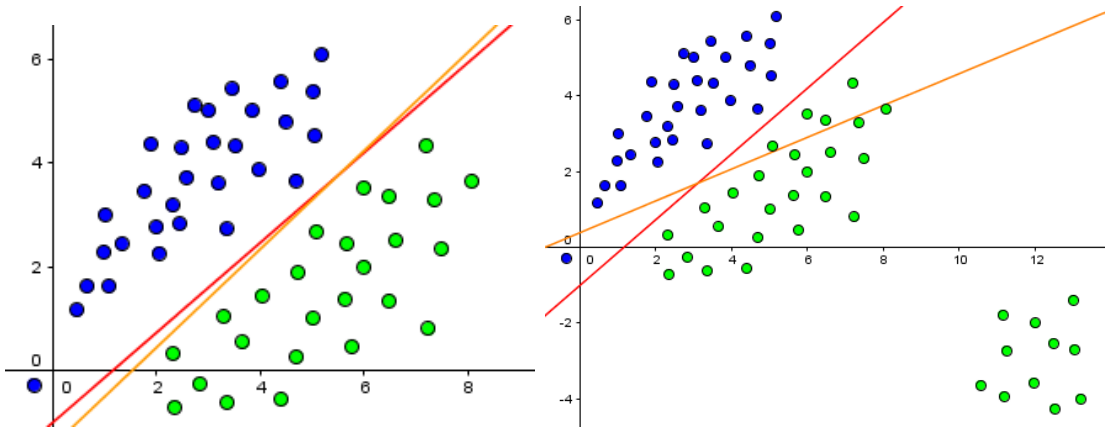


Figura 4: El dibuix de l'esquerra mostra les dades de dos classes (blau i verd) juntament amb la frontera de decisió trobada per mínims quadrats (recta taronja) i també pel model de regressió logística (recta vermella). El dibuix de la dreta mostra els resultats corresponents obtinguts a l'afegir punts de dades addicionals a la part inferior dreta del diagrama, on es veu que els mínims quadrats és molt sensible als valors atípics, a diferència de la regressió logística.

6.1.2 Distribució asimptòtica i test de Wald

S'indica per $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, .., \widehat{\beta}_p)'$ l'estimació dels paràmetres. Aplicant la teoria asimptòtica dels estimadors màxim versemblants, la matriu d'informació de Fisher és $I_\beta = \mathrm{X}'\mathrm{V}\mathrm{X}$, sent

$$\mathrm{V} = \begin{bmatrix} p(\mathrm{x}_1)(1-p(\mathrm{x}_1)) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & p(\mathrm{x}_n)(1-p(\mathrm{x}_n)) \end{bmatrix}$$

La distribució asimptòtica de $\widehat{\beta}$ és normal multivariant $N_{p+1}(\beta, \mathrm{I}_\beta^{-1})$. En particular, la distribució asimptòtica del paràmetre $\widehat{\beta}_i$ és normal $N(\beta_i, \mathrm{var}(\widehat{\beta}_i))$, on $\mathrm{var}(\widehat{\beta}_i)$ és el corresponent element diagonal de la matriu inversa I_β^{-1}.

El test de Wald per la significació de β_i, utilitza l'estadístic

$$z = \widehat{\beta}_i / \sqrt{\mathrm{var}(\widehat{\beta}_i)}$$

amb distribució asimptòtica $N(0,1)$, o bé z^2 amb distribució khi-quadrat amb un grau de llibertat.

Si es desitja estudiar la significació de tots els paràmetres de regressió, el test de Wald calcula $\omega = \widehat{\beta}' \mathrm{I}_\beta \widehat{\beta}$, amb distribució asimptòtica khi-quadrat amb $p+1$ graus de llibertat sota la hipòtesi nul·la $\beta = 0$.

6.1.3 Ajust del model

L'ajust del model en regressió logística s'obté estimant els paràmetres per màxima versemblança L del model i utilitzant l'estadístic de desviació:

$$D = -2 \ln L(\text{model de regressió})$$

Es pot interpretar D de la següent manera:

$$D = -2 \ln \frac{L(\text{model de regressió})}{L(\text{model saturat})}$$

on el model saturat és el que posseeix tants paràmetres com observacions:

$$L(\text{model saturat}) = \prod_{i=1}^{n} y_i^{y_i} \times (1 - y_i)^{1-y_i} = 1$$

Es suposa ara que es vol estudiar la significació d'una o diverses covariables. En particular, la significació d'un coeficient de regressió: $H_0 : \beta_i = 0$. Utilitzant la desviació D es calcularà

$$\begin{aligned} G &= D(\text{model sense les variables}) - D(\text{model amb les variables}) = \\ &= -2 \ln \frac{L(\text{model sense les variables})}{L(\text{model amb les variables})} \end{aligned}$$

Si es vol estudiar la significació de k variables, llavors la distribució asimptòtica de G és khi-quadrat amb k graus de llibertat.

6.1.4 Corba ROC

(Curva ROC, 2015, Wikipedia). L'anàlisi ROC (*Receiver Operating Characteristic*) és una metodologia desenvolupada per analitzar un sistema de decisió. Les corbes ROC s'utilitzen, per exemple, en l'àmbit sanitari per avaluar la qualitat d'un procediment diagnòstic. Aquest anàlisi treballa amb les corbes de sensibilitat i d'especificitat. Es suposa que la població consisteix en individus que posseixen un tumor, que pot ser maligne (succés A), o benigne (contrari de A), aleshores:

- Es diu sensibilitat a la corba $Se(t) = P(p(\mathrm{x}) > t|y = 1)$, $0 \leq t \leq 1$, que és la proporció d'individus als quals es detecta tumor maligne.
- Es diu especificitat a la corba $Es(t) = P(p(\mathrm{x}) < t|y = 0)$, $0 \leq t \leq 1$, que és la proporció d'individus als quals es detecta tumor benigne.

La corba ROC resumeix les dues corbes de sensibilitat i especificitat variant la probabilitat de tall. És la corba que resulta de representar els punts $(1 - Es(t), Se(t))$, $0 \leq t \leq 1$, és a dir, 1-Especificitat a l'eix OX, i la sensibilitat a l'eix OY. La corba ROC està per sobre de la diagonal, i com més s'allunya de la diagonal, millor és la discriminació. A la Figura 5 veiem una representació de la corba ROC.

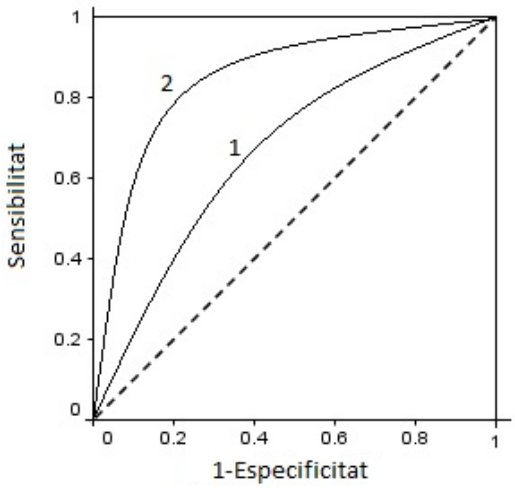


Figura 5: Corba ROC que representa les corbes 1-Especificitat i Sensibilitat. La corba 2 indicaria que les dades tenen millor capacitat de discriminació que la corba 1.

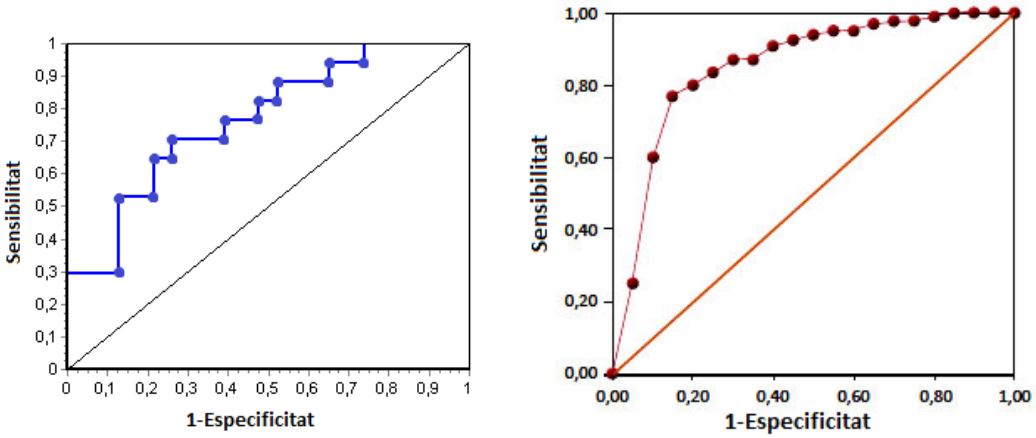
Quan la corba coincideixi amb la diagonal, resultarà que:

$$Se(t) = P(p(\mathrm{x}) > t|y = 1) = 1 - Es(t) = P(p(\mathrm{x}) > t|y = 0)$$

Aleshores, no és possible distingir entre les dues poblacions. És a dir, s'obtindria que la funció discriminant logística $L_g(\mathrm{x}) = \ln[p(\mathrm{x})/(1 - p(\mathrm{x}))]$ té exactament la mateixa distribució tant si $y = 1$ com si $y = 0$.

L'àrea per sota de la corba ROC és sempre major o igual que 0.5. Un valor a partir de 0.8 es considera com que la discriminació és bona. Un valor a partir de 0.9 es consideraria com molt bó. La discriminació seria perfecta si l'àrea val 1.

A continuació, es tenen dos gràfics que mostren les corbes ROC de dos tests diagnòstics hipotètics on cada punt de les corbes corresponen a un possible punt de tall del test diagnòstic corresponent, i ens informa de la seva respectiva Sensibilitat (eix Y) i 1-Especificitat (eix X).



6.2 Anàlisi discriminant basat en distàncies

(Cuadras et al, 1997). Els mètodes descrits anteriorment funcionen bé amb variables quantitatives o quan es coneix la densitat. Però, moltes vegades les variables són binàries, categòriques o mixtes. Suposant que sempre és possible definir una distància entre observacions, és possible donar una versió de l'anàlisi discriminant utilitzant només distàncies.

6.2.1 La funció de proximitat

Sigui Ω una població, X un vector aleatori amb valors en $F \subset R^p$ i densitat $f(x_1, ..., x_p)$, δ una funció de distància entre les observacions de X. Llavors, es

defineix la variabilitat geométrica com:

$$V_\delta(\mathrm{X}) = \frac{1}{2}\int_F \delta^2(\mathrm{x},\mathrm{y}) f(\mathrm{x}) f(\mathrm{y})\, dxdy$$

que és el valor esperat de les distàncies al quadrat entre observacions independents de X.

Sigui ω un individu de Ω i $\mathrm{x} = (x_1, ..., x_p)'$ les observacions de X sobre ω. La funció de proximitat de ω a Ω en relació amb X és:

$$\phi^2_\delta(\mathrm{x}) = E\left[\delta^2(\mathrm{x},\mathrm{X})\right] - V_\delta(\mathrm{X}) = \int_F \delta^2(\mathrm{x},\mathrm{t}) f(\mathrm{t})\, dt - V_\delta(\mathrm{X}) \tag{6.2}$$

Teorema 6.2.1.1. Es suposa que existeix una representació de (F,δ) en un espai L (Euclidià o de Hilbert) $(F,\delta) \longrightarrow L$, amb un producte escalar $< .,. >$ i una norma $\|\mathrm{z}\|^2 =< \mathrm{z},\mathrm{z} >$, tal que $\delta^2(\mathrm{x},\mathrm{y}) = \|\psi(\mathrm{x}) - \psi(\mathrm{y})\|^2$, on $\psi(\mathrm{x}), \psi(\mathrm{y}) \in L$ són les imatges de x, y. Es verifica:

- $V_\delta(\mathrm{X}) = E(\|\psi(\mathrm{X})\|^2) - \|E(\psi(\mathrm{X}))\|^2$
- $\phi^2_\delta(\mathrm{x}) = \|\psi(\mathrm{x}) - E(\psi(\mathrm{X}))\|^2$

A partir d'aquest teorema, es pot afirmar que la variabilitat geomètrica és una variància generalitzada, i que la funció de proximitat mesura la distància d'un individu a la població.

6.2.2 La regla discriminant DB

Siguin Ω_1, Ω_2 dues poblacions, δ una funció distància que és formalment la mateixa en cada població, però pot tenir diferents versions δ_1, δ_2, quan estem en Ω_1, Ω_2, respectivament. Si les poblacions són normals $N_p(\mu_i, \Sigma_i)$, $i = 1, 2$, i es consideren les distàncies de Mahalanobis

$$\delta^2_i(\mathrm{x},\mathrm{y}) = (\mathrm{x}-\mathrm{y})'\Sigma_i^{-1}(\mathrm{x}-\mathrm{y}),\ i = 1, 2$$

l'únic que canvia és la matriu Σ. S'observa que δ depèn del vector aletori X, que tindrà diferent distribució en Ω_1 i Ω_2.

Mitjançant (6.2), es trobaran les funcions de proximitat ϕ^2_1, ϕ^2_2, corresponents a Ω_1, Ω_2. Sigui ω un individu que es vol classificar, amb valors $\mathrm{x} = \mathrm{X}(\omega)$.

La regla de classificació basada en distàncies (DB, distance-based) és:

$$\begin{cases} \text{Si } \phi_1^2(\mathrm{x}) \leqslant \phi_2^2(\mathrm{x}) & \text{assignem } \omega \text{ a } \Omega_1 \\ \text{En cas contrari} & \text{assignem } \omega \text{ a } \Omega_2 \end{cases}$$

A partir del teorema 6.2.1.1, es compleix

$$\phi_i^2(\mathrm{x}) = \|\psi(\mathrm{x}) - E_{\Omega_i}(\psi(\mathrm{X}))\|^2,\ i = 1, 2$$

i, per tant, la regla DB assigna ω a la població més próxima.

6.2.3 La regla DB comparada amb altres

Els discriminadors lineal i quadràtic són casos particulars de la regla DB.

1. Si les poblacions són $N_p(\mu_1, \Sigma_1), N_p(\mu_2, \Sigma_2)$ i δ^2 és la distància de Mahalanobis entre observacions $\delta^2(\mathrm{x}, \mathrm{y}) = (\mathrm{x} - \mathrm{y})'\Sigma^{-1}(\mathrm{x} - \mathrm{y})$, aleshores les funcions de proximitat són $\phi_i^2(\mathrm{x}) = (\mathrm{x} - \mu_i)'\Sigma^{-1}(\mathrm{x} - \mu_i)$ i el discriminador lineal és

$$L(\mathrm{x}) = \frac{1}{2}\left[\phi_2^2(\mathrm{x}) - \phi_1^2(\mathrm{x})\right]$$

2. Si les poblacions són $N_p(\mu_1, \Sigma_1), N_p(\mu_2, \Sigma_2)$ i δ_i^2 és la distància de Mahalanobis més una constant

$$\begin{cases} \delta_i^2(\mathrm{x}, \mathrm{y}) = (\mathrm{x} - \mathrm{y})'\Sigma_i^{-1}(\mathrm{x} - \mathrm{y}) + \log|\Sigma_i|/2, & \mathrm{x} \neq \mathrm{y} \\ \delta_i^2(\mathrm{x}, \mathrm{y}) = 0, & \mathrm{x} = \mathrm{y} \end{cases}$$

   aleshores el discriminador quadràtic és

$$Q(\mathrm{x}) = \frac{1}{2}\left[\phi_2^2(\mathrm{x}) - \phi_1^2(\mathrm{x})\right]$$

3. Si δ és la distància ordinària entre observacions, la regla DB equival a utilitzar el discriminador euclidià

$$E(\mathrm{x}) = \left[\mathrm{x} - \tfrac{1}{2}(\mu_1 - \mu_2)\right]'(\mu_1 - \mu_2)$$

6.2.4 La regla DB en el cas de mostres

En el cas pràctic no es disposa de les densitats $f_1(\mathrm{x}), f_2(\mathrm{x})$, sinó de dos mostres de mides n_1, n_2 de les variables $\mathrm{X} = (X_1, ..., X_p)$ en les poblacions Ω_1, Ω_2. Sigui $\Delta_1 = (\delta_{ij}(1))$ la matriu $n_1 \times n_1$ de distàncies entre les mostres de la primera població, i $\Delta_2 = (\delta_{ij}(2))$ la matriu $n_2 \times n_2$ de distàncies entre les mostres de la segona població. S'indiquen les representacions euclidianes de les mostres com:

$$\begin{cases} \mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_{n_1} & \text{mostra de } \Omega_1 \\ \mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_{n_1} & \text{mostra de } \Omega_2 \end{cases}$$

és a dir, $\delta_{ij}(1) = \delta_E(\mathrm{x}_i, \mathrm{x}_j), \delta_{ij}(2) = \delta_E(\mathrm{y}_i, \mathrm{y}_j)$.

Les estimacions de les variables geomètriques són:

$$\widehat{V}_1 = \frac{1}{2n_1^2} \sum_{i,j=1}^{n_1} \delta_{ij}^2(1), \qquad \widehat{V}_2 = \frac{1}{2n_2^2} \sum_{i,j=1}^{n_2} \delta_{ij}^2(2)$$

Sigui ω un individu, $\delta_i(1)$, $i = 1, ..., n_1$, les distàncies als n_1 individus de Ω_1 i $\delta_i(2)$, $i = 1, ..., n_2$, les distàncies als n_2 individus de Ω_2. Si x són les coordenades de ω quan es suposa que és de Ω_1, i anàlogament y, les estimacions de les funcions de proximitat són

$$\widehat{\phi}_1^2(\mathrm{x}) = \frac{1}{n_1} \sum_{i=1}^{n_1} \delta_i^2(1) - \widehat{V}_1, \qquad \widehat{\phi}_2^2(\mathrm{y}) = \frac{1}{n_2} \sum_{i=1}^{n_2} \delta_i^2(2) - \widehat{V}_2$$

La regla DB en el cas de mostres és

$$\begin{cases} \text{Si } \widehat{\phi}_1^2(\mathrm{x}) \leq \widehat{\phi}_2^2(\mathrm{y}) & \text{s'assigna } \omega \text{ a } \Omega_1 \\ \text{En cas contrari} & \text{s'assigna } \omega \text{ a } \Omega_2 \end{cases}$$

7 Exemples d'aplicacions en R

En aquest apartat s'aplica la teoria explicada en els capítols anteriors a diferents exemples amb dades reals. S'utilitza el programa R per implementar les funcions necessàries a l'hora de fer l'anàlisi discriminant, on el codi d'aquests es troba a l'annex d'aquest treball. El capítol consta de tres exemples, a dos d'ells es fa la classificació en dues poblacions i en el tercer es té una discriminació en el cas de més de dues poblacions.

Per poder fer l'anàlisi i obtenir els resultats d'aquest, es necessita tenir un conjunt comú d'entrenament i de test, per aquest motiu es divideixen les dades dels exemples que es mostren a continuació en dues parts:

- El 60% de les dades totals s'assignen a l'entrenament dels models seleccionats, on es reajusten els paràmetres necessaris.
- El 40% de les dades totals restants s'utilitzen per al testeig i validació de la capacitat de predicció dels models.

Per mesurar la qualitat dels resultats de les classificacions s'utilitza la matriu de confusió. Cada columna d'aquesta matriu representa el nombre de prediccions de cada classe, mentre que cada fila representa cada categoria en la classe real. Amb això es podrà veure si el sistema està confonent dues classes.

7.1 Classificació en dues classes

7.1.1 Package *ElemStatLearn*: SAheart

Les dades SAheart que s'utilitzen en aquest apartat es troben al package *ElemStatLearn* del programa R. És una mostra dels homes en una regió d'alt risc de malalties cardíaques de Western Cape a Sud Àfrica. Les dades s'han pres d'un conjunt de dades més gran[1] que conté 462 observacions on es mesuren les 10 variables següents:

[1] Aquest conjunt de dades està descrit a Rousseauw et al, 1983, South African Medical Journal.

Variable	Definició de la variable
sbp	Pressió arterial sistòlica
tobacco	Tabac acumulat en Kg
ldl	Colesterol de les lipoproteïnes de baixa densitat
adiposity	Adipositat (vector numèric)
famhist	Antecedents familiars de malalties del cor amb dos factors: Absent i Present
typea	Comportament tipus-A
obesity	Obesitat (vector numèric)
alcohol	Consum actual d'alcohol
age	Edat d'inici
chd	Reacció malaltia cardíaca coronària

Taula 1: Variables corresponents a les dades SAheart.

Classificació lineal per mínims quadrats

Una vegada feta la divisió de les dades en dues parts (*train* i *test*) s'ajusta el model lineal per mínims quadrats amb les dades d'entrenament utilitzant la funció **lm** de R.

A partir dels resultats que es mostren a la Taula 2, s'observa que les variables significatives[2] del model que millor expliquen que una persona tingui una malaltia cardíaca o no (variable *chd*) són el tabac acumulat de la persona (*tobacco*), la presència de malalties al cor en familiars (*famhist*), el comportament del tipus A (*typea*) i el factor d'obesitat (*obesity*).

El valor del coeficient de determinació[3] R^2 és de 0.2405 que és més proper a 0 que a 1, la qual cosa vol dir que aquest model no és gaire bó per explicar la presència de malalties cardíaques als homes d'aquesta regió.

[2]El p-valor d'aquestes variables és menor a 0.05.

[3]Mesura quina proporció de la variabilitat total de la variable independent *chd* es veu explicada pel model.

Variable	**Coeficient "No"**	**Coeficient "Si"**	***p*-valor**
constant	1.4718787	-0.4718787	4.32e-08
sbp	-0.0014161	0.0014161	0.284909
tobacco	-0.0232002	0.0232002	0.000145
ldl	-0.0198592	0.0198592	0.155454
adiposity	-0.0092048	0.0092048	0.141987
famhistPresent	-0.1298728	0.1298728	0.018890
typea	-0.0106754	0.0106754	8.50e-05
obesity	0.0231084	-0.0231084	0.016887
alcohol	0.0006646	-0.0006646	0.575449
age	-0.0047264	0.0047264	0.060184
$\mathbf{R}^2$	0.2405	$\mathbf{R}^2$ **ajustat**	0.215

Taula 2: Resultats del model obtinguts per mínims quadrats.

S'analitza ara la matriu de confusió (Taula 3) resultant d'aquest model per mesurar la qualitat dels resultats obtinguts. Com es pot veure el percentatge d'observacions ben classificades ha sigut d'un $67\% = ((97 + 27)/184) \cdot 100$, i l'error de classificació és d'un $33\% = ((22 + 38)/184) \cdot 100$. El valor de l'error de classificació és baix, encara que amb altres discriminadors es pot baixar més aquest valor.

Predicció / Reals	No	Si	Total
No	97	22	119
Si	38	27	65
Total	135	49	184

Taula 3: Matriu de confusió.

Discriminador lineal de Fisher

El següent discriminador que s'utilitza és el discriminador lineal de Fisher a partir de la funció **lda** del package *MASS* de R. Els coeficients que s'obtenen al realitzar aquesta funció es mostren a la Taula 4.

Variable	**Coeficient**
sbp	0.009343471
tobacco	0.071723532
ldl	0.194603181
adiposity	-0.016381495
famhist	0.890616833
typea	0.033528875
obesity	-0.045344312
alcohol	-0.009916705
age	0.039003451

Taula 4: Coeficients del discriminador lineal de Fisher.

Per mesurar la qualitat dels resultats obtinguts s'analitza la matriu de confusió resultant (Taula 5). El percentatge d'observacions ben classificades és $70\% = ((98+31)/184) \cdot 100$, i l'error de classificació és d'un $30\% = ((19+36)/184) \cdot 100$. El valor de l'error de classificació ha baixat comparat amb l'obtingut per mínims quadrats. També, el percentatge d'observacions ben classificades ha augmentat, aleshores, es pot dir que amb aquest model s'han classificat més observacions correctament, en concret, 5 observacions més.

Reals \ Predicció	No	Si	Total
No	98	19	117
Si	36	31	67
Total	134	50	184

Taula 5: Matriu de confusió.

Discriminador quadràtic

Utilitzant el discriminador quadràtic amb la funció **qda** del package *MASS* de R s'obté una nova matriu de confusió (Taula 6). El percentatge d'observacions ben classificades és de $71\% = ((100 + 30)/184) \cdot 100$, i l'error de classificació és

d'un 29% = $((19 + 35)/184) \cdot 100$. El valor de l'error de classificació ha baixat si es compara amb l'obtingut per mínims quadrats i per Fisher. El percentatge d'observacions ben classificades ha augmentat, aleshores, es pot dir que amb aquest model s'han classificat més observacions correctament, en concret, 1 observació més que amb Fisher.

Reals \ Predicció	No	Si	Total
No	100	19	119
Si	35	30	65
Total	135	49	184

Taula 6: Matriu de confusió.

Regressió logística

A continuació, s'usa la regressió logística amb les dades SAheart i resulten les estimacions de les variables amb els corresponents p-valors que mesuren la significació dels coeficients (Taula 7). S'observa que les variables que resulten significatives són el tabac acumulat (*tobacco*), el colesterol (*ldl*), la presència de malalties cardíaques en familiars (*famhist*), el comportament tipus A (*typea*) i l'edat d'inici (*age*). Aleshores, s'eliminen les restants variables que no són rellevants i s'ajusta novament el model per regressió logística obtenint els resultats de la Taula 8.

Per últim, s'analitza la matriu de classificació (Taula 9) i es veu que el percentatge d'observacions ben classificades és 71% = $((99 + 31)/184) \cdot 100$, i l'error de classificació és d'un 29% = $((18 + 36)/184) \cdot 100$. El valor de l'error de classificació ha baixat si es compara amb l'obtingut pel discriminador lineal de Fisher. El percentatge d'observacions ben classificades ha augmentat, aleshores, es pot dir que amb aquest model s'han classificat més observacions correctament, en concret, 1 observació més que amb Fisher. Si es comparen els resultats amb el discriminador quadràtic, els percentatges resulten iguals.

Variable	**Coeficient**	***p*-valor**
constant	-6.635288	0.000104
sbp	0.009613	0.215686
tobacco	0.073551	0.017822
ldl	0.222067	0.009221
adiposity	-0.013623	0.705192
famhistPresent	0.959274	0.001528
typea	0.044715	0.008097
obesity	-0.054487	0.317948
alcohol	-0.011163	0.082221
age	0.051903	0.001283

Taula 7: Resultats de la regressió logística.

Variable	**Coeficient**	***p*-valor**
constant	-6.81967	7.3e-08
tobacco	0.06248	0.030528
ldl	0.18793	0.012834
famhistPresent	0.92510	0.001739
typea	0.04110	0.012401
age	0.05124	0.000167

Taula 8: Resultats de la regressió logística amb les variables significatives.

Reals \ Predicció	No	Si	Total
No	99	18	117
Si	36	31	67
Total	135	49	184

Taula 9: Matriu de confusió.

7.1.2 Package *ElemStatLearn*: Spam

Aquí s'utilitzen les dades spam que també es troben al package *ElemStatLearn* de R. Es tracta de classificar correus segons siguin spam o missatge. Conté 4601 observacions i 58 variables on la última variable denota si l'email es considera spam (1) o no (0), és a dir, que és correu comercial no sol·licitat pel correu electrònic. La major part de variables indiquen si una paraula o caràcter particular apareix amb freqüència en el correu. Les variables 55-57 són els atributs d'execució que mesuren la longitud de les seqüències de lletres majúscules consecutives.

S'ha aplicat el mateix procediment de classificació que s'ha utilitzat amb les dades SAheart, amb els mateixos discriminadors.

Classificació lineal per mínims quadrats

Amb la classificació per mínims quadrats s'obté un model on el coeficient de determinació és $R^2 = 0.5626$, que vol dir que la variable que determina si el correu és spam o missatge es veu explicada un 56% per les variables explicatives. No és un percentage molt elevat, però es pot dir que explica bastant de la variable dependent.

Per mesurar millor la qualitat d'aquests resultats, es pot analitzar la matriu de classificació exposada a la Taula 10. El percentatge de correus electrònics ben clasifcats és d'un 89%, molt elevat, i el percentatge d'error de classificació és d'un 11%. Per tant, l'error de classificació és bastant baix, però amb els següents discriminadors es pot baixar més aquest percentatge.

Predicció / Reals	Email	Spam	Total
Email	1055	51	1106
Spam	144	590	734
Total	1199	641	1840

Taula 10: Matriu de confusió.

Discriminador lineal de Fisher

La matriu de confusió que resulta de fer la classificació aplicant el discriminador lineal de Fisher és la que es mostra en la Taula 11. Si es calcula el percentatge de

correus electrònics ben classificats s'obté un 90% i el percentatge d'error de classificació és d'un 10%. Comparant aquest resultat amb l'obtingut amb el discriminador anterior, l'error de classificació ha baixat i, en concret, s'han classificat correctament 5 correus més.

Reals \ Predicció	Email	Spam	Total
Email	1045	49	1094
Spam	141	605	746
Total	1186	654	1840

Taula 11: Matriu de confusió.

Regressió logística

La matriu de classificació que s'obté quan es realitza la regressió logística és la que es mostra a la Taula 12. El percentatge d'emails ben classificats ha pujat a un 93% i l'error de classificació ha baixat fins un 7%. Amb aquest procediment s'han classificat més observacions que amb els discriminadors per mínims quadrats i Fisher. En concret, s'han classificat 66 emails més que amb el discriminador lineal de Fisher. Per tant, aquesta classificació és bastant bona amb només un 7% d'error.

Reals \ Predicció	Email	Spam	Total
Email	1042	52	1094
Spam	72	674	746
Total	1114	726	1840

Taula 12: Matriu de confusió.

7.2 Classificació en més de dues classes

7.2.1 Package *car*: Skulls

En aquest exemple s'aplica l'anàlisi canònica de poblacions explicat al capítol 4 d'aquest treball. Recordem que és una tècnica amb l'objectiu de representar diversos

grups d'individus de forma òptima a partir d'uns eixos canònics ortogonals. Això s'aconsegueix de forma que la dispersió entre aquests grups sigui màxima en relació a la dispersió dins de cada grup.

L'exemple tracta de dades sobre cranis d'egipcis masculins de cinc èpoques històriques diferents.[4] Les dades consten de 150 observacions, on es mesuren 30 cranis de cada període de temps, i les 5 variables següents:

Variable	**Definició de la variable**
epoch	És un factor que correspon a l'any aproximat de formació del crani. Trobem cinc anys diferents: 150, -200, -1850, -3300, -4000
mb	És la màxima amplitud del crani
bh	És l'altura basi-bregmàtica del crani
bl	És la longitud basi-alveolar del crani
nh	És l'altura nasal del crani

Taula 13: Variables corresponents a les dades Skulls, on les quatre últimes variables són les mesures biomètriques dels cranis.

En primer lloc, es realitza un MANOVA per contrastar la diferència de mitjanes entre els nivells del factor o poblacions. La hipòtesi nul·la és

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

és a dir, que els cranis no han canviat amb el temps. Els resultats obtinguts per aquesta funció es resumeixen en la següent taula:

	Df	Wilks	aprox F	num Df	den Df	Pr(>F)
epoch	4	0.66359	3.909	16	434.45	7.01×10^{-7}

L'estadístic de Wilks és 0.66359 i el p-valor és 7.01×10^{-7} que és menor a 0.05, aleshores es pot concloure que les mitjanes no són totes iguals i que els cranis han canviat amb el temps. Per tant, justifica l'anàlisi canònica de poblacions.

[4]Les dades "skulls"es poden trobar a la pàgina web següent:
http://lib.stat.cmu.edu/DASL/Datafiles/EgyptianSkulls.html

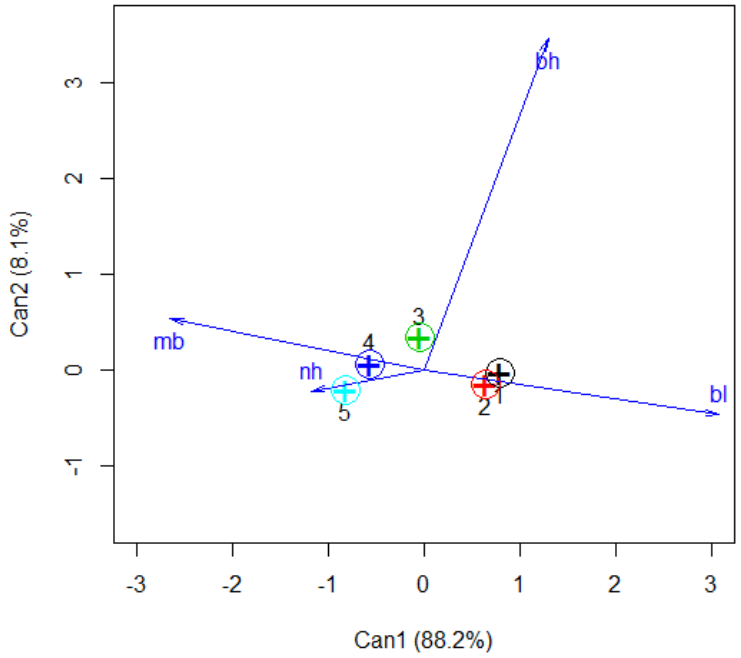


Figura 6: Gràfic on es veu la separació de les poblacions en forma cronològica.

Ara, es realitza l'anàlisi canònica discriminant amb el package *candisc* de R. Per fer això, es necessita com objecte principal un model lineal. En el gràfic de la Figura 6 es representen les variables amb un biplot que explica millor les diferències entre les poblacions que es separen de forma cronològica. Però, aquest no és un autèntic gràfic de l'anàlisi canònica de poblacions, on els cercles han de ser regions de confiança amb les dades sense escalar. Amb els resultats de *skulls.can1* es disposa dels coeficients sense escalar que proporcionen les variables canòniques i, així, es pot calcular les mitjanes de cada població (Figura 7).

Per últim, es determinen els cercles de confiança per un nivell de confiança del 90 %. Per fer això es calculen els radis de cada cercle segons la mida de la mostra de cada població (Figura 7). D'acord amb aquest resultat, s'observa que la proximitat entre les poblacions 1 i 2 és bastant alta, i passa el mateix amb les poblacions 4 i 5, i amb la 3 i 4. En canvi, les poblacions 1 i 5 es troben bastant separades. Tot això indica que com més avancen els anys, més canvia la forma dels cranis. És a dir, els anys transcorreguts entre diferents èpoques determinen el grau de canvi dels cranis. A més anys, més canvi i viceversa.

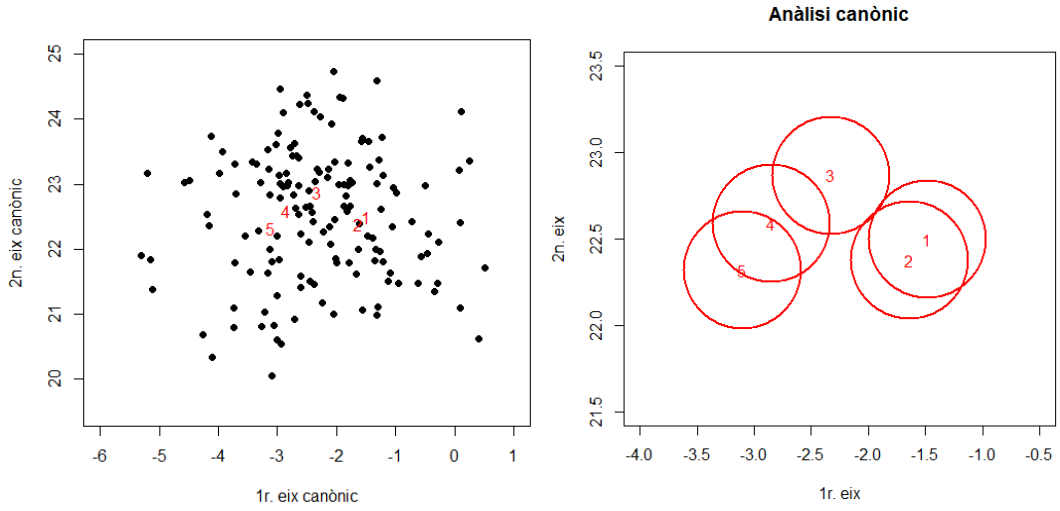


Figura 7: Al gràfic de l'esquerra hi ha la representació de les mitjanes de cada població i a la figura de la dreta, la representació canònica de les 5 poblacions.

D'altra banda, es poden aplicar els discriminadors utilitzats als exemples anteriors per veure com resulta la classificació en aquest cas. A partir de les matrius de confusió resultants, es pot concloure que la classificació feta usant aquests discriminadors és subòptima ja que els percentatges de cranis ben classificats són molt baixos amb percentatges d'error bastant elevats. Per exemple, en el cas de la discriminació per mínims quadrats el percentatge d'error és d'un 75% (Taula 14), en el cas del discriminador lineal de Fisher, d'un 72% (Taula 15) i amb el discriminador quadràtic és d'un 77% (Taula 16).

Predicció / Reals	1	2	3	4	5
1	4	3	4	1	1
2	5	1	3	1	2
3	2	3	3	1	0
4	0	2	3	4	2
5	1	0	7	4	3

Taula 14: Matriu de confusió resultant per mínims quadrats.

Reals \ Predicció	1	2	3	4	5
1	6	2	0	3	1
2	7	3	0	3	0
3	2	2	2	7	2
4	1	0	1	3	4
5	4	1	0	3	3

Taula 15: Matriu de confusió resultant per Fisher.

Reals \ Predicció	1	2	3	4	5
1	3	5	3	2	0
2	2	2	3	1	4
3	2	1	3	2	1
4	1	0	5	3	2
5	0	1	6	5	3

Taula 16: Matriu de confusió utilitzant el discriminador quadràtic.

8 Conclusions

Des del començament d'aquest treball, s'ha fet un estudi de diversos discriminadors basats en l'Estadística multivariant clàssica: el discriminador lineal de Fisher, el discriminador quadràtic o la regla de Bayes. Amb aquest estudi s'ha fet possible tenir una idea bastant completa, tant teòrica com pràctica, d'alguns métodes de classificació estadística.

S'ha observat que hi ha discriminadors que fan una millor classificació que uns altres on els percentatges d'errors són més baixos. Per exemple, amb els resultats obtinguts a la part d'aplicacions en R, es veu que el mètode per mínims quadrats fa una pitjor classificació amb percentgaes d'errors més elevats que amb altres discriminadors. A més, en el cas de la discriminació en més de dos grups és preferible fer la classificació mitjançant l'anàlisi canònica de poblacions. Pel que fa a la regressió logística, s'ha arribat a veure quines són les variables rellevants per fer una bona classificaió de les observacions.

D'altra banda, hi ha moltes altres tècniques de classificació, sobretot no lineals, que solen agrupar-se sota el nom d'Aprenentatge Automàtic i Mineria de Dades, tals com Xarxes Neuronals, Arbres de Classificació, Màquines de Vectors Suport, que queden fora de l'abast del present treball.

Referències

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, **4**, (pp. 179-224). New York: Springer.

[2] Cuadras, C. M.; Fortiana, J.; Oliva, F. (1997). *The Proximity of an Individual to a Population with Applications in Discriminant Analysis*, **14**, (pp. 117-136). Springer International.

[3] Cuadras, C. M. (2014). *Nuevos Métodos de Análisis Multivariante*. Barcelona: CMC Editions.

[4] Curva ROC. (2015, 30 de novembre). Wikipedia, La enciclopedia libre. Data de consulta: 16 de gener de 2016. Disponible a:

`https://es.wikipedia.org/w/index.php?title=Curva_ROC&oldid=87439625`

[5] Duda, R.O.; Hart, P. E.; Stork, D. G. (2001). *Pattern Classification* (2a ed.), **2**, (pp. 20-83). Canadà: Wiley & Sons.

[6] Fisher, R. A. (1936). The use of multiple measuraments in taxonomic problems. *Annals of Eugenics*, **7**, (pp. 179-188).

[7] Mardia, K. V.; Kent, J. T.; Bibby, J. M. (1979). *Multivariate Analysis*, **11**, (pp. 300-332). Londres: Academic Press.

[8] McCullagh, P.; Nelder, J. (1989). *Generalized Linear Models* (2a ed.). Boca Raton: Chapman and Hall/CRC.

[9] Peña, D. (2002). *Análisis de datos multivariantes*. Madrid: McGraw-Hill Interamericana de España, S.A.U.

Annex

A continuació es detallen els codis del programa R que s'han utilitzat per l'elaboració dels càlculs i gràfics exposats al capítol 7.

Dades SAheart

```
#MÍNIMS QUADRATS
library(ElemStatLearn)
require(ElemStatLearn)
data(SAheart)
#Preparem les dades SAheart
y<-SAheart$chd
Y<-cbind((y=="0"),(y=="1"))*1
colnames(Y)<-c("No","Si")
SAheart2<-data.frame(SAheart[!names(SAheart) %in% c("chd")],Y)
#Subdividim el conjunt en dues parts
n<-nrow(SAheart2)
ntrain<-ceiling(0.6*n)
ntest<-n-ntrain
Itrain<-sample(1:n,ntrain,replace=FALSE)
SAheart2.train<-SAheart2[Itrain,]
SAheart2.test<-SAheart2[-Itrain,]
#Ajustem el model lineal
SAheart2.lm1<-lm(cbind(No,Si)~.,data=SAheart2.train)
#Predicció
SAheart2.test.to.predict<-SAheart2.test[,-c(10,11)]
Yhat<-predict(SAheart2.lm1,newdata<-SAheart2.test.to.predict)
#Càlcul de la matriu de confusió
y.test<-y[-Itrain]
yhat<-as.factor(apply(Yhat,1,which.max))
C<-table("True"=y.test,"Predicted"=yhat)
```

```
#FISHER
library(ElemStatLearn)
data(SAheart)
SAheart$famhist<-as.numeric(SAheart$famhist)
SAheart.lda<-SAheart
SAheart.lda$chd<-as.factor(SAheart.lda$chd)
n<-nrow(SAheart)
ntrain<-ceiling(0.60*n)
Itrain<-sample(1:n,ntrain,replace=FALSE)
n<-nrow(SAheart)
ntrain<-ceiling(0.60*n)
Itrain<-sample(1:n,ntrain,replace=FALSE)
SAheart.train<-SAheart[Itrain,]
SAheart.test<-SAheart[-Itrain,]
library(MASS)
SAheart.lda1<-lda(chd~.,data=SAheart.train)
SAheart.pred<-predict(SAheart.lda1,newdata=SAheart.test)
C<-table("True"=SAheart.test$chd,"Predicted"=SAheart.pred$class)
#DISCRIMINADOR QUADRÀTIC
library(ElemStatLearn)
require(ElemStatLearn)
data(SAheart)
SAheart$famhist<-as.numeric(SAheart$famhist)
SAheart.qda<-SAheart
SAheart.qda$chd<-as.factor(SAheart.qda$chd)
n<-nrow(SAheart)
ntrain<-ceiling(0.60*n)
Itrain<-sample(1:n,ntrain,replace=FALSE)
SAheart.train<-SAheart[Itrain,]
SAheart.test<-SAheart[-Itrain,]
library(MASS)
```

```
SAheart.qda1<-qda(chd~.,data=SAheart.train)
SAheart.pred<-predict(SAheart.qda1,newdata=SAheart.test)
C<-table("True"=SAheart.test$chd,"Predicted"=SAheart.pred$class)
#REGRESSIÓ LOGÍSTICA
library(ElemStatLearn)
data(SAheart)
n<-nrow(SAheart)
ntrain<-ceiling(0.60*n)
Itrain<-sample(1:n,ntrain,replace=FALSE)
n<-nrow(SAheart)
ntrain<-ceiling(0.60*n)
Itrain<-sample(1:n,ntrain,replace=FALSE)
SAheart.train<-SAheart[Itrain,]
SAheart.test<-SAheart[-Itrain,]
SAheart.logit1<-glm(chd~.,data=SAheart.train,family=binomial)
summary(SAheart.logit1)
step(SAheart.logit1)
SAheart.logit2<-glm(chd ~ tobacco + ldl + famhist + typea +
age,data=SAheart.train,family=binomial)
SAheart.pred<-predict(SAheart.logit1,newdata=SAheart.test,type="response")
SAheart.pred.crisp<-1*(SAheart.pred>=0.5)
C<-table("True"=SAheart.test$chd,"Predicted"=SAheart.pred.crisp)
```

Dades Spam

```
#MÍNIMS QUADRATS
library(ElemStatLearn)
require(ElemStatLearn)
data(spam)
y<-spam$spam
Y<-cbind((y=="email"),(y=="spam"))*1
```

```
colnames(Y)<-c("Email","Spam")
spam2<-data.frame(spam[!names(spam) %in% c("spam")],Y)
n<-nrow(spam2)
ntrain<-ceiling(0.6*n)
ntest<-n-ntrain
Itrain<-sample(1:n,ntrain,replace=FALSE)
spam2.train<-spam2[Itrain,]
spam2.test<-spam2[-Itrain,]
spam2.lm1<-lm(cbind(Email,Spam)~.,data=spam2.train)
spam2.test.to.predict<-spam2.test[,-c(58,59)]
Yhat<-predict(spam2.lm1,newdata<-spam2.test.to.predict)
y.test<-y[-Itrain]
yhat<-as.factor(apply(Yhat,1,which.max))
C<-table("True"=y.test,"Predicted"=yhat)
#FISHER
library(ElemStatLearn)
data(spam)
n<-nrow(spam)
ntrain<-ceiling(0.60*n)
Itrain<-sample(1:n,ntrain,replace=FALSE)
n<-nrow(spam)
ntrain<-ceiling(0.60*n)
Itrain<-sample(1:n,ntrain,replace=FALSE)
spam.train<-spam[Itrain,]
spam.test<-spam[-Itrain,]
library(MASS)
spam.lda1<-lda(spam~.,data=spam.train)
spam.pred<-predict(spam.lda1,newdata=spam.test)
C<-table("True"=spam.test$spam,"Predicted"=spam.pred$class)
#REGRESSIÓ LOGÍSTICA
library(ElemStatLearn)
```

```
data(spam)
n<-nrow(spam)
ntrain<-ceiling(0.60*n)
Itrain<-sample(1:n,ntrain,replace=FALSE)
n<-nrow(spam)
ntrain<-ceiling(0.60*n)
Itrain<-sample(1:n,ntrain,replace=FALSE)
spam.train<-spam[Itrain,]
spam.test<-spam[-Itrain,]
spam.logit1<-glm(spam~.,data=spam.train,family=binomial)
spam.pred<-predict(spam.logit1,newdata=spam.test,type="response")
spam.pred.crisp<-1*(spam.pred>=0.5)
C<-table("True"=spam.test$spam,"Predicted"=spam.pred.crisp)
```

Dades Skulls

```
library(heplots)
data(Skulls)
skulls<-Skulls
skulls$epoch<-as.factor(as.numeric(Skulls$epoch))
attach(skulls)
skulls.manova <- manova(cbind(mb,bh,bl,nh) ~ epoch)
summary(skulls.manova, test="Wilks") # test="Pillai" or "Hotelling"
or "Roy"
library(candisc)
skulls.mod <- lm(cbind(mb,bh,bl,nh) ~ epoch)
Anova(skulls.mod, test="Wilks") # Manova
skulls.can1 <- candisc(skulls.mod, term="epoch")
plot(skulls.can1, xlim=c(-3,3),ylim=c(-1,3),conf=0.90, type="n")
scores <- as.matrix(skulls[,-1]) %*% skulls.can1$coeffs.raw
plot(scores[,1],scores[,2],xlim=c(-6,2),ylim=c(19,25), xlab=
"1r. eix canònic",ylab="2n. eix canònic",pch=16)
```

```
medias<-aggregate(skulls[,-1],skulls["epoch"],mean)
Medias <- as.matrix(medias[,-1])
scores.medias <- Medias %*% skulls.can1$coeffs.raw
text(scores.medias[,1],scores.medias[,2],1:5,pch=15,col="red")
resumen <- table(epoch)
g <- length(resumen) #nombre de poblacions
p <- dim(skulls[,-1])[2] #nombre de variables
n <- as.vector(resumen) #mida de les mostres de cada població
radios <- function(g,p,n,conf.level=0.95) {
N <- sum(n)
F <- qf(conf.level,p,N-g-p+1)
sqrt(F*(N-g)*p/((N-g-p+1)*n))
}
r <- radios(g,p,n,0.90)
plot.new()
plot.window(xlim=c(-4,-0.5),ylim=c(21.5,23.5))
axis(1)
axis(2)
box()
title(main="Anàlisi canònic", xlab="1r. eix", ylab="2n. eix")
text(scores.medias[,1],scores.medias[,2],labels=levels(epoch),col="red")
symbols(scores.medias[,1],scores.medias[,2],circles=r,inches=FALSE,
add=T,lwd=2,fg="red")
```