

ESQUEMES D'ANALISI MULTIVARIANT AB-1

MATRIU DE DADES

$$X = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Nn} \end{pmatrix} \quad n = \text{no. variable} \\ N = \text{no. individus}$$

(1)

VECTOR DE MITJANES

$$\bar{x} = (\bar{x}_1, \dots, \bar{x}_n)' \quad (2) \quad \bar{x}_j = \frac{1}{N} \sum_{i=1}^N x_{ij} \quad (3)$$

$\bar{x} = \frac{1}{N} X' 1$

(4)

$$1 = (1 \ 1 \ \dots \ 1)'$$

MATRIU DE COVARIANCES

$$S = (s_{ij})$$

$$s_{ij} = \frac{1}{N} \sum_{h=1}^N (x_{hi} - \bar{x}_i)(x_{hj} - \bar{x}_j) = \frac{1}{N} \sum_{h=1}^N x_{hi}x_{hj} - \bar{x}_i\bar{x}_j$$

$$S = \frac{1}{N} X' X - \bar{x} \bar{x}'$$

$S = \frac{1}{N} X' H X$

$$H = I - \frac{1}{N} 11'$$

és la matrui centradora.

MATRÍU DE CORRELACIÓNS

$$r(x_i, x_j) = \frac{s_{ij}}{s_i s_j}$$

$$s_i = \sqrt{s_{ii}} \quad s_j = \sqrt{s_{jj}}$$

$$D_s = \text{diag}(s_1, s_2, \dots, s_n)$$

$$D_s^{-1} = \text{diag}(\frac{1}{s_1}, \dots, \frac{1}{s_n})$$

$$R = D_s^{-1} S D_s^{-1}$$

Relació matricial entre les matrícies de correlacions i de covariances.

TRANSFORMACIÓ LINEAL

$$X(n \times n) \xrightarrow{T} Y(n \times p)$$

$$Y = X T \quad T(n \times p)$$

Vector mitjanes:

Matrícies covariances:

$$\bar{y} = T' \bar{x}$$

$$S_y = T' S T$$

Matrícies correlacions: a) $R_y = R_x \quad \left\{ \begin{array}{l} \text{No canvia} \\ \text{si } n=p \text{ i} \\ \text{T és diagonal} \end{array} \right.$

$$b) R_y = I_p$$

I_p és la matrícia identitat d'ordre p

$\left\{ \begin{array}{l} \text{Si T és la transformació} \\ \text{definida per les components} \\ \text{principals.} \end{array} \right.$

DISTRIBUCIONS MULTIVARIANTS

$\mathbf{X} = (X_1, \dots, X_n)'$ vector aleatori, o sigui, n variables aleatories estudiades conjuntament

Funció de distribució:

$$F(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n)$$

Funció de densitat (cas absolutament contínuu):

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f(t_1, \dots, t_n) dt_1 \dots dt_n$$

Notació vectorial:

$$F(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f(\mathbf{t}) d\mathbf{t}$$

↑ (funció de densitat)

$$\mathbf{x} = (x_1, \dots, x_n) \quad \mathbf{t} = (t_1, \dots, t_n) \quad d\mathbf{t} = dt_1 \dots dt_n$$

DISTRIBUCIONS MARGINALS i CONDICIONADES

$$\mathbf{x} = (x_1, x_2) \quad x_1 \text{ té } k \text{ components}$$

x_2 té $n-k$ components

Densitat condicional de x_1

$$f_1(x_1) = \int_{-\infty}^{+\infty} f(x_1, x_2) dx_2$$

$$\text{Exemple: } n=3 \quad k=1 \quad x_1 = x_1 \quad x_2 = (x_2, x_3)$$

$$f_1(x_1) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x_1, x_2, x_3) dx_2 dx_3$$

DISTRIBUCIONS CONDICIONADES

$$f(x_2 | x_1 = x_1^o) = \frac{f(x_1^o, x_2)}{f_1(x_1^o)}$$

Densitat de x_2 condicionat a que sabem $x_1 = x_1^o$.

Independència estocàstica:

$$f(x_2 | x_1) = f_2(x_2)$$

$$f(x_1, x_2) = f_1(x_1) f_2(x_2)$$

MOMENTS POBLACIONALS

$$g(x) = g(x_1 \dots x_n) \quad \text{funció escalar}$$

$$E\{g(x)\} = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

Esperança matemàtica

$$G(x) = (g_{ij}(x)) \quad \text{funció matricial}$$

$$E\{G(x)\} = (E\{g_{ij}(x)\})$$

VARIABLE COMPOSTA

$$\alpha = (a_1 \dots a_n)'$$

$$Y = \alpha' X = a_1 X_1 + \dots + a_n X_n \quad \left\{ \begin{array}{l} \text{Combinació lineal} \\ \text{de les variables observables} \end{array} \right.$$

Aquestes variables són fonamentals en Anàlisi Multivariant. Segons la tècnica utilitzada reben diferents denominacions: Funcions discriminants, variables canòniques, components de grandària i forma, eixos principals, etc.

MITJANA POBLACIONAL

$$\mu = E(\mathbf{x}) = (E(x_1) \dots E(x_n))' = (\mu_1 \dots \mu_n)'$$

Propietats

$$E(A\mathbf{x} + b) = A\mu + b$$

$$E(\mathbf{x} + \mathbf{y}) = E(\mathbf{x}) + E(\mathbf{y})$$

$$\mu_a = E(y) = a'\mu \quad \text{ni } Y = a_1x_1 + \dots + a_nx_n = a'X \quad \text{és variable composta}$$

MATRÍU DE COVARIANCES

$$\Sigma = E\{(x - \mu)(x - \mu)'\} = (\sigma_{ij})$$

Propietats:

$$\Sigma = E(\mathbf{x}\mathbf{x}') - \mu\mu'$$

$$\text{var}(a'X) = a'\Sigma a \quad \text{ni } Y = a'X \quad \text{és variable composta}$$

MATRÍU DE CORRELACIONS

$$\rho_{ij} = \rho(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}} \sqrt{\sigma_{jj}}}$$

$$R_\rho = D_\sigma^{-1} \sum D_\sigma^{-1}$$

$$D_\sigma = \text{diag}(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{nn}})$$

DISTANCIAS ESTADÍSTICAS

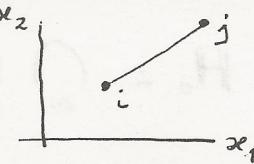
$$(\Omega, \delta) \longrightarrow (\mathbb{V}, d)$$

Espai experimental (dades) \rightarrow Espai geomètric (model de representació)

PROPIETAT DE LA DISTÀNCIA

$$\delta_{ij}^2 = (x_1 - y_1)^2 + \dots + (x_m - y_m)^2$$

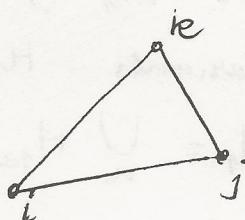
Distància Euclidiana

INTERPRETACIÓ GEOMÈTRICA

P.4

$$\delta_{ij} \leq \delta_{ik} + \delta_{jk}$$

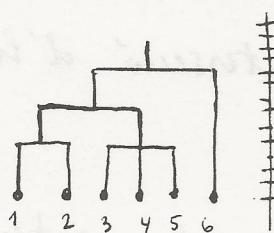
Desigualtat triangular : i, j, k



P.6

$$\delta_{ij} \leq \max \{ \delta_{ik}, \delta_{jk} \}$$

Desigualtat ultramètrica : i, j, k

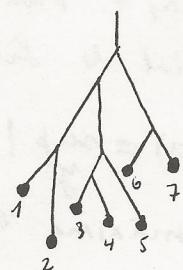


dendrograma
(Classificació fenotípica)

P.7

$$\delta_{ij} + \delta_{kl} \leq \max \{ \delta_{ik} + \delta_{jl}, \delta_{il} + \delta_{jk} \}$$

Desigualtat additiva : i, j, k, l (minimum 4 punts)



arbre additiu

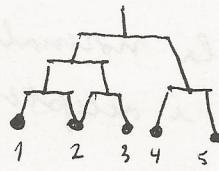
(Classificació filogenètica)

P.11

$$i \leq j \leq k \Rightarrow \max \{ \delta_{ij}, \delta_{jk} \} \leq \delta_{ik}$$

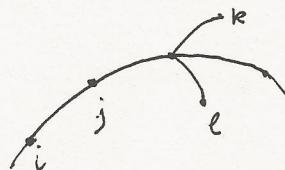
Desigualtat de Robinson : i, j, k

$i \leq j \leq k$ vol dir que els podem ordenar en relació a un paràmetre (exemple: temps)



arbre piramidal

(Seriació)



DISTANCIA EUCLIDIANA

Siguim x_1, x_2, \dots, x_m variables

Coordenades individu i: $x_i = (x_{i1}, \dots, x_{in})'$ (1)

Coordenades individu j: $x_j = (x_{j1}, \dots, x_{jn})'$ (2)

La distància² Euclidiana entre i, j és

$$d_{ij}^2 = (x_i - x_j)' (x_i - x_j) = \sum_{k=1}^n (x_{ik} - x_{jk})^2 \quad (3)$$

És una distància útil com a representació final de les dades.

Com a distància inicial té dos inconvenients:

a) No és invariant per a canvis d'escala

b) Considera a les variables independents (incorrelacionades).

DISTANCIA EUCLIDIANA NORMALITZADA (K. Pearson)

Siguim $\sigma_k^2 = \text{var}(X_k)$ ⁽⁴⁾ les variàncies de les variables

a) La distància normalitzada és defineix

$$k_{ij}^2 = \sum_{k=1}^n \frac{(x_{ik} - x_{jk})^2}{\sigma_k^2} \quad (5)$$

És invariant per a canvi d'escala de mesura de les variables.

b) Si tenim dues poblacions $N_m(\mu_1, \Sigma), N_m(\mu_2, \Sigma)$, la distància normalitzada entre μ_1, μ_2 és

$$k_{ij}^2 = (\mu_1 - \mu_2)' [\text{diag}(\Sigma)]^{-1} (\mu_1 - \mu_2) \quad (6)$$

ALTRES DISTANCIAS VARIABLES CONTINUES

Minkowski $\left(\sum_{k=1}^n (x_{ik} - x_{jk})^q \right)^{1/q}$ $q > 0$ (7)

Canberra

$$\sum_{k=1}^n \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|} \quad (8)$$

DISTANCIA DE MAHALANOBIS

Sigui una població de mitjana μ i matríc de covariàncies Σ (exemple $N_n(\mu, \Sigma)$)

Aquesta distància té tres versions:

a) Entre dos individus i, j

$$M_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)' \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j) \quad (1)$$

b) Entre un individu i , la població de mitjana μ

$$M_i^2 \mu = (\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu)$$

c) Entre dues poblacions de mitjanes μ_i, μ_j

$$M_{\mu_i \mu_j}^2 = (\mu_i - \mu_j)' \Sigma^{-1} (\mu_i - \mu_j)$$

Aquesta distància té interessants propietats

- 1) És invariant per canvis d'escala i qualsevol transformació lineal de les variables $\mathbf{Y} = \mathbf{T}\mathbf{X}$
- 2) Té en compte la redundància entre les variables (afegint variables molt correlacionades no augmenta)
- 3) Si $\mathbf{X} = (x_1, \dots, x_m)$ està incorrelacionat amb $\mathbf{Y} = (y_1, \dots, y_n)$ alleshores $M_{m+n}^2 = M_m^2 + M_n^2$

- 2) Si les poblacions són $N_n(\mu_1, \Sigma), N_m(\mu_2, \Sigma)$ i prenem com estimació de la distància

$$\hat{M}^2(\mu_1, \mu_2) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \hat{\Sigma}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

alleshores, si $\mu_1 = \mu_2$, es verifica

$$F = \frac{N_1 N_2 (N_1 + N_2 - 1)}{n(N_1 + N_2)(N_1 + N_2 - 2)} \hat{M}^2 \sim F_{n, N_1 + N_2 - n + 1}$$

(n = no. variable, N_1, N_2 = grandàries mòtges poblacions 1 i 2).

DISTANCIA ENTRE DISTRIBUCIONS MULTINOMIALS

Suposem n caràcters exclusius A_1, A_2, \dots, A_n amb diferents probabilitats

$$\begin{array}{c} A_1 \ A_2 \ \dots \ A_n \\ \text{Població 1: } p_1 \ p_2 \ \dots \ p_n \\ \text{Població 2: } q_1 \ q_2 \ \dots \ q_n \end{array} \quad \sum_{i=1}^n p_i = \sum_{i=1}^n q_i = 1$$

Aplicant Mahalanobis s'obté la distància de Balakrishnan-Sanghvi

$$d_{12}^2 = 2 \sum_{i=1}^n \frac{(p_i - q_i)^2}{(p_i + q_i)}$$

Aplicant criteris geomètrics s'obté la distància de Bhattacharyya (coneguda en Genètica com de Cavalli-Sforza)

$$\begin{array}{ll} \text{ARC} & \text{CORDA} \\ d_{12}^2 = \arccos \left(\sum_{i=1}^n \sqrt{p_i q_i} \right) & d_{12}^2 = \sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2 \end{array}$$

COM SABER SI UNA DISTANCIA ES EUCLIDIANA

Sigui $\Omega = \{1, 2, \dots, n\}$ un conjunt finit i $s_{ij} = s(i, j)$ una distància. Es diu que la matríg de distàncies $n \times n$ $\Delta = (s_{ij})$ és euclidiana si \exists n punts P_1, P_2, \dots, P_n de \mathbb{R}^m tals que

$$s_{ij} = d(P_i, P_j)$$

on $d(\cdot, \cdot)$ significa distància euclidiana m -dimensional. Amb altres paraules, si podem representar els elements $d(\cdot, \cdot)$ com a punts de \mathbb{R}^m , es a dir, utilitzant coordenades.

TEOREMA. Sigui $A = (a_{ij})$ la matríg $n \times n$, on $a_{ij} = -\frac{1}{2} s_{ij}^2$,

$H = I_m - \frac{1}{n} J_n$ la matríg centradora de dades i calculem $B = HAH$.

Aleshores $\Delta = (s_{ij})$ és una matríg de distàncies euclidianes si i només

si B és definida positiva i la dimensió és $m = \text{rang}(B)$.

A més, si trobem $X(n \times m)$ tal que $B = X'X$ aleshores les coordenades son les files de X .

$$\begin{matrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 2 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & x_{n1} & x_{n2} & \dots & x_{nm} \end{matrix} = X$$

COEFICIENTS DE SIMILARITAT: CAS DE
VARIABLES BINARIES

Propiedades de algunos coeficientes de similaridad para variables binarias.

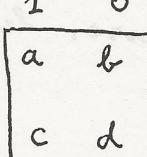
SIMILARIDAD	AUTOR	RANGO	$S \geq 0$	METRICA	EUCLIDEA
$\frac{a}{b+c}$	Kulczynski	$0, \infty$	Si		
$\frac{a}{a+b+c+d}$	Russell y Rao	0,1	Si	Si(Si)	Si
$\frac{a}{a+b+c}$	Jaccard	0,1	Si	Si(Si)	Si
$\frac{a+d}{a+b+c+d}$	Sokal y Michener	0,1	Si	Si(Si)	Si
$\frac{a}{a+2(b+c)}$	Anderberg	0,1	Si	Si(Si)	Si
$\frac{a+d}{a+2(b+c)+d}$	Rogers y Tanimoto	0,1	Si	Si(Si)	Si
$\frac{a}{a+\frac{1}{2}(b+c)}$	Sorensen	0,1	Si	Si (No)	Si
$\frac{a+d}{a+\frac{1}{2}(b+c)+d}$	Sneath y Sokal	0,1	No	Si (No)	No
$\frac{a-(b+c)+d}{a+b+c+d}$	Harman	-1,1	Si	Si (Si)	Si
$\frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c})$	Kulczynski	0,1	No	No (No)	No
$\frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{c+d} + \frac{d}{b+d})$	Anderberg	0,1	No	No (No)	No
$\frac{a}{\sqrt{(a+b)(a+c)}}$	Ochiai	0,1	Si	Si (No)	Si
$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$		0,1	Si	Si (No)	Si
$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	Pearson	-1,1	Si	Si (No)	Si
$\frac{ad-bc}{ad+bc}$	Yule	-1,1	No	No (No)	No

Notas: 1) $S \geq 0$ significa que la matriz de similaridades es (semi)definida positiva.

2) La propiedad métrica se refiere a la distancia $d_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$ y a la distancia $\bar{d}_{ij} = 1 - s_{ij}$ (entre paréntesis).

Ejemplo: Sorensen es métrica para d_{ij} pero no para \bar{d}_{ij} .

3) Ninguna de las distancias \bar{d}_{ij} es euclídea.

i) 

$$\rightarrow s_{ij} = \frac{a+d}{a+b+c+d} \rightarrow (\text{Exemple de similaritat de Sokal i Michener})$$

Similaritat entre els individus i, j .

DOS TEOREMES FONAMENTALS : PRIMER TEOREMA

Importants mètodes multivariants es basen en aquests dos teoremes.

Siguin A i B dues matrius $n \times n$, $A \geq 0$, $B > 0$. Direm que $v_i = (v_{1i}, \dots, v_{ni})'$

és vector propi normalitzat de A respecte de B de valor propi λ_i si

$$\left. \begin{array}{l} Av_i = \lambda_i B v_i \\ v_i' B v_i = 1 \end{array} \right\} \quad i=1, \dots, m \quad m = \text{rang}(A)$$

Indiquem

$$V = \begin{pmatrix} v_{11} & \dots & v_{1m} \\ v_{21} & \dots & v_{2m} \\ \vdots & & \\ v_{n1} & \dots & v_{nm} \end{pmatrix}$$

la matriu $n \times m$ dels vectors propis de A sobre B . Supsem ordenats de manera que $\lambda_1 \geq \dots \geq \lambda_m$

TEOREMA 1. Si A i B poden interpretar-se com dues "matrius de covariances" de n v.a. X_1, X_2, \dots, X_n i definim les variables compostes

$$Y_i = v_{1i} X_1 + \dots + v_{ni} X_n = v_i' X$$

on v_i és vector propi de A respecte B de valor propi λ_i , aleshores

1) Y_1, Y_2, \dots, Y_n són simultàniament ortogonals (incorrelacionades) respecte A i B , és a dir,

$$\text{cov}_A(Y_i, Y_j) = \text{cov}_B(Y_i, Y_j) = 0 \quad i \neq j = 1, \dots, m$$

(Recordem que $\text{cov}_A(Y_i, Y_j) = v_i' A v_j$)

2) Són variables unitàries respecte a B i de variàncies respectivament màximes en relació a A , aquelles ^{variancias} ~~són els~~ valors propis:

$$\text{var}_A(Y_1) = \lambda_1 \geq \text{var}_A(Y_2) = \lambda_2 \geq \dots \geq \text{var}_A(Y_m) = \lambda_m$$

$$\text{var}_B(Y_1) = \dots = \text{var}_B(Y_m) = 1$$

APLICACIONS: Anàlisi Components Principals ($A = \Sigma$, $B = I$)

Anàlisi Canònic de Poblacions (A = "covariància entre grups", $B = \Sigma$)

Anàlisi Factorial

Anàlisi Multivariant de la variància

DOS TEOREMES FONAMENTALS: SEGON TEOREMA

Suposem que k elements (individus, espècies, marques de cotxes, etc.) tenen una coordenades en relació a una v.a. $x_1 x_2 \dots x_m$

$$\begin{matrix} & x_1 & x_2 & \dots & x_n \\ 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 2 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & & & & \\ k & x_{k1} & x_{k2} & \dots & x_{km} \end{matrix} = X$$

Aquestes coordenades poden també ésser mitjanes de k poblacions.

Siguin B una "matrícula de covariàncies" i considerem la distància

$$M^2(i,j) = (x_i - x_j)^T B^{-1} (x_i - x_j) \quad (\text{Mahalanobis})$$

Considerem d variables y_1, \dots, y_d

$$y_i = v_{1i} x_1 + \dots + v_{ni} x_n$$

ortonormals, vol dir: $\text{var}_B(y_i) = 1$, $\text{cov}_B(y_i, y_j) = 0$, $i \neq j$.

Si $V = (v_{ij})$ conté els coeficients de les variables y_i , podem introduir un canvi de coordenades $X \rightarrow Y$

$$Y = X V = \begin{pmatrix} y_{11} & \dots & y_{1d} \\ y_{21} & \dots & y_{2d} \\ \dots & & \dots \\ y_{n1} & \dots & y_{nd} \end{pmatrix}$$

i la distància² entre i, j passa a ser euclídeana d -dimensional

$$M^2(i,j) \rightarrow d^2(i,j) = \sum_{h=1}^d (y_{ih} - y_{jh})^2$$

Quina projecció de dimensió n a dimensió d és la millor?

Siguin $A = \bar{X}^T \bar{X}$ on \bar{X} és la matrícula X centrada.
Si y_1, y_2, \dots, y_k s'obtenen aplicant el Teorema 1, direm que són les variables canòniques.

TEOREMA 2. La dispersió en dimensió reduïda d , mesurada per la suma de distàncies² entre cada parella d'elements

$$\sum_{i,j=1}^k D_d^2(i,j) = \sum_{i,j=1}^k \sum_{h=1}^d (y_{ih} - y_{jh})^2$$

és màxima si prenem les d primeres variables canòniques. Aquesta dispersió màxima és igual a: $2 \sum_{h=1}^k (\lambda_1 + \dots + \lambda_d)$. (Nota: $d \ll n$, usualment $d = 2, 3$).

APLICACIONS: Representació de dades per A. Components Principals

Anàlisi de Coordenades Principals

Anàlisi Canònica de Poblacions

Anàlisi Canònica d'efectes principals d'un disseny factorial

anàlisi discriminant