DENIS ENĂCHESCU

TEHNICI STATISTICE DE DATA MINING

Editura Universităţii din Bucureşti

DENIS ENĂCHESCU

.....

TEHNICI STATISTICE DE DATA MINING

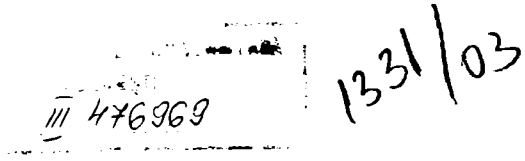
DENIS ENĂCHESCU

TEHNICI STATISTICE DE DATA MINING

Note de curs

EDITURA UNIVERSITĂȚII DIN BUCUREȘTI
2003

Referenți științifici: **prof. dr. Ion VĂDUVA**
prof. dr. Monica DUMITRESCU



Tiparul s-a executat sub c-da nr. 1114/2003 la
Tipografia Editurii Universității din București

Șos. Panduri, 90-92, București – 050663; Telefon/Fax: 410.23.84
E-mail: editura@unibuc.ro
Internet: www.editura.unibuc.ro

Descrierea CIP a Bibliotecii Naționale a României
ENĂCHESCU, DENIS
Tehnici statistice de Data Mining / Denis Enăchescu – București: Editura Universității din București, 2003
Bibliografie
ISBN 973-575-814-8

519.22

INTRODUCERE

Suntem copleşiţi de date - date ştiinţifice, date medicale, date demografice, date financiare, date de marketing. Oamenii nu mai au timp să se uite la aceste date. Atenţia umană a devenit o resursă importantă, astfel încât trebuiesc găsite căi de a analiza datele automat, de a le clasifica automat, de a le sintetiza automat, de a descoperi automat tendinţe în date şi de a caracteriza automat aceste tendinţe. Acest "minerit în date" în vederea găsirii automate de cunoştinţe şi informaţii interesante/noi, este astăzi unul dintre cele mai active şi interesante domenii de cercetare. Cercetătorii din domeniile bazelor de date, statisticii matematice, intelingenţei artificiale şi vizualizării computerizate sunt implicaţi şi contribuie la dezvoltarea acestui domeniu.

Lucrarea de faţă prezintă tehnicile clasice "împrumutate" din statistica matematică de noul domeniu - am numit aici Data Mining; este vorba, mai precis, de tehnici de statistică exploratorie multidimensionale.

Statistica descriptivă permite reprezentarea vie şi asimilabilă a informaţiilor statistice prin simplificare şi schematizare. Statistica descriptivă multidimensională este generalizarea naturală a cazului în care informaţiile privesc mai multe variabile şi/sau dimensiuni.

Trecerea la multidimensional implică însă o schimbare calitativă importantă. Într-adevăr, se spune despre microscop sau despre aparatul radiografic că nu sunt numai intrumente de descriere ci şi instrumente de observaţie, de explorare şi de cercetare. Prin metodele de statistică exploratorie multidimensională realitatea nu este doar simplificată pentru că este complexă, ci şi explorată pentru că este ascunsă. Munca de pregătire şi de codificare a datelor, regulile de interpretare şi validare furnizate de tehnicile furnizate în cazul multidimensional, nu au simplitatea întâlnită în statistica descriptivă elementară. Nu este vorba doar de a prezenta ci şi de a analiza, a descoperi, uneori de a verifica şi dovedi, eventual de a testa anumite ipoteze.

Această lucrare conţine prelegerile ţinute studenţilor de la specializările INFORMATICĂ, şi MATEMATICĂ APLICATĂ ale Facultăţii de Matematică şi Informatică a Universităţii din Bucureşti începând cu anul universitar 1995/1996, în cadrul unor cursuri opţionale organizate anual sau semestrial în funcţie de solicitări.

Numărul metodelor ce permit descrierea şi explorarea tabelelor rectangulare de date statistice (tabele de măsurători-observaţii, tabele de contingenţă, tabele de prezenţă-absenţă, sau tabele de incidenţă) este destul de mare. Metodele reţinute pentru a fi prezentate au fost alese în funcţie de posibilităţile pe care le au de a manipula tabele voluminoase, în funcţie de transpareţa funcţionării lor, în funcţie de calitatea inserţiei în evantaiul metodelor ce sunt în mod real aplicabile şi aplicate.

Două mari familii de metode răspund la aceste exigenţe:

- [capitolul 1]: *metodele factoriale* bazate pe căutarea axelor principale (analiza în componente principale şi analiza corespondenţelor simple şi multiple sunt metodele factoriale cele mai utilizate) care, produc în principal, vizualizări grafice plane sau spaţiale ale obiectelor cercetate;
- [capitolul 2]: *metodele de clasificare* care produc agregări în clase de obiecte sau în familii de clase ierarhizate, obţinute în urma unor calcule algoritmice. Obiectele cercetate sunt grupate, pornind de la vectorii care le descriu, în maiera cea mai puţin arbitrară.

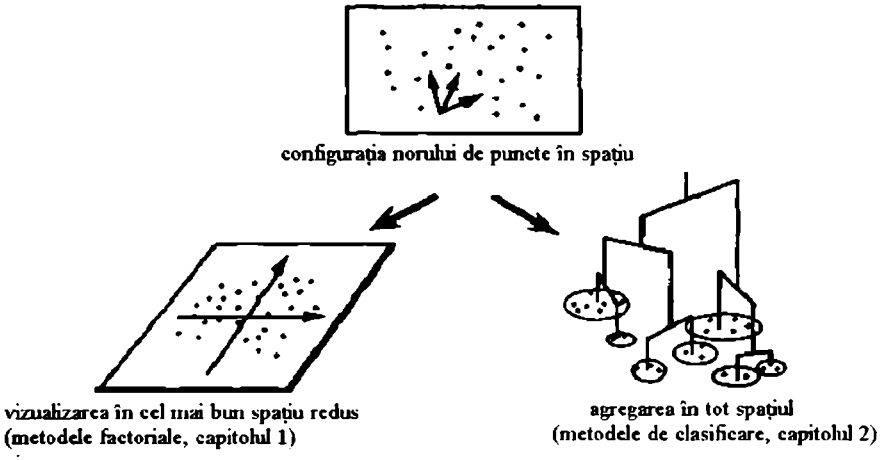


Figura 1 Cele două mari familii de metode ale statisticii exploratorii multidimensionale

Punctele de vedere furnizate de cele două tipuri de metode sunt în esenţă complementare. Vom insista asupra acestei complementarităţi care se manifestă de altfel la mai multe niveluri, fie că este vorba de posibilitatea de a înţelege structuri diverse, fie că este vorba de a ajuta lectura rezultatelor obţinute.

- [capitolul 3]: *metodele explicative uzuale* vor lămuri pe utilizator asupra vocaţiei specifice fiecărei metode (este vorba de analiza discriminantă şi de metodele de segmentare) cât şi asupra legăturilor cu metodele statisticii exploratorii (descrise în primele două capitole). Acest evantai de tehnici acoperă o parte importantă a aplicaţiilor potenţiale ale statisticii.

Nu există totuşi o metodologie generală de articulare în practică a metodelor exploratorii de bază (metodele prezentate în capitolele 1 şi 2) şi metodele explicative uzuale (prezentate în capitolul 3). Fiecare aplicaţie implică, în funcţie de domeniu şi problemă, o muncă originală de codificare şi selecţie a metodelor particulare aplicate. În plus, trebuie să fim conştienţi de faptul că metodele prezentate sunt eficiente în special în cazul datelor nestructurate sau amorfe (în care informaţia a priori asupra acestora este săracă).

Trebuie menționat faptul că există o literatură bogată privind tematica acestei lucrări. Bibliografia ataşată constă numai dintr-o selecție a lucrărilor pe care autorul le-a consultat şi care pot fi găsite cu uşurință în bibioteci.

Metodele prezentate au un pronunțat carcater matematic-aplicativ. Studenți, practicieni şi cercetători din toate disciplinele ce trebuiesc să analizeze şi să prelucreze volume mari de date multidimensionale, vor găsi în aceasta lucrare metodele de bază necesare.

Intenția autorului este de a continua dezvoltarea materialului prezentat aici într-o ediție următoare; în consecință observațiile şi sugestiile sunt bine venite.

Autorul

CUPRINS

1. METODE DE ANALIZĂ FACTORIALĂ

Metodele factoriale îşi propun să furnizeze reprezentări sintetice ale unor mulţimi mari de valori numerice, în general sub forma unor vizualizări grafice. Pentru aceasta, se urmăreşte reducerea dimensiunilor tabelului de date prin reprezentarea asociaţiilor între indivizi şi variabile în spaţii de dimensiuni mici. Distanţele între liniile sau între coloanele unui tabel dreptunghiular de valori numerice pot fi întotdeauna calculate dar nu este posibilă vizualizarea imediată a acestora (reprezentările geometrice asociate implicând, în general, spaţii de dimensiuni superioare lui trei); este necesar să procedăm la transformări şi aproximări pentru a obţine o reprezentare plană.

Metodele factoriale vor căuta, în consecinţă, subspaţii de dimensiuni mici (unu, doi sau trei) care aproximează cel mai bine norul de puncte-indivizi sau cel de puncte-variabile astfel încât vecinităţile măsurate în aceste spaţii să reflecte cât mai exact proximităţile reale. Se obţine astfel un spaţiu de reprezentare, spaţiul factorial. Geometria norilor de puncte şi calculul proximităţilor sau a distanţelor care decurg de aici diferă în funcţie de natura liniilor şi coloanelor tabelului analizat.

Coloanele tabelelor dreptunghiulare de date pot fi variabile continue sau variabile nominale sau categorii în cazul tabelelor de contingenţă. Liniile pot fi indivizi sau categorii. Natura informaţiilor, codificarea, specificitatea domeniului de aplicaţie vor introduce variante în cadrul metodei factoriale.

În cele ce urmeaza vor fi prezentate trei tehnici fundamentale:

- *analiza în componente principale* (secţiunea 1.2) se apalică tabelelor de tip "variabile-indivizi" unde coloanele sunt variabile numerice continue şi liniile sunt indivizi, observaţii, obiecte, etc. Proximităţile între variabile se interpretează în termeni de corelaţii; proximităţile între indivizi se interpretează în termeni de similitudini globale ale valorilor observate.
- *analiza corespondenţelor simple* (secţiunea 1.3) se aplică tabelelor de contingenţă, adică tabelelor ce conţin numărul indivizilor care posedă concomitent două modalităţi a două variante nominale. Aceste tabele au particularitatea că atât liniile cât şi coloanele lor joacă un rol identic în analiza corespondenţelor simple. Analiza furnizează reprezentări ale asociatiilor între liniile şi coloanele tabelelor bazate pe o distanţă între profile (care sunt vectori de frecvenţe condiţionate) cunoscută sub numele de distanţa χ^2.
- *analiza corespondenţelor multiple* (secţiunea 1.4) este o extindere a domeniului aplicaţiilor analizei corespondenţelor simple având totuşi proceduri de calcul şi reguli de interpretare specifice. Ea face obiectul unei menţiuni particulare datorită numărului mare de aplicaţii la care se pretează. Analiza

corespondențelor multiple este în mod deosebit adaptată la descrierea tabelelor mari de variabile nominale cum sunt fișierele de anchete socio-economice sau cele medicale. Liniile acestor tabele sunt, în general, indivizi sau observații (pot exista câteva mii); coloanele sunt modalități ale variabilelor nominale cel mai adesea modalitățile răspunsurilor la întrebări.

Oricare ar fi tipul tabelului de date, toate tehnicile factoriale au un nucleu comun prezentat în secțiunea 1.1 sub forma unor *preliminarii matematice*.

1.1 PRELIMINARII MATEMATICE

1.1.1 CONCEPTE METRICE ÎNTR-UN SPAȚIU EUCLIDIAN

Fie X mulțime oarecare, $X \neq \varnothing$.

Definiția 1.1-1 O *metrică* pe mulțimea X este o aplicație $d : X \times X \to \mathbb{R}$, care satisface următoarele axiome :

a) $d(x,y) = d(y,x)$, $(\forall)\, x, y \in X$;

b) $d(x,y) \geq 0$; $(\forall)\, x, y \in X$;

c) $d(x,y) = 0 \Leftrightarrow x = y$

d) $d(x,y) \leq d(x,z) + d(z,y)$, $(\forall)\, x, y, z \in X$ (inegalitatea triunghiului).

Definiția 1.1-2 Dacă $d : X \times X \to \mathbb{R}$ aplicație care satisface axiomele (a)-(c) și în plus este satisfăcută axioma

d') $d(x,y) \leq \max\left(d(x,z), d(z,y)\right)$, $(\forall)\, x, y, z \in X$

atunci d se numește *ultrametrică* pe X.

Observație. d ultrametrică implică d metrică.

Definiția 1.1-3 Un *spațiu (ultra)metric* este o pereche (X,d), unde X este o mulțime nevidă și d este o ultra(metrică) pe X.

Definiția 1.1-4 O *pseudometrică* pe X este o aplicație $d : X \times X \to \mathbb{R}$ care satisface :

a) $d(x,y) = d(y,x)$, $(\forall)\, x, y \in X$;

b) $d(x,y) \geq 0$ $(\forall)\, x, y \in X$;

c) $d(x,x) = 0$ $(\forall)\, x \in X$.

O mulțime nevidă înzestrată cu o psudometrică se numește *spațiu pseudometric*.

Observație. Într-o altă terminologie, legată de problema de clasificare, o pseudoemtrică se numește și *coeficient de disimilaritate*.

Definiția 1.1-5 O pseudometrică care satisface în plus axioma

(d) $d(x,y)=0 \Rightarrow x=y$

se numește *semimetrică*.

Observație. Pentru oricare din spațiile considerate mai sus $d(x,y)$ se va numi distanța dintre x și y în spațiul (X,d).

Fie K un corp comutativ, $X \neq \varnothing$ înzestrată cu o operație internă (adunare)

$$X \times X \ni (x,y) \to x+y \in X,$$

și cu operația de înmulțire cu scalari,

$$K \times X \ni (a,x) \to ax \in X.$$

Definiția 1.1-6 X se numește *spațiu vectorial (spațiu liniar)* peste K dacă:

(1) $(x+y)+z=x+(y+z)$, $(\forall)\, x,y,z \in X$;

(2) $\exists\, 0 \in X$ a.î. $(\forall)\, x \in X$, $x+0=x$;

(3) $(\forall)\, x \in X$ $\exists(-x) \in X$ a.î. $x+(-x)=0$;

(4) $x+y=y+x$, $(\forall)\, x,y \in X$;

(5) $1x=x$, $(\forall)\, x \in X$;

(6) $a(bx)=(ab)x$, $(\forall)\, a,b \in K$ și $\forall\, x \in X$;

(7) $(a+b)x=ax+bx$, $(\forall)\, a,b \in K$ și $\forall\, x \in X$;

(8) $a(x+y)=ax+ay$, $(\forall)\, a \in K$ și $x,y \in X$.

Fie X un spațiu vectorial peste $\mathbb{R}$ sau $\mathbb{C}$.

Definiția 1.1-7 Se numește *produs scalar* pe X o funcție de două variabile $(\cdot,\cdot): X \times X \to K$ pentru care sunt satisfăcute următoarele axiome:

(1) $(x,y)=\overline{(y,x)}$, $(\forall)\, x,y \in X$;

(2) $(ax,y)=a(x,y)$, $(\forall)\, x,y \in X$ și $a \in K$;

(3) $(x+y,z)=(x,z)+(y,z)$, $(\forall)\, x,y,z \in X$;

(4) $(x,x) \geq 0$, $(\forall)\, x \in X$;

(5) $(x,x)=0 \Leftrightarrow x=0$.

Observație. Dacă X este spațiu vectorial peste $\mathbb{R} \Rightarrow X$ se numește spațiu vectorial real.

Definiția 1.1-8 Se numește *spațiu euclidian* – un spațiu vectorial finit dimensional. Spațiul $\mathbb{R}^n$ înzestrat cu produsul scalar

$$(x,y)=\sum_{i=1}^{n}x_iy_i$$

este un spațiu euclidian.

Dacă un element din $\mathbb{R}^n$ se scrie ca un vector coloană $\mathbf{x}=\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix}$ atunci produsul scalar se mai scrie $(\mathbf{x},\mathbf{y})=\mathbf{x}'\mathbf{y}$.

Observație. În $\mathbb{R}$ se mai pot defini şi alte produse scalare.

Doi vectori $\mathbf{x},\mathbf{y}$ se numesc *ortogonali (perpendiculari)* dacă $(\mathbf{x},\mathbf{y})=0$.

Definiția 1.1-9 O *normă* pe un spațiu vectorial X definit peste corpul K este o funcțională $\|\cdot\|: X\to\mathbb{R}$, pentru care sunt verificate axiomele :

(1) $\|x\|\geq 0$, $(\forall)\,x\in X$ (pozitivă) ;

(2) $\|x\|=0\Leftrightarrow x=0$ (pozitiv definită) ;

(3) $\|ax\|=|a|\cdot\|x\|$, $(\forall)\,a\in K$ şi $x\in X$ (absolut omogenă) ;

(4) $\|x+y\|\leq\|x\|+\|y\|$, $(\forall)\,x,y\in X$ (subadditivă).

Definiția 1.1-10 Un spațiu vectorial înzestrat cu o normă se numeşte *spațiu normat.*

Observație. Orice spațiu euclidian este normat în raport cu norma indusă de produsul scalar $\|\mathbf{x}\|=(\mathbf{x},\mathbf{x})^{1/2}$.

La rândul său norma induce o distanță

$$d(\mathbf{x},\mathbf{y})=\|\mathbf{x}-\mathbf{y}\|=(\mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y})^{1/2}.$$

Rezultă că orice spațiu euclidian poate fi înzestrat cu o metrică generată de produsul scalar.

1.1.2 OPERATORI LINIARI

Fie X un spațiu vectorial de dimensiune n. Considerăm o bază $\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n$ în X şi fie $U: X\to X$, un operator liniar.

$U\mathbf{e}_i$ este un vector din $X\Rightarrow$ se poate scrie ca o combinație liniară de vectorii bazei, adică :

$$U\mathbf{e}_i=\sum_{j=1}^{n}A_{ij}\mathbf{e}_j,\quad i=\overline{1,n}.$$

Coeficienții A_{ij} determină o matrice **A** de dimensiune (n,n) numită *matricea operatorului* U *în baza* $\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n$.

Fie X spațiu euclidian și U operator liniar, $U: X \to X$. Se poate arăta că există U' astfel încât

$$(U\mathbf{x},\mathbf{y}) = (\mathbf{x},U'\mathbf{y}), \quad \mathbf{x},\mathbf{y} \in X.$$

Operatorul U' se numește *adjunctul* lui U.

Matricea operatorului U' în orice bază ortogonală a spațiului X este transpusa matricei operatorului U în acea bază.

Un operator se numește *autoadjunct* dacă $U' = U$. Matricea unui operator autoadjunct este simetrică.

1.1.3 VALORI ȘI VECTORI PROPRII

Fie X un spațiu vectorial și $U: X \to X$.

Definiția 1.1-11 Un subspațiu X_0 al lui X se numește *invariant în raport cu operatorul* U, dacă $(\forall)\,\mathbf{x} \in X_0 \Rightarrow U\mathbf{x} \in X_0$, adică $U(X_0) \subseteq X_0$.

Observație. Un rol deosebit îl joacă subspațiile invariante de dimensiune 1. Ele se numesc direcții invariante (direcții proprii).

Definiția 1.1-12 $\lambda \in \mathbb{R}$ se numește *valoarea proprie* a operatorului U dacă $\exists\,\mathbf{x} \in X, \mathbf{x} \neq 0$ astfel încât

$$U\mathbf{x} = \lambda\mathbf{x},$$

iar $\mathbf{x}$ se numește *vector propriu* corespunzător valorii proprii λ.

Mulțimea valorilor proprii ale lui U se numește *spectrul* lui U.

Definiția 1.1-13 Mulțimea tuturor vectorilor proprii ai unui operator U corespunzător valorii proprii λ, la care se adaugă vectorul nul se numește *subspațiul propriu* al lui U, corespunzător lui λ.

Propoziția 1.1-1 (Demidovitch & Maron, 1973) *Vectorii proprii* $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^n$ *ai unui operator* U, *corespunzând valorilor proprii distincte* $\lambda_1,\lambda_2,\ldots,\lambda_n$, *sunt liniar independenți.*

Observație. Din propoziție rezultă că într-un spațiu n-dimensional orice operator U nu poate avea mai mult de n vectori proprii cu valori proprii distincte.

Propoziția 1.1-2 (Demidovitch & Maron, 1973) *Subspațiul propriu al unui operator liniar* U, *corespunzător unei valori proprii* λ *este un spațiu invariant al lui* U.

Propoziția 1.1-3 (Demidovitch & Maron, 1973) *Dacă U este un operator autoadjunct acționând pe un spațiu euclidian și λ este o valoare proprie a lui U, atunci $\exists \mathbf{x}$, vector unitar astfel încât*

$$\lambda = (U\mathbf{x}, \mathbf{x}) \quad \|\mathbf{x}\| = 1.$$

Propoziția 1.1-4 (Demidovitch & Maron, 1973) *Orice operator autoadjunct U acționând pe un spațiu euclidian n-dimensional V are n vectori proprii unitari liniar independenți, ortogonali doi câte doi.*

1.1.4 POLINOMUL CARACTERISTIC

Fie $\mathbf{A}$ matricea operatorului liniar U într-o bază fixată. Dacă E este operatorul identitate, operatorul $U - \lambda E$ va avea în această bază matricea $\mathbf{A} - \lambda \mathbf{I}$, unde $\mathbf{I}$ este matricea identitate. Dacă $\mathbf{x}$ este un vector propriu al lui U, corespunzător valorii proprii λ, atunci :

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = 0$$

iar $\mathbf{x}$ se mai numește vector propriu al lui $\mathbf{A}$.

Dacă $\mathbf{A}$ este matrice (n,n) atunci egalitatea de mai sus reprezintă un sistem omogen de n ecuații cu n necunoscute. Sistemul admite o soluție nenulă dacă și numai dacă $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$. Membrul stâng al acestei ecuații în λ se numește *polinomul caracteristic* al matricei $\mathbf{A}$. Oricărei rădăcini a acestei ecuații îi corespunde cel puțin un vector propriu al operatorului liniar U. Cum ecuația are cel puțin o rădăcină (reală sau complexă) rezultă că *un operator liniar are cel puțin un vector propriu.*

Fie $\mathbf{A}$ matricea operatorului U într-o bază fixată $\mathbf{e}$ și $\mathbf{A}'$ matricea aceluiași operator într-o altă bază $\mathbf{f}$. Operatorul $U - \lambda E, \lambda \in \mathbb{R}$ va avea în baza $\mathbf{e}$ matricea $\mathbf{A} - \lambda \mathbf{I}$, iar în baza $\mathbf{f}$, matricea $\mathbf{A}' - \lambda \mathbf{I}$. Cum determinantul matricei unui operator nu depinde de alegerea bazei, rezultă:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \det(\mathbf{A}' - \lambda \mathbf{I}).$$

Propoziția 1.1-5 (Demidovitch & Maron, 1973) *Polinomul caracteristic al unui operator este invariant în raport cu alegerea bazei.*

Observație. Din Propoziția 1.1-5 rezultă că toate conceptele spectrale (spectrul, ordinele de multiplicitate ale valorilor proprii) sunt invariante la o transformare a bazei.

Dacă $\mathbf{A}$ este matricea unui operator U în baza $\mathbf{e}_1, \ldots, \mathbf{e}_n$ și $\mathbf{A}'$ este matricea aceluiași operator în baza $\mathbf{f}_1, \ldots, \mathbf{f}_n$ atunci (un calcul simplu) arată că

$\mathbf{A}' = \mathbf{B}^{-1}\mathbf{A}\mathbf{B}$.

Două matrici $\mathbf{A}$ şi $\mathbf{A}'$ între care există o asemenea egalitate se numesc *matrici asemenea (similare).*

Din relaţia de mai sus

$\Rightarrow \mathbf{B}\mathbf{A}' = \mathbf{A}\mathbf{B} \Rightarrow \det(\mathbf{B}\mathbf{A}') = \det(\mathbf{A}\mathbf{B}) \Rightarrow \det\mathbf{B}\det\mathbf{A}' = \det\mathbf{A}\det\mathbf{B}$.

Cum $\det\mathbf{B} \neq 0 \Rightarrow \det\mathbf{A}' = \det\mathbf{A}$, adică determinantul matricei unui operator nu depinde de alegerea bazei. Rezultă:

Propoziţia 1.1-6 (Demidovitch & Maron, 1973) *Determinantul matricei unui operator este un invariant în raport cu alegerea bazei spaţiului respectiv.*

Fie $\mathbf{A}, \mathbf{B}$ două matrici asemenea. $\Rightarrow \exists\, \mathbf{C}$ matrice astfel încât

$\mathbf{B} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}$.

Putem scrie aşadar succesiv :

$$\det(\mathbf{B} - \lambda\mathbf{I}) = \det(\mathbf{C}^{-1}\mathbf{A}\mathbf{C} - \lambda\mathbf{I})$$
$$= \det(\mathbf{C}^{-1}(\mathbf{A} - \lambda\mathbf{I})\mathbf{C}) = \det\mathbf{C}^{-1}\det(\mathbf{C}^{-1}(\mathbf{A} - \lambda\mathbf{I})\mathbf{C})\det\mathbf{C}.$$
$$= \det(\mathbf{A} - \lambda\mathbf{I})$$

Rezultă că λ valoare proprie a lui $\mathbf{B} \Leftrightarrow \lambda$ valoare proprie a lui $\mathbf{A}$. Am demonstrat următoarea:

Propoziţia 1.1-7 *Două matrici asemenea au aceleaşi valori proprii.*

Pentru a aduce polinomul caracteristic la o formă convenabilă, îl scriem explicit

$$P(\lambda) = \begin{vmatrix} A_{11} - \lambda & A_{12} + 0 & & A_{1n} + 0 \\ A_{21} + 0 & A_{22} - \lambda & & A_{2n} + 0 \\ \cdots & & & \cdots \\ A_{n1} + 0 & A_{n2} + 0 & \cdots & A_{nn} - \lambda \end{vmatrix}$$

Propoziţia 1.1-8 *Polinomul caracteristic* $P(\lambda)$ *al matricei* $\mathbf{A}$ *se poate scrie*

$$P(\lambda) = (-\lambda)^n + I_1(-\lambda)^{n-1} + \ldots + I_{n-1}(-\lambda) + I_n$$

unde I_k *este suma minorilor principali de ordinul* k *ai determinantului matricei* $\mathbf{A}$.

Observații.

1. Coeficientul I_1 al lui $(-\lambda)^{n-1}$ coincide cu : $I_1 = \operatorname{tr}\mathbf{A}$. Termenul liber I_n, este determinantul $\mathbf{A}$. Coeficientul I_k al lui $(-\lambda)^{n-k}$ este suma minorilor principali de ordinul k.
2. $P(\lambda) = (-1)^n\left(-\lambda^n - I_1\lambda^{n-1} + \ldots + (-1)I_n\right)$. Notând cu m_i ordinul de multiplicitate al rădăcinii λ_i şi ținând cont de relațiile între rădăcini şi coeficienți rezultă $I_n = \det\mathbf{A} = \prod_{i=1}^{p}(\lambda_i)^{m_i}$ şi $I_1 = \operatorname{tr}\mathbf{A} = \sum_{i=1}^{p} m_i\lambda_i$ unde $m_i > 0, i = \overline{1,p}, m_j = 0$ dacă $j > p$.
3. Deoarece $I_n = \det\mathbf{A}$ este un invariant, rezultă că şi produsul valorilor proprii ale unui operator este un invariant (nu depinde de alegerea bazei).
4. Deoarece două matrici asemenea au valori proprii identice rezultă că matricile asemenea au determinanții şi urma identice.

1.1.5 BAZA VECTORILOR PROPRII

Propoziția 1.1-9

a) O matrice reală simetrică are toate valorile proprii reale.

b) *Vectorii proprii corespunzând la valorile proprii distincte sunt ortogonali.*

Propoziția 1.1-10 (Demidovitch & Maron, 1973) *Matricea unui operator într-o bază formată din vectorii săi proprii este diagonală şi elementele diagonale sunt valori proprii ale operatorului.*

Demonstrație Fie $\mathbf{A}'$ o matrice (n,n) care se obține din $\mathbf{A}$ prin intermediul unei transformări de similarități, adică

$$\mathbf{A}' = \mathbf{B}^{-1}\mathbf{A}\mathbf{B},$$

unde $\mathbf{B}$ este matricea transformării. Condiția ca matricea $\mathbf{A}'$ să fie diagonală se scrie :

$$\mathbf{A}' = \mathbf{B}^{-1}\mathbf{A}\mathbf{B} = \begin{pmatrix} \lambda_1 & & 0 \\ & & \\ 0 & & \lambda_n \end{pmatrix}$$

de unde se obține imediat

$$\mathbf{A}\mathbf{B} = \mathbf{B}\cdot\begin{pmatrix} \lambda_1 & & 0 \\ & & \\ 0 & & \lambda_n \end{pmatrix}.$$

Urmează că :

$$\sum_k A_{ik}B_{kj} = B_{ij}\lambda_j, \quad i,j = \overline{1,n}.$$

Fixând indicele j obținem n ecuații :

$$\sum_k A_{ij}B_{kj} = \lambda_j B_{ij}, \quad i = \overline{1,n}$$

Fie acum vectorul $\mathbf{b}^j = \begin{pmatrix} B_{1j} \\ \vdots \\ B_{nj} \end{pmatrix}$ format cu elementele coloanei j a matricei $\mathbf{B}$.

Egalitățile de mai sus se pot scrie

$$\mathbf{A} \cdot \mathbf{b}^j = \lambda_j \cdot \mathbf{b}^j,$$

deci $\mathbf{b}^j$ este vector propriu al matricei $\mathbf{A}$. Rezultă deci că matricea transformată $\mathbf{A}$ este diagonală dacă matricea $\mathbf{B}$ a transformării este aleasă astfel încât coloanele sale să fie vectori proprii ai matricei inițiale $\mathbf{A}$. Se poate arăta că o astfel de matrice există dacă toate valorile proprii ale lui $\mathbf{A}$ sunt diferite.

□

Propoziția 1.1-11 *O matrice* $\mathbf{A}$ *poate fi adusă la forma diagonală prin intermediul unei transformări de similaritate*

$$\mathbf{A}' = \mathbf{B}^{-1}\mathbf{A}\mathbf{B}$$

dacă valorile proprii ale lui $\mathbf{A}$ *sunt distincte.*

Propoziția 1.1-12 *O matrice* $\mathbf{A}$ *poate fi adusă la forma diagonală prin intermediul unei transformări de similaritate.*

1.1.6 FORME PĂTRATICE

Definiția 1.1-14 O *formă biliniară* pe un spațiu vectorial X este o aplicație $F : X \times X \to X$, liniară în ambele argumente. Dacă $\dim X = n$ și $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ - bază în X, atunci forma biliniară F se poate scrie :

$$F(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i y_j$$

unde $F(\mathbf{e}_i,\mathbf{e}_j) = A_{ij}, \quad i,j = \overline{1,n}$. Coeficienții A_{ij} sunt elementele unei matrice pătrate $\mathbf{A}$, numită *matricea formei biliniare* F, în baza $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$.

Să observăm că relația de definiție se mai poate scrie :

$$F(\mathbf{x},\mathbf{y}) = \mathbf{x}'\mathbf{A}\mathbf{y}.$$

Definiția 1.1-15 O formă biliniară se numește *simetrică* dacă

$$F(\mathbf{x},\mathbf{y}) = F(\mathbf{y},\mathbf{x}) \quad (\mathbf{x},\mathbf{y} \in X).$$

Observație. Matricea unei forme bilinire simetrice este simetrică.

Definiția 1.1-16 O formă biliniară pe X în care $\mathbf{y} = \mathbf{x}$ se numește *formă pătratică* pe X.

$F(\mathbf{x},\mathbf{y})$ se numește *formă biliniară polară* a formei $F(\mathbf{x},\mathbf{x})$.

Propoziția 1.1-13 *Forma polară* $F(\mathbf{x},\mathbf{y})$ *este unic determinată de forma ei pătratică.*

Într-o bază fixată forma pătratică F se scrie :

$$F(\mathbf{x},\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}.$$

Definiția 1.1-17 Forma pătratică $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ și matricea $\mathbf{A}$ se numesc *pozitiv semidefinite* dacă:

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \geq 0, \quad (\forall)\,\mathbf{x} \in X$$

și *pozitiv definite* dacă :

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0, \quad (\forall)\,\mathbf{x} \in X, \quad \mathbf{x} \neq 0.$$

Observații.

1. Condiția ca $\mathbf{A}$ să fie pozitiv definită se mai scrie $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = 0 \Rightarrow \mathbf{x} = 0$.
2. Produsul scalar este o formă biliniară simetrică corespunzătoare unei forme pătratice pozitiv definite. Rezultă că produsul scalar se poate exprima sub forma $(\mathbf{x},\mathbf{y}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{y}$, unde $\mathbf{A}$ este o matrice simetrică, pozitiv definită.

Distanța indusă de o normă generată de un produs scalar se va scrie

$$d^2(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2 = (\mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y})$$

și deci distanța are forma

$$d^2(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^{\mathrm{T}}\mathbf{A}(\mathbf{x}-\mathbf{y}).$$

Pentru diferite alegeri, obținem diferite tipuri de distanțe. Astfel, dacă $\mathbf{A}$ este matricea unitate, obținem distanța euclidiană

$$d^2(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}(x_i - y_i)^2$$

unde $x_1, x_2, \ldots, x_n$ sunt componentele vectorului $\mathbf{x}$ în baza considerată.

Propoziția 1.1-14 *Dacă* $\mathbf{A}$ *este o matrice pozitiv semidefinită, atunci* $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = 0 \Rightarrow \mathbf{A}\mathbf{x} = 0$.

Propoziția 1.1-15 *Fie* $\mathbf{A}$ *pozitiv semidefinită. Matricea* $\mathbf{A}$ *este pozitiv definită* $\Leftrightarrow$ *este nesingulară. În acest caz şi matricea* $\mathbf{A}^{-1}$ *este pozitiv definită.*

Propoziția 1.1-16 *Dacă matricea* $\mathbf{A}(n,n)$ *este simetrică şi pozitiv semidefinită atunci,* $(\forall)\,\mathbf{B}(n,m)$, *matricea* $\mathbf{B}^{\mathrm{T}}\mathbf{AB}$ *este simetrică şi pozitiv semidefinită.*

Dacă rang $\mathbf{B} = m$ *şi* $\mathbf{A}$ *este pozitiv definită, atunci* $\mathbf{B}^{\mathrm{T}}\mathbf{AB}$ *este pozitiv definită.*

Observație. Dacă $\mathbf{A}$ este pozitiv definită şi rang $\mathbf{B} = m \Rightarrow \mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B}$ este pozitiv definită şi deci inversabilă.

Propoziția 1.1-17 *Matricea* $\mathbf{A}$ *este pozitiv definită* $\Leftrightarrow$ *toți minorii săi principali sunt pozitivi, adică :*

$$a_{11} > 0, \quad \det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \ldots, \det \mathbf{A} > 0.$$

$\mathbf{A}$ *este pozitiv semidefinită* $\Leftrightarrow$ *minorii principali sunt nenegativi.*

Propoziția 1.1-18 *Fie* $\mathbf{A}$ *simetrică.* $\mathbf{A}$ *este pozitiv semidefinită* $\Leftrightarrow$ *valorile sale proprii sunt nenegative.*

Propoziția 1.1-19 *Fie* $\mathbf{A}$ *simetrică.* $\mathbf{A}$ *este pozitiv definită* $\Leftrightarrow$ *toate va.orile sale proprii sunt pozitive.*

1.1.7 DERIVAREA. PROPRIETĂȚI EXTREMALE ALE FORMELOR PĂTRATICE

Definiția 1.1-18 Dacă funcția $f : \mathbb{R}^n \to \mathbb{R}$ este derivabilă parțial în raport cu toate variabilele $x_1, \ldots, x_n$ în punctul $\mathbf{x}$, punctul $\nabla f(\mathbf{x})$ definit prin

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

se numeşte *gradientul funcției* f în punctul $\mathbf{x}$.

Definiția 1.1-19 Fie $A \subseteq \mathbb{R}^n$ o mulțime nevidă și $f : A \to \mathbb{R}$. Funcția f se numește *diferențiabilă Fréchet* în punctul $\mathbf{x}_0 \in A$ dacă există o funcțională liniară $F : \mathbb{R}^n \to \mathbb{R}$ astfel încât

$$\lim_{\|\mathbf{h}\| \to 0} \frac{\left\| f\left(\mathbf{x}^0 + \mathbf{h}\right) - f\left(\mathbf{x}^0\right) - F\left(\mathbf{h}\right) \right\|}{\|\mathbf{h}\|} = 0 .$$

Propoziția 1.1-20 *Dacă $A \subseteq \mathbb{R}^n$ și funcția $f : A \to \mathbb{R}^n$ este diferențiabilă Fréchet în punctul $\mathbf{x}^0$, atunci există o unică funcțională liniară F cu proprietatea*

$$\lim_{\|\mathbf{h}\| \to 0} \frac{\left\| f\left(\mathbf{x}^0 + \mathbf{h}\right) - f\left(\mathbf{x}^0\right) - F\left(\mathbf{h}\right) \right\|}{\|\mathbf{h}\|} = 0 .$$

Definiția 1.1-20 Dacă funcția $f : A \to \mathbb{R}, A \subset \mathbb{R}^n$ este diferențiabilă Fréchet în punctul $\mathbf{x}^0$, funcționala care verifică egalitatea din definiția Definiția 1.1-19 se numește *derivata Fréchet* a funcției f în $\mathbf{x}^0$ și se notează $\mathrm{d} f\left(\mathbf{x}^0\right)$, iar valoarea ei în punctul h, $F\left(\mathbf{h}\right) = \mathrm{d} f\left(\mathbf{x}^0\right)\left(\mathbf{h}\right)$ se numește *diferențiala funcției* f în $\mathbf{x}^0$ cu creșterea $\mathbf{h}$.

Propoziția 1.1-21 *Dacă funcția $f : \mathbb{R}^n \to \mathbb{R}$ este diferențiabilă în punctul $\mathbf{x}^0$. atunci f este derivabilă parțial în raport cu toate variabilele din $\mathbf{x}^0$ și are loc egalitatea :*

$$\mathrm{d} f\left(\mathbf{x}^0\right)\left(\mathbf{h}\right) = \left(\nabla f\left(\mathbf{x}^0\right), \mathbf{h}\right) = \sum_{i=1}^{n} \frac{\partial f\left(\mathbf{x}\right)}{\partial x_i} h_i . \qquad (\forall)\, \mathbf{h} \in \mathbb{R}^n .$$

Observație. Dacă f este diferențiabilă în $\mathbf{x}^0$, atunci derivata $\mathrm{d} f\left(\mathbf{x}^0\right)$ a lui f în punctul $\mathbf{x}^0$ se poate reprezenta prin gradientul lui f în $\mathbf{x}^0$, adică :

$$\mathrm{d} f\left(\mathbf{x}^0\right) = \nabla f\left(\mathbf{x}^0\right) = \begin{pmatrix} \dfrac{\partial f\left(\mathbf{x}^0\right)}{\partial x_1} \\ \vdots \\ \dfrac{\partial f\left(\mathbf{x}^0\right)}{\partial x_n} \end{pmatrix}.$$

În cele ce urmează vom utiliza pentru $\nabla f\left(\mathbf{x}^0\right)$ și notația $\dfrac{\partial f}{\partial x}\left(\mathbf{x}^0\right)$.

Definițiile și rezultatele de mai sus se extind ușor pentru o funcție $g:\mathbb{R}^n \to \mathbb{R}^m$. În acest caz derivata în punctul $\mathbf{x}^0$ se reprezintă printr-o matrice

$$\frac{\partial g}{\partial x}\left(\mathbf{x}^0\right) = \nabla g\left(\mathbf{x}^0\right) = \begin{pmatrix} \left(\nabla g_1\left(\mathbf{x}^0\right)\right)^T \\ \left(g_m\left(\mathbf{x}^0\right)\right)^T \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1\left(\mathbf{x}^0\right)}{\partial x_1} & \frac{\partial g_1\left(\mathbf{x}^0\right)}{\partial x_n} \\ \cdots & \cdots \\ \frac{\partial g_m\left(\mathbf{x}^0\right)}{\partial x_1} & \cdots \quad \frac{\partial g_m\left(\mathbf{x}^0\right)}{\partial x_n} \end{pmatrix}$$

Definiția 1.1-21 Fie $I \subseteq \mathbb{R}$ și $f: I \to \mathbb{R}$. Prin *derivata funcției* f *în punctul* $\mathbf{x}^0$ *în raport cu matricea* $\mathbf{A}(n,m)$ se înțelege matricea

$$\frac{\partial f\left(\mathbf{x}^0\right)}{\partial A} = \begin{pmatrix} \frac{\partial f\left(\mathbf{x}^0\right)}{\partial a_{11}} & \cdots & \frac{\partial f\left(\mathbf{x}^0\right)}{\partial a_{1m}} \\ \cdots & & \cdots \\ \frac{\partial f\left(\mathbf{x}^0\right)}{\partial a_{n1}} & \cdots & \frac{\partial f\left(\mathbf{x}^0\right)}{\partial a_{nm}} \end{pmatrix}. n \times n$$

Propoziția 1.1-22 *Dacă* $\mathbf{x}$ *și* $\mathbf{y} \in \mathbb{R}^n$ *și* $\mathbf{M}$ *este matrice atunci :*

a) $\frac{\partial}{\partial x}\left(\mathbf{x}^T\mathbf{y}\right) = \frac{\partial}{\partial x}\left(\mathbf{y}^T\mathbf{x}\right) = \mathbf{y}$;

b) $\frac{\partial}{\partial y}\left(\mathbf{x}^T\mathbf{M}\mathbf{y}\right) = \mathbf{M}^T\mathbf{x}$;

c) $\frac{\partial}{\partial x}\left(\mathbf{x}^T\mathbf{M}\mathbf{y}\right) = \mathbf{M}\mathbf{x} + \mathbf{M}^T\mathbf{x}$;

d) $\frac{\partial}{\partial x}\left(\mathbf{x}^T\mathbf{M}\mathbf{y}\right) = \mathbf{M}\mathbf{y}$;

e) $\frac{\partial}{\partial x}\left(\mathbf{M}\mathbf{x}\right) = \mathbf{M}$;

f) $\frac{\partial}{\partial M}\left(\mathbf{x}^T\mathbf{M}\mathbf{y}\right) = \mathbf{x}\mathbf{y}^T$.

Observație. Dacă $\mathbf{M}$ este matrice simetrică c) atunci $\frac{\partial}{\partial x}\left(\mathbf{x}^T\mathbf{M}\mathbf{x}\right) = 2\mathbf{M}\mathbf{x}$.

Dacă $\mathbf{M}$ matrice unitate atunci $\frac{\partial}{\partial x}\left(\mathbf{x}^T\mathbf{x}\right) = 2\mathbf{x}$.

Fie $F(\mathbf{x},\mathbf{x})=\mathbf{x}^T\mathbf{A}\mathbf{x}$, $\mathbf{x}\in\mathbb{R}^n$ o formă pătratică simetrică. Considerăm valorile formei pătratice F pe sfera unitate, adică pentru acei $\mathbf{x}$ pentru care $\|\mathbf{x}\|^2=(\mathbf{x},\mathbf{x})=1$. Ne interesează care din punctele sferei unitate sunt puncte staționare pentru F, adică verifică ecuația $\frac{\partial}{\partial x}F(\mathbf{x},\mathbf{x})=0$. Punctele de extrem se vor găsi printre punctele staționare. Problema determinării punctelor staționare este o problemă de extrem condiționat, pentru rezolvarea căreia vom folosi metoda multiplicatorilor lui Lagrange. Restricția $\|\mathbf{x}\|^2=1$ se mai scrie :

$$g(\mathbf{x})=1-\mathbf{x}^T\mathbf{x}=0$$

deci avem problema :

$$\begin{cases}F(\mathbf{x},\mathbf{x})\to\min\\ g(\mathbf{x})=0\end{cases}$$

Conform metodei lui Lagrange, construim funcția $L:\mathbb{R}^n\times\mathbb{R}\to\mathbb{R}$,

$$L(\mathbf{x},\lambda)=F(\mathbf{x},\mathbf{x})+\lambda g(\mathbf{x})$$

care se mai scrie și

$$L(\mathbf{x},\lambda)=\mathbf{x}^T\mathbf{A}\mathbf{x}-\lambda(\mathbf{x}^T\mathbf{x}-1).$$

Condițiile necesare ca punctul $(\mathbf{x}^0,\lambda^0)$ să fie un punct de extrem cu legături sunt :

$$\frac{\partial L(\mathbf{x}^0,\lambda^0)}{\partial x}=0,\quad \frac{\partial L(\mathbf{x}^0,\lambda^0)}{\partial\lambda}=0.$$

Deoarece $\mathbf{A}$ este matrice simetrică prima ecuație ne dă :

$$2\mathbf{A}\mathbf{x}-2\lambda\mathbf{x}=0\Rightarrow\mathbf{A}\mathbf{x}=\lambda\mathbf{x}.$$

Rezultă că :

Propoziția 1.1-23 *Vectorii sferei unitate care sunt vectorii proprii ai matricei* $\mathbf{A}$ *asociate unei forme pătratice simetrice*

$$F(\mathbf{x},\mathbf{x})=\mathbf{x}^T\mathbf{A}\mathbf{x},$$

reprezintă puncte staționare ale lui F.

Valorile formei pătratice în punctele staționare sunt date de :

$$F(\mathbf{x},\mathbf{x})=\mathbf{x}^T\mathbf{A}\mathbf{x}=\mathbf{x}^T\lambda\mathbf{x}=\lambda\|\mathbf{x}\|^2=\lambda.$$

Rezultă că valoarea formei pătratice $F(\mathbf{x},\mathbf{x})=\mathbf{x}^T\mathbf{A}\mathbf{x}$ într-un punct staționar $\mathbf{x}$ este egală cu valoarea proprie corespunzătoare a matricei $\mathbf{A}$ a formei pătratice.

În particular, maximul (minimul) formei pătratice simetrice $F(\mathbf{x},\mathbf{x})$ pe sfera unitate este egal cu cea mai mare (cea mai mică) valoare proprie a lui $\mathbf{A}$.

Vectorul propriu corespunzând celei mai mari valori proprii este un vector ce porneşte din origine şi trece prin punctul de pe sfera unitate în care este atins maximul.

1.2 ANALIZA ÎN COMPONENTE PRINCIPALE (ACP)

Inventată de Karl Pearson în 1901 şi introdusă în statistica matematică de Harold Hotelling în 1933, analiza în componente principale a început să fie utilizată efectiv odată cu apariţia şi extinderea calculatoarelor electronice.

Analiza în componente principale, ACP, poate fi prezentată din diverse puncte de vedere:

- pentru statisticianul clasic, analiza în componente principale înseamnă a estima, pornind de la un eşantion dat, axele principale ale elipsoidului indicator al unei distribuţii normale multidimensionale. Aceasta este prezentarea iniţială a lui Hotelling urmată apoi de manualele clasice de analiză multivariată, cum este cazul lucrării fundamentale ale lui Anderson, 1958);
- pentru psihologi, analiza în componente principale este un caz particular de analiză factorială utilizată în psihometrie (cazul dispersiilor nule sau egale; conform Harman, 1967);
- in fine, pentru al analiştii de date, ACP este o tehnică de reprezentare a datelor cu un caracter optimal din punct de vedere al unor criterii algebrice sau geometrice şi care este utilizată, în general, fără vreo referire la o ipoteză de natură statistică nici la un model particular. Acest punct de vedere, foarte răspândit la ora actuală (şi adoptat în cele ce urmează), este poate cel mai vechi fiind punctul de vedere adoptat de Pearson. Desigur în prezentatrea lui Pearson nu este vorba de analiza în componente principale aşa cum este ea expusă astăzi dar ideile esenţiale ale metodei pot fi deja întrevăzute laacest autor (o discuţie mai largă asupra acestui subiect se găseşte în articolul de sinteză a lui Rao, 1964).

Analiza în componente principale este utilizată pentru a pune în evidenţă:

- sistemul de relaţii existent între variabile (asocierea sau opoziţia lor);
- reprezentarea indivizilor în raport cu variabilelele observate (indivizi care prezintă caracteristici comune sau antagoniste).

ACP se aplică tabelelor cu două dimensiuni care „încrucişează" indivizi cu variabile numerice continue (sau care pot fi considerate continue). După proveniența variabilelor, trei mari categorii de tabele pot face obiectul unui demers de ACP şi anume:

1) tabelele de măsurători: variabilele sunt obţinute în urma unui sondaj sau recensământ şi sunt cantitative,

2) tabelele de note: variabilele sunt obținute în urma unor notații. Notele sunt variabile calitative care pot fi în general asimilate cu variabilele cantitative;
3) tabelele de ranguri: variabilele sunt obținute în urma unor clasamente și sunt variabile calitative ordinale care pot fi transformate în variabile continue.

Analiza în componente principale prezintă numeroase variante după transformările aduse tabelului de date: norul de puncte-indivizi poate fi centrat sau nu, redus sau nu. Dintre variante, analiza în componente principale normată (central-redusă) este cea mai utilizată.

1.2.1 DATELE ȘI CARACTERISTICILE LOR

Se presupune că dispunem de observații asupra a p variabile continue măsurate pe n indivizi. Valorile sunt „listate" într-un tabel de n linii și p coloane; notăm cu $\mathbf{X} = \left(x_{ij}\right)_{i=1,n}^{j=1,p}$ matricea asociată tabelului, unde x_{ij} este valoarea luată de variabila j măsurată pe individul i.

O variabilă este identificată prin vectorul-coloană j a tabelului $\mathbf{X}$ (notație $\mathbf{x}_j$) iar un individ prin vectorul-linie i (notație $\mathbf{e}_i'$).

Dacă datele nu au fost culese în urma unui sondaj aleator cu probabilități egale atunci fiecărui individ i i se atribuie o pondere[1] p_i conform importanței pe care o are în studiul întreprins.

Definiția 1.2-1 Se numește *matrice* (sau *metrică*) *de ponderi* matricea $\mathbf{D} = \text{diag}\left(p_1, \ldots, p_n\right)$ unde $p_i > 0 \quad (\forall) i = \overline{1,n}$ și $\sum_i p_i = 1$.

În cazul indivizilor echiponderați $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$ unde $\mathbf{I}_n$ este matricea identitate de dimensiune n.

Să notăm că $\mathbf{x}_j$ poate fi interpretat ca o selecție de volum n asupra variabilei j și că în acest context:

- *media de selecție* a variabilei j este
$$\text{m}\left(\mathbf{x}_j\right) \equiv \bar{x}_j = \sum_i p_i x_{ij},$$
- *dispersia de selecție* a variabilei j este
$$\text{s}^2\left(\mathbf{x}_j\right) \equiv s_j^2 = \sum_i p_i \left(x_{ij} - \bar{x}_j\right)^2,$$
- *covarianța de selecție* a variabilelor j și j' este

[1] Termenii de pondere sau masă sunt utilizați cu același sens în statistică și desemnează adesea frecvențele relative sau probabilitățile *a priori*.

$$\operatorname{cov}\left(\mathbf{x}_j, \mathbf{x}_{j'}\right) \equiv v_{jj'} = \sum_i p_i \left(x_{ij} - \bar{x}_j\right)\left(x_{ij'} - \bar{x}_{j'}\right),$$

– *coeficientul de corelație de selecție* a variabilelor j și j' este

$$\operatorname{cor}\left(\mathbf{x}_j, \mathbf{x}_{j'}\right) \equiv r_{jj'} = \frac{v_{jj'}}{s_j s_{j'}}.$$

Definiția 1.2-2 Se numește *punct mediu* (*centru de greutate*) al norului de puncte-indivizi $\{\mathbf{e}_i\}_{i=1}^n$ vectorul $\mathbf{g}' = \left(\bar{x}_1, \ldots, \bar{x}_p\right)$.

Se observă că:

$\mathbf{g} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$ unde $\mathbf{1}_n' = (1, \ldots, 1) \in \mathbb{R}^n$.

Definiția 1.2-3 Se numește *tabel centrat* asociat lui $\mathbf{X}$ matricea

$$\mathbf{Y} = \left(y_{ij}\right)_{i=\overline{1,n}}^{j=\overline{1,p}} \quad \text{unde} \quad y_{ij} = x_{ij} - \bar{x}_j.$$

Se numește *tabel centrat-redus* asociat lui $\mathbf{X}$ matricea

$$\mathbf{Z} = \left(z_{ij}\right)_{i=\overline{1,n}}^{j=\overline{1,p}} \quad \text{unde} \quad z_{ij} = \frac{y_{ij}}{s_j}.$$

Cu acestea, următoarele relații sunt imediate:

Lema 1.2-1

a) $\mathbf{Y} = \mathbf{X} - \mathbf{1}_n\mathbf{g}' = \left(\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'\mathbf{D}\right)\mathbf{X}$.

b) $\mathbf{Z} = \mathbf{Y}\,\mathbf{D}_{1/s}$ unde $\mathbf{D}_{1/s} = \operatorname{diag}\left(\frac{1}{s_1}, \ldots, \frac{1}{s_p}\right)$.

c) Matricea de varianță-covarianță asociată tabelului $\mathbf{X}$ *este* $\mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{X} - \mathbf{g}\mathbf{g}' = \mathbf{Y}'\mathbf{D}\mathbf{Y}$.

d) *d) Matricea de corelație asociată tabelului* $\mathbf{X}$ *este* $\mathbf{R} = \mathbf{D}_{1/s}\mathbf{V}\,\mathbf{D}_{1/s} = \mathbf{Z}'\mathbf{D}\mathbf{Z}$.

Observație. Relația $\mathbf{X}'\mathbf{D}\mathbf{X} = \sum_{i=1}^n p_i \mathbf{e}_i\mathbf{e}_i'$ este o formulă utilă implementării pe calculator a metodei căci evită introducerea în memoria RAM a întregii matrici $\mathbf{X}$.

Un individ $\mathbf{e}_i$ este definit de p coordonate corespunzând valorilor celor p variabile măsurate pe acest individ; îl putem considera ca un element dintr-un spațiu vectorial $\mathcal{F} \subseteq \mathbb{R}^p$, numit *spațiul indivizilor*. Mulțimea celor n indivizi formează un „nor de puncte-indivizi" în $\mathcal{F}$, cu $\mathbf{g}$ centrul de greutate al norului. Se dotează acest spațiu cu o metrică care să permită definirea distanței între indivizi.

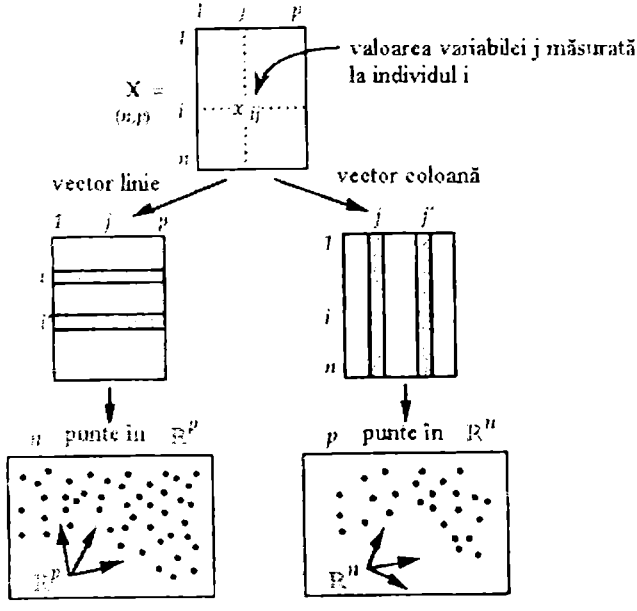


Figura 1.2-1 Principiul reprezentării geometrice

Fie $\mathbf{M} \in \mathcal{M}_{p,p}(\mathbb{R})$ o matrice simetrică, pozitiv definită, de dimensiune p, cu coeficienţi reali.

Definiţia 1.2-4 Se numeşte *matricea produsului scalar între indivizi* matricea

$$\mathbf{W} = \left(w_{ij}\right)_{i,j=1}^{n} \quad \text{unde} \quad w_{ij} = \left\langle \mathbf{e}_i, \mathbf{e}_j \right\rangle$$

cu $\left\langle \mathbf{e}_i, \mathbf{e}_j \right\rangle_{\mathbf{M}} \overset{\text{def}}{=} \mathbf{e}_i' \mathbf{M} \mathbf{e}_j$ este produsul scalar pe spaţiul $\mathcal{F}$ definit de metrica $\mathbf{M}$.

Se observă că:

$$\mathbf{W} = \mathbf{X}\mathbf{M}\mathbf{X}'$$

şi că distanţa dintre doi indivizi $\mathbf{e}_i$ şi $\mathbf{e}_j$ din spaţiul $\mathcal{F}$ este dată de relaţia

$$d^2\left(\mathbf{e}_i, \mathbf{e}_j\right) = \left\langle \mathbf{e}_i - \mathbf{e}_j, \mathbf{e}_i - \mathbf{e}_j \right\rangle_{\mathbf{M}} = \left\| \mathbf{e}_i - \mathbf{e}_j \right\|_{\mathbf{M}}^2 .$$

În teorie alegerea metricii $\mathbf{M}$ depinde de utilizator singurul care poate preciza metrica adecvată. În practică metricile cele mai uzuale în ACP sunt:

- $\mathbf{M} = \mathbf{I}_p$ ce induce produsul scalar uzual şi distanţa euclidiană;

- $\mathbf{M} = \mathbf{D}_{1/s^2}$. Utilizarea acestei metrici revine la adimensionalizarea variabilelor măsurate căci fiecare valoare este împărţită cu abaterea standard de selecţie a variabilei corespunzătoare (x_{ij}/s_j).

Metrica $\mathbf{M} = \mathbf{I}_p$ dă fiecărei variabile aceeaşi importanţă independent de dispersia sa; utilizarea ei va privilegia variabilele cu dispersie mare pentru care diferenţele între indivizi sunt mari şi va neglija diferenţele între celelalte variabile.

Metrica $\mathbf{M} = \mathbf{D}_{1/s^2}$ echilibrează influenţa variabilelor transformându-le în variabile cu dispersia de selecţie unitară.

Observaţie. Dacă $\mathbf{M} = \mathrm{diag}(m_1,\ldots,m_p)$ atunci $\mathrm{d}^2(\mathbf{e}_i,\mathbf{e}_j) = \sum_{k=1}^{p} m_k (x_{ik} - x_{jk})^2$ iar coeficienţii $\{\sqrt{m_k}\}_{k=\overline{1,p}}$ pot fi consideraţi ca ponderi ale variabilelor $\mathbf{x}_j$ în distanţa dintre indivizi.

Lema 1.2-2. *Matricea produsului scalar între indivizi poate fi întotdeauna exprimată în funcţie de metrica* $\mathbf{I}_p$.

Demonstraţie. Într-adevăr, dacă $\mathbf{M}$ este simetrică şi pozitiv definită atunci ea poate fi scrisă ca $\mathbf{M} = \mathbf{T}'\mathbf{T}$ (conform algoritmului lui Cholesky; vezi Demidovitch & Maron, 1973). Atunci $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_{\mathbf{M}} = \mathbf{e}_i' \mathbf{M} \mathbf{e}_j = \mathbf{e}_i' \mathbf{T}' \mathbf{T} \mathbf{e}_j = (\mathbf{T}\mathbf{e}_i)' (\mathbf{T}\mathbf{e}_j) = (\mathbf{T}\mathbf{e}_i)' \mathbf{I}_p (\mathbf{T}\mathbf{e}_j)$ ceeace înseamnă că $\mathbf{W} = (\mathbf{X}\mathbf{T}')\mathbf{I}_p(\mathbf{T}\mathbf{X}')$ adică matricea produsului scalar al tabelului $\mathbf{X}\mathbf{T}'$ faţă de metrica $\mathbf{M} = \mathbf{I}_p$.

[]

Corolarul 1.2-1. *Utilizarea metricii* $\mathbf{M} = \mathbf{D}_{1/s^2}$ *pentru tabelul* $\mathbf{Y}$ *revine la folosirea metricii* $\mathbf{M} = \mathbf{I}_p$ *pentru tabelul centrat-redus* $\mathbf{Z}$.

Reamintim că ipoteza fundamentală a unui demers ACP este aceea că întreaga informaţie este conţinută în distanţele dintre punctele-indivizi ai norului. Acest lucru justifică introducerea noţiunii de inerţie totală[2].

Definiţia 1.2-5. Se numeşte *inerţie totală* (*globală*) a norului de puncte-indivizi $\{\mathbf{e}_i\}_{i=1}^{n}$ media ponderată a pătratelor distanţelor de la punctele-indivizi la centrul de greutate al norului, adică:

[2] Termenul inerţie este împrumutat din mecanică şi este sinonim, în acest context, cu termenul statistic dispersie.

$$I_g = \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{g})' \mathbf{M} (\mathbf{e}_i - \mathbf{g}) = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{g}\|_{\mathbf{M}}^2 .$$

Prin analogie, inerția într-un punct oarecare $\mathbf{a} \in \mathbb{R}^p$ se definește ca $I_{\mathbf{a}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{a}\|_{\mathbf{M}}^2$.

Proprietățile inerției globale, puse în evidență de enunțurile de mai jos, sunt utile în demersul ce urmează.

Propoziția 1.2-1. *(formula lui Huyghens) Inerția față de un punct satisface următoarea relație:* $I_{\mathbf{a}} = I_{\mathbf{g}} + (\mathbf{g} - \mathbf{a})' \mathbf{M} (\mathbf{g} - \mathbf{a}) = I_{\mathbf{g}} + \|\mathbf{g} - \mathbf{a}\|_{\mathbf{M}}^2$.

Demonstrație. Într-adevăr

$$\begin{aligned} I_{\mathbf{a}} &= \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{a})' \mathbf{M} (\mathbf{e}_i - \mathbf{a}) = \sum_{i=1}^{n} p_i \left[(\mathbf{e}_i - \mathbf{g}) + (\mathbf{g} - \mathbf{a}) \right]' \mathbf{M} \left[(\mathbf{e}_i - \mathbf{g}) + (\mathbf{g} - \mathbf{a}) \right] \\ &= \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{g})' \mathbf{M} (\mathbf{e}_i - \mathbf{g}) + \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{g})' \mathbf{M} (\mathbf{g} - \mathbf{a}) + \\ &\quad + \sum_{i=1}^{n} p_i (\mathbf{g} - \mathbf{a})' \mathbf{M} (\mathbf{e}_i - \mathbf{g}) + \sum_{i=1}^{n} p_i (\mathbf{g} - \mathbf{a})' \mathbf{M} (\mathbf{g} - \mathbf{a}). \end{aligned}$$

Se observă că primul termen al sumei este $I_{\mathbf{g}}$ că produsul $(\mathbf{g} - \mathbf{a})' \mathbf{M} (\mathbf{g} - \mathbf{a})$ nu depinde de i, că $\sum_i p_i = 1$ și că produsele $(\mathbf{e}_i - \mathbf{g})' \mathbf{M} (\mathbf{g} - \mathbf{a})$ și $(\mathbf{g} - \mathbf{a})' \mathbf{M} (\mathbf{e}_i - \mathbf{g})$ sunt scalare. Cu aceasta egalitatea de mai sus devine

$$I_{\mathbf{a}} = I_{\mathbf{g}} + 2 \left[\left(\sum_i p_i \mathbf{e}_i' \mathbf{M} \mathbf{g} - \mathbf{g}' \mathbf{M} \mathbf{g} \right) + \left(\mathbf{g}' \mathbf{M} \mathbf{a} - \sum_i p_i \mathbf{e}_i' \mathbf{M} \mathbf{a} \right) \right] + \|\mathbf{g} - \mathbf{a}\|_{\mathbf{M}}^2 .$$

Se notează cu $\mathbf{b}' = (\mathbf{M}\mathbf{g})' = (b_1, \ldots, b_p)$ și reamintind că $g_j = \sum_{i=1}^{n} p_i x_{ij}$ rezultă

$$\sum_i p_i \mathbf{e}_i' \mathbf{M} \mathbf{g} - \mathbf{g}' \mathbf{M} \mathbf{g} = \sum_i p_i \sum_{j=1}^{p} x_{ij} b_j - \sum_{j=1}^{p} g_j b_j = \sum_{j=1}^{p} b_j \left(\sum_i p_i x_{ij} \right) - \sum_{j=1}^{p} g_j b_j = 0 .$$

Analog $\mathbf{g}' \mathbf{M} \mathbf{a} - \sum_i p_i \mathbf{e}_i' \mathbf{M} \mathbf{a} = 0$.

□

Corolarul 1.2-2. *Pentru un nor de puncte-indivizi dat,* $\mathbf{g}$ *centrul de greutate al norului minimizează inerția totală.*

Lema 1.2-3. *Inerția totală este media pătratelor distanțelor dintre punctele-indivizi, adică:*

$$2I_{\mathbf{g}} = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \left\|\mathbf{e}_i - \mathbf{e}_j\right\|_{\mathbf{M}}^2 .$$

Demonstrație. Se aplică formula lui Huyghens pentru fiecare punct-individ, apoi se adună cele n relații.

$$p_1 I_{\mathbf{e}_1} = p_1 I_{\mathbf{g}} + p_1 \left\|\mathbf{e}_1 - \mathbf{g}\right\|_{\mathbf{M}}^2$$
$$p_2 I_{\mathbf{e}_2} = p_2 I_{\mathbf{g}} + p_2 \left\|\mathbf{e}_2 - \mathbf{g}\right\|_{\mathbf{M}}^2$$
$$\vdots$$
$$p_n I_{\mathbf{e}_n} = p_n I_{\mathbf{g}} + p_n \left\|\mathbf{e}_n - \mathbf{g}\right\|_{\mathbf{M}}^2$$

$$\sum_{j=1}^{n} p_j I_{\mathbf{e}_j} = \sum_{j=1}^{n} p_j I_{\mathbf{g}} + \sum_{j=1}^{n} p_j \left\|\mathbf{e}_j - \mathbf{g}\right\|_{\mathbf{M}}^2 \quad \Rightarrow \quad \sum_{j=1}^{n} p_j \sum_{i=1}^{n} p_i \left\|\mathbf{e}_i - \mathbf{e}_j\right\|_{\mathbf{M}}^2 = I_{\mathbf{g}} + I_{\mathbf{g}}.$$

□

Lema 1.2-4.

a) $I_{\mathbf{g}} = \mathrm{tr}(\mathbf{MV}) = \mathrm{tr}(\mathbf{VM})$, *unde cu* $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$ *s-a notat urma matricii* $\mathbf{A} \in \mathcal{M}_{n,n}(\mathbb{R})$.

b) Dacă centrul de greutate al norului este în originea axelor de coordonate, adică $\mathbf{g} = \mathbf{0}$, *atunci* $I_{\mathbf{g}} = \mathrm{tr}(\mathbf{WD}) = \mathrm{tr}(\mathbf{DW})$.

Demonstrație. a) Într-adevăr

$$\mathrm{tr}(\mathbf{MV}) = \mathrm{tr}(\mathbf{MY'DY}) = \sum_{i=1}^{n} \mathbf{M}\mathbf{y}_i\, p_i\, \mathbf{y}_i' = \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{g})' \mathbf{M} (\mathbf{e}_i - \mathbf{g}) = I_{\mathbf{g}} .$$

Analog $\mathrm{tr}(\mathbf{VM}) = I_{\mathbf{g}}$.

b) Dacă $\mathbf{g} = \mathbf{0}$ atunci $I_{\mathbf{g}} = \sum_{i=1}^{n} p_i \mathbf{e}_i' \mathbf{M} \mathbf{e}_i$. Pe de altă parte

$$\mathrm{tr}(\mathbf{WD}) = \mathrm{tr}(\mathbf{XMX'D}) = \sum_{i=1}^{n} \mathbf{e}_i' \mathbf{M} \mathbf{e}_i p_i = I_{\mathbf{g}} = \sum_{i=1}^{n} p_i \mathbf{e}_i' \mathbf{M} \mathbf{e}_i = \mathrm{tr}(\mathbf{DXMX'}) = \mathrm{tr}(\mathbf{DW}) .$$

□

Observații.

1. Dacă $\mathbf{M} = \mathbf{I}_p$ inerția este egală cu suma dispersiilor de selecție a celor p variabile.

2. Dacă $\mathbf{M} = \mathbf{D}_{1/s^2}$ atunci $I_g = \mathrm{tr}\left(\mathbf{D}_{1/s^2}\mathbf{V}\right) = \mathrm{tr}\left(\mathbf{D}_{1/s}\mathbf{V}\mathbf{D}_{1/s}\right) = \mathrm{tr}(\mathbf{R}) = \sum_{j=1}^{p} r_{jj} = \sum_{j=1}^{p} 1 = p$, din
3. Lema 1.2-4. Inerția este, în acest caz, egală cu numărul variabilelor şi nu depinde de valorile acestora.

Fiecare variabilă $\mathbf{x}_j$ poate fi considerată ca un vector a unui spațiu vectorial $\mathcal{E} \subseteq \mathbb{R}^n$ numit *spațiul variabilelor*. Mulțimea celor p variabile formează un „nor de puncte-variabile" în $\mathcal{E}$. Metrica utilizată în spațiul variabilelor este dată de matricea $\mathbf{D}$ (matricea diagonală a ponderilor indivizilor). Cu acestea se observă:

Lema 1.2-5 *Dacă variabilele sunt centrate atunci:*

a) produsul scalar indus de metrica $\mathbf{D}$ *este egal cu covarianța de selecție dintre cele două variabile necentrate;*

b) norma („lungimea") unei variabile este egală cu abaterea standard de selecție a variabilei necentrate;

c) unghiul dintre două variabile este egal cu coeficientul de corelație liniară de selecție al variabilelor necentrate.

Demonstrație. Într-adevăr:

a) $$\left\langle \mathbf{y}_j, \mathbf{y}_k \right\rangle_{\mathbf{D}} = \mathbf{y}_j' \mathbf{D} \mathbf{y}_k = \sum_{i=1}^{n} p_i \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right) = \mathrm{cov}\left(\mathbf{x}_j, \mathbf{x}_k\right).$$

b) $$\left\|\mathbf{y}_j\right\|_{\mathbf{D}}^2 = \left\langle \mathbf{y}_j, \mathbf{y}_j \right\rangle_{\mathbf{D}} = \mathbf{y}_j' \mathbf{D} \mathbf{y}_j = \sum_{i=1}^{n} p_i \left(x_{ij} - \bar{x}_j\right)^2 = \mathrm{s}^2\left(\mathbf{x}_j\right).$$

c) Fie θ_{jk} unghiul dintre variabilele $\mathbf{y}_j$ şi $\mathbf{y}_k$. Atunci

$$\cos\left(\theta_{jk}\right) = \frac{\left\langle \mathbf{y}_j, \mathbf{y}_k \right\rangle_{\mathbf{D}}}{\left\|\mathbf{y}_j\right\|_{\mathbf{D}} \left\|\mathbf{y}_k\right\|_{\mathbf{D}}} = \frac{v_{jk}}{s_j s_k} = \mathrm{cor}\left(\mathbf{x}_j, \mathbf{x}_k\right).$$

□

Corolarul 1.2-3

a) Mediile de selecție ale variabilelor $\left\{\mathbf{y}_j\right\}_{j=1}^{p}$ *sunt nule, dispersiile de selecție sunt egale cu dispersiile de selecție ale variabilelor* $\left\{\mathbf{x}_j\right\}_{j=1}^{p}$ *şi coeficienții de corelație de selecție sunt egali cu coeficienții de corelație de selecție a variabilelor* $\left\{\mathbf{x}_j\right\}_{j=1}^{p}$.

b) *Mediile de selecție ale variabilelor* $\{\mathbf{z}_j\}_{j=1}^{p}$ *sunt nule, dispersiile de selecție sunt unitare și coeficienții de corelație liniară de selecție sunt egali cu coeficienții de corelație liniară de selecție a variabilelor* $\{\mathbf{x}_j\}_{j=1}^{p}$.

Din cele de mai sus rezultă:

Lema 1.2-6 $\quad d^2(\mathbf{z}_j, \mathbf{z}_k) = 2(1 - r_{jk})$.

Demonstrație.

$$d^2(\mathbf{z}_j, \mathbf{z}_k) = \langle \mathbf{z}_j - \mathbf{z}_k, \mathbf{z}_j - \mathbf{z}_k \rangle_{\mathbf{D}}' = \sum_{i=1}^{n} p_i (z_{ij} - z_{ik})^2 = \sum_i p_i z_{ij}^2 + \sum_i p_i z_{ik}^2 - 2\sum_i p_i z_{ij} z_{ik} .$$

Conform corolarului de mai sus $\sum p_i z_{ij}^2 = s^2(\mathbf{z}_j) = 1 = s^2(\mathbf{z}_k) = \sum p_i z_{ik}^2$ și $\sum p_i z_{ij} z_{ik} = \mathrm{cor}(\mathbf{z}_j, \mathbf{z}_k) = r_{jk}$ ceeace implică relația din enunț.

□

Sistemul de proximități dintre două puncte-variabile din $\mathcal{E}$ indus de relația din Lema 1.2-6 este familiar statisticianului:

- două variabile puternic corelate sunt foarte apropiate una de cealaltă (căci $r_{jk} \approx 1$ implică $d^2(\mathbf{z}_j, \mathbf{z}_k) \approx 0$) sau din contră foarte depărtate (căci $r_{jk} \approx -1$ implică $d^2(\mathbf{z}_j, \mathbf{z}_k) \approx 4$) după cum relația liniară care le leagă este directă sau inversă;
- două variabile ortogonale sunt la distanță medie (căci $r_{jk} \approx 0$ implică $d^2(\mathbf{z}_j, \mathbf{z}_k) \approx 2$).

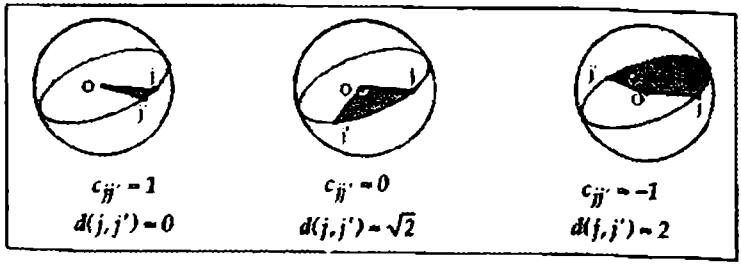


Figura 1.2-2 Corelațiile și distanțele între punctele-variabile

Proximitatea între două puncte-variabile se interpretează deci în termeni de corelații.

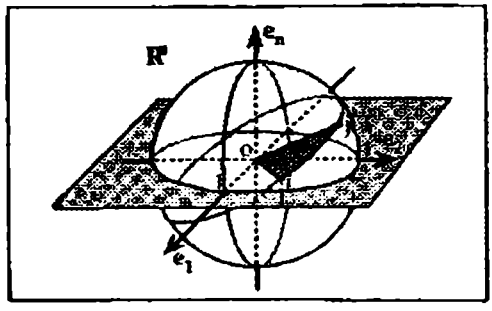


Figura 1.2-3 Sistemul de proximităţi între două puncte-variabile

Din Corolarul 1.2-3 punctul a) rezultă că toate punctele-variabile se află pe hipersfera de rază 1 centrată în originea axelor. Această hipersferă se numeşte *sfera de corelaţie*.

Planurile în care vor fi proiectate variabilele intersectează sfera după cercurile diametrale (cercuri de rază 1), numite *cercuri de corelaţie* în interiorul cărora se află proiecţiile punctelor-variabile.

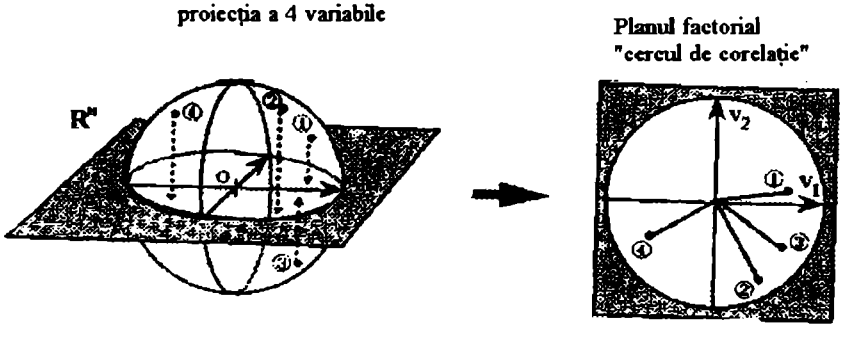


Figura 1.2-4 Reprezentarea sferei şi cercului de corelaţie

Observaţie Operaţia de centrare a tabelului **X** are în spaţiile $\mathbb{R}^p$ şi $\mathbb{R}^n$ interpretări geometrice diferite.

În $\mathbb{R}^p$ această transformare echivalează cu o translaţie a originii axelor în centrul de greutate (punctul mediu) al norului;

În $\mathbb{R}^n$ această transformare este o proiecţie pe hiperplanul ce trece prin originea axelor şi este ortogonal pe dreapta ce trece prin originea axelor şi are ca parametrii

directori $\{p_i\}_{i=1}^n$. Matricea $\mathbf{P} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'\mathbf{D}$ asociată acestei transformări este idempotentă ($\mathbf{P}^2 = \mathbf{P}$) şi $\mathbf{M}$-simetrică ($\mathbf{P}'\mathbf{M} = \mathbf{M}\mathbf{P}$), cu $\mathbf{M} = \mathbf{I}_n$. Ea este matricea proiecţiei $\mathbf{M}$-ortogonale pe subspaţiul general de vectorii coloană liniari independenţi ai matricii $\mathbf{Y}$. Coordonatele acestor vectori satisfac relaţia $\sum_i p_i y_{ij} = 0 \quad (\forall) j = \overline{1,p}$ care reprezintă ecuaţia unui hiperplan în $\mathbb{R}^n$ ce trece prin originea axelor şi are ca normală în punctul $\mathbf{0}_n$ dreapta de parametrii directori $\{p_i\}_{i=1}^n$. Dacă $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$ atunci hiperplanul este ortogonal pe prima bisectoare.

Definiţia 1.2-6 (după Dazy&Le Bazic, 1996) Se numeşte *studiu* un triplet $(\mathbf{Y}, \mathbf{M}, \mathbf{D})$ unde

$\mathbf{Y}$ este matricea centrată asociată tabelului de date indivizi-variabile;
$\mathbf{M}$ este o metrică în spaţiul vectorial $\mathcal{F}$;
$\mathbf{D}$ este metrica ponderilor în spaţiul vectorial $\mathcal{E}$.

Studiul este caracterizat de două „obiecte":

- matricea $\mathbf{W} = \mathbf{Y}\mathbf{M}\mathbf{Y}'$ a produsului scalar între indivizi;
- matricea $\mathbf{V} = \mathbf{Y}'\mathbf{D}\mathbf{Y}$ de varianţă-covarianţă a variabilelor centrate.

1.2.2 ANALIZA GENERALĂ, DESCOMPUNEREA ÎN VALORI SINGULARE

S-a arătat mai sus cum liniile şi coloanele unui tabel dreptunghiular permit definirea norilor de puncte.

Poziţia punctelor în nor este dată de mulţimea distanţelor între toate punctele şi determină *forma norului*. Forma norului este cea care caracterizează natura şi intensitatea relaţiilor între indivizi (liniile) şi între variabile (coloanele) şi relevă structurile de informaţii conţinute în date. De exemplu, un nor de puncte alungit uniform de-a lungul unei drepte traduce existenţa unei relaţii liniare dominante între puncte în timp ce o formă parabolică ilustrează existenţa unei relaţii neliniare iar o formă sferică indică, mai degrabă, absenţa unei relaţii.

O modalitate de a reda vizual forma unui nor este aceea de a-l proiecta pe o dreaptă, sau mai bine pe un plan, minimizând deformările pe care această proiecţie le implică; aceasta este esenţa analizei generale. În cele ce urmează se va prezenta detaliat programul enunţat.

Matricea $\mathbf{W} = \mathbf{Y}\mathbf{M}\mathbf{Y}'$ este o matrice simetrică de dimensiune n al cărui termen general $w_{ij} = \mathbf{e}_i'\mathbf{M}\mathbf{e}_j$ este un produs scalar între indivizii i şi j. Indivizii aparţin unui spaţiu vectorial euclidian $(\mathcal{F}, \mathbf{M})$ de dimensiune p (căci sunt p variabile).

Definiția 1.2-7 Se numește *imaginea euclidiană a indivizilor* asociați produselor scalare w_{ij} un nor compus din n puncte $A_1,\ldots,A_n$ și dintr-un punct O din $\mathcal{F}$ astfel încât aceste puncte să reconstituie produsele scalare w_{ij}, adică

$$\left\langle \overline{OA_i}, \overline{OA_j} \right\rangle = w_{ij} \quad (\forall) i,j = \overline{1,n}$$

unde produsul scalar $\langle \circ, \circ \rangle$ este definit de metrica euclidiană $\mathbf{I}_p$.

Matricea $\mathbf{V} = \mathbf{Y}'\mathbf{D}\mathbf{Y}$ este o matrice simetrică de dimensiune p al cărui termen general $v_{ij} = \mathbf{y}_i'\mathbf{D}\mathbf{y}_j$ este un produs scalar dintre variabilele i și j. Variabilele aparțin unui spațiu vectorial euclidian $(\mathcal{E}, \mathbf{D})$ de dimensiune n (căci sunt n indivizi).

Definiția 1.2-8 Se numește *imaginea euclidiană a variabilelor* asociată produselor scalare v_{ij} un nor compus din p puncte $B_1,\ldots,B_p$ și dintr-un punct O din $\mathcal{E}$ astfel încât aceste puncte să reconstituie produsele scalare v_{ij}, adică

$$\left\langle \overline{OB_i}, \overline{OB_j} \right\rangle = v_{ij} \quad (\forall) i,j = \overline{1,p}$$

unde produsul scalar $\langle \circ, \circ \rangle$ este definit de metrica euclidiană $\mathbf{I}_n$.

Dacă dimensiunea spațiului vectorial este egală cu 3 atunci imaginea euclidiană a unui nor de puncte poate fi vizualizată. Dacă dimensiunea spațiului este strict superioară lui 3 atunci acest lucru devine imposibil; în acest caz trebuie căutată o imagine euclidiană aproximativă. Să notăm că există o infinitate de imagini euclidiene ale aceluiași nor de puncte. Două imagini euclidiene sunt echivalente dacă ele reconstituie aceleași produse scalare.

1.2.2.1 Analiza norului de puncte-indivizi

Să ne plasăm, mai întâi, în spațiul $\mathcal{F} \subseteq \mathbb{R}^p$ al indivizilor în care tabelul $\mathbf{Y}$ poate fi reprezentat ca un nor de n puncte-indivizi centrați în punctul mediu al norului și ale căror p coordonate reprezintă liniile lui $\mathbf{Y}$. Principiul metodei ACP constă în reprezentarea aproximativă a norului de puncte-indivizi într-un subspațiu de dimensiune mult mai mică (de regulă egală cu 2); se pleacă deci de la o imagine euclidiană dintr-un spațiu afin de dimensiune p și se ajunge la o imagine euclidiană într-un spațiu afin de dimensiune $q << p$[3].

Demersul de mai sus se realizează prin proiecția punctelor-indivizi pe un subspațiu $\mathcal{F}_q$ de dimensiune q obținut astfel încât *media pătratelor distanțelor între proiecții să fie maximă* sau, ținând cont de

[3] Dacă rg($\mathbf{Y}$)=q atunci problema aproximării este practic rezolvată. Într-adevăr este suficient să găsim o bază a subspațiului vectorial de dimensiune q din $\mathbf{R}^p$ ce conține norul de puncte-indivizi și să calculăm coordonatele punctelor în noua bază. Vom putea astfel reconstitui cei np coeficienți ai tabelului $\mathbf{Y}$ pornind de la cei $qp+nq=(n+p)q$ coeficienți definiți mai sus.

Lema 1.2-3, *inerția norului proiectat pe* $\mathcal{F}_q$ *să fie maximă* sau, în fine, *deformarea distanțelor prin proiecție să fie minimă.*

Cu notațiile de mai sus, problema ce trebuie rezolvată se formulează astfel:

„*Să se găsească* $\mathcal{H} \equiv \mathcal{F}_q$ *astfel încât* $\max_{(\mathcal{H})} \sum_{i=1}^{n} \mathbf{d}^2(\mathbf{y}_i, \mathbf{0})$ "[4]

Soluția problemei este dată de următoarea teoremă:

Teorema 1.2-1 *Subspațiul de dimensiune q pe care se proiectează optim (în sensul celor mai mici pătrate) cele n puncte din* $\mathbb{R}^p$ *este generat de primii q vectori proprii ai matricii* $\mathbf{A} = \mathbf{V}\mathbf{M} \in \mathcal{M}_{p,p}(\mathbb{R})$ *corespunzători valorilor proprii* $\lambda_1 > \lambda_2 > \ldots > \lambda_q$.

Demonstrație. Să notăm cu $\{P_1,\ldots,P_n\}$ proiecțiile pe $\mathcal{H}$ ale punctelor $\{A_1,\ldots,A_n\}$ și să observăm că:

$$\overline{OA_i}^2 = \overline{OP_i}^2 + \overline{A_iP_i}^2 \quad i = \overline{1,n},$$

conform teoremei lui Pitagora, sau

$$\sum_{i=1}^{n} \overline{A_iP_i}^2 = \sum_{i=1}^{n} \overline{OA_i}^2 - \sum_{i=1}^{n} \overline{OP_i}^2 \quad (1).$$

Cum $\overline{OA_i}^2$, $i = \overline{1,n}$ sunt fixe (norul de puncte-indivizi este dat) a minimiza deformările produse prin proiecții este echivalent cu a minimiza suma ponderată a pătratelor distanțelor de la punctele $\{A_1,\ldots,A_n\}$ la subspațiul $\mathcal{H}$, adică a afla $\min \sum_{i=1}^{n} p_i \overline{A_iP_i}^2$ sau, conform relației (1), $\max \sum_{i=1}^{n} p_i \overline{OP_i}^2$.

Fie **a** un vector **M**-normat din $\mathbb{R}^p$ (adică $\mathbf{a}'\mathbf{M}\mathbf{a} = 1$). Coordonata proiecției P_i a punctului A_i pe dreapta $\Delta_{\mathbf{a}}$ având ca suport pe **a** este $OP_i = \mathbf{y}_i'\mathbf{M}\mathbf{a}$. Coordonatele tuturor punctelor P_i pe $\Delta_{\mathbf{a}}$ sunt $\mathbf{YMa}$ de unde rezultă că:

$$\sum_i p_i \overline{OP_i}^2 = \mathbf{a}'\mathbf{M}\mathbf{Y}'\mathbf{D}\mathbf{Y}\mathbf{M}\mathbf{a} = \mathbf{a}'\mathbf{M}\mathbf{V}\mathbf{M}\mathbf{a} = \mathbf{a}'\mathbf{M}\mathbf{A}\mathbf{a}.$$

Așadar, dacă $\mathcal{H} \equiv \Delta_{\mathbf{a}}$ atunci găsirea lui $\mathcal{H}$ s-a redus la următoarea problemă de programare pătratică cu restricții liniare:

$$\begin{cases} \max_{(\mathbf{a})} \{\mathbf{a}'\mathbf{M}\mathbf{A}\mathbf{a}\} \\ \mathbf{a}'\mathbf{M}\mathbf{a} = 1 \end{cases}.$$

---

[4] Dacă se lucrează pe tabelul **X** atunci problema se formulează astfel: „*Să se găsească* $\mathcal{H} \equiv \mathcal{F}_q$ *astfel încât* $\max_{(\mathcal{H})} \sum_{i=1}^{n} \mathbf{d}^2(\mathbf{e}_i, \mathbf{g})$ "

Pentru a rezolva problema de mai sus se utilizează metoda multiplicatorilor lui Lagrange. Fie deci lagrangeanul $\mathcal{L} = \mathbf{a}'\mathbf{M}\mathbf{A}\mathbf{a} - \lambda(\mathbf{a}'\mathbf{M}\mathbf{a} - 1)$ cu λ multiplicator Lagrange. Rezultă

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}} = 2\mathbf{M}\mathbf{A}\mathbf{a} - 2\lambda\mathbf{M}\mathbf{a}$$ căci $\mathbf{MA}$ este o matrice simetrică. Dar

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}} = 0 \quad \Rightarrow \quad \mathbf{M}\mathbf{A}\mathbf{a} = \lambda\mathbf{M}\mathbf{a} \quad (2).$$

Înmulțind la stânga relația (2) cu $\mathbf{a}'$ și ținând cont că $\mathbf{a}$ este $\mathbf{M}$-normat rezultă

$$\lambda = \mathbf{a}'\mathbf{M}\mathbf{A}\mathbf{a}.$$

Valoarea parametrului λ este deci maximul căutat. Cum matricea $\mathbf{M}$ este pozitiv definită ea este deci inversabilă și înmulțind la stânga cu $\mathbf{M}^{-1}$ relația (2) se obține

$$\mathbf{A}\mathbf{a} = \lambda\mathbf{a}$$

adică $\mathbf{a}$ este vector propriu al matricii $\mathbf{A}$ corespunzând celei mai mari valori proprii λ (dacă aceasta este unică); să le notăm cu $\mathbf{a}_1$ respectiv λ_1.

Să căutăm vectorul $\mathbf{a}_2$ din $\mathbb{R}^p$, $\mathbf{M}$-normat și $\mathbf{M}$-ortogonal pe $\mathbf{a}_1$ (adică $\mathbf{a}_2'\mathbf{M}\mathbf{a}_2 = 1$ si $\mathbf{a}_1'\mathbf{M}\mathbf{a}_2 = 0$) care maximizează forma pătratică $\mathbf{a}_2'\mathbf{M}\mathbf{A}\mathbf{a}_2$. Analog cu demersul de mai sus, se anulează derivatele lagrangeanului

$$\mathcal{L} = \mathbf{a}_2'\mathbf{M}\mathbf{A}\mathbf{a}_2 - \lambda_2(\mathbf{a}_2'\mathbf{M}\mathbf{a}_2 - 1) - \mu_2\mathbf{a}_2'\mathbf{M}\mathbf{a}_1.$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}_2} = 0 \qquad \Rightarrow \qquad 2\mathbf{M}\mathbf{A}\mathbf{a}_2 - 2\lambda_2\mathbf{M}\mathbf{a}_2 - \mu_2\mathbf{M}\mathbf{a}_1 = 0.$$

Înmulțind cu $\mathbf{a}_1'$, la stânga, relația de mai sus se obține

$$\mathbf{a}_1'\mathbf{M}\mathbf{A}\mathbf{a}_2 - \lambda_2\mathbf{a}_1'\mathbf{M}\mathbf{a}_2 - \mu_2\mathbf{a}_1'\mathbf{M}\mathbf{a}_1 = 0$$

sau

$$\lambda_2\mathbf{a}_1'\mathbf{M}\mathbf{a}_2 - \mu_2 = 0 \quad \Rightarrow \quad \mu_2 = 0.$$

Rămâne, ca în cazul precedent

$$\mathbf{M}\mathbf{A}\mathbf{a}_2 = \lambda_2\mathbf{M}\mathbf{a}_2$$

ceeace implică că $\mathbf{a}_2$ este al doilea vector al matricii $\mathbf{A}$ relativ la a doua valoare proprie λ_2, dacă aceasta este unică.

Demonstrația se repetă analog pentru ceilalți vectori $\mathbf{M}$-normați $\mathbf{a}_k \in \mathbb{R}^p$, $k \le q$, $\mathbf{M}$-ortogonali cu vectorii $\mathbf{a}_j$ găsiți înainte ($\mathbf{a}_k'\mathbf{M}\mathbf{a}_j = 0$ pentru $j < k$) și care maximizează forma pătratică $\mathbf{a}_k'\mathbf{M}\mathbf{A}\mathbf{a}_k$. Se obține $\mathbf{M}\mathbf{A}\mathbf{a}_k = \lambda_k\mathbf{M}\mathbf{a}_k$ și cum $\mathbf{M}$ este inversabilă $\mathbf{A}\mathbf{a}_k = \lambda_k\mathbf{a}_k$.

□

Observaţii

- Teorema 1.2-1 poate fi demonstrată folosind formula proiectorului **M**-ortogonal pe $\mathcal{H}$ (conform Saporta, 1990) sau bazându-se pe descompunerea $\mathbf{M} = \mathbf{T}'\mathbf{T}$ (conform cu Lebart et al., 1995).
- Cum **A** este o matrice **M**-simetrică, pozitiv definită, cu coeficienţi reali, valorile sale proprii sunt reale şi pozitive (conform cu Demidovitch&Maron, 1973). Vectorii proprii ai matricii **A** sunt **M**-ortonormaţi.

Definiţia 1.2-9 Matricea **A** se numeşte *matricea inerţiei*.

Definiţia 1.2-10 Imaginea euclidiană a norului de puncte-indivizi obţinută prin proiecţia pe subspaţiul $\mathcal{H}$ dat de Teorema 1.2-1 se numeşte *imaginea euclidiană a punctelor -indivizi asociaţi* aproximaţiei de ordinul q al produselor scalare.

Lema 1.2-7 $I_g = \mathrm{tr}(\mathbf{A}) = \sum_{j=1}^{p} \lambda_j$.

Demonstraţie. Matricea inerţiei **A** este reală şi **M**-simetrică. Atunci, conform Demidovitch&Maron, 1973,

$$\mathbf{A} = \mathbf{L}\Lambda\mathbf{L}^{-1}$$

cu **L** matricea vectorilor proprii corespunzători valorilor proprii $\lambda_1,\ldots,\lambda_p$ ai matricii **A** şi $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_p)$. Cu acestea

$$\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{L}\Lambda\mathbf{L}^{-1}) = \mathrm{tr}(\mathbf{L}\mathbf{L}^{-1}\Lambda)$$

căci $\mathrm{tr}(\mathbf{BC}) = \mathrm{tr}(\mathbf{CB})$ dacă produsele **BC** şi **CB** au sens. Rezultă

$$\mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\Lambda) = \mathrm{tr}(\mathrm{diag}(\lambda_1,\ldots,\lambda_p)) = \sum_{j=1}^{p} \lambda_j$$

dar din

Lema 1.2-4 rezultă $I_g = \mathrm{tr}(\mathbf{A})$.

□

Definiţia 1.2-11 Se numesc *axe principale de inerţie* vectorii proprii $\mathbf{a}_j$ ai matricii de inerţie **A**, **M**-normaţi.

Definiţia 1.2-12 Se numeşte *factor principal* asociat axei principale $\mathbf{a}_j$ şi se notează cu $\mathbf{u}_j$ forma liniară din $\mathbb{R}^p$ definită de relaţia $\mathbf{u}_j = \mathbf{M}\,\mathbf{a}_j$.

Lema 1.2-8 *Factorii principali* $\{\mathbf{u}_j\}_{j=1}^{p}$ *sunt vectorii proprii ai matricii* **MV** *asociați valorilor proprii* $\{\lambda_j\}_{j=1}^{p}$ *ai matricii* **A**.

Demonstrație. Într-adevăr

$$\mathbf{MVu}_j = \mathbf{MVMa}_j = \mathbf{MAa}_j = \lambda_j \mathbf{Ma}_j = \lambda_j \mathbf{u}_j$$

$$\mathbf{u}_j'\mathbf{M}^{-1}\mathbf{u}_k = \mathbf{a}_j'\mathbf{MM}^{-1}\mathbf{Ma}_k = \mathbf{a}_j'\mathbf{Ma}_k = \delta_{jk}$$

□

Definiția 1.2-13 Se numește *plan factorial principal* subsațiul $\mathcal{F}_2$ generat de vectorii $\{\mathbf{u}_1, \mathbf{u}_2\}$.

Definiția 1.2-14 Se numește *componentă principală* asociată factorului principal $\mathbf{u}_j$ și se notează cu $\mathbf{c}_j$ forma liniară din $\mathbb{R}^n$ definită de relația $\mathbf{c}_j = \mathbf{Yu}_j$.

Observație Din definiție $\mathbf{c}_j$ este proiecția **M**-ortogonală a indivizilor pe axa principală $\mathbf{a}_j$.

Lema 1.2-9 *Componentele principale* $\{\mathbf{c}_j\}_{j=1}^{n}$ *sunt vectorii proprii ai matricii* **WD** *asociați valorilor proprii* $\{\lambda_j\}_{j=1}^{p}$ *ai matricii* **A**. *Componentele principale sunt* **D***-ortogonale (deci necorelate).*

Demonstrație.

$$\mathbf{WDc}_j = \mathbf{YMY'DYMa}_j = \mathbf{YMVMa}_j = \mathbf{YMAa}_j$$
$$= \lambda_j \mathbf{YMa}_j = \lambda_j \mathbf{Yu}_j = \lambda_j \mathbf{c}_j$$

$$\mathbf{c}_j'\mathbf{Dc}_k = \mathbf{u}_j'\mathbf{Y'DYu}_k = \mathbf{u}_j'\mathbf{Vu}_k = \mathbf{a}_j'\mathbf{MVMa}_k$$
$$= \mathbf{a}_j'\mathbf{MAa}_k = \mathbf{a}_j'\mathbf{M}(\lambda_k \mathbf{a}_k) = \lambda_k \mathbf{a}_j'\mathbf{Ma}_k = \lambda_k \delta_{jk}$$

□

Lema 1.2-10

a) Mediile de selecție ale componentelor principale sunt nule (pe datele centrate și centrat-reduse).

b) Dispersia de selecție a componentei principale $\mathbf{c}_j$ *este* λ_j *- valoarea proprie a matricii inerției* **A** *pentru* $(\forall)\, j = \overline{1, p}$.

Demonstrație. Într-adevăr, cum $\mathbf{c}_j = \mathbf{Yu}_j$ atunci

$$\mathbf{m}(\mathbf{c}_j)=\sum_{i=1}^{n}p_i c_{ij}=\sum_{i=1}^{n}p_i\sum_{k=1}^{p}y_{ik}u_{kj}=\sum_{k=1}^{p}\left(\sum_{i=1}^{n}p_i y_{ik}\right)u_{kj}=\sum_{i=1}^{p}\mathbf{m}(\mathbf{y}_k)u_{kj}=0$$

conform **Corolarul 1.2-3**.

Analog, dacă $\mathbf{c}_j=\mathbf{Z}\mathbf{u}_j$.

b) $$s^2(\mathbf{c}_j)=\mathbf{c}_j'\mathbf{D}\mathbf{c}_j=\mathbf{u}_j'\mathbf{Y}'\mathbf{D}\mathbf{Y}\mathbf{u}_j=(\mathbf{a}_j'\mathbf{M})\mathbf{V}(\mathbf{M}\mathbf{a}_j)$$
$$=\mathbf{a}_j'\mathbf{M}\mathbf{A}\mathbf{a}_j=\mathbf{a}_j'\mathbf{M}(\lambda_j\mathbf{a}_j)=\lambda_j\mathbf{a}_j'\mathbf{M}\mathbf{a}_j=\lambda_j.$$

□

Propoziția 1.2-2

a) Componentele principale sunt combinații liniare de variabilele inițiale, de dispersie maximă și care satisfac restricțiile $\mathbf{u}_j'\mathbf{M}^{-1}\mathbf{u}_j=1$.

b) În cazul unei ACP normate, componentele principale $\{\mathbf{c}_i\}_{i=\overline{1,p}}$ *asociate valorilor proprii* $\{\lambda_i\}_{i=\overline{1,p}}$ *ale matricii* $\mathbf{A}$ *sunt variabilele cele mai „legate" de variabilele inițiale* $\mathbf{z}_1,\ldots,\mathbf{z}_p$ *în sensul că suma pătratelor coeficienților de corelație* $\{\mathrm{cor}(\mathbf{c}_j,\mathbf{z}_k)\}_{k=1}^{p}$ *este maximă, pentru oricare* $j=\overline{1,p}$.

Demonstrație. *a)* Să considerăm o combinație liniară de variabilele inițiale $\mathbf{x}_1,\ldots,\mathbf{x}_p$; fie aceasta $\mathbf{c}=\sum_{j=1}^{p}u_j\mathbf{x}_j$ sau vectorul $\mathbf{c}=\mathbf{X}\mathbf{u}$. Ne propunem să găsim pe $\mathbf{u}'=(u_1,\ldots,u_p)$ astfel încât

$$\begin{cases}\max\limits_{(u)} s^2(\mathbf{c})\\ \mathbf{u}'\mathbf{M}^{-1}\mathbf{u}=1\end{cases}\Leftrightarrow\begin{cases}\max\limits_{(u)}\mathbf{c}'\mathbf{D}\mathbf{c}\\ \mathbf{u}'\mathbf{M}^{-1}\mathbf{u}=1\end{cases}\Leftrightarrow\begin{cases}\max\limits_{(u)}\mathbf{u}'\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{u}\\ \mathbf{u}'\mathbf{M}^{-1}\mathbf{u}=1\end{cases}$$

Soluția problemei de programare pătratică cu restricții liniare de mai sus este, conform unui raționament analog cu cel din Teorema 1.2-1 vectorul propriu $\mathbf{u}_1$ al matricii $\mathbf{MV}$ asociat celei mai mari valori proprii λ_1 (cum $\mathbf{MV}$ este simetrică și pozitiv definită λ_1 există, este real și strict pozitiv). Dar $\mathbf{u}_1$ este, conform definiției, chiar factorul principal rezultat dintr-o ACP pe tabelul $\mathbf{X}$ iar valoarea maximă a funcției este λ_1.

b) Să reamintim mai întâi că, în cazul unei ACP normate, $\mathbf{X}\to\mathbf{Z}$ și $\mathbf{M}=\mathbf{I}_p$. Cu acestea

$$\mathrm{cor}^2(\mathbf{c},\mathbf{z}_j)=\frac{\mathrm{cov}^2(\mathbf{c},\mathbf{z}_j)}{s^2(\mathbf{c})s^2(\mathbf{z}_j)}=\frac{(\mathbf{c}'\mathbf{D}\mathbf{z}_j)^2}{s^2(\mathbf{c})}.$$

$$\sum_{j=1}^{p}\mathrm{cor}^2(\mathbf{c},\mathbf{z}_j)=\frac{1}{s^2(\mathbf{c})}\sum_{j=1}^{p}(\mathbf{c}'\mathbf{D}\mathbf{z}_j)(\mathbf{c}'\mathbf{D}\mathbf{z}_j)'=\frac{1}{s^2(\mathbf{c})}\mathbf{c}'\mathbf{D}\left(\sum_{j=1}^{p}\mathbf{z}_j\mathbf{z}_j'\right)\mathbf{D}\mathbf{c}$$

şi cum $\sum_{j=1}^{p}\mathbf{z}_j\mathbf{z}_j'=\mathbf{Z}\mathbf{Z}'$ rezultă $\sum_{j=1}^{p}\mathrm{cor}^2(\mathbf{c},\mathbf{z}_j)=\frac{\mathbf{c}'\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{c}}{\mathbf{c}'\mathbf{D}\mathbf{c}}$. Problema s-a redus la a găsi $\max_{(\mathbf{c})}\frac{\mathbf{c}'\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{c}}{\mathbf{c}'\mathbf{D}\mathbf{c}}$.

Să remarcăm că $\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D}$ şi $\mathbf{D}$ sunt matrici reale, simetrice şi de ordin n. Un punct de extrem al câtului de mai sus se obţine anulând derivata sa ceeace implică

$$\frac{(\mathbf{c}'\mathbf{D}\mathbf{c})(2\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{c})-(\mathbf{c}'\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{c})(2\mathbf{D}\mathbf{c})}{(\mathbf{c}'\mathbf{D}\mathbf{c})^2}=0.$$

Din $(\mathbf{c}'\mathbf{D}\mathbf{c})\,\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{c}=(\mathbf{c}'\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{c})\mathbf{D}\mathbf{c}$ rezultă $\mathbf{D}^{-1}(\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D})\mathbf{c}=\left(\frac{\mathbf{c}'\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{c}}{\mathbf{c}'\mathbf{D}\mathbf{c}}\right)\mathbf{c}$. $\mathbf{c}$ este deci vectorul propriu al matricii $\mathbf{Z}\mathbf{Z}'\mathbf{D}$ asociat valorii proprii $\lambda=\left(\frac{\mathbf{c}'\mathbf{D}\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{c}}{\mathbf{c}'\mathbf{D}\mathbf{c}}\right)$.

Maximul este deci atins dacă această valoare proprie este cea mai mare.

Din ipoteză $\mathbf{c}$ este o combinaţie liniară de variabile iniţiale, adică $\mathbf{c}=\mathbf{Z}\mathbf{u}$. Înlocuind în relaţia de mai sus se obţine

$$\mathbf{Z}\mathbf{Z}'\mathbf{D}\mathbf{Z}\mathbf{u}=\lambda\mathbf{Z}\mathbf{u}$$

şi cum $\mathbf{Z}'\mathbf{D}\mathbf{Z}=\mathbf{R}\ \Rightarrow\ \mathbf{Z}\mathbf{R}\mathbf{u}=\lambda\mathbf{Z}\mathbf{u}$ iar $\mathbf{Z}$ este de rang p rezultă $\mathbf{R}\mathbf{u}=\lambda\mathbf{u}$ adică $\mathbf{u}$ este vectorul propriu al matricii $\mathbf{R}$ asociat valorii proprii maxime. În ACP normat $\mathbf{A}=\mathbf{R}$ şi axele principale coincid cu factorii principali, deci $\mathbf{c}=\mathbf{Z}\mathbf{u}$ este chiar componenta principală obţinută prin proiecţia indivizilor pe axa principală $\mathbf{a}\equiv\mathbf{u}$.

□

Un rezumat al elementelor principale ce intervin într-o ACP pe norul de puncte-indivizi se găseşte în tabelul de mai jos.

Elemente principale	**Definiţie**	**Proprietăţi**	**Relaţii**
Axe principale: $\mathbf{a}\in\mathbb{R}^p$	$\mathbf{V}\mathbf{M}\mathbf{a}=\lambda\mathbf{a}$	$\mathbf{M}$-ortonormate	
Factori principali: $\mathbf{u}\in(\mathbb{R}^p)^*$	$\mathbf{u}=\mathbf{M}\mathbf{a}$	$\mathbf{M}^{-1}$-ortonormaţi	$\mathbf{M}\mathbf{V}\mathbf{u}=\lambda\mathbf{u}$
Componente principale: $\mathbf{c}\in\mathbb{R}^n$	$\mathbf{c}=\mathbf{Y}\mathbf{u}$ sau $\mathbf{c}=\mathbf{Z}\mathbf{u}$	$\mathbf{D}$-ortogonale $s^2(\mathbf{c})=\lambda$	$\mathbf{W}\mathbf{D}\mathbf{c}=\lambda\mathbf{c}$ şi analoaga

Tabelul 1.2-1 Proprietăţile elementelor principale dintr-o ACP pe norul de puncte-indivizi.

1.2.2.2 Analiza norului de puncte-variabile

Să considerăm acum spațiul $\mathcal{E} \subseteq \mathbb{R}^n$ al variabilelor în care tabelul **Y** poate fi reprezentat ca un nor de p puncte-variabile ale căror n coordonate reprezintă coloanele lui **Y**.

Principiul metodei ACP în acest caz este identic cu cel utilizat pentru reprezentarea norului de puncte-indivizi şi constă în găsirea axelor principale şi a subspațiului afin $\mathcal{E}_q$ generat de aceste axe, de dimensiune q din $\mathbb{R}^n$ care aproximează optim norul de puncte-variabile. Aceasta înseamnă să fie maximizată media pătratelor distanțelor dintre cele p proiecții pe $\mathcal{E}_q$, adică de rezolvat problema de programare pătratică cu restricții liniare

$$\begin{cases} \max\limits_{(\mathbf{b})} \mathbf{b'DYMY'Db} \\ \mathbf{b'Db} = 1 \end{cases}$$

Teorema 1.2-1 arată că **b** este vectorul propriu al matricii $\mathbf{B} = \mathbf{YMY'D}$ (**D**-simetrică, reală) corespunzând celei mai mari valori proprii μ. Ecuația axei factoriale **b** din $\mathbb{R}^n$ este:

$$\begin{cases} \mathbf{YMY'Db} = \mu\mathbf{b} \\ \mathbf{b'Db} = 1 \end{cases}$$

ecuația factorului principal **v** din $\left(\mathbb{R}^n\right)^*$ este **v=Db** şi ecuația componentei principale **d** din $\mathbb{R}^p$ este $\mathbf{d} = \mathbf{Y'v}$ sau $\mathbf{d} = \mathbf{Z'v}$. Analog ca în cazul norului de puncte-indivizi se poate enunța:

Lema 1.2-11

a) Factorii principali $\mathbf{v}_i \in \left(\mathbb{R}^n\right)^*, i = \overline{1,n}$ *sunt* $\mathbf{D}^{-1}$ *ortonormați şi satisfac relațiile* $\mathbf{DYMY'v}_i = \mu_i \mathbf{v}_i$.

b) *Componentele principale* $\mathbf{d}_i \in \mathbb{R}^n, i = \overline{1,n}$ *sunt* **M***-ortogonale, au dispersia de selecție egală cu* μ_i *şi satisfac relațiile* $\mathbf{X'DXMd}_i = \mu_i \mathbf{d}_i$.

Demonstrație. a) Într-adevăr

$$\mathbf{v}_i' \mathbf{D}^{-1} \mathbf{v}_j = \mathbf{b}_i' \mathbf{DD}^{-1} \mathbf{b}_j = \mathbf{b}_i' \mathbf{Db}_j = \delta_{ij} \text{ şi}$$

$$\mathbf{DYMY'v} = \mathbf{DYMY'Db} = \mu\mathbf{Db} = \mu\mathbf{v}.$$

b) Într-adevăr

$$\mathbf{d}_i' \mathbf{Md}_j = \mathbf{v}_i' \mathbf{YMY'v}_j = \mathbf{b}_i' \mathbf{D}\left(\mathbf{YMY'Db}_j\right) = \mathbf{b}_i' \mathbf{D}\left(\mu_j \mathbf{b}_j\right) = \mu_j \left(\mathbf{b}_i' \mathbf{Db}_j\right) = \mu_j \delta_{ij}$$

$$\mathbf{X'DXMd} = \mathbf{X'DXMX'v} = \mathbf{X'D}\left(\mathbf{XMX'Db}\right) = \mathbf{X'D}\left(\mu\mathbf{b}\right) = \mu\mathbf{d}$$

$$s^2(\mathbf{d}) = \mathbf{d'Md} = \mathbf{v'XMX'v} = \mathbf{b'DXMX'Db} = \mathbf{b'D}\left(\mu\mathbf{b}\right) = \mu\mathbf{b'Db} = \mu.$$

□

Definiția 1.2-15 Se numește *cerc de corelație principal* subspațiul $\mathcal{E}_2$ generat de vectorii $\{\mathbf{v}_1, \mathbf{v}_2\}$.

În cazul ACP normat norul de puncte-variabile aflându-se pe hipersfera de corelație planul factorial o va intersecta după un cerc diametral (vezi și **Corolarul 1.2-3** și observația 4).

Un rezumat al elementelor principale ce intervin într-o ACP pe norul de puncte-variabile se găsește în tabelul de mai jos.

Elemente principale	**Definiție**	**Proprietăți**	**Relații**
Axe principale: $\mathbf{b} \in \mathbb{R}^n$	$\mathbf{YMY'Db} = \mu\mathbf{b}$	$\mathbf{D}$-ortonormale	
Factori principali: $\mathbf{v} \in (\mathbb{R}^n)^*$	$\mathbf{v} = \mathbf{Db}$	$\mathbf{D}^{-1}$-ortonormați	$\mathbf{DYMY'v} = \mu\mathbf{v}$
Componente principale: $\mathbf{d} \in \mathbb{R}^p$	$\mathbf{d} = \mathbf{Y'v}$ sau $\mathbf{d} = \mathbf{Z'v}$	$\mathbf{M}$-ortogonale $s^2(\mathbf{d}) = \mu$	$\mathbf{X'DXMd} = \mu\mathbf{d}$ și analoaga

Tabelul 1.2-2 Proprietățile elementelor principale dintr-o ACP pe norul de puncte-variabile.

1.2.2.3 Relații de tranziție între cele două spații

Se observă că, din punct de vedere numeric, o analiză în componente principale a unui studiu se reduce la calculul primelor q valori și vectori proprii asociați ai matricilor $\mathbf{Y'DYM} \in \mathcal{M}_{p,p}(\mathbb{R})$ și $\mathbf{YMY'D} \in \mathcal{M}_{n,n}(\mathbb{R})$. O întrebare naturală este următoarea: există o relație între elementele principale dintr-o ACP pe spațiul $(\mathcal{F}, \mathbf{M})$ și elementele principale dintr-o ACP pe spațiul $(\mathcal{E}, \mathbf{D})$? Răspunsul la această întrebare este dat de următoarea:

Propoziția 1.2-3 *(relația de tranziție între spațiul indivizilor și spațiul variabilelor) Toate valorile proprii ale matricilor* $\mathbf{Y'DYM}$ *și* $\mathbf{YMY'D}$ *nenule sunt egale (cu același ordin de multiplicitate eventual) și pentru* $\lambda_j \neq 0$ *sunt adevărate următoarele relații de tranziție între cele două spații* $\mathcal{F} \subseteq \mathbb{R}^p$ *și* $\mathcal{E} \subseteq \mathbb{R}^n$*:*

$$\begin{cases} \mathbf{b}_j = \frac{1}{\sqrt{\lambda_j}}\mathbf{YMa}_j = \frac{1}{\sqrt{\lambda_j}}\mathbf{Yu}_j = \frac{1}{\sqrt{\lambda_j}}\mathbf{c}_j \\ \qquad\qquad\qquad\qquad\qquad\qquad j = \overline{1, \mathrm{rg}(\mathbf{Y'Y})}. \\ \mathbf{a}_j = \frac{1}{\sqrt{\lambda_j}}\mathbf{Y'Db}_j = \frac{1}{\sqrt{\lambda_j}}\mathbf{Y'v}_j = \frac{1}{\sqrt{\lambda_j}}\mathbf{d}_j \end{cases}$$

Demonstrație. În $\mathbb{R}^p$ există relația

$$\mathbf{Y'DYM}\mathbf{a}_j = \lambda_j \mathbf{a}_j \qquad (1)$$

iar în $\mathbb{R}^n$ relația

$$\mathbf{YMY'D}\mathbf{b}_j = \mu_j \mathbf{b}_j \qquad (2).$$

Înmulțind la stânga egalitatea (1) cu **YM** se obține

$$\mathbf{YMY'D}\left(\mathbf{YM}\mathbf{a}_j\right) = \lambda_j\left(\mathbf{YM}\mathbf{a}_j\right) \qquad (3)$$

relație care arată că oricărui vector propriu $\mathbf{a}_j$ a lui **Y'DYM** corespunzător unei valori proprii $\lambda_j \neq 0$ îi corespunde un vector propriu $\mathbf{YM}\mathbf{a}_j$ al matricii **YMY'D** relativ la aceeaşi valoare proprie λ_j. Cum cu μ_1 a fost notată valoarea proprie maximă a matricii **YMY'D** rezultă, în mod necesar, că $\lambda_1 \leq \mu_1$.

Înmulțind la stânga egalitatea (2) cu **Y'D** se obține

$$\left(\mathbf{Y'DYM}\right)\left(\mathbf{Y'D}\mathbf{b}_j\right) = \mu_j\left(\mathbf{Y'D}\mathbf{b}_j\right) \qquad (4)$$

relație care arată că oricărui vector propriu $\mathbf{b}_j$ a lui **YMY'D** corespunzător unei valori proprii $\mu_j \neq 0$ îi corespunde un vector propriu $\mathbf{Y'D}\mathbf{b}_j$ al matricii **Y'DYM** relativ la aceeaşi valoare proprie μ_j. Cum cu λ_1 a fost notată valoarea proprie maximă a matricii **Y'DYM** rezultă că $\mu_1 \leq \lambda_1$ ceeace arată, în final, că $\lambda_1 = \mu_1$.

Analog se poate arăta că toate valorile proprii ne-nule ale celor două matrici **Y'DYM** şi **YMY'D** sunt egale (cu acelaşi ordin de multiplicitate eventual), adică:

$$\lambda_j = \mu_j \neq 0 \quad j = \overline{1, \mathrm{rg}\left(\mathbf{Y'Y}\right)}$$
$$\lambda_j = 0 \quad j = \overline{\mathrm{rg}\left(\mathbf{Y'Y}\right)+1, p}\ ^{5}.$$
$$\mu_i = 0 \quad j = \overline{\mathrm{rg}\left(\mathbf{Y'Y}\right)+1, n}$$

(se poate arăta uşor, având în vedere proprietățile matricilor **M** şi **D**, că $\mathrm{rg}\left(\mathbf{Y'DYM}\right) = \mathrm{rg}\left(\mathbf{Y'Y}\right) = \mathrm{rg}\left(\mathbf{YMY'D}\right) = \mathrm{rg}\left(\mathbf{YY'}\right)$).

Revenind la relația (3) se observă că aceasta este verificată de orice vector de forma $\mathbf{b} = k\,\mathbf{YM}\mathbf{a}$, cu k constantă ce se determină din condiția de **D**-ortonormalitate a lui **b**. Într-adevăr:

$$1 = \mathbf{b'Db} = k^2\,\mathbf{a'MY'DYM}\mathbf{a} = k^2\,\mathbf{a'M}\left(\lambda\mathbf{a}\right) = k^2\,\lambda\,\mathbf{a'Ma} = k^2\,\lambda$$

ceeace implică $k = \dfrac{1}{\sqrt{\lambda}}$ deci $\mathbf{b} = \frac{1}{\sqrt{\lambda}}\mathbf{YM}\mathbf{a}$ dacă $\lambda \neq 0$.

[5] În relațiile alăturate ca şi în cele ce urmează se utilizează convenția: în *Relație(j)*, $j = \overline{a,b}$ dacă $a>b$ atunci *Relație(j)* nu există.

Analog, relaţia (4) este verificată de orice vector de forma $\mathbf{a} = k\,\mathbf{Y}'\mathbf{D}\mathbf{b}$ cu k constantă ce se determină din condiţia de **M**-ortonormalitate a lui **a**. Se obţine $k = \frac{1}{\sqrt{\mu}}$ deci $\mathbf{b} = \frac{1}{\sqrt{\mu}}\mathbf{Y}\mathbf{M}\mathbf{a} = \frac{1}{\sqrt{\lambda}}\mathbf{Y}\mathbf{M}\mathbf{a}$ pentru $\mu = \lambda \neq 0$.

□

Observaţii

1. Propoziţia 1.2-3 demonstrează că este suficient să calculăm valorile şi vectorii proprii ai matricii cu dimensiunea cea mai mică iar apoi, prin relaţiile de tranziţie, să obţinem elementele pincipale din celălalt spaţiu. Cum, în general, numărul de variabile este mai mic decât numărul de indivizi, adică $p<n$, este suficient ca analiza în componente principale să se efectueze pe norul de puncte-indivizi elementele principale pentru norul de puncte-variabile obţinându-se prin relaţiile de tranziţie.
2. Coordonalele punctelor pe o axă factorială în $\mathbb{R}^p$ sunt proporţionale cu componentele axei factoriale din $\mathbb{R}^n$ corespunzătoare aceleiaşi valori proprii şi reciproc. Într-adevăr $\mathbf{c} = \mathbf{X}\mathbf{u}$ şi $\mathbf{d} = \mathbf{X}'\mathbf{v}$ şi ţinând cont de relaţiile de tranziţie $\mathbf{c} = \sqrt{\lambda}\mathbf{b}$ şi $\mathbf{d} = \sqrt{\lambda}\mathbf{a}$.

Referitor la analiza în componente principale trebuie să remarcăm:

3. Orientarea axelor factoriale este arbitrară deoarece vectorii proprii sunt determinaţi modulo semnul lor. Acest lucru nu împietează asupra formei norului, adică a distanţelor între puncte.
4. Analiza în componente principale nu pune în evidenţă decât legăturile liniare între variabile. Un coeficient de corelaţie slab între două variabile semnifică doar că acestea sunt independente liniar în timp ce poate exista o relaţie de ordin superior lui 1 (relaţie neliniară).
5. Coordonata unui punct-variabilă $\mathbf{z}_k$ pe axa $\mathbf{b}_j$ este mai mică sau egală cu 1 în valoare absolută (nefiind altceva decât coeficientul de corelaţie al variabilei cu factorul $\mathbf{v}_j$ considerat ca o variabilă artificială ale cărui coordonate sunt date de cele n proiecţii ale indivizilor pe această axă, conform relaţiilor de tranziţie). În plus, în cazul datelor centrat-reduse, $\sum_{j=1}^{p} \mathrm{cor}^2\left(\mathbf{z}_k, \mathbf{v}_j\right) = \mathbf{a}'\mathbf{M}\mathbf{a} = 1$.

1.2.2.4 Reconstituirea datelor iniţiale

Metodele de analiză factorială rezidă toate pe reprezentarea geometrică a unei proprietăţi a matricilor dreptunghiulare şi anume *descompunerea în valori singulare*. Descompunerea a fost obţinută de Eckart şi Young în 1936 pentru matrici dreptunghiulare şi generalizează lucrările lui Sylvester din 1889 relativ la matrici

pătratice; Gifi, 1990 menționează, relativ la această problematică, și lucrările lui Beltrami din 1873 și independent ale lui Jordan din 1874.

Descompunerea în valori singulare semnifică, în principal, că, în condiții destul de generale, o matrice dreptunghiulară poate fi reprezentată în mod unic ca o „sumă optimală" (în sensul minimului celor mai mici pătrate) de matrici de rang 1 (produse de matrici coloană cu matrici linie). În cazul nostru, pornind de la relația $\mathbf{c}=\mathbf{Yu}$, înmulțind la dreapta membrii egalității cu $\mathbf{u}'\mathbf{M}^{-1}$ și sumând după numărul de axe[6] se obține $\mathbf{Y}\left\{\sum_{j=1}^{p}\mathbf{u}_j\mathbf{u}_j'\mathbf{M}^{-1}\right\}=\sum_j\mathbf{c}_j\mathbf{u}_j'\mathbf{M}^{-1}$. Dar $\sum_{j=1}^{p}\mathbf{u}_j\mathbf{u}_j'\mathbf{M}^{-1}=\mathbf{I}_p$ căci $\mathbf{u}_j$ sunt $\mathbf{M}^{-1}$-ortonormați deci

$$\mathbf{Y}=\sum_{j=1}^{p}\mathbf{c}_j\mathbf{u}_j'\mathbf{M}^{-1}.$$

Relația de mai sus se numește *formula de reconstituire* a tabelului de date $\mathbf{Y}$ pornind de la componentele și factorii principali. Analog, se poate reconstitui tabelul $\mathbf{X}$ și de asemenea

$$\mathbf{MV}=\sum_{j=1}^{p}\lambda_j\mathbf{u}_j\mathbf{u}_j'\mathbf{M}^{-1}$$

și

$$\mathbf{VM}=\sum_{j=1}^{p}\lambda_j\mathbf{a}_j\mathbf{a}_j'\mathbf{M}.$$

Dacă $\mathbf{M}=\mathbf{I}$, adică în cazul metricii euclidiene, axele principale coincid cu factorii principali și, conform formulelor de tranziție, se obține formula de reconstituire $\mathbf{Y}=\sum_{j=1}^{p}\mathbf{c}_j\mathbf{u}_j'=\sum_{j=1}^{p}\sqrt{\lambda_j}\mathbf{v}_j\mathbf{u}_j'$ cu $\mathbf{v}_j$ vectori proprii normați ai matricii $\mathbf{YY}'$ și $\mathbf{u}_j$ vectori proprii normați ai matricii $\mathbf{Y}'\mathbf{Y}$.

Dacă în formula de mai sus sumarea se face doar după primii $q<p$ termeni atunci se obține cea mai bună aproximare a lui $\mathbf{Y}$ printr-o matrice de rang q în sensul celor mai mici pătrate (desigur dacă în sumarea de mai sus valorile proprii sunt ordonate descrescător). Să observăm că, privite doar din acest punct de vedere metodele de analiză factorială se reduc la metode de compresie a datelor.

[6] Unii vectori proprii $\mathbf{b}$ pot să corespundă unei valori proprii nule; ei sunt atunci aleși astfel încât să completeze baza ortonormată formată din axele precedente.

1.2.3 INTERPRETAREA ŞI CALITATEA REZULTATELOR UNEI ACP

ACP construieşte variabile noi, artificiale şi reprezentări grafice ce permit vizualizarea relaţiilor între variabile şi a eventualelor grupe de indivizi şi de variabile. Interpretarea rezultatelor este o fază delicată ce trebuie întreprinsă respectând următoarele etape:

- studiul calităţii reprezentărilor în planurile factoriale;
- interpretarea rezultatelor pornind de la datele utilizate în ACP (interpretarea „internă");
- interpretarea rezultatelor pornind de la indivizi şi / sau variabile suplimentare care nu au fost utilizate în construirea reprezentărilor ACP (interpretarea „externă");
- reprezentarea simultană a indivizilor şi variabilelor ce fac obiectul ACP.

1.2.3.1 Calitatea reprezentărilor în planurile factoriale

Axele factoriale permit obţinerea celei mai bune vizualizări aproximative (în sensul celor mai mici pătrate) a distanţelor dintre indivizi, de o parte şi între variabile, de cealaltă parte. În acest sens primul demers care se impune este legat de măsurarea calităţii acestei aproximări.

Se observă că, dacă ultimele p-q valori proprii ordonate în prealabil descrescător, ale matricii $\mathbf{Y}$ sunt considerate „neglijabile" atunci, conform descompunerii în valori singulare $\mathbf{Y} \approx \mathbf{Y}^* = \sum_{j=1}^{q} \sqrt{\lambda_j}\, \mathbf{v}_j \mathbf{u}'_j$ în cazul metricii euclidiene. Aceasta înseamnă că cei np coeficienţi ai matricii $\mathbf{Y}$ pot fi reprezentaţi doar prin cei $q(n+p)$ termeni ai sumei de mai sus, ceeace reprezintă, din punct de vedere numeric, un câştig important dacă $q<<p$. Cu acestea o măsură naturală a calităţii aproximării este dată de raportul:

$$\tau_q = \frac{\sum_i \sum_j p_i p_j \left(y_{ij}^*\right)^2}{\sum_i \sum_j p_i p_j y_{ij}^2}$$

sau încă

$$\tau_q = \frac{\mathrm{tr}\left(\left(\mathbf{Y}^*\right)' \mathbf{D}\mathbf{Y}^*\right)}{\mathrm{tr}\left(\mathbf{Y}'\mathbf{D}\mathbf{Y}\right)} = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\sum_{j=1}^{q} \lambda_j}{I_g}$$

conform
Lema 1.2-7.

Raportul $\tau_q \leq 1$ se numeşte *rata inerţiei* (sau *procentul de dispersie* datorat primilor q factori. Interpretarea sa ca măsură a calităţii numerice a aproximării este destul de

clară dar semnificația sa statistică este delicată. Intr-adevăr, din punct de vedere statistic, interpretarea rapotului τ_q comportă două aspecte:

- alegerea numărului de axe principale;
- găsirea intervalului de încredere pentru dispersia coordonantelor punctelor-indivizi pe axa principală corespunzătoare.

Principalul scop al ACP constând în reducerea dimensiunii spațiului indivizilor, alegerea numărului de axe principale ce trebuie reținut, adică a lui q, este o problemă importantă care, din păcate, nu are o soluție riguroasă. Să remarcăm, mai înainte de toate, că reducerea dimensiunii nu este posibilă decât dacă există o redondanță între variabilele $\mathbf{x}_1,...,\mathbf{x}_p$; dacă acestea sunt independente, ceeace este un rezultat important în sine, ACP va fi ineficientă în reducerea dimensiunii. Există mai multe proceduri care să ghideze alegerea numărului de axe (vezi Lebart et al., 1995). În cele ce urmează ne vom opri asupra:

a) regulilor empirice, şi
b) criteriilor bazate pe anumite proprietăți statistice ale valorilor proprii.

a-Reguli empirice

Regulile empirice se bazează pe forma secvenței de valori proprii; două reguli, atribuite lui Cattell şi respectiv Kaiser vor fi citate cu titlu istoric.

Regula „cotului" (sau *the scree-test* introdusă de Cattell în 1966 (vezi, de exemplu, Saporta, 1990) constă în studiul histogramei valorilor proprii ordonate descrescător în vederea decelării unei schimbări de pantă urmând a fi reținute acele valori proprii, deci număr de axe, aflate la stânga punctului „de discontinuitate" observat.

Fundamentarea criteriului cotului este dată de observația empirică că valorile proprii descresc regulat dacă datele sunt puțin structurate (variabilele nu sunt prea corelate între ele); se poate deci presupune că a intervenit un factor de structurare de fiecare dată când diagrama valorilor proprii prezintă o schimbare evidentă de pantă.

Al doilea criteriu empiric este cel enunțat de Kaiser în 1961 (vezi, de exemplu, Saporta, 1990) care recomandă reținerea acelor valori proprii superioare mediei tuturor valorilor proprii (să remarcăm, conform

Lema 1.2-7 şi observației 2) din secțiunea 1.1.1, că în cazul ACP normate media valorilor este 1). Datorită simplității sale acest criteriu este foarte răspândit şi implementat drept criteriu standard în majoritatea pachetelor de programe de analiză factorială.

b-Criterii bazate pe proprietățile statistice ale valorilor proprii

Lucrările relative la studiul distribuției valorilor şi vectorilor proprii cât şi lucrările relative la comportamentul asimptotic al acestor elemente sunt în număr mare dar puține rezultate sunt practic utilizabile. Cu excepția mențiunilor explicite toate rezultatele ce vor fi prezentate presupun că observațiile, în număr de n, urmează o lege normală p-

dimensională $\mathbf{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Bartlett în 1951 propune o metodă pentru testarea egalității a *p-q* valori proprii ale matricilor $\boldsymbol{\Sigma}$ sau **R**. Lawley în 1956 aprofundează studiul la cazul celor mai mici *p-q* valori proprii ale lui $\boldsymbol{\Sigma}$. Anderson, 1963 generalizează aceste rezultate şi determină legile limită ale valorilor proprii fără să presupună în mod necesar că valorile teoretice corespunzătoare sunt distincte. El demonstrează în particular, pentru a testa egalitatea celor mai mici *r* valori proprii $\hat{\lambda}_j$ ale matricii de covarianță de selecție corectate $\mathbf{V}^* = \frac{n}{n-1}\mathbf{V}$ că statistica

$$X^2 = nr\log\frac{\left(\frac{1}{r}\right)\sum_{j=p-r+1}^{p}\hat{\lambda}_j}{\left(\prod_{j=p-r+1}^{p}\hat{\lambda}_j\right)^{1/r}}$$

este asimptotic distribuită χ^2 cu $\frac{r(r+1)}{2}-1$ grade de libertate.

Legat de găsirea intervalului de încredere pentru dispersia coordonatelor punctelor-indivizi pe axa principală reamintim că aceasta este egală cu valoarea proprie corespunzătoare (conform Lema 1.2-10). Dacă valorile teoretice λ_j ale lui $\boldsymbol{\Sigma}$ sunt distincte, T.W.Anderson a arătat că $\sqrt{n-1}\left(\hat{\lambda}_j-\lambda_j\right)$ converge către o lege normală $N\left(0,2\lambda_j^2\right)$. Se deduce imediat că intervalul de încredere cu pragul 95% este:

$$\hat{\lambda}_j\left(1-1{,}96\sqrt{2/(n-1)}\right) < \lambda_j < \hat{\lambda}_j\left(1+1{,}96\sqrt{2/(n-1)}\right).$$

Lungimea intervalului este o indicație asupra stabilității valorii proprii față de fluctuațiile eşantionului presupus repartizat gaussian. Intersecția intervalelor a două valori proprii consecutive sugerează deci egalitatea acestor valori proprii. Axele corespunzătoare sunt atunci definite modulo o rotație ceeace permite utilizatorului să evite interpretarea unei axe instabile după acest criteriu.

O îmbunătățire a criteriului lui Kaiser este dată de Enăchescu & Enăchescu, 2000. Aceştia demonstrează că, în cazul analizei în componente principale normate $\hat{\lambda}_i$ este semnificativ mai mare ca unu dacă

$$\hat{\lambda}_i > 1+2\sqrt{\frac{p-1}{n-1}}.$$

Generalizări ale rezultatelor asimptotice ale lui T.W.Anderson la cazul ne-gaussian se pot găsi, printre alții, în Davis, 1977 fără însă o utilizare practică.

Intervalele de încredere ale lui Anderson se referă atât la valorile proprii ale matricilor de covarianță cât şi la valorile proprii ale matricilor de corelație. Simulările întreprinse au arătat că rezultatele obținute sunt în general „prudente": procentul de acoperire al adevăratei valori proprii este cel mai adesea superior pragului de

semnificație anunțat (Lebart et al., 1995). În orice caz, natura asimptotică a rezultatelor ca și ipoteza subiacentă de normalitate fac ca acestea să aibă doar un caracter indicativ.

Concluzionând asupra calității reprezentărilor în planurile factoriale vom spune că rata inerției definește „puterea explicativă" a factorilor; ea reprezintă partea din dispersia totală datorată celor q factori reținuți. Această apreciere trebuie să țină cont atât de numărul de indivizi cât și de numărul de variabile; o rată de inerție, relativ la o axă, de 10% poate fi o valoare importantă dacă tabelul posedă 100 de variabile și poate fi o valoare neglijabilă dacă nu sunt decât 10 variabile. Rata inerției este deci o măsură pesimistă a calității proiecției imaginii euclidiene a indivizilor. Rata inerției este în plus o măsură globlă a calității reprezentării în planul factorial; ea trebuie completată cu alte măsuri, locale, ale calității acestei reprezentări.

Printre măsurile locale cele mai „populare" se numără cea a cosinusului pătrat a unghiului dintre vectorul cu originea în proiecția centrului de greutate a norului și cu vârful în punctul-individ și planul factorial. Fundamentarea teoretică a utilizării acestei măsuri se bazează pe faptul că distanțele între puncte se deformează prin proiecție cu atât mai puțin cu cât punctele sunt mai apropiate de planul în care sunt proiectate (desigur, cazul în care punctele se află pe o dreaptă paralelă cu planul de proiecție este neinteresant în acest context). Valoarea acestei măsuri este dată de următoarea:

Lema 1.2-12 *Calitatea reprezentării unui punct-individ A_i în planul factorial principal este*

$$\text{cal}(i) = \frac{c_{1i}^2 + c_{2i}^2}{\sum_{j=1}^{p} c_{ji}^2}.$$

Demonstrație Fie A_i punctul considerat și P_i proiecția sa în planul factorial principal. Conform definiției componentelor principale, în $\mathcal{F}$, A_i are coordonatele $(c_{1i}, c_{2i}, \ldots, c_{pi})'$ iar P_i are coordonatele $(c_{1i}, c_{2i})'$.

Cosinusul unghiului dintre $\overrightarrow{GA_i}$ și planul factorial principal este cosinusul unghiului dintre $\overrightarrow{GA_i}$ și $\overrightarrow{GP_i}$, notat cu θ. În $\Delta A_i P_i G$ dreptunghic în P_i (din construcție) $\cos^2\theta = \frac{GP_i^2}{GA_i^2}$ și cum, conform teoremei lui Pitagora, $\overline{GP_i}^2 = c_{1i}^2 + c_{2i}^2$ rezultă

$$\cos^2\theta = \frac{c_{1i}^2 + c_{2i}^2}{\sum_{j=1}^{p} c_{ji}^2}.$$

□

Un mod mai bun de a afla dacă o observație este bine reprezentată într-un subspațiu este de a da o interpretare statistică pătratului distanței la acel subspațiu. Astfel pentru observații repartizate normal, inerția globală este o sumă ponderată de p variabile independente repartizate χ_1^2, adică $I_g = \sum_{j=1}^{p} \lambda_j \chi_{1,j}^2$. Cum, în această ipoteză $\mathrm{E}\left[I_g\right] = \sum_{j=1}^{p} \lambda_j$ și $\mathrm{D}^2\left[I_g\right] = 2\sum_{j=1}^{p} \lambda_j^2$ iar pe de altă parte distanța de la un punct-individ la planul factorial principal este

$$\mathrm{d}^2\left(A_i, \left\{\mathbf{w} \in \mathbb{R}^2 \,|\, \mathbf{w} = a_1\mathbf{u}_1 + a_2\mathbf{u}_2\right\}\right) = \sum_{j=3}^{p} c_{ji}^2 = \sum_{j=3}^{p} \lambda_j \frac{c_{ji}^2}{\lambda_j}$$

o modalitate de a da o semnificație statistică acestei distanțe este de a o compara cu o combinație liniară de χ_1^2. Utilizând intervalele de încredere de tip 2σ se poate conchide că punctele aflate față de planul factorial principal la o distanță mai mare de

$$\sum_{j=3}^{p} \lambda_j + 2\sqrt{2\sum_{j=3}^{p} \lambda_j^2}$$

sunt prost reprezentate în acest subspațiu cu o probabilitate de 95% (conform Enăchescu & Enăchescu, 2000).

Datorită egalității $\lambda_j = \mu_j$ (conform

Propoziția 1.2-3) măsura globală a calității proiecției imaginii euclidiene a norului de puncte-variabile este tot τ_q cu aceleași observații ca pentru norul de puncte-indivizi.

În ceeace privește măsurile locale trebuie să remarcăm că, în cazul punctelor variabile interesează unghiurile dintre proiecțiile vectorilor cu vârfurile în aceste puncte și nu proximitatea proiecțiilor în planul factorial principal (cercul de corelație în cazul ACP normate).

1.2.3.2 Interpretarea „internă"

Metoda cea mai naturală de a da o semnificație unei componente principale $\mathbf{c}$ este de a o corela cu variabilele inițiale $\mathbf{x}_j$. În acest sens se vor calcula coeficienții de corelație liniară $\mathrm{cor}(\mathbf{c},\mathbf{x}_j)$ și se vor pune în evidență coeficienții cu valori absolute mari. Valorile acestor coeficienți sunt date de următoarea:

Lema 1.2-13 *În cazul unei ACP normate* $\mathrm{cor}\left(\mathbf{c}, \mathbf{z}_j\right) = \sqrt{\lambda}\,\mathbf{u}_j$.

Demonstrație Din definiție $\text{cor}(\mathbf{c},\mathbf{z}_j)=\frac{\text{cov}(\mathbf{c},\mathbf{z}_j)}{s(\mathbf{c})s(\mathbf{z}_j)}=\frac{\text{cov}(\mathbf{c},\mathbf{z}_j)}{\sqrt{\lambda}}$ (conform Corolarul 1.2-3 şi Lema 1.2-10). Dar $\text{cov}(\mathbf{c},\mathbf{z}_j)=\mathbf{c}'\mathbf{D}\mathbf{z}_j=\mathbf{u}'\mathbf{Z}'\mathbf{D}\mathbf{z}_j$ adică $\mathbf{z}'_j\mathbf{D}\mathbf{Z}$ este al *j*-lea coeficient al vectorului $(\mathbf{Z}'\mathbf{D}\mathbf{Z})\mathbf{u}$. Cum din definiție $\mathbf{Z}'\mathbf{D}\mathbf{Z}=\mathbf{R}$ şi cum $\mathbf{R}\mathbf{u}=\lambda\mathbf{u}$ (Lema 1.2-8) rezultă $\text{cor}(\mathbf{c},\mathbf{z}_j)=\sqrt{\lambda}\mathbf{u}_j$.

□

Corolarul 1.2-4 *Cercul de corelație principal este în spațiul variabilelor corespondentul exact al planului factorial principal.*

Demonstrație Într-adevăr în ACP normată coordonatele proiecției unui punct-variabilă B_j sunt (d_{1j},d_{2j}) care, conform formulelor de tranziție, sunt egale cu $\left(\sqrt{\lambda_1}a_{1j},\sqrt{\lambda_2}a_{2j}\right)$. Dar în cazul unei ACP normate axele principale coincid cu factorii principali deci $(d_{1j},d_{2j})=\left(\sqrt{\lambda_1}a_{1j},\sqrt{\lambda_2}a_{2j}\right)=\left(\sqrt{\lambda_1}u_{1j},\sqrt{\lambda_2}u_{2j}\right)=\left(\text{cor}(\mathbf{c}_1,\mathbf{z}_j),\text{cor}(\mathbf{c}_2,\mathbf{z}_j)\right)$ conform Lema 1.2-13.

□

A spune că $\mathbf{c}_1$ este foarte corelată cu o variabilă $\mathbf{x}_j$ semnifică că indivizii cu o coordonată pozitivă mare pe axa unu sunt caracterizați de o valoare a lui $\mathbf{x}_j$ net superioară mediei (căci originea axelor principale este în centrul de greutate al norului de puncte-indivizi). Reciproc, dacă indivizii nu sunt anonimi pot ajuta la interpretarea axelor şi componentelor principale (vor fi evidențiați, de exemplu, indivizii opuşi de-a lungul unei axe).

O măsură naturală a *contribuției unui punct-individ la o axă factorială* este raportul dintre dispersia individului şi dispersia întregii axe. Din Lema 1.2-10 se cunoaşte faptul că $\sum_{i=1}^{n}p_ic_{ji}^2=\lambda_j$ deci contribuția individului *i* la axa principală *j* este $\text{cr}_j(i)=\frac{p_ic_{ji}^2}{\lambda_j}$. Când indivizii sunt anonimi (adică au toți ponderile $p_i=\frac{1}{n}$) contribuțiile „cr" nu aduc mai multe informații decât coordonatele acestora. Dacă cei *n* indivizi au aceeaşi pondere *1/n*, inerția unui punct variază direct proporțional cu distanța la centrul de greutate. Indivizii care contribuie determinant la inerția axei sunt cei mai depărtați de punctul mediu şi lectura coordonatelor factoriale sau vizualizarea graficului sunt suficiente pentru a interpreta factorii în acest caz. Prezentarea indivizilor în planul factorial permite să apreciem repartiția lor şi să reperăm zonele de densități mai mari sau mai slabe. Ca o recomandare generală se va considera importantă contribuția care depăşeşte ponderea p_i a individului (sau ¼ în cazul indivizilor anonimi). Dacă *p* şi *n* sunt mari atunci componentele principale sunt deseori considerate ca fiind selecții asupra unor variabile

aleatoare repartizate normal de medie zero şi dispersie λ. În acest caz $\frac{c_{ji}^2}{\lambda_j}$ este distribuită χ_1^2 şi o contribuţie mai mare decât $3{,}84/n$ poate fi considerată semnificativă cu un prag de încredere de 95% (conform Enăchescu & Enăchescu, 2000).

Considerarea contribuţiilor, când acestea nu sunt excesive, ajută la interpretarea axelor. În mod normal şi aceasta în special pentru primele axe factoriale, nu este de dorit ca un individ să aibă o contribuţie excesivă căci acesta poate constitui un factor de instabilitate (omiterea individului poate modifica profund rezultatele analizei). În cazul unui sondaj (indivizi anonimi) contribuţia excesivă a unui individ este adesea cauzată de erori de preluare a datelor. Pentru a pune în evidenţă aceste anomalii (şi evident pentru a le elimina) C. şi D.Enăchescu recomană următorul test empiric în cazul unei ACP normate:

dacă pătratul distanţei de la un punct-individ la centrul de greutate al norului este mai mare decât $p+2\sqrt{2\sum_{i=1}^{p}\lambda_i^2}$ *atunci observaţa respectivă poate fi considerată o valoare aberantă.*

Într-adevăr, dacă observaţiile sunt normal distribuite, I_g este o sumă ponderată de p variabile repartizate χ_1^2 cu media $\sum_{i=1}^{p}\lambda_i = p$ (datorită datelor centrat-reduse) şi dispersia $2\sum_{i=1}^{p}\lambda_i^2$. Considerând intervalul de încredere de 95% pentru I_g se obţine marginea din recomandarea de mai sus.

Dacă observaţiile sunt independente atunci λ_i estimate pe baza acestor observaţii sunt de medie 1 şi satisfac egalitatea $\sum_{i=1}^{p}\lambda_i^2 = p+2\sum_{i>j} r_{ij}^2$. Deoarece media pătratului coeficientului de corelaţie între două variabile normale independente este $1/n-1$ rezultă că $\mathrm{E}\left(\sum_{j=1}^{p}\lambda_i^2\right) = p+\frac{p(p-1)}{n-1}$. Revenind la marginea pentru valori aberante găsită mai sus, în cazul independenţei observaţiilor, o putem rafina înlocuind-o cu $p+2\sqrt{2p\left(1+\frac{p-1}{n-1}\right)}$ sau $p+2{,}8\sqrt{p}$ pentru n mare.

Analiza unui nor de variabile facându-se pornind din origine, variabilele pot fi toate situate de aceeaşi parte a unei axe factoriale. O astfel de dispoziţie apare atunci când toate variabilele sunt corelate pozitiv între ele. În acest caz $\mathbf{c}_1$, prima componentă

principală defineşte un *factor de talie*. Conform teoremei lui Perron[7] (vezi, de exemplu, Demidovitch & Maron, 1973) $\mathbf{c}_1$ este atunci corelată pozitiv cu toate variabilele şi indivizii sunt ordonaţi pe prima axă principală crescător după mediile $\left\{\frac{1}{p}\sum_{j=1}^{p} y_{ij}\right\}_{i=1}^{n}$

Ortogonalitatea axelor face să nu existe decât un singur factor de „talie". A doua componentă principală diferenţiază atunci indivizii de „talie" comparabilă şi această componentă se va numi *factor de formă*.

1.2.3.3 Interpretarea „externă": variabile şi indivizi suplimentari

Interpretările interne au dezavantajul că sunt tautologice: se explică un rezultat cu ajutorul datelor care au servit la obţinerea lui. Riscul care apare într-un astfel de caz este acela de a confunda un artefact introdus de metodă cu un fenomen semnificativ. Din contră, dacă se găseşte o corelaţie puternică între o componentă principală şi o variabilă care nu a fost utilizată în analiză, caracterul probant al fenomenului va fi mult mai ridicat. De unde practica curentă de a împărţi în două mulţimea variabilelor: o parte din variabile, numite *active*, vor fi utilizate pentru determinarea axelor principale şi cealaltă parte a variabilelor numite *pasive* sau *suplimentare* sau *ilustrative*, care vor fi corelate a posteriori cu componentele principale. În plus, variabilele active, definite într-un spaţiu şi utilizate la calculul planurilor factoriale, trebuie să formeze un ansamblu omogen ca textură (trebuie, adică, să aibă aceeaşi natură) pentru ca distanţele între elemente să aibă un sens. Pentru a interpreta similitudinile între elemente acestea trebuiesc să fie omogene şi în conţinut, adică să privească o aceeaşi temă; se compară obiectele după un anumit punct de vedere şi nu utilizând fără discernământ toate atributele cunoscute şi adesea disparate. Variabilele suplimentare nu sunt însă supuse acestor condiţii de omogenitate.

Un tratament analog se poate aplica şi mulţimii indivizilor distingând între *indivizi activi* şi *indivizi suplimentari* care nu participă la calculul matricilor de covarianţă / corelaţie). Indivizii suplimentari permit verificarea ne-tautologică a ipotezelor formulate asupra indivizilor activi după o ACP.

Se notează cu $\mathbf{Y}^{+} \in \mathcal{M}_{I,s}(\mathbb{R})$ cele s variabile (coloane) continue ilustrative şi cu $\mathbf{Y}_{+} \in \mathcal{M}_{t,p}(\mathbb{R})$ cei t indivizi (linii) suplimentari. După eventuala normare a datelor suplimentare, coordonatele noilor variabile pe axa j sunt componentele vectorului

$$\left(\mathbf{Y}^{+}\right)' \mathbf{v}_j \text{ sau } \left(\mathbf{Z}^{+}\right)' \mathbf{v}_j$$

iar coordonatele noilor indivizi pe axa j sunt componentele vectorului

$$\left(\mathbf{Y}_{+}\right) \mathbf{u}_j \text{ sau } \left(\mathbf{Z}_{+}\right) \mathbf{u}_j .$$

[7] Dacă o matrice pătratică şi simetrică are toţi coeficienţii pozitivi atunci valoarea sa proprie cea mai mare în modul este pozitivă, rădăcină simplă a ecuaţiei caracteristice şi i se asociază un vector propriu cu componente pozitive.

Dacă variabila suplimentară este nominală transformarea de mai sus nu mai poate fi aplicată. În această situație, analiza unei variabile nominale suplimentare nu se mai face în $\mathbb{R}^n$ ci în $\mathbb{R}^p$. Fiecare modalitate a variabilei nominale este reprezentată în spațiul indivizilor prin centrul de greutate al subnorului de puncte-indivizi care au ales respectiva modalitate.

1.2.3.4 Reprezentarea simultană

Analiza norului de variabile este dedusă din analiza norului de indivizi: reprezentarea variabilelor pe axele factoriale în $\mathbb{R}^n$ ajută la interpretarea axelor factoriale în $\mathbb{R}^p$ și reciproc. Cei doi nori nu folosesc însă același reper ceeace face imposibilă reprezentarea simultană a indivizilor și variabilelor. Astfel:

- în spațiul $\mathbb{R}^p$, reprezentarea norului de n puncte-indivizi se face în reperul $\{\mathbf{G}, \mathbf{u}_1, \ldots, \mathbf{u}_p\}$. Reprezentarea indivizilor în planul factorial furnizează cea mai bună vizualizare aproximativă a distanțelor între indivizi. Vecinătatea indivizilor în planul factorial se interpretează în termeni de similitudini de comportament față de variabilele observate;
- în spațiul $\mathbb{R}^n$, reprezentarea norului de p puncte-variabile se face în reperul $\{\mathbf{O}, \mathbf{v}_1, \ldots \mathbf{v}_n\}$. Reprezentarea variabilelor în cercul de corelație furnizează o sinteză grafică a matricii de corelație. Vecinătatea variabilelor în planul cercului de corelație se interpretează în termeni de corelații.

Față de cele de mai sus suprapunerea celor două planuri factoriale este lipsită de sens; **trebuie să ne ferim a interpreta distanța dintre un punct-individ și un punct-variabilă deoarece aceste puncte nu fac parte nici din același nor, nici din același spațiu și nici nu sunt reprezentate în același reper**.

Dacă însă se consideră în loc de puncte-variabile direcții de variabile în $\mathbb{R}^p$, atunci se pot reprezenta simultan, în acest spațiu, atât punctele-indivizi cât și vectorii reprezentând variabilele.

În spațiul $\mathbb{R}^p$ al celor n puncte-indivizi, după transformarea tabelului de date, dispunem de două sisteme de axe:

- vechile axe unitare $\{\mathbf{e}_1, \ldots \mathbf{e}_p\}$ corespunzând celor p variabile înainte de analiză și reprezentând sistemul de axe de referință pentru coordonatele inițiale ale indivizilor (cu $\mathbf{e}'_j = (0, .., 1_j, .., 0) \quad j = \overline{1, p}$);
- noile axe unitare $\{\mathbf{u}_1, \ldots \mathbf{u}_p\}$ formate din axele factoriale.

Posibilitatea unei reprezentări simultane rezidă în acest context în proiecția, ca individ suplimentar, a vechii axe $\mathbf{e}_j$ pe noua axă $\mathbf{u}_k$. Coordonata proiecției lui $\mathbf{e}_j$ pe $\mathbf{u}_k$ este $\mathbf{e}'_j \mathbf{u}_k = u_{kj}$. Este astfel posibil să se reprezinte în $\mathbb{R}^p$ direcțiile date de variabilele inițiale pe planul factorial al norului de indivizi; aceste direcții pot fi materializate prin

vectori unitari. Aceşti vectori constituie reperul originar în care a fost construit norul de indivizi. Aceşti vectori sunt deci ortogonali doi câte doi (este acum evident faptul că această reprezentare a variabilelor este diferită de reprezentarea norului de variabile descrisă mai sus). Ceeace se va numi *reprezentare simultană* este deci proiectarea reperului ortonormat al axelor de origine în planul factorial al norului de indivizi.

Se reaminteşte că, în $\mathbb{R}^n$, în metrica euclidiană, coordonata variabilei j pe axa k este egală cu coeficientul de corelaţie (conform formulei de tranziţie) între variabilă şi factor şi este $d_{kj} = \sqrt{\lambda_k} u_{kj}$. Cei doi nori de variabile nu coincid; ei diferă unul de celălalt, pe fiecare axă, prin coeficientul de dilataţie $\sqrt{\lambda_k}$.

În cazul reprezentării simultane, care este de fapt o reprezentare în $\mathbb{R}^n$, distanţa dintre două variabile nu se interpretează în termeni de corelaţie deoarece este vorba de extremităţile unor vectori ortonormaţi (distanţă egală cu $\sqrt{2}$ în spaţiul complet). Interpretarea distanţei între două variabile (în termeni de corelaţie) nu se poate face decât în $\mathbb{R}^n$ (să observăm totuşi că norul proiectat al extremităţilor vectorilor unitari din $\mathbb{R}^p$ şi norul extremităţilor vectorilor variabile în $\mathbb{R}^n$ au, în general, forme asemănătoare mai ales dacă vectorii proprii sunt comparabili, deci dilatările sunt puţin deformante). Ţinând cont de aceste consideraţii, este licit să comparăm, în reprezentarea simultană, poziţia a doi indivizi faţă de ansamblul variabilelor, sau poziţia a două variabile faţă de ansamblul indivizilor. Astfel, direcţia unei variabile defineşte zone pentru indivizi: de o parte indivizii ce iau valori mari pentru această variabilă şi, în partea opusă, indivizii care iau valori mici. Ne vor interesa distanţele între indivizi în direcţia variabilei. La intersecţia axelor se găsesc valorile medii ale tuturor variabilelor.

1.2.4 ANALIZE NEPARAMETRICE

Metodele de analiză neparametrică nu diferă de ACP decât printr-o transformare preliminară a datelor. Aceste metode sunt recomandate atunci când datele preliminare sunt heterogene, dau rezultate foarte robuste şi se pretează la interpretări simple în termeni statistici.

1.2.4.1 Analiza rangurilor

În analiza rangurilor tabelul iniţial este transformat în tabel de ranguri. Observaţia i a variabilei j constă, în acest caz, într-un clasament q_{ij}: este rangul observaţiei i în ordonarea crecătoare a celor n observaţii iniţiale. În aceste condiţii, distanţa între două variabile $\mathbf{q}_j$ şi $\mathbf{q}_k$ este definită de formula: $d^2(\mathbf{q}_j, \mathbf{q}_k) = \frac{1}{n(n-1)(n+1)} \sum_{i=1}^{n} (q_{ij} - q_{ik})^2$

(recunoaştem în această formulă complementul faţă de 1 a *coeficientului de corelaţie Spearman*.

Utilizarea rangurilor este justificată în următoarele contexte:

- datele inițiale sunt ele însile un clasament, în care caz acest tip de analiză se impune;
- scările de măsură a variabilelor pot fi atât de diferite încât operația de reducere practicată de analiza în componente principale normate nu este suficientă. În plus, operația de normare, nu reduce, de exemplu, nesimetria distribuțiilor. În fine, atunci când este mai interesant sintetizarea unei familii de clasamente decât a unei mulțimi foarte eterogene de măsurători;
- ipotezele *a priori* făcute implicit asupra măsurătorilor sunt mult mai slabe și în consecință mai puțin arbitrare: legea de repartiție a distanțelor este acum neparametrică; dispunem deci de praguri de încredere care nu mai depind decât de ipoteza de continuitate asupra distribuțiilor observațiilor, mai plauzibilă decât cea de normalitate;
- în fine, reprezentările obținute sunt robuste, puțin sensibile la existența valorilor aberante, ceeace este adeseori o calitate apreciabilă.

Regulile de interpretare se deduc din cele ale analizei în componente principale deaorece aceasta este analiza ce se aplică după operația de transformare în ranguri (să notăm că, în acest caz, nu este necesară reducerea tabelului de date deoarece toate rangurile au aceeași dispersie). Proximitatea între două variabile se interpretează în termeni de corelație a rangurilor: două variabile sunt apropiate dacă prezintă clasamente asemănătoare ale observațiilor inițiale; două variabile sunt depărtate dacă prezintă clasamente practic opuse ale observațiilor inițiale. Două observații vor fi apropiate dacă au ranguri similare pentru fiecare variabilă. Să mai notăm că, în reprezentarea simultană, se poate avea o idee asupra întregului clasament al observațiilor pentru o variabilă examinându-se pozițiile respective ale acestei variabile și mulțimea observațiilor.

În fine, caracterul neparametric al reprezentării obținute permite efectuarea de teste de validare asupra valorilor proprii. Distribuția valorilor proprii obținute din analiza unui tabel de ranguri nu depinde decât de parametrii n și p, numărul de linii și de coloane al tabelului. Este posibil să procedăm la o listare a pragurilor de încredere a valorilor proprii.

1.2.4.2 Analiza în componente robuste

Criteriul de ajustare al celor mai mici pătrate este în mod particular adaptat distribuției normale. În cazul unei distribuții uniforme (cazul analizei rangurilor) acesta tinde să dea o importanță excesivă observațiilor extreme. Pentru ca analiza să fie mai robustă distribuția uniformă a rangurilor este „normalizată".

Fie cea de a k-a observație din n observații ordonate crescător și fie F funcția de repartiție normală. Se înlocuiește observația de rang k prin valoarea y_k dată de transformarea $y_k = \mathrm{F}^{-1}\left(\frac{k}{n+1}\right)$ unde F^{-1} este inversa funcției de repartiție normală.

Pentru n mare, transformarea este echivalentă cu înlocuirea celei de a k observații cu media celei de a k observații într-un eşantion ordonat de n valori normale.

1.2.5 ALTE METODE DERIVATE

Numeroase tehnici sunt direct derivate din analiza în componente principale; variantele neparametrice din paragraful precedent sunt un astfel de exemplu.

Unele prezentări ale analizei de corespondență consideră această metodă ca o analiză în componente principale particulară. Aceasta este posibil dacă se tratează cele două spații (al liniilor şi al coloanelor) separat, dar nu aceasta este optica aleasă aici. Acest tratament separat maschează unul din aporturile metodologice fundamentale ale analizei factoriale descriptive. Analiza în componente principale, fie că este vorba de analiza normată sau nenormată, analizează indivizii în raport cu *centrul lor de greutate* şi variabilele în raport cu *originea axelor*. Această asimetrie de tratament corespunde la domenii de aplicație specifice şi induce reguli de interpretare particulare. Descompunerea în valori singulare (sau încă analiza generală, sau teorema lui Eckart & Young) formează miezul teoretic comun al celor două metode.

Vom cita printre metodele derivate *analiza parțială a corelațiilor* sau *analiza cu variabile instrumentale* (Rao, 1964). În acest caz se urmăreşte nu numai eliminarea efectelor eterogenității variabilelor (prin centrarea şi reducerea lor) ci şi a efectelor celorlalte variabile printr-o regresie multiplă prealabilă. Analiza logaritmică (Kazmierczak, 1985) este o analiză în componente principale nenormate a tabelului (dublu centrat pe linii şi pe coloane) valorilor inițiale logaritmate. Această variantă posedă proprietăți de stabilitate şi robustețe interesante.

În fine, alte tehnici cum ar fi regresia pe componente principale sau clasificarea pe factori sunt mai degrabă tehnici complementare decât derivate.

1.2.6 ALTE DEMERSURI

Descompunerea în valori singulare este o proprietate a tuturor matricilor dreptunghiulare. Ea se bazează pe distanțe euclidiene, adică pe forme pătratice pozitiv definite şi pe aproximări ale spațiilor vectoriale prin minimizarea unui criteriu legat de distanțe. Sunt posibile şi alte demersuri care modifică tipul de distanță, sau natura subspațiilor sau amândouă. Desigur în acest caz multe din proprietățile matematice simple ale analizei bazate pe metrica euclidiană nu se mai regăses: unicitatea descompunerii, simetria rolurilor jucate de linii şi de coloane, simplitatea formulelor de reconstrucție, poziționarea naturală a variabilelor suplimentare. Alte criterii de aproximare pot fi totuşi utile. În locul metodei celor mai mici pătrate $\min\left\{\sum \mathbf{e}_i^2\right\}$ (norma „L_2") se poate utiliza, de exemplu metoda celor mai mici valori absolute $\min\left\{\sum|\mathbf{e}_i|\right\}$ (norma „L_1") care induce distanța „city-block" (pentru contribuții la acest punct de vedere se recomandă, printre altele, culegerea editată de Dodge, 1987).

Într-un spirit puţin diferit Meyer, 1994 enunţă un algoritm pentru a aproxima (în sensul celor mai mici pătrate, adică în L_2) o matrice de distanţe de tip L_p cu o matrice de disimilaritate dată.

Pentru a studia anumite tabele de contingenţă, în speţă tabelele de schimb, Domenges&Volle, 1979 propun utilizarea *distanţei lui Hellinger*: $d^2(x,y) = \sum\left(\sqrt{x_i} - \sqrt{y_i}\right)^2$ („analiza factorială sferică").

În fine, fără a schimba nici metrica nici criteriul de aproximare, se pot aproxima alte suprafeţe decât hiperplanele. Astfel, în cazul analizei în componente principale normate care este, în spaţiul $\mathbb{R}^n$ analiza punctelor situate pe o sferă, Falissard, 1995 propune aproximarea unei hipersfere.

1.3 ANALIZA CORESPONDENŢELOR SIMPLE (ACS)

Prezentată sub acest nume şi dezvoltată în Franţa de J. P. Benzécri (1969), metoda are precursori pe Guttman (1941) şi Hayushi (1956).

Analiza corespondenţelor este o metodă adaptată tabelelor de contingenţă permiţând studiul relaţiilor existente între două sau mai multe variabile nominale (discrete).

Distingem între:

- analiza corespondenţelor simple (ACS) în cazul studiului relaţiilor între două variabile nominale;
- analiza corespondenţelor multiple (ACM) în cazul studiului relaţiilor între mai multe variabile nominale.

Definiţia 1.3-1 Se numeşte *tabel de contingenţă* (sau *de dependenţă*, sau *încrucişat*) un tabel ale cărui linii, respectiv coloane desemnează două partiţii ale aceleiaşi mulţimi, partiţii date de modalităţile a două variabile nominale.

Fie $\mathcal{X}$ şi $\mathcal{Y}$ variabile nominale cu n respectiv p modalităţi descriind o mulţime de k indivizi.

Fie $\mathbf{K}$ - tabelul de contingenţă cu n - linii, p - coloane şi elementele k_{ij} = numărul de indivizi având simultan modalitatea i a variabilei X şi modalitatea j a variabilei Y.

Se notează

$$k_{i\cdot} = \sum_j k_{ij} \; ; \; k_{\cdot j} = \sum_i k_{ij} \; ; \; k = \sum_{i\cdot j} k_{ij}$$

şi cu

$$f_{ij} = \frac{k_{ij}}{k} \; ; \; f_{i\cdot} = \sum_i f_{ij} \; ; \; f_{\cdot j} = \sum_i f_{ij}$$

frecvențele relative (**observație** $\sum_{i,j} f_{ij} = 1$).

Grafic tabelul se prezintă

$\mathcal{X} \setminus \mathcal{Y}$	y_1	$y_2 \cdots$	y_j	$\cdots y_p$	
x_1			$\vdots$		$k_{1\cdot}$
x_2			$\vdots$		$k_{2\cdot}$
$\vdots$			$\vdots$		$\vdots$
x_i			k_{ij}		$k_{i\cdot}$
$\vdots$					$\vdots$
x_n					$k_{n\cdot}$
	$k_{\cdot 1}$	$k_{\cdot 2} \cdots$	$k_{\cdot j}$	$\cdots k_{\cdot p}$	k

Două lecturi sunt posibile, după cum este privilegiată una sau alta dintre variabile: pe linii, cu frecvențele $\left\{\frac{f_{ij}}{f_{i\cdot}}\right\}_{j=\overline{1,p}}^{i=\overline{1,n}}$, respectiv pe coloane, cu frecvențele $\left\{\frac{f_{ij}}{f_{\cdot j}}\right\}_{j=\overline{1,p}}^{i=\overline{1,n}}$

1.3.1 SCHEMA GENERALĂ A ACS

Analiza corespondențelor simple revine la efectuarea unei analize generale a unui nor de puncte ponderate într-un spațiu cu o metrică specială.

1.3.1.1 Geometria norilor și elementele de bază

Fie **F** matricea $n \times p$ a frecvențelor relative;

$\mathbf{D}_n = \text{diag}(f_{i\cdot})$ matricea $n \times n$ cu diagonala principală conținând marjele liniilor

$\mathbf{D}_p = \text{diag}(f_{\cdot j})$ matricea $p \times p$ cu diagonala principală conținând marjele coloanelor.

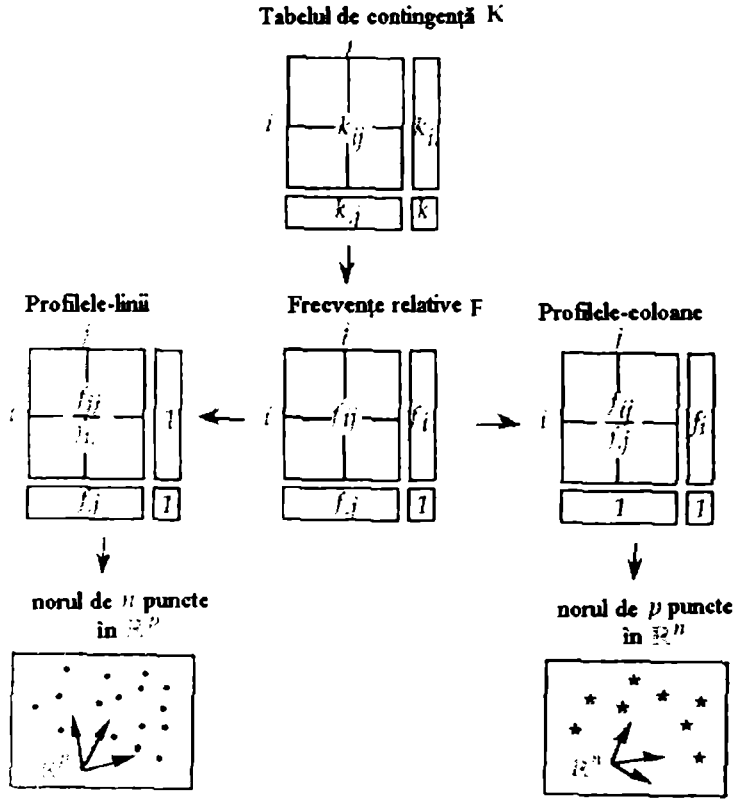


Figura 1.3-1 Transformările tabelului de contingență

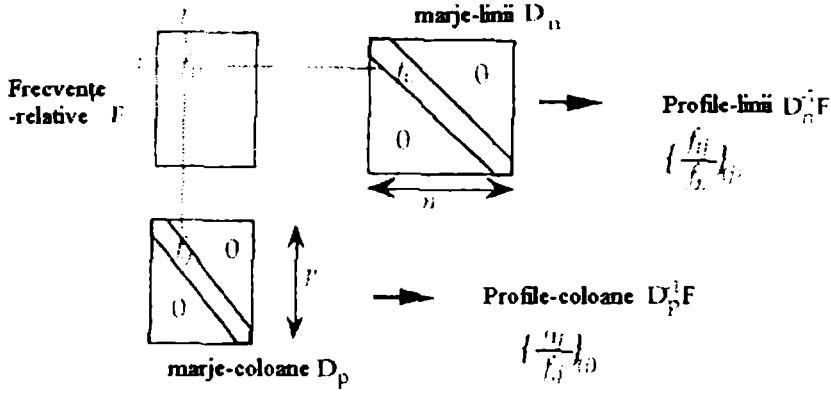


Figura 1.3-2 Frecvențe, marje, profile

1.3.1.2 Alegerea distanței și a metricii

Este firesc să ne gândim la distanța euclidiană între profilurile linie, respectiv profilurile coloană:

$$d^2(i,i') = \sum_j \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$

și analoaga.

Această distanță favorizează coloanele care au o masă $f_{\cdot j}$ importantă. Pentru a remedia acest lucru cât și din alte considerente (discutate mai jos) se ponderează fiecare diferență cu inversa masei coloanei obținându-se distanța χ^2

$$d_\chi^2(i,i') = \sum_j \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$

și analoaga

$$d_\chi^2(j,j') = \sum_i \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2 .$$

Propoziția 1.3-1 *Distanța χ^2 este invariantă la agregarea liniilor, respectiv a coloanelor cu același profil.*

Demonstrație

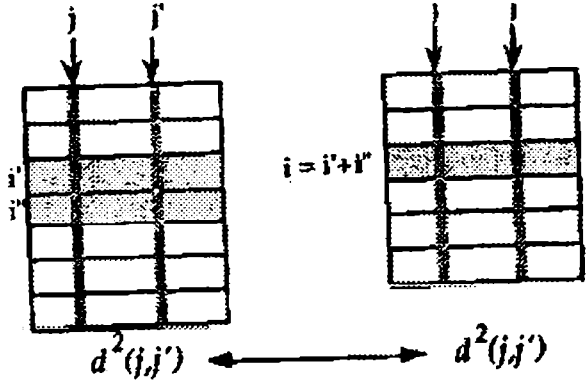


Figura 1.3-3 Echivalența distribuțională: invarianța distanțelor între coloane față de agregarea liniilor

$$d^2(j,j')=\sum_{i=1}^{i_1-1}\frac{1}{f_{i\cdot}}\left(\frac{f_{ij}}{f_{\cdot j}}-\frac{f_{ij'}}{f_{\cdot j'}}\right)^2+\frac{1}{f_{i_1}}\left(\frac{f_{i_1j}}{f_{\cdot j}}-\frac{f_{i_1j'}}{f_{\cdot j'}}\right)^2+\frac{1}{f_{i_2\cdot}}\left(\frac{f_{i_2j}}{f_{\cdot j}}-\frac{f_{i_2j'}}{f_{\cdot j'}}\right)^2+\sum_{i=i_2+1}^{n}\frac{1}{f_{i\cdot}}\left(\frac{f_{ij}}{f_{\cdot j}}-\frac{f_{ij'}}{f_{\cdot j'}}\right)^2$$

$$d_r^2(j,j')=\sum_{i=1}^{i_1-1}\frac{1}{f_{i\cdot}}\left(\frac{f_{ij}}{f_{\cdot j}}-\frac{f_{ij'}}{f_{\cdot j'}}\right)^2+\frac{1}{f_{i_0\cdot}}\left(\frac{f_{i_0j}}{f_{\cdot j}}-\frac{f_{i_0j'}}{f_{\cdot j'}}\right)^2+\sum_{i=i_2+1}^{n}\frac{1}{f_{i\cdot}}\left(\frac{f_{ij}}{f_{\cdot j}}-\frac{f_{ij'}}{f_{\cdot j'}}\right)^2.$$

Dar $\frac{f_{i_1j}}{f_{i_1\cdot}}=\frac{f_{i_2j}}{f_{i_2\cdot}}=r_j,(\forall)j=\overline{1,p}$ căci liniile au acelaşi profil. Pe de altă parte, prin agregarea liniilor i_1 şi $i_2\Rightarrow$

$$n_{i_1j}+n_{i_2j}=n_{i_0j},(\forall)j=\overline{1,p}\Rightarrow\begin{cases}f_{i_1j}+f_{i_2j}=f_{i_0j},(\forall)j\\ f_{i_1\cdot}+f_{i_2\cdot}=f_{i_0\cdot}\end{cases}\Rightarrow\frac{f_{i_0j}}{f_{i_0\cdot}}=r_j$$

căci $n_{i_1j}=n_{i_1\cdot}r_j, n_{i_2j}=n_{i_2\cdot}r_j$ şi $\frac{f_{i_0j}}{f_{i_0\cdot}}=\frac{f_{i_1j}+f_{i_2j}}{f_{i_1\cdot}+f_{i_2\cdot}}=\frac{n_{i_1j}+n_{i_2j}}{n_{i_1\cdot}+n_{i_2\cdot}}=\frac{r_j\left(n_{i_1\cdot}+n_{i_2\cdot}\right)}{\left(n_{i_1\cdot}+n_{i_2\cdot}\right)}=r_j$.

Aşadar :

$$A(i_1)=\frac{1}{f_{i_1\cdot}}\left(\frac{f_{i_1j}}{f_{\cdot j}}-\frac{f_{i_1j'}}{f_{\cdot j'}}\right)^2=f_{i_1\cdot}\left[\left(\frac{f_{i_1j}}{f_{i_1\cdot}}\right)\cdot\frac{1}{f_{\cdot j}}-\left(\frac{f_{i_1j'}}{f_{i_1\cdot}}\right)\cdot\frac{1}{f_{\cdot j'}}\right]^2=f_{i_1\cdot}\left[\frac{r_j}{f_{\cdot j}}-\frac{r_{j'}}{f_{\cdot j'}}\right]^2=f_{i_1\cdot}B$$

$$A(i_2)=\frac{1}{f_{i_2\cdot}}\left(\frac{f_{i_2j}}{f_{\cdot j}}-\frac{f_{i_2j'}}{f_{\cdot j'}}\right)^2=f_{i_2\cdot}\left[\left(\frac{f_{i_2j}}{f_{i_2\cdot}}\right)\cdot\frac{1}{f_{\cdot j}}-\left(\frac{f_{i_2j'}}{f_{i_2\cdot}}\right)\cdot\frac{1}{f_{\cdot j'}}\right]^2=f_{i_2\cdot}\left[\frac{r_j}{f_{\cdot j}}-\frac{r_{j'}}{f_{\cdot j'}}\right]^2=f_{i_2\cdot}B$$

$$\Rightarrow A(i_1)+A(i_2)=f_{i_1\cdot}B+f_{i_2\cdot}B=\left(f_{i_1\cdot}+f_{i_2\cdot}\right)B=f_{i_0\cdot}B$$

$$A(i_0)=\frac{1}{f_{i_0\cdot}}\left(\frac{f_{i_0j}}{f_{\cdot j}}-\frac{f_{i_0j'}}{f_{\cdot j'}}\right)^2=f_{i_0\cdot}\left[\left(\frac{f_{i_0j}}{f_{i_0\cdot}}\right)\cdot\frac{1}{f_{\cdot j}}-\left(\frac{f_{i_0j'}}{f_{i_0\cdot}}\right)\cdot\frac{1}{f_{\cdot j'}}\right]^2=f_{i_0\cdot}\left[\frac{r_j}{f_{\cdot j}}-\frac{r_{j'}}{f_{\cdot j'}}\right]=f_{i_0\cdot}B$$

$$\Rightarrow A(i_1)+A(i_2)=A(i_0)\Rightarrow d^2(j,j')=d_r^2(j,j')$$

Analog pentru invarianţa distanţei între liniile profil la agregarea coloanelor.

□

Observaţii

a) Proprietatea demonstrată în propoziţia de mai sus se numeşte *principiul echivalenţei distribuţiilor*. Distanţa euclidiană nu are această proprietate. Distanţa Hollering posedă această proprietate.

b) Echivalenţa distribuţională permite agregarea a două modalităţi ale aceleiaşi variabile cu profile identice (ceeace face ca în $\mathbb{R}^p$ ele să se confunde) într-o nouă modalitate cu o pondere sumată fără însă a afecta prin aceasta nici distanţele între modalităţile variabilei, nici distanţele între modalităţile celeilalte variabile.

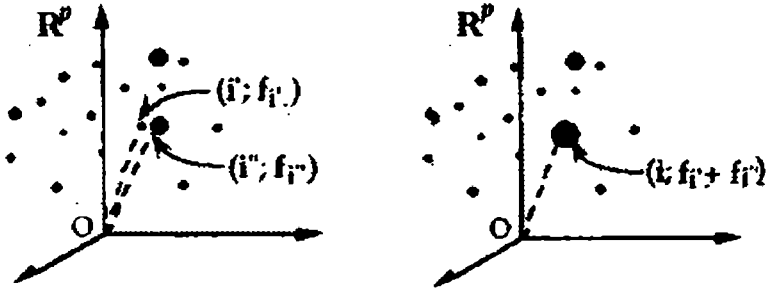


Figura 1.3-4 Echivalența distribuțională: puncte-linii confundate

Din punct de vedere practic aceasta proprietate este fundamentală deoarece garantează o oarecare invarianță a rezultatelor față de nomenclatura aleasă pentru construcția modalităților unei variabile, cu condiția regrupării modalităților asemănătoare. Nu se pierde astfel informația prin agregarea unor clase și nu se câștigă informație prin divizarea claselor omogene.

c) Metrica spațiului $\mathbb{R}^p$, respectiv $\mathbb{R}^n$ este în acest caz $\mathbf{M} = \mathbf{D}_p^{-1}$, respectiv $\mathbf{M} = \mathbf{D}_n^{-1}$.

d) Cum profilurile-linie, respectiv profilurile-coloane au mase $\{f_{i\cdot}\}_{i=\overline{1,n}}$, respectiv $\{f_{\cdot j}\}_{j=\overline{1,p}}$, matricile de pondere sunt $\mathbf{N} = \mathbf{D}_n$, respectiv $\mathbf{N} = \mathbf{D}_p$.

Tabelul 1.3-1 Tabel recapitulativ cu elementele de bază ale unei ACS

Norul de n puncte-linie în spațiul $\mathbb{R}^p$	Elemente de bază	Norul de p puncte-coloană în spațiul $\mathbb{R}^n$
$\mathbf{X} = \mathbf{D}_n^{-1}\mathbf{F} = \left\{ \frac{f_{ij}}{f_{i\cdot}} \right\}_{i=\overline{1,n}}^{j=\overline{1,p}}$	Matricea $\mathbf{X}$ (tabelul)	$\mathbf{X} = \mathbf{D}_p^{-1}\mathbf{F}' = \left\{ \frac{f_{ij}}{f_{\cdot j}} \right\}_{j=\overline{1,p}}^{i=\overline{1,n}}$
$\mathbf{M} = \mathbf{D}_p^{-1}$ $d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)$	Metrica și distanța	$\mathbf{M} = \mathbf{D}_n^{-1}$ $d^2(j,j') = \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)$
$\mathbf{N} = \mathbf{D}_n$	Ponderea (Masa)	$\mathbf{N} = \mathbf{D}_p$

masa liniei i : $f_{i\cdot}$		masa coloanei j : $f_{\cdot j}$

Lema 1.3-1

a) *Centrul de greutate al profilurilor*

linie este $\mathbf{X}_{G_l} = \left(f_{\cdot 1},\ldots,f_{\cdot p}\right)'$ *coloană este* $\mathbf{X}_{G_c} = \left(f_{1\cdot},\ldots,f_{n\cdot}\right)'$

b) *Inerția globală a norului de puncte-linie, respectiv puncte-coloană măsoară ecartul față de legea empirică* f_{ij} *și* $f_{i\cdot}f_{\cdot j}$.

Demonstrație

a) Din definiție $\mathbf{X}_G = \mathbf{X}'\cdot\mathbf{N}\cdot\mathbf{1}$ cu ponderi normate $\mathbf{X}_{G_l} = \left(\mathbf{D}_n^{-1}\mathbf{F}\right)'\mathbf{D}_n\cdot\mathbf{1}_{n+1} = \begin{pmatrix} f_{\cdot 1} \\ f_{\cdot 2} \\ \vdots \\ f_{\cdot p} \end{pmatrix}$;

$$\mathbf{X}_{G_c} = \left(\mathbf{D}_p^{-1}\mathbf{F}\right)' D_p\cdot\mathbf{1}_{p+1} = \begin{pmatrix} f_{1\cdot} \\ f_{2\cdot} \\ \vdots \\ f_{n\cdot} \end{pmatrix}$$

b) Din definiție $I_{G_l} = \sum_i p_{i\cdot}d^2\left(i,G_l\right)$ respectiv $I_{G_c} = \sum_j p_{\cdot j}d^2\left(j,G_c\right)$

$$I_{G_l} = \sum_i f_{i\cdot}d_\chi^2\left(i,G_l\right) = \sum_i\sum_j f_{i\cdot}\cdot\frac{1}{f_{\cdot j}}\left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j}\right)^2 = \sum_i\sum_j\frac{\left(f_{ij} - f_{i\cdot}f_{\cdot j}\right)^2}{f_{i\cdot}f_{\cdot j}}$$

respectiv

$$I_{G_c} = \sum_j f_{\cdot j}d_\chi^2\left(j,G_c\right) = \sum_j\sum_i f_{\cdot j}\frac{1}{f_{i\cdot}}\left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot}\right)^2 = \sum_i\sum_j\frac{\left(f_{ij} - f_{i\cdot}f_{\cdot j}\right)^2}{f_{i\cdot}f_{\cdot j}}$$

Reamintim că două variabile aleatoare discrete luând n, respectiv p valori, cu distribuția de probabilitate comună $\left\{p_{ij}\right\}_{i=\overline{1,n}}^{j=\overline{1,p}}$ și distribuțiile marginale $\left\{p_{i\cdot}\right\}$ respectiv $\left\{p_{\cdot j}\right\}$ sunt independente $\Leftrightarrow p_{ij} = p_{i\cdot}p_{\cdot j}, (\forall)i,j$ ceea ce se traduce în termeni de estimații empirice a acestor distribuții în

$f_{ij} = f_{i\cdot}f_{\cdot j}$

Statistica testului

H_0 : $p_{ij} = p_{i\cdot}p_{\cdot j}$ $\quad (\forall)i,j$

$$H_A: \quad (\exists) i_1 \quad p_{i_1 j} \neq p_{i_1 \cdot} p_{\cdot j}$$

este $X^2 = k \cdot \sum_i \sum_j \frac{\left(f_{ij} - f_{i\cdot} f_{\cdot j}\right)^2}{f_{i\cdot} f_{\cdot j}}$ care, conform demonstrației lui K. Pearson $\sim \chi^2_{(n-1)(p-1)}$ dacă volumul de selecție pe baza căruia au fost estimate f_{ij}, adică k, tinde la ∞. Aceasta este motivația pentru care distanța folosită în ACS se numește χ^2 și măsoară cât de "independente" statistic sunt liniile față de coloanele tabelului de contingență **K** (și reciproc).

1.3.1.3 Criteriul de maximizat și matricea de diagonalizat

Dorim să reprezentăm grafic proximitatea între profile. Ne plasăm, pe rând, în cele două spații în centrul de greutate al norului corespunzător. Este o particularitate a ACS, în comparație cu ACP, echivalența dintre analiza generală și tabloul necentrat (adică cu originea O) și tabloul centrat (adică cu originea în G) cu condiția să neglijăm, în primul caz, axa factorială care unește pe O cu G (această axă este asociată valorii proprii egală cu unu, numită valoare proprie trivială). Pentru simplificarea calculelor vom întreprinde analiza generală pe tabloul necentrat în $\mathbb{R}^p$ spațiul profilurilor-linie.

Conform celor anterioare

$$\begin{cases} \max\limits_{u} \left\{ \sum\limits_i f_{i\cdot} d^2(i,0) \right\} \\ \mathbf{u}'\mathbf{D}_p^{-1}\mathbf{u} = 1 \end{cases}$$

$\Rightarrow \mathbf{u}$ vector propriu al matricii $\mathbf{S} = \mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1}$ asociat $\begin{bmatrix} \mathbf{u}'\mathbf{M}\mathbf{u} = 1 \\ \mathbf{X}'\mathbf{N}\mathbf{X}\mathbf{M}\mathbf{u} = \lambda \mathbf{u} \\ \mathbf{\Psi} = \mathbf{X}\mathbf{M}\mathbf{u} \end{bmatrix}$ celei mai mari valori proprii $\lambda \neq 1$.

Analog, în $\mathbb{R}^n$

$$\begin{cases} \max\limits_{v} \left\{ \sum\limits_j f_{\cdot j} d^2(j,0) \right\} \\ \mathbf{v}'\mathbf{D}_p^{-1}\mathbf{v} = 1 \end{cases} \quad \begin{bmatrix} \mathbf{v}'\mathbf{M}\mathbf{v} = 1 \\ \mathbf{X}\mathbf{N}\mathbf{X}'\mathbf{M}\mathbf{v} = \lambda \mathbf{v} \\ \boldsymbol{\varphi} = \mathbf{X}'\mathbf{M}\mathbf{v} \end{bmatrix}$$

$\Rightarrow \mathbf{v}$ vector propriu al matricii $\mathbf{T} = \mathbf{F}\mathbf{D}_p^{-1}\mathbf{F}'\mathbf{D}_n^{-1}$ asociat celei mai mari valori proprii $\lambda \neq 1$.

Propoziția 1.3-2 *ACS pe tabelul centrat este echivalentă cu ACS pe tabelul necentrat.*

Demonstrație Pentru fixarea ideilor să raționăm în $\mathbb{R}^p$. Se observă că

$$\mathbf{x}'_{G_l} \cdot \underbrace{\mathbf{D}_p^{-1}}_{M} \mathbf{x}_{G_l} = 1 \text{ căci } \mathbf{D}_p^{-1}\mathbf{x}_{G_l} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{matrix} \uparrow \\ p \\ \downarrow \end{matrix} \qquad \sum_{j=1}^{p} f_{\cdot j} = \mathbf{x}'_{G_l} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = 1$$

$$\mathbf{S}\mathbf{x}_{G_l} = \mathbf{x}_{G_l} \text{ căci } s_{jj'} = \sum_i \frac{f_{ij} f_{ij'}}{f_{i\cdot} f_{\cdot j'}}, \text{ iar } \sum_{j'} s_{jj'} \cdot \mathbf{x}_{G_l \cdot j'} = \sum_{j'} \sum_i \frac{f_{ij} f_{ij'}}{f_{i\cdot} f_{\cdot j'}} \cdot f_{\cdot j'}$$

$$= \sum_i \frac{f_{ij}}{f_{i\cdot}} \sum_{j'} f_{ij'} = f_{\cdot j} = \mathbf{x}_{G_l \cdot j}$$

altfel spus $\mathbf{x}_{G_l}$ este vector propriu $\mathbf{M}$-normat al matricii $\mathbf{S}$ asociat valorii proprii $\lambda_1 = 1$. Să-l notăm cu $\mathbf{u}_1 = \mathbf{x}_{G_l}$. Din construcția spațiului H

$$\mathbf{u}_1' \mathbf{M} \mathbf{u}_\alpha = 0 \qquad \alpha = \overline{2, p}$$

unde

$$\begin{cases} \mathbf{u}_\alpha' \mathbf{M} \mathbf{u}_\alpha = 1 \\ \mathbf{S}\mathbf{u}_\alpha = \lambda_\alpha \cdot \mathbf{u}_\alpha \end{cases}$$

Se notează cu $\mathbf{S}^\circ$ matricea obținută prin centrarea tabelului $\mathbf{X}$. Se observă că

$$\mathbf{S}^\circ = \mathbf{S} - \mathbf{x}_{G_l} \cdot \mathbf{x}'_{G_l} \cdot \mathbf{D}_p^{-1} = \mathbf{S} - \mathbf{u}_1 \cdot \mathbf{u}_1' \mathbf{M}$$

$$\mathbf{S}^\circ \mathbf{u}_\alpha = \mathbf{S}\mathbf{u}_\alpha - \mathbf{u}_1 \cdot \mathbf{u}_1' \mathbf{M} \mathbf{u}_\alpha = \mathbf{S}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha \quad , \quad \alpha = \overline{2, p}$$

$$\mathbf{S}^\circ \mathbf{u}_1 = \mathbf{S}\mathbf{u}_1 - \mathbf{u}_1 \cdot \mathbf{u}_1' \mathbf{M} \mathbf{u}_1 = \mathbf{u}_1 - \mathbf{u}_1 = \mathbf{0} = 0 \cdot \mathbf{u}_1 .$$

Aşadar $\mathbf{u}^\circ_\alpha = \mathbf{u}_{\alpha+1}$ şi $\lambda^\circ_\alpha = \lambda_{\alpha+1}$, $\alpha = \overline{1, p-1}$

$\mathbf{u}^\circ_p = \mathbf{u}_1$ şi $\lambda^\circ_p = 0$ şi $\lambda_1 = 1$

Aşadar în $\mathbb{R}^p$, analog în $\mathbb{R}^n$, ACS pe tabloul centrat cu termenul general $\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j}$ este echivalentă cu ACS pe tabloul cu termenul general $\frac{f_{ij}}{f_{i\cdot}}$.

□

Observații

a) În ACS punctele sunt conținute în hiperplanul $\mathcal{H}$ de dimensiune $p-1$ (pentru $\mathbb{R}^p$) datorită faptului că $\sum_i \frac{f_{ij}}{f_{i\cdot}} = 1 \quad (\forall) i = \overline{1,n}$

b) i) Cum $\sum_j x_{G_l, j} = \sum_j f_{\cdot j} = 1 \Rightarrow G_l \in \mathcal{H}$

ii) Cum $\mathbf{x}'_{G_l}\mathbf{M}\mathbf{x}_{G_l} = 1 \Rightarrow G_l$ se află la distanță 1 de origine. Cum $\left\langle OG_l, \mathbf{x}_{G_l} \right\rangle_M = 0$ căci $\left(\mathbf{x} - \mathbf{x}_{G_l}\right)' \mathbf{M}\mathbf{x}_{G_l} = \sum_j x_l - \sum_j x_{G_l,j} = 0$ căci $\mathbf{x} \in \mathcal{H}$ deci $\sum_j x_j = 1$.

iii) $\Rightarrow OG_l \perp \mathcal{H}$

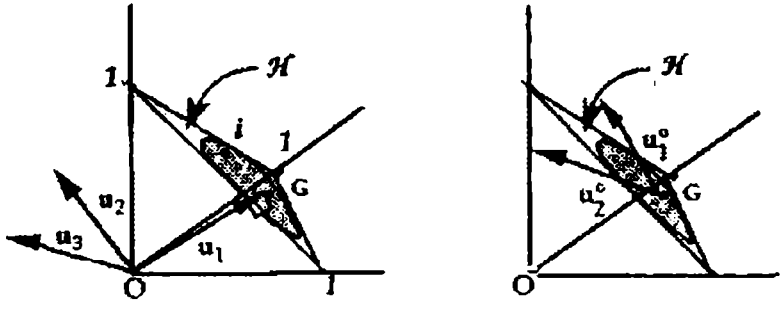


Figura 1.3-5 Analiza în $\mathbb{R}^3$

ACS în raport cu originea axelor inițiale

ACS în raport cu centrul de greutate al norului

În analiza în raport cu originea, prima direcție $\mathbf{u}_1$ este axa ce leagă originea de centrul de greutate al norului și este ortonormală pe $\mathcal{H}$. Inerția proiectată pe această axă este 1, egală cu distanța dintre O și G_l deoarece toate punctele norului se proiectează pe această axă în același punct G_l. Următoarele $p-1$ axe $\left(\mathbf{u}_2, \mathbf{u}_3, ..., \mathbf{u}_p\right)$ conținute în $\mathcal{H}$ constituie o bază definind direcții de inerție maximă ale norului. Ele coincid cu primele $p-1$ axe ale ACS în raport cu G_l și $\left(\mathbf{u}^\circ_1, \mathbf{u}^\circ_2, ..., \mathbf{u}^\circ_{p-1}\right)$. În această analiză, a p-a axă corespunde lui $\mathbf{u}_1 = OG_l$ și nu indică nici o direcție în $\mathcal{H}$ deoarece nu este conținută în $\mathcal{H}$. Inerția sa (valoarea proprie asociată) este nulă.

1.3.1.4 Axele factoriale

Presupunem că $p << n$. Conform analizei generale:

în $\mathbb{R}^p$		în $\mathbb{R}^n$
	Matricea de diagonalizat	
$\mathbf{S} = \mathbf{F}'\mathbf{D}_n\mathbf{F}\mathbf{D}_p$		$\mathbf{T} = \mathbf{F}\mathbf{D}_p^{-1}\mathbf{F}'\mathbf{D}_n^{-1}$
	Axele factoriale	

$$\mathbf{S}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha \qquad \mathbf{T}\mathbf{v}_\alpha = \lambda_\alpha \mathbf{v}_\alpha$$

Coordonatele factoriale

$$\boldsymbol{\psi}_\alpha = \mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1}\mathbf{u}_\alpha \qquad \boldsymbol{\varphi}_\alpha = \mathbf{D}_p^{-1}\mathbf{F}'\mathbf{D}_n^{-1}v_\alpha$$

$$\psi_{\alpha i} = \sum_j \frac{f_{ij}}{f_{i\cdot}f_{\cdot j}}u_{\alpha j} \qquad \varphi_{\alpha i} = \sum_i \frac{f_{ij}}{f_{i\cdot}f_{\cdot j}}v_{\alpha i}$$

Lema 1.3-2 *Coordonatele factoriale sunt variabile cu media empirică 0 şi dispersia empirică* λ_α.

Demonstraţie

$$\sum_i f_{i\cdot}\psi_{\alpha i}^2 = \boldsymbol{\psi}_\alpha'\mathbf{D}_n\boldsymbol{\psi}_\alpha = \mathbf{u}_\alpha'\mathbf{D}_p^{-1}\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{D}_n\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1}\mathbf{u}_\alpha = \mathbf{u}_\alpha'\mathbf{D}_p^{-1}\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1}\mathbf{u}_\alpha = \mathbf{u}_\alpha'\mathbf{D}_p^{-1}\mathbf{S}\mathbf{u}_\alpha =$$

$$= \mathbf{u}_\alpha'\mathbf{D}_p^{-1}\lambda_\alpha\mathbf{u}_\alpha = \lambda_\alpha \underbrace{\mathbf{u}_\alpha'\mathbf{D}_p^{-1}\mathbf{u}_\alpha}_{1} = \lambda_\alpha$$

Analog pentru $\sum_j f_{\cdot j}\varphi_{\alpha j}^2 = \lambda_\alpha$

Datorită echivalenţei dintre ACS necentrată şi ACS centrată

$$\sum_i f_{i\cdot}\psi_{\alpha i} = \sum_i f_{i\cdot}\sum_j\left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j}\right)\frac{1}{f_{\cdot j}}u_{\alpha j} = \sum_i f_{i\cdot}\sum_j \frac{f_{ij}}{f_{i\cdot}f_{\cdot j}}u_{\alpha j} - \sum_i f_{i\cdot}\sum_j u_{\alpha j} =$$

$$= \sum_j\left(\sum_i f_{ij}\right)\frac{u_{\alpha j}}{f_{\cdot j}} - \sum_j u_{\alpha j} = 0$$

□

1.3.1.5 Relaţiile între cele două spaţii

Analiza generală a arătat că matricile **S** şi **T** au aceleaşi valori proprii nenule egale cu 0, λ_α, şi că între vectorii proprii normaţi $\mathbf{u}_\alpha$ ai lui **S** asociaţi lui λ_α şi vectorii proprii normaţi $\mathbf{v}_\alpha$ a lui **T** asociaţi aceleiaşi valori proprii există relaţiile :

$$\mathbf{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{F}\mathbf{D}_p^{-1}\mathbf{u}_\alpha \quad ; \quad \mathbf{u}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{v}_\alpha$$

Înlocuind în formulele coordonatelor factoriale :

$$\boldsymbol{\psi}_\alpha = \sqrt{\lambda_\alpha}\mathbf{D}_n^{-1}\mathbf{v}_\alpha \text{ (pe componentele } \psi_{\alpha i} = \frac{\sqrt{\lambda_\alpha}}{f_{i\cdot}}v_{\alpha i}) \Rightarrow \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{D}_n\boldsymbol{\psi}_\alpha = \mathbf{v}_\alpha$$

respectiv $\boldsymbol{\varphi}_\alpha = \sqrt{\lambda_\alpha}\mathbf{D}_p^{-1}\mathbf{u}_\alpha$ (pe componentele $\varphi_{\alpha j} = \frac{\sqrt{\lambda_\alpha}}{f_{\cdot j}}u_{\alpha j}$) $\Rightarrow \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{D}_p\boldsymbol{\varphi}_\alpha = \mathbf{u}_\alpha$

care înlocuite în formulele coordonatelor factoriale dau formulele quasi-baricentrice

$$\boldsymbol{\psi}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{D}_n^{-1}\mathbf{F}\boldsymbol{\varphi}_\alpha \qquad \boldsymbol{\varphi}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{D}_p^{-1}\mathbf{F}\boldsymbol{\psi}_\alpha$$

$$\psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{j=1}^{p}\frac{f_{ij}}{f_{i\cdot}}\varphi_{\alpha j} \qquad \varphi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}}\sum_{i=1}^{n}\frac{f_{ij}}{f_{\cdot j}}\psi_{\alpha i}$$

Astfel, modulo coeficientul de dilatație $\frac{1}{\sqrt{\lambda_\alpha}}$, proiecțiile punctelor unui nor sunt, pe o axă, coordonatele baricentrice ale proiecțiilor punctelor celuilalt nor.

Matricea cu termenul general $\left\{\frac{f_{ij}}{f_{i\cdot}}\right\}_{i=\overline{1,n}}^{j=\overline{1,p}}$ ce exprimă coordonatele unui punct i pe baza tuturor punctelor j este matricea profilurilor linie.

Lema 1.3-3 *Valorile proprii sunt subunitare* ($\lambda_\alpha \le 1, (\forall)\alpha$).

Demonstrație Din $\sqrt{\lambda_\alpha}\psi_{\alpha i} = \sum_j \frac{f_{ij}}{f_{i\cdot}}\varphi_{\alpha j} \Rightarrow$

$$\min_j\{\varphi_{\alpha j}\}\underbrace{\sum_j \frac{f_{ij}}{f_{i\cdot}}}_{1} \le \sqrt{\lambda_\alpha}\psi_{\alpha i} \le \max_j\{\varphi_{\alpha j}\}\underbrace{\sum_j \frac{f_{ij}}{f_{i\cdot}}}_{1}$$

$$\max_i\left(\sqrt{\lambda_\alpha}\psi_{\alpha i}\right) \le \max_j\left(\varphi_{\alpha j}\right)$$

Analog

$$\max_j\left(\sqrt{\lambda_\alpha}\varphi_{\alpha j}\right) \le \max_i\left(\psi_{\alpha i}\right)$$

Cum $\lambda_\alpha \ge 0$, $\max_j\left(\sqrt{\lambda_\alpha}\varphi_{\alpha j}\right) \le \max_j\left(\varphi_{\alpha j}\right) \Rightarrow \lambda_\alpha \le 1$.

□

Relațiile quasi-baricentrice justifică reprezentarea simultană a liniilor şi a coloanelor.

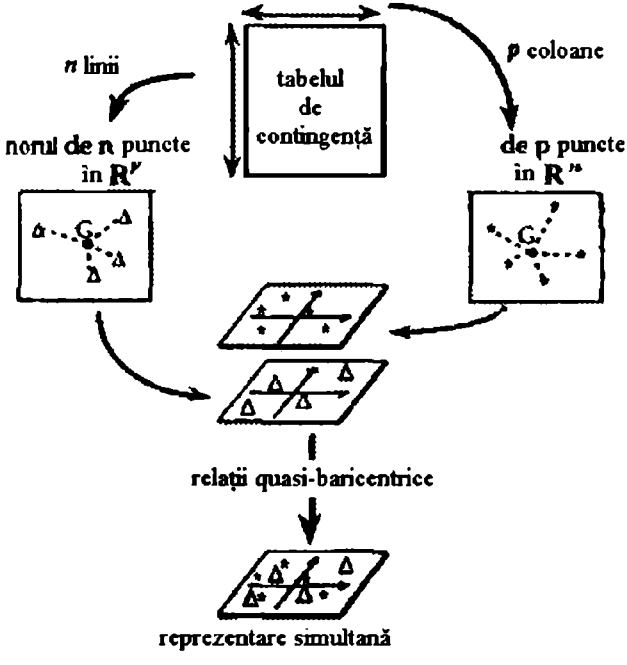


Figura 1.3-6 Schema reprezentării simultane

Rămâne în continuare valabilă observația de la ACP legată de faptul că distanța dintre un punct-linie şi un punct-coloană este lipsită de sens deoarece acestea sunt în spații diferite. ACS oferă totuşi posibilitatea de a poziționa şi interpreta un punct dintr-un nor în raport cu punctele din celălalt nor.

1.3.2 REGULI DE INTERPRETARE A REZULTATELOR

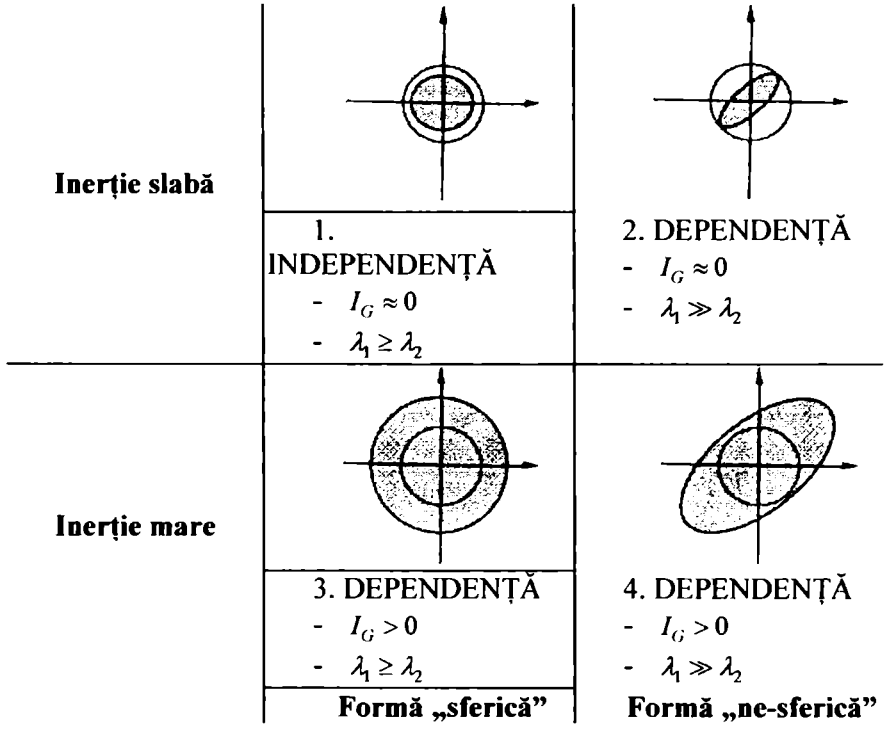
- **Inerția**

Măsurând distanța de la independența statistică, $I_G \approx 0$ şi $\lambda_1 \geq \lambda_2$ semnifică puncte grupate în jurul lui G într-o formă aproximativ circulară (nu există direcție privilegiată) generată de profile independente statistic.

$\lambda_1 \to 1$ semnifică o dicotomie a punctelor.

$\lambda_1, \lambda_2 \to 1$ semnifică 3 sub-nori.

Dacă $\lambda_1, \lambda_2, \ldots, \lambda_{p-1} \to 1$ atunci există o corespondență aproape biunivocă între modalitățile variabilelor



Să considerăm câteva forme clasice de nori de puncte pentru a arăta cum poate fi reorganizat tabelul de date corespunzător pornind de la proiecția acestora.

În cazul norului de puncte împărțit în doi sub-nori, tabelul de date poate fi reorganizat prin ordonarea coordonatelor liniilor şi coloanelor pe primul factor. Se obține schematic:

Figura 1.3-7 Norul de puncte împărțit în două

Pot exista situații în care analiza separată a celor doi sub-nori definiți de tabelele corespunzătoare (I_1, J_1) şi (I_2, J_2) să fie interesantă.

În cazul norului de puncte împărțit în trei sub-nori, tabelul de date poate fi reorganizat analog prin permutarea liniilor şi coloanelor; el poate face de asemenea obiectul unor ACS separate.

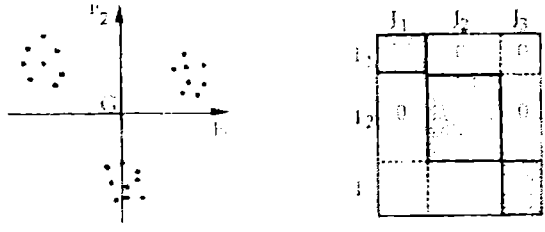


Figura 1.3-8 Norul de puncte împărțit în trei

Se poate întâlni situația în care norul de puncte are o formă parabolică. Permutând liniile şi coloanele, tabelul poate fi reordonat sub forma unei matrici diagonale relativ încărcate:

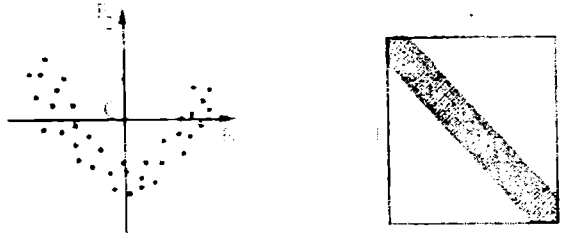


Figura 1.3-9 Efectul Guttman şi structura posibilă a tabelului

Această situație pune în evidență *efectul Guttman* care corespunde unei redondanțe a celor două variabile: cunoaşterea liniei *i* permite deducerea coloanei *j*. Toată informația este dată aproape în totalitate de primul factor.

Matricea asociată tabelului nu este, totuşi, de rang 1 şi dispunem de *p-1* factori; al doilea factor este o funcție de ordinul doi de primul factor, al treilea factor este o funcție de ordinul trei, etc. Informația dată de axele de rang superior traduce acelaşi fenomen, totuşi examinarea celui de al doilea factor rafinează interpretarea primului factor (conform lui van Rijckevorsel, 1987).

În general efectul Guttman apare atunci când variabilele sunt ordonate (variabile continue transformate în variabile nominale). O axă (adesea prima) opune valorile extreme iar o altă axă opune valorile intermediare valorilor extreme. Efectul Guttman

pune în evidență, uneori, o structură neinteresantă care poate fi interesantă dacă forma parabolică nu este perfectă. Punctele de ruptură sunt, în acest caz, interesante.

- **Inerția (dispersia) explicată de un factor**

$\lambda_\alpha = \sum_i f_i \psi_{\alpha i}^2 \Rightarrow \mathrm{Cr}_\alpha(i) = \frac{f_i \psi_{\alpha i}^2}{\lambda_\alpha}, (\forall) i = \overline{1,n}$ în $\mathbb{R}^p$ contribuția elementului i la axa α.

Analog $\mathrm{Cr}_\alpha(j) = \frac{f_j \varphi_{\alpha j}^2}{\lambda_\alpha}, (\forall) j = \overline{1,p}$ în $\mathbb{R}^n$

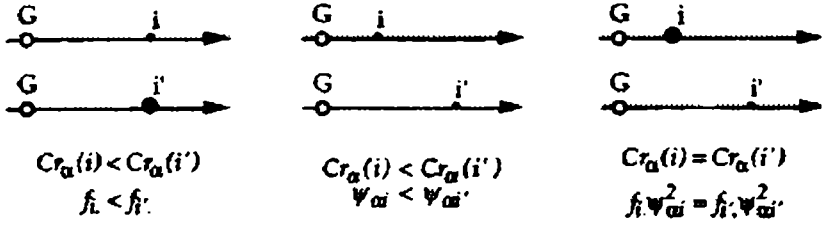


Figura 1.3-10 Contribuția la axa α : trei situații posibile

- **Calitatea reprezentării unui punct**

Din definiție $d_\alpha^2(i,G) = \psi_{\alpha i}^2$. Cum în ACS punctele se află în spațiul $\mathcal{H}$ de dimensiune $p-1 \Rightarrow \sum_\alpha d_\alpha^2(i,G) = d^2(i,G)$.

Un punct i din $\mathbb{R}^p$ poate fi mai aproape sau mai departe de axa α. Proximitatea între două puncte proiectate pe axa α este cu atât mai bine reflectată cu cât aceste puncte sunt mai apropiate de axa pe care sunt proiectate.

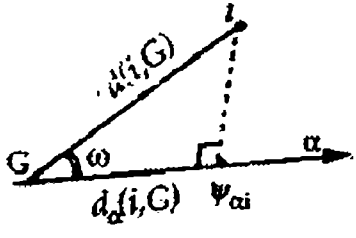
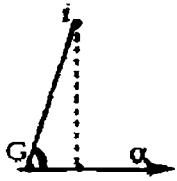


Figura 1.3-11 Proiecția punctului i pe axa α

Calitatea reprezentării unui punct i pe axa α poate fi evaluată de:

$$\cos_{\alpha}^{2}(i)=\frac{d_{a}^{2}(i,G)}{d^{2}(i,G)}.$$

Această cantitate, numită *cosinusul pătrat*, reprezintă contribuția relativă a factorului α la poziția punctului i. Cu cât cosinusul pătrat este mai aproape de 1 cu atât proiecția punctului este mai aproape de poziția acestuia în spațiu.



$Cos_{\alpha}^{2}(i) \approx 0$

i este bine reprezentat pe axa α

$Cos_{\alpha}^{2}(i) \approx 1$

i este bine reprezentat pe axa α

Figura 1.3-12 Calitatea reprezentării unui punt i pe axa α

Din definiție $\Rightarrow \sum_{\alpha} \cos_{\alpha}^{2}(i)=1, (\forall) i$ puncte active.

Cosinul pătrat pentru un element ilustrativ este subunitar dacă acesta aparține lui $\mathbb{R}^{p}$.

În ACS elementele active aparțin lui $\mathbb{R}^{p-1}$.

1.4 ANALIZA CORESPONDENȚELOR MULTIPLE (ACM)

Analiza corespondențelor multiple –ACM- este o generalizare posibilă a analizei de corespondență.

Numele apare într-o lucrare a lui Lebart (1975) dar principiile metodei urcă până la Guttman (1941), Burt (1950), Hayashi (1956).

Sub numele de *Homogeneity Analysis* este dezvoltată de echipa lui J. de Leeuw începând cu 1973, iar sub numele de *Dual Scaling* de Nishisato (1980).

Se notează cu:

- s – numărul întrebărilor puse la n – indivizi;
- p_q - numărul modalităților întrebării q, $q = \overline{1,s}$;
- $\mathbf{R} = (r_{iq})_{i=\overline{1,n}}^{q=\overline{1,s}}$ tabelul de date condensat, unde r_{iq} = numărul modalității întrebării q aleasă de individul i, deci $r_{iq} \le p_q$.

Ipoteză fundamentală: *Modalitățile fiecărei întrebări se exclud reciproc, iar o modalitate este obligatoriu aleasă.*

Exemplu: la întrebarea: *Starea dvs. civila este?* cu modalitățile

1. celibatar 2. căsătorit sau trăind marital
3. văduv 4. divorțat 5. nu răspund

există cinci modalități de răspuns ce satisfac ipoteza fundamentală.

Un astfel de tabel nu este exploatabil; sumele pe linie şi pe coloane nu au nici un sens. Variabilele trebuies recodate.

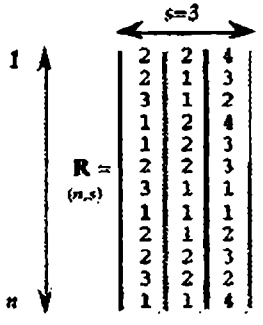


Figura 1.4-1 Tabel de date sub formă codificată condensată

În acest sens se notează cu $p = \sum_{q=1}^{s} p_q$ numărul total de modalități ale celor s întrebări şi, se construieşte, pornind de la **R,** tabelul

$$\mathbf{Z} = \left[\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_q, \ldots, \mathbf{Z}_s \right]$$

cu n linii si p coloane ce descrie cele s răspunsuri ale celor n indivizi printr-un codaj binar.

Z se obține din **R** astfel

$$z_{ij.q} = \begin{cases} 1 & \text{dacă } r_{iq} = l \\ 0 & \text{în rest} \end{cases}.$$

În notația de mai sus $\mathbf{Z}_q$ este un tabel $n \times p_q$ fiecare linie conținând $p_q - 1$ zerouri şi un singur unu.

Definiția 1.4-1 Z se numeşte *tabel disjunctiv complet.*

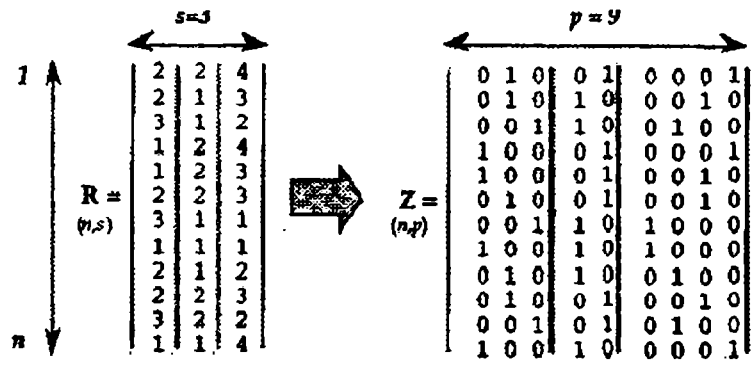


Figura 1.4-2 Construcţia tabelului disjunctiv complet

Marjele tabelului **Z** sunt:

$$z_{i.} = \sum_{j=1}^{p} z_{ij,q} = s\,,$$

$z_{.j} = \sum_{i=1}^{n} z_{ij,q} =$ -numărul de indivizi care au ales modalitatea j a întrebării q.

Rezultă $n = \sum_{j=1}^{p_q} z_{.j} = z_q$ şi $z = \sum_{i=1}^{n} z_{i.} = \sum_{q=1}^{s} z_q = \sum_{i=1}^{n}\sum_{j=1}^{p} z_{ij} = ns$ -efectivul total.

1.4.1 TABELUL DE CONTINGENŢĂ BURT

Definiţia 1.4-2 $\mathbf{B} = \mathbf{Z}' \cdot \mathbf{Z}$ se numeşte *tabelul de contingenţă Burt* asociat tabelului disjunctiv complet **Z**.

Termenul general se scrie: $b_{jj'} = \sum_{i=1}^{n} z_{ij} z_{ij'}$ cu $j, j' = \overline{1, p}$.

Marjele sunt: $b_{j} = \sum_{j'=1}^{p} b_{jj'} = s \cdot z_{.j}$.

Efectivul total: $b = \sum_{j=1}^{p} b_{j} = s^2 \cdot n$.

Tabelul **B** este format din s^2 blocuri unde se disting:

- blocurile de tip $\mathbf{Z}'_q \cdot \mathbf{Z}_{q'}$ indexate de (q, q'), de dimensiune $p_q \times p_{q'}$ care se obţin prin „încrucişarea" răspunsurilor la întrebările q şi q';

- blocurile de tip $\mathbf{Z}'_q \cdot \mathbf{Z}_{q'}$ obținute prin „încrucișarea" răspunsurilor la aceeași întrebare. Este o matrice $p_q \times p_q$ diagonală căci două modalități ale unei aceeași întrebări nu pot fi alese simultan (datorită ipotezei fundamentale). Termenii de pe diagonală sunt efectivele $\{ z_j \}$ ale modalităților întrebării q.

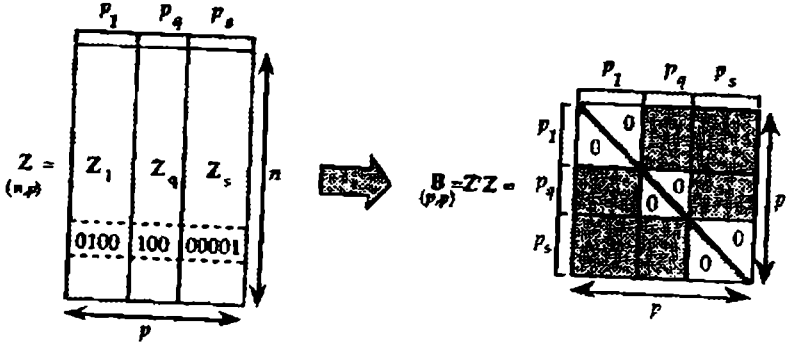


Figura 1.4-3 Construcția tabelului Burt pornind de la tabelul disjunctiv complet **Z**.

Se notează cu **D** matricea diagonală $p \times p$ definită de relațiile

$$d_{jj} = b_{jj} = z_j$$

$$d_{jj'} = 0 \quad (\forall) \quad j \neq j' \qquad j, j' = \overline{1, p}$$

Matricea **D** poate fi de asemenea considerată ca fiind formată din s^2 blocuri. Numai cele s matrici diagonale $\mathbf{D}_q = \mathbf{Z}'_q \mathbf{Z}_q \quad (q = 1, \ldots, s)$ ce formează blocurile diagonale ale lui **B** sunt matrici nenule.

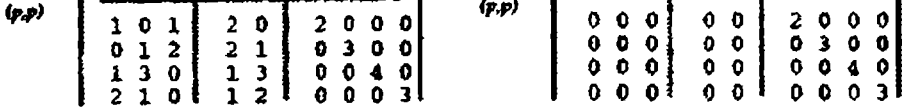


Figura 1.4-4 Tabloul Burt **B** și matricea diagonală **D** asociată (datele sunt din figurile **Figura 1.4-1** și **Figura 1.4-2**)

1.4.2 PRINCIPIILE ACM

Analiza corespondențelor multiple este analiza corespondențelor simple a unui tabel disjunctiv complet.

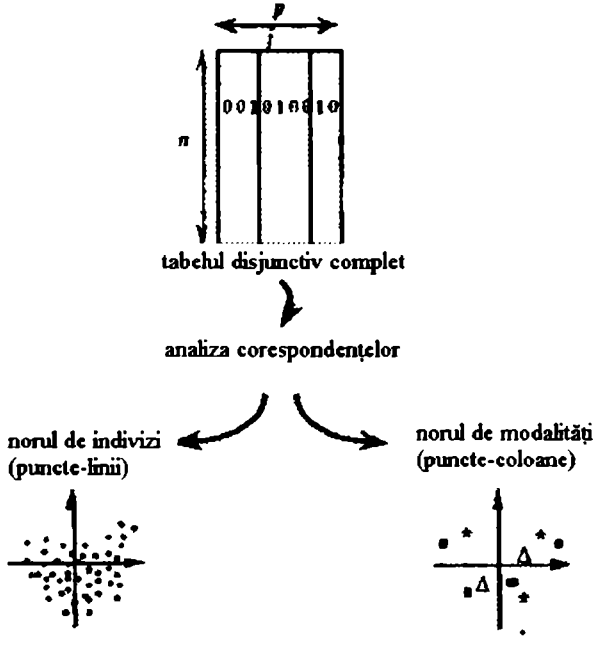


Figura 1.4-5 Analiza de corespondență multiplă

În consecință:

- se aplică aceleași transformări tabelului de date pentru obținerea profilurilor linie/coloană;
- aceleași ponderi ale punctelor funcție de profilurile marginale;
- aceeași distanță, distanța χ^2.

Așadar, indivizii sunt toți afectați de o pondere identică egală cu $m_i = \frac{1}{n}$, $i = \overline{1,n}$.

Fiecare modalitate j este ponderată de frecvența sa $m_j = \frac{z_j}{ns}$.

În $\mathbb{R}^n$ distanța χ^2 între modalități, pe un tabel disjunctiv se scrie

$$d^2(j,j') = \sum_{i=1}^{n} n \left(\frac{z_{ij}}{z_j} - \frac{z_{ij'}}{z_{j'}} \right)^2$$

şi este nulă dacă modalităţile j şi j' sunt alese de aceeaşi indivizi; în plus modalităţile de efectiv scăzut (alese de puţini indivizi) sunt depărtate faţă de celelalte modalităţi.

În $\mathbb{R}^p$ distanta χ^2 între indivizi, pe un tabel disjunctiv se scrie

$$d^2(i,i') = \frac{1}{s}\sum_{j=1}^{p}\frac{n}{z_{\cdot j}}\left(z_{ij} - z_{i'j}\right)^2$$

şi este nulă dacă indivizii i şi i' au ales aceleaşi modalităţi; ei sunt cu atât mai depărtaţi cu cât au răspuns mai diferit.

În plus, trebuie observat, că o modalitate j intervine în distanţa dintre indivizi cu atât mai mult cu cât masa sa este mai mică.

Reluând rezultatele analizei de corespondenţă şi notaţiile adoptate avem:

$\mathbf{F} = \frac{1}{ns}\mathbf{Z}$ cu termenul general $f_{ij} = \frac{z_{ij}}{ns}$;

$\mathbf{D}_p = \frac{1}{ns}\mathbf{D}$ cu termenul general $f_{\cdot j} = \delta_{ij}\frac{z_{\cdot j}}{ns}$;

$\mathbf{D}_n = \frac{1}{n}\mathbb{I}_n$ cu termenul general $f_{i\cdot} = \frac{\delta_{ij}}{n}$.

Pentru a găsi axele factoriale $\mathbf{u}_\alpha$ se diagonalizează matricea:

$$\mathbf{S} = \mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1} = \frac{1}{s}\mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1}$$

cu termenul general (atenţie, s fără indice desemnează numărul de întrebări)

$$s_{jj'} = \frac{1}{s \cdot z_{\cdot j'}}\sum_{i=1}^{n} z_{ij} z_{ij'}.$$

În $\mathbb{R}^p$, ecuaţia celei de-a α-a axă factorială $\mathbf{u}_\alpha$ este

$$\frac{1}{s}\mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha.$$

Ecuaţia celui de al α-lea factor $\varphi_\alpha = \mathbf{D}^{-1}\mathbf{u}_\alpha$ (componentă principală) este

$$\frac{1}{s}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z}\varphi_\alpha = \lambda_\alpha \varphi_\alpha$$

Analog, ecuaţia celui de al α-lea factor ψ_α în $\mathbb{R}^n$ este

$$\frac{1}{s}\mathbf{Z}\mathbf{D}^{-1}\mathbf{Z}'\psi_\alpha = \lambda_\alpha \psi_\alpha$$

Factorii φ_α şi ψ_α (de normă λ_α) reprezintă coordonatele punctelor linie şi a punctelor coloană pe axa factorială α.

Relaţiile de tranziţie între factorii φ_α şi ψ_α sunt:

$$\varphi_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{D}^{-1}\mathbf{Z}'\psi_\alpha; \qquad \psi_\alpha = \frac{1}{s\sqrt{\lambda_\alpha}}\mathbf{Z}\varphi_\alpha.$$

Coordonatele factoriale ale individului i pe axa α sunt date de:

$$\psi_{\alpha,i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^{p} \frac{z_{ij}}{z_{i.}} \varphi_{\alpha,j} = \frac{1}{s\sqrt{\lambda_\alpha}} \sum_{j \in p(i)} \varphi_{\alpha,j}$$

unde $p(i)$ desemnează mulțimea modalităților alese de individul i.

Corolar 1.4-1 *Modulo coeficientul* $\frac{1}{\sqrt{\lambda_\alpha}}$ *individul i se găseşte proiectat în planul factorial principal în centrul de greutate (punctul de coordonate media aritmetică) al modalităților pe care le-a ales.*

Analog, coordonatele factoriale ale modalității j pe axa α sunt date de:

$$\varphi_{\alpha,j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{n} \frac{z_{ij}}{z_{.j}} \psi_{\alpha,i} = \frac{1}{z_{.j}\sqrt{\lambda_\alpha}} \sum_{i \in n(j)} \psi_{\alpha,i}$$

unde $n(j)$ desemnează mulțimea indivizilor care au ales modalitatea j.

Observație. În formulele de mai sus modalitățile/indivizii nu sunt ponderați; coordonatele sunt simple medii aritmetice.

Norul modalităților din $\mathbb{R}^n$ poate fi descompus în s submulțimi, a q-a submulțime – subnor- corespunzând mulțimii p_q a modalităților variabilei q.

Corolar 1.4-2 *Centrele de greutate al celor s submulțimi ale norului modalităților din* $\mathbb{R}^n$ *coincid cu centrul de greutate al norului global.*

Demonstrație Într-adevăr, coordonatele punctelor subnorului relativ la variabila q sunt coloanele lui $\mathbf{Z}_q\mathbf{D}_q^{-1}$ iar elementele de pe diagonala principală a lui $\frac{1}{n}\mathbf{D}_q$ sunt masele relative ale celor p_q puncte ale subnorului.

Deoarece $\sum_{j \in p(q)} z_{ij} = 1$ a i-a componenta a centrului de gravitate a subnorului este

$$G_{q,i} = \sum_{j \in p(q)} \frac{d_{jj}}{n} \cdot \frac{z_{ij}}{d_{jj}} = \frac{1}{n} = G_i$$

rezultă că $G_{q,i}$ nu depinde de q ($p(q)$ desemnează mulțimea modalităților variabilei nominale q).

□

Observație

1. Dacă tabelul **Z** nu este complet disjunctiv (adică, dacă pentru cel puțin un individ nici o modalitate a unei întrebări nu a fost aleasă), modalitățile acelei variabile nu mai sunt centrate pe centrul de greutate al norului global.
2. Codificarea disjunctivă completă permite transformarea unei variabile continue într-o variabilă nominală a cărei modalități sunt clase ordonate. În această situație este util să se traseze traiectoria care poate sugera legături neliniare între această variabilă si axele factoriale.

Coordonatele modalităților în $\mathbb{R}^n$ sunt coloanele tabelului $\mathbf{ZD}^{-1}$; ele (coloanele, adică) generează un subspațiu a cărui dimensiune este rangul lui $\mathbf{ZD}^{-1}$, deci rangul lui $\mathbf{Z}=\left[\mathbf{Z}_1,\mathbf{Z}_2,...,\mathbf{Z}_q,...,\mathbf{Z}_s\right]$. Reamintim că toate subspațiile generate de coloanele lui $\mathbf{Z}_q$, $q=\overline{1,s}$ au în comun prima bisectoare (căci $\sum_{j\in p(q)} z_{ij}=1$). Rangul maxim a lui **Z** este deci:

$$p_1+(p_1-1)+...+(p_s-1)=p-s+1$$

Rangul maxim al matricii de diagonalizat $\mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z}$ va fi deci $p-s+1$. Dar in analiza norului în raport cu originea O, prima bisectoare este vectorul propriu corespunzând valorii proprii 1.

În analiza în raport cu centrul de greutate G vor fi găsite deci $p-s$ valori proprii nenule. Alegând o bază în suportul norului, ne putem restrânge la a căuta valorile proprii ale unei matrici de ordin $p-s$.

1.4.3 CALCULUL INERȚIEI

Distanța de la o modalitate *j* şi centrul de greutate *G* este

$$d^2(j,G)=(j-G)'\,\mathbf{D}_n^{-1}(j-G)=n\sum_{i=1}^{n}\left(\frac{z_{ij}}{z_{.j}}-\frac{1}{n}\right)^2$$

$$=n\left[\sum_{i=1}^{n}\frac{z_{ij}^2}{z_{.j}^2}-2\frac{1}{n}\sum_{i=1}^{n}\frac{z_{ij}}{z_{.j}}+\frac{1}{n^2}\sum_{i=1}^{n}1\right]$$

$$=n\left[\frac{1}{z_{.j}^2}\sum_{i}z_{ij}^2-2\cdot\frac{1}{n}\cdot\frac{1}{z_{.j}}\sum_{i}z_{ij}+\frac{1}{n^2}\cdot n\right]$$

$$=n\left(\frac{1}{z_{.j}}-2\cdot\frac{1}{n}+\frac{1}{n}\right)=\frac{n}{z_{.j}}-1$$

căci $z_{ij}^2=z_{ij}$ şi $\sum_i z_{ij}=z_{.j}$

Inerția $I(j)$ a unei modalități *j* este, prin definiție:

$$I(j)=m_j d^2(j,G)\ \text{cu}\ m_j=\frac{z_{.j}}{ns}$$

rezultă $I(j)=\frac{1}{s}\left(1-\frac{z_{.j}}{n}\right)$.

Corolar 1.4-3 *Inerția unei modalități este cu atât mai mare cu cât efectivul acestei modalități $z_{.j}$ (numărul de indivizi care au ales-o) este mai mic.*

Maximul $\frac{1}{s}$ va fi atins pentru modalitățile de efectiv nul. În consecință, se va evita, în momentul codificării, introducerea unor modalități susceptibile a fi alese de puțini indivizi tocmai pentru a nu introduce perturbații în primele axe factoriale.

Inerția $I(q)$ a unei întrebari q este , prin definiție

$$I(q)=\sum_{j=1}^{p_q} I(j)=\frac{1}{s}\left(p_q-1\right).$$

Corolar 1.4-4 *Inerția unei întrebări este cu atât mai mare cu cât numărul de modalități asociat, p_q, este mai mare. Minimul $\frac{1}{s}$ este atins de întrebările cu doar două modalitățide răspuns. În consecință, dacă se doreşte ca toate întrebările să joace un rol aproximativ egal atunci se va echilibra sistemul de întrebări (variabilele vor fi „decupate" într-un număr egal de modalități).*

Inerția totală este

$$I=\sum_{q=1}^{s} I(q)=\sum_{j=1}^{p}\frac{z_{.j}}{ns}d^2(j,G)=\frac{p}{s}-1 \text{ căci } \sum_{q=1}^{s} p_q=p.$$

În particular $I = 1$ dacă toate întrebările au două modalități de răspuns (adica $p = 2s$).

În consecință, depinzând exclusiv de numărul de întrebări şi de modalitățile asociate acestora, inerția globală nu are în cazul ACM (ca şi în cazul ACP normat, de altfel) nici o semnificație statistică (căci nu depinde de legătura între variabile).

1.4.4 REGULI DE INTERPRETARE

A spune că există afinități între răspunsuri este acelaşi lucru cu spune că există indivizi care au ales simultan toate sau aproape toate aceleaşi răspunsuri.

Analiza corespondențelor multiple pune atunci în evidență tipuri de indivizi care au profile asemănătoare din punct de vedere al atributelor alese spre a-i descrie. Ținând cont de distanțele între elementele tabelului disjunctiv complet şi a relațiilor baricentrice particulare:

- proximitatea între indivizi semnifică faptul că au ales global aceleaşi modalităţi ca răspuns la întrebările puse
- proximitatea între modalităţi ale unor întrebări diferite semnifică că ele au fost alese ca răspuns de grupe de indivizi asemănători (căci aşa cum s-a demonstrat mai sus, ele corespund centrelor de greutate ale acelor grupe de indivizi);
- proximitatea între modalităţile aceleiaşi întrebări semnifică faptul că grupele de indivizi care le-au ales sunt asemănătoare (din construcţie modalităţile unei aceleiaşi variabile se exclud).

Regulile de interpretare a rezultatelor (coordonate, contribuţii, cosinus pătrat) privind elementele active ale unei ACM sunt asemănătoare cu cele corespunzătoare unei ACS. În plus, se poate calcula contribuţia unei variabile-întrebări la factorul α sumând contribuţiile modalităţilor acestora la factorul respectiv:

$$Cr_\alpha(q) = \sum_{j \in p(q)} Cr_\alpha(j) = \sum_{j=1}^{p_q} \frac{z_{.j}}{ns} \cdot \frac{\varphi_{\alpha,j}^2}{\lambda_\alpha} = \frac{1}{ns\lambda_\alpha} \sum_{i=1}^{p_q} z_{.j} \varphi_{\alpha,j}^2$$

1.4.5 PRINCIPII DE TRANSFORMARE A VARIABILEI CONTINUE ÎN VARIABILĂ DISCRETĂ

Variabilele continue pentru a fi active într-o ACM trebuiesc transformate în variabile nominale (discrete). În acest proces apar următoarele probleme:

- câte clase trebuie alese şi cum;
- unde trebuie plasate marginile claselor.

Din rezultatele de mai sus au reieşit următoarele cerinţe: constituirea de modalităţi de efective comparabile şi decuparea variabilelor astfel încât să avem un număr de modalităţi comparabile. Din practică, un nr. de 4-8 modalităţi par să acopere majoritatea aplicaţiilor.

În consecinţă, este vorba de a găsi un compromis între un decupaj acceptabil tehnic din punct de vedere al principiilor de mai sus şi un decupaj care exhibă cel mai bine informaţia ce trebuie reţinută; în concluzie nu se poate recurge la algoritmi „orbi" pentru a elabora un decupaj satisfăcător. Astfel, se poate reţine o modalitate cu un efectiv scăzut dacă aceasta este importantă pentru studiu; analog, pentru a selecţiona bornele claselor unei variabile continue se va respecta, mai degrabă, pragurile naturale în contextul studiului său, reieşite ca semnificative după examenul histogramei, decât decupajul în clase de mase egale dar (uneori) inadecvate.

Transformarea variabilelor continue în variabile nominale duce la pierderea unei părţi din informaţia brută dar prezintă unele avantaje:

- utilizarea simultană a variabilelor nominale şi continue în ACM;
- validarea a posteriori a datelor permiţând observarea eventualelor clase contigue;
- punerea în evidenţă a eventualelor legături neliniare între variabilele continue. Asupra acestui ultim aspect vom insista puţin

Dându-se p variabile continue $x^1, x^2, ..., x^p$ ACP caută o combinație liniară de dispersie maximală

$$\max V\left(\sum_{j=1}^{p} u_j x^j\right)$$

Dacă se urmărește punerea în evidență a unor relații neliniare se vor căuta transformări funcționale $f^1\left(x^1\right), ..., f^p\left(x^p\right)$ ale variabilelor astfel încât să se realizeze

$$\max V\left(\sum_{j=1}^{p} f^j\left(x^j\right)\right)$$

Numărul de indivizi fiind finit trebuie să ne limităm la transformări funcționale alese într-o mulțime finită.

Să alegem pe f^j funcții scară (constante pe porțiuni). Se cunoaște că aceste funcții permit aproximarea oricărei funcții continue (teorema lui Weierstrass).

Concret, vom împărți intervalul de variatie a lui x^j in m^j clase. $f^j\left(x^j\right)$ va fi deci o funcție cu valorile $a_1, a_2, ..., a_{m_j}$ pe intervalele de decupaj ce se explicitează sub forma unei combinații liniare de funcții indicator ale intervalului de decupaj având coeficienții $a_1, a_2, ..., a_{m_j}$.

Criteriul $\max V\left(\sum f^j\left(x^j\right)\right)$ este identic cu $\max V\left(\sum_j \mathbf{Z}_j a_j\right)$.

Soluția este dată de primele componente ale ACM pe tabelul $\mathbf{Z} = \left[\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_q, ..., \mathbf{Z}_s\right]$.

Sub rezerva de a avea suficiente observații în fiecare clasă se poate astfel utiliza, pentru evidențierea unor legături neliniare, în locul unei ACP pe tabelul **X** o ACM pe tabelul **Z** obținut din **X** ca mai sus.

1.4.6 VALORI–TEST PENTRU MODALITĂȚI SUPLIMENTARE

Coordonata factorială $\varphi_{\alpha,j}$ a unei modalități j pe axa α este modulo coeficientul $\frac{1}{\sqrt{\lambda_\alpha}}$, media aritmetică a coordonatelor $\psi_{\alpha i}$ a indivizilor care au ales această modalitate ca răspuns, adică $\varphi_{\alpha,j} = \frac{1}{z_j \sqrt{\lambda_\alpha}} \sum_{i \in n(j)}^{n} \psi_{\alpha,i}$.

Să presupunem că o modalitate suplimentară j a fost aleasă de n_j indivizi $\left(n_j = z_j\right)$. Ne propunem să testăm dacă această modalitate a fost aleasă întâmplător sau alegerea ei are o semnificație.

Fie ipoteza H_o: „cei n_j indivizi au fost aleşi aleator din eşantionul de n indivizi" (alegerea este presupusă fără revenire).

În ipoteza H_o media coordonatelor $\psi_{\alpha i}$ a celor n_j indivizi este o variabilă aleatoare $x_{\alpha_j} = \frac{1}{n_j} \sum_{i \in n(j)} \psi_{\alpha i}$ de $E[x_{\alpha i}] = 0$ şi $D^2_{H_o}[x_{\alpha_j}] = \frac{n-n_j}{n-1} \cdot \frac{\lambda_\alpha}{n_j}$ repartizată hipergeometric.

Rezultă $E[\varphi_{\alpha i}] = 0$ şi $D^2_{H_o}[\varphi_{\alpha i}] = \frac{n-n_j}{n-1} \cdot \frac{1}{n_j}$.

Definiţia 1.4-3 $t_{\alpha i} = \sqrt{n_j \frac{n-n_j}{n-1}} \cdot \varphi_{\alpha_j}$ se numeşte *valoare-test* şi măsoară în număr de ecarturi-tip distanţa între modalitatea j, adică qvasi-baricentrul celor n_j indivizi, şi originea axei factoriale α.

Conform teoremei limită-centrală, distribuţia lui $t_{\alpha i}$ tinde la o $N(0,1)$.

Astfel, poziţia unei modalităţi este interesantă într-o direcţie α dată, dacă sub-norul a cărui baricentru este, ocupă o zonă apropiată de această axă şi destul de depărtată de centrul de greutate global în direcţia axei.

Valoarea-test este un criteriu care permite o apreciere rapidă a poziţiei, „semnificativă" sau nu, unei modalităţi pe o axă. Se consideră, în general, ca ocupând o poziţie semnificativă modalităţile a căror valoare- test, în modul, este mai mare sau egală cu 2, ceea ce corespunde unui prag de semnificaţie de 95%.

Propoziţia 1.4-1 *Analiza corespondenţelor aplicată unui tabel disjunctiv complet* **Z** *este echivalentă cu analiza tabelului Burt asociat, în sensul că produce aceeaşi factori.*

Demonstraţie φ_α este al α-lea vector propriu –factor al unei ACS pe un tabel **Z**– al matricii $\mathbf{S} = \frac{1}{s}\mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z} = \frac{1}{s}\mathbf{D}^{-1}\mathbf{B}$, adică $\mathbf{S}\varphi_\alpha = \lambda_\alpha \varphi_\alpha$.

Pentru ACS-ul tabelului **B** asociat lui **Z**, tabelul frecvenţelor relative **F** este

$$\mathbf{F} = \frac{1}{ns^2}\mathbf{B} \text{ şi } \mathbf{D}_p = \mathbf{D}_n = \frac{1}{ns}\mathbf{D}.$$

Matricea de diagonalizat este

$$\mathbf{S}^* = \frac{1}{s^2}\mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}\mathbf{B} \Rightarrow \mathbf{S}^* = \mathbf{S}^2$$

$$\frac{1}{s}\mathbf{D}^{-1}\mathbf{B}\varphi_\alpha = \lambda_\alpha\varphi_\alpha \;\Big|\times\frac{1}{s}\mathbf{D}^{-1}\mathbf{B} \Rightarrow \frac{1}{s^2}\mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}\mathbf{B}\varphi_\alpha = \lambda_\alpha\frac{1}{s}\mathbf{D}^{-1}\mathbf{B}\varphi_\alpha$$

$$= \lambda_\alpha \cdot \lambda_\alpha\varphi_\alpha = \lambda_\alpha^2\varphi_\alpha$$

Rezultă $\mathbf{S}^*\varphi_\alpha = \lambda_\alpha^2\varphi_\alpha$.

Factorii celor două analize sunt deci coliniare în $\mathbb{R}^p$ dar valorile proprii asociate diferă; cele rezultate din analiza lui **D**, notate λ_B sunt egale cu pătratul celor rezultate din analiza lui **Z**, adică $\lambda_B = \lambda^2$.

□

Factorul φ_α rezultat din analiza lui **Z** şi reprezentând coordonatele factoriale ale modalităţilor, are ca normă pe λ_α, in timp ce factorul corespunzând analizei lui **B**, notat φ_{B_α}, are ca normă pe λ_α^2.

Corolar 1.4-5 *Relaţia care leagă cele două sisteme de coordonate factoriale este* $\varphi_{B_\alpha} = \varphi_\alpha \sqrt{\lambda_\alpha}$.

2. METODE DE CLASIFICARE

Tehnicile de clasificare automată sunt destinate să producă grupări de linii sau de coloane ale unui tabel; este vorba, cel mai adesea, de obiecte sau indivizi descrişi printr-un număr de variabile sau de caractere.

Circumstanţale utilizării acestor metode sunt analoage cu cele ale metodelor de analiză factorială descrise în capitolul 1: utilizatorul se găseşte în faţa unui tabel rectangular de valori numerice. Acest tabel poate fi un tabel de variabile continue, un tabel de contingenţă sau un tabel de prezenţă-absenţă (valori de zero sau unu după cum un individ sau un obiect posedă sau nu un anumit caracter sau atribut). În anumite aplicaţii utilizatorul poate dispune de un tabel pătrat simetric de similarităţi sau de distanţă.

Există mai multe familii de algoritmi de clasificare: algoritmi ce conduc direct la *partiţii* cum sunt metodele de agregare în jurul centrilor mobili; *algoritmi ascendenţi* (sau algoritmi care construiesc clasele prin aglomerarea succesivă a câte două obiecte şi care furnizează o ierarhie de repatiţii de obiecte; în fine, *algoritmi descendenţi* (sau divizivi) care procedează prin dihotomii succesive ale mulţimii obiectelor şi care furnizează o ierarhie de partiţii. Ne vom limita în această lucrare la primele două tehnici de clasificare:

- grupările se *pot* face prin căutarea directă a unei partiţii afectând elementele la centrii provizori ai claselor, apoi prin recentrarea claselor şi agregarea iterativă a elementelor. Este vorba de tehnicile de *agregare în jurul centrilori mobili* tehnici înrudite cu metoda *norilori dinamici* sau metoda *k-means*, metode gratifiante în cazul tabelelor mari (secţiunea 2.1).
- grupările se pot face prin aglomerarea progresivă a elementelor două câte două. Este cazul clasificării ascendente ierarhice cu agregare după mai multe criterii. În lucrare sunt prezentate tehnica "saltului minimal", echivalentă dintr-un anumit punct de vedere, cu căutarea arborelui minimal şi tehnica agregării "după disperie", interesantă prin compatibilitatea rezultaelor sale cu unele rezultate din analiza factorială (secţiunea 2.2).

Aceste tehnice prezintă avantaje diferite, dar pot fi utilizate şi împreună. Este astfel posibilă o strategie de clasificare bazată pe un *algoritm mixt* bine adaptat partiţionării mulţimilor formate din mii de indivizi (secţiunea 2.3).

Metodele de clasificare sau de tipologie (ştiinţa care le studiază se numeşte *taxonomie*) au ca scop regruparea indivizilor înr-un număr restrâns de clase omogene. Este vorba deci, spre deosebire de demersul analizei factoriale de a descrie datele procedând la o reducere a numărului de indivizi (faţă de o reducere a numărului de variabile).

În cele ce urmează se vor avea în vedere doar metodele de clasificare automată; clasele vor fi obținute pe baza algoritmilor formalizați și nu prin metode subiective sau vizuale ce fac apel la inițiativa practicianului!

2.1 GENERALITĂȚI

În taxonomie informația utilă se prezintă sub forma unui tabel n x n conținând distanțele sau disimilaritățile dintre cei n indivizi de clasificat.

Reamintim, de aceea, că

Definiția 2.1-1 Fie E mulțimea celor n obiecte de clasificat, se numește distanță o funcție $d: E \times E \to \mathbb{R}_+$ cu proprietățile:

$d(i,j) = d(j,i)$ (simetrică);

$d(i,j) \geq 0$ (pozitivă);

$d(i,j) = 0 \Leftrightarrow i = j$ (idempotentă);

$d(i,j) \leq d(i,k) + d(k,j)$ (tranzitivă).

Pentru ca o distanță să fie euclidiană ea trebuie să fie generată de un produs scalar.

Când datele sunt prezentate sub forma unui tabel $\mathbf{X}$ de n indivizi cu p caracteristice numerice, cele mai des utilizate distanțe sunt:

- distanța euclidiană clasică, cu metrica $\mathbf{M} = \mathbf{I}$;
- distanța euclidiană cu metrica $\mathbf{M} = \mathbf{D}_{1/s^2}$;
- distanța Mahanalobis, $\mathbf{M} = \mathbf{V}^{-1}$;
- distanța L_1 în care $d(i,j) = \sum_k \left| x_i^k - x_j^k \right|$;
- distanța Minkowski, L_p, în care $d(i,j) = \left(\sum_k \left(x_i^k - x_j^k \right)^q \right)^{1/q}$

Definiția 2.1-2 Se numește *similaritate* o funcție $s: E \times E \to \mathbb{R}_+$, cu proprietățile:

$s(i,j) = s(j,i)$ (simetrică);

$s(i,j) \geq 0$ (pozitivă);

$s(i,i) \geq s(i,j)$ (nu există un individ mai asemănător decât el însuși).

Definiția 2.1-3 Se numește *disimilaritate* o funcție $d: E \times E \to \mathbb{R}_+$, cu proprietățile:

$d(i,j) = d(j,i)$ (simetrie);

$d(i,j) \geq 0$ (pozitivă);

$d(i,i) = 0$.

O situație frecvent întâlnită este cea în care datele se prezintă sub forma următoare: n indivizi sunt descriși prin prezența/absența a p caracteristici (datele inițiale sunt deci de formă binară).

Datele binare sunt „compactate" în n numere ce caracterizează fiecare cuplu de indivizi (deci 4 tabele n x n) astfel:

a) numărul de caracteristici comune;
b) numărul de caracteristici posedate de i dar nu de j;
c) numărul de caracteristici posedate de j dar nu de i;
d) numărul de caracteristici neposedate nici de i şi nici de j;

Atenție! Cu toate că logic a) şi d) sunt complementare cele două numere nu joacă acelaşi rol pentru datele reale; de exemplu, faptul că două vegetale nu cresc în acelşi loc, nu înseamnă în mod necesar că sunt asemănătoare.

Pe baza acestor 4 tabele se construieşte tabelul de similaritate sau prin complementare faţa de 1, tabelul de disimilaritate, utilizând diferiţi indici:

Jaccard: $\frac{a}{a+b+c}$;

Dice: $\frac{2a}{2a+b+c}$;

Ochiai: $\frac{a}{\sqrt{(a-b)(a+c)}}$;

Russel şi Rao: $\frac{a}{a+b+c+d}$;

Rogers şi Tanimoto: $\frac{a+d}{a+d+2(b+c)}$ etc.

2.2 ASPECTE COMBINATORII ALE CLASIFICĂRII

La prima vedere s-ar putea crede, deoarece E –multimea indivizilor de clasificat– este finită ($\text{card}(E)=n<\infty$), că problema clasificării este relativ facilă: se generează toate partiţiile posibile iar apoi se alege aceea/acelea care satisface/satisfac un criteriu de optimalitate dat.

Din păcate acest algoritm nu poate fi implementat încă în practică deoarece, chiar un calculator ce poate trata un milion de partiţii pe secundă are nevoie de 126 de mii de ani pentru a putea genera toate partiţiile unei mulţimi de numai 25 de indivizi!

Va trebui deci, în marea majoritate a situaţiilor să ne mulţumim cu soluţii aproximative.

Se noteaza cu $P_{n,k}$ numărul de partiţii în k clase a unei mulţimi de n elemente (numărul lui Stirling de speţa a doua).

Se observă uşor că:

$$P_{n,1} = 1 = P_{n,n}, \quad P_{n,n-1} = \frac{n(n-1)}{2}$$

$$P_{n,2} = 2^{n-1} - 1.$$

Se demonstrează prin inducție că:

$$P_{n,k} = P_{n-1,k-1} + kP_{n-1,k}.$$

Se poate de asemenea arăta că:

$$P_{n,k} = \frac{1}{k!}\sum_{i=1}^{k} C_k^i (-1)^{k-i} \cdot i^n.$$

şi deci, când $n \to \infty$, $P_{n,k} \approx \frac{k^n}{n!}$.

Se notează cu

$$P_n = \sum_{k=1}^{n} P_{n,k}$$

numărul total de partiții al unei mulțimi de n elemente (numerele lui Bell).

Dacă se convine ca $P_0 = 1$ atunci, se poate arăta, prin inducție, că

$$P_n = P_0 + (n-1)P_1 + C_{n-1}^2 P_2 + \ldots + P_{n-1} \text{ şi că } P_n = \frac{1}{e}\sum_{k=1}^{\infty} \frac{k^n}{n!}.$$

2.3 METODE DE CLASIFICARE NEIERARHICE

Aceste metode permit clasificarea rapidă a unor mulțimi destul de mari optimizând local un criteriu de tip inerție.

Se presupune că:

- cei n indivizi sunt puncte dintr-un spațiu euclidian $\subset \mathbb{R}^p$;
- se dorește clasificarea indivizilor în k clase.

Scopul fiecărei clasificări fiind acela de a obține clase cât mai omogene, iar statistic omogenitatea fiind caracterizată de dispersie, rezultă că o clasă va fi cu atât mai omogenă cu cât inerția norului de puncte ce o alcătuiește este mai mică.

Fie deci $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k$ centrele de greutate ale celor k clase, atunci

- inerția clasei C_i este:

$$I_i = \sum_{j=1}^{\text{card}(C_i)} p_j d^2(j, \mathbf{g}_i), \ (\forall)\ i = \overline{1,k}$$

cu p_j ponderea individului j;

- inerția intraclase este:

$$I_W = \sum_{j=1}^{k} P_j I_j$$

cu P_j ponderea clasei j;

- inerția interclase este:

$$I_B = \sum_{j=1}^{k} P_j d^2(\mathbf{g}_j, \mathbf{g})$$

cu $\mathbf{g}$ centrul de greutate al întregului nor de n indivizi.

Cum conform principiului lui König-Huyghens, inerția totala a norului este

$$I = I_B + I_W$$

un criteriu vizual de clasificare pentru a avea în medie clase omogene, constă în a căuta acea partiție în k clase pentru care inerția intraclase este minimă, deci inerția interclase este maximă.

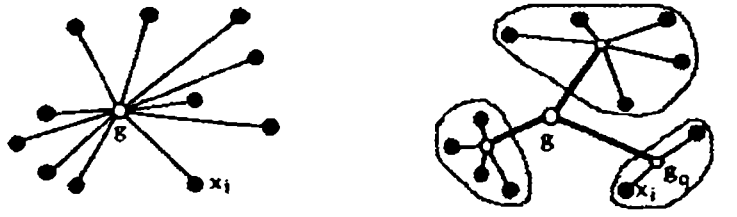


Figura 2.3-1 Descompunerea inerției conform principiului lui Huyghens

Trebuie să remarcăm, mai întâi, că acest criteriu presupune cunoașterea à priori a numărului de clase şi că nu este posibilă compararea a două partiții cu număr de clase diferit căci, cea mai bună partiție de k clase va avea o inerție intraclase superioară oricărei partiții de $k+1$ clase, iar la limită, cea mai bună partiție este cea trivială, în care fiecare individ formează o clasă (în acest caz $I_W = 0$ căci fiecare individ este propriul său centru de greutate).

2.3.1 METODA CENTRELOR MOBILE (A LUI FORGY)

Fie E o mulțime de n indivizi caracterizați de p variabile. Vom presupune spațiul $\mathbb{R}^p$ ce conține norul de n puncte-indivizi dotat cu o distanță corespunzătoare, notată d (adesea distanța euclidiană uzuală sau distanța χ^2). Se dorește constituirea a k clase. Etapele algoritmului sunt următoarele:

Pasul 1: Se aleg, în general aleator, k puncte distincte din E. Fie acestea $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$.

Se inițializează:

$j=0$ contorul de numărare al iterațiilor;

$I_W^{(j)}=\infty$ inerția intraclase (un număr foarte mare, dat).

Pasul 2: Se împarte mulțimea E in k clase astfel:

Pentru fiecare i, cu $i=\overline{1,k}$,

$E_{c_i}=\left\{\mathbf{e}\in E / d(\mathbf{e},\mathbf{c}_i)<d(\mathbf{e},\mathbf{c}_m), m=\overline{1,k}, m\neq i\right\}$.

Cazul inegalității se rezolvă prin tragere la sorți, în sensul că $\mathbf{e}$ este asignat aleator acelor partiții pentru care $d(\mathbf{e},\mathbf{c}_{i_1})=d(\mathbf{e},\mathbf{c}_{i_2})=\ldots=d(\mathbf{e},\mathbf{c}_{i_e})$.

Dacă $\text{card}(E_{c_i})=0$ atunci se generează aleator un nou centru $\mathbf{c}_i$.

Geometric, fiecare clasă este un domeniu poliedral convex determinat de hiperplanele mediatoare pe segmentele $\overline{\mathbf{c}_i\mathbf{c}_m}$ cu $m\neq i$ şi $m=\overline{1,k}$.

Pasul 3: Se calculează centrele de greutate ale partiției $\{E_{c_i}\}_{i=\overline{1,k}}$ şi se notează cu $\{\mathbf{g}_i\}_{i=\overline{1,k}}$.

Se calculează $I_W^{(j+1)}$ = inerția intraclase ale partiției $\{E_{c_i}\}_{i=\overline{1,k}}$.

Pasul 4: Dacă $j>N$ (N – numărul total de iterații admis, dat)

sau

$\left|I_W^{(j+1)}-I_W^{(j)}\right|\leq\varepsilon$ (ε -pragul sub care ameliorarea inerției intraclase este considerată nesemnificativă, dat)

atunci STOP;

altfel

$\mathbf{c}_i=\mathbf{g}_i$, $i=\overline{1,k}$;

$j=j+1$;

salt la Pasul 2.

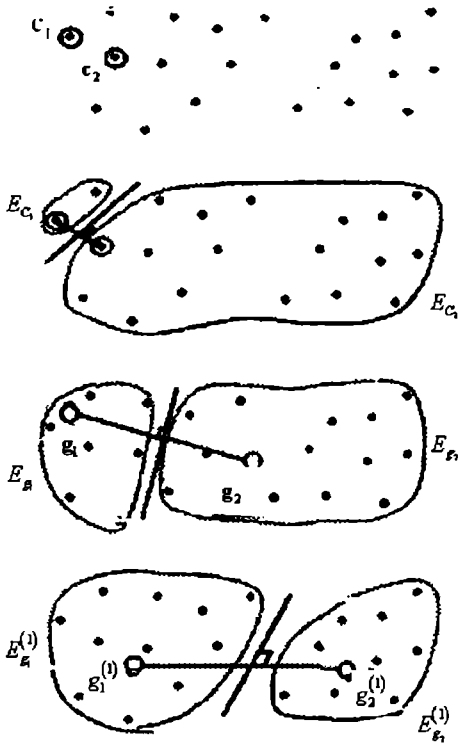


Figura 2.3-2 Etapele algoritmului lui Forgy

Propoziţia 2.3-1 *Algoritmul converge într-un număr finit de paşi, altfel spus,* $I_W^{(j+1)} \le I_W^{(j)}$ *şi* $j < \infty$.

Demonstraţie Având în vedere că algoritmul este iterativ este suficient să demonstrăm inegalitatea pentru $j=1$, iar pentru simplificarea calculelor presupunem că ponderile indivizilor sunt egale cu p iar ponderile claselor cu P.

Atunci, trebuie demonstrat că $I_W^{(2)} \le I_W^{(1)}$.

Conform algoritmului

$\{E_{c_i}\}$ este partiţia având punctele fiecărei clase grupate cât mai aproape de $\{\mathbf{c}_i\}$ şi cu centrele de greutate $\{\mathbf{g}_i^{(1)}\}$, deci

$$I_W^{(1)} = \sum_{i=1}^{k} P \sum_{j \in E_{c_i}} p d^2\left(j, \mathbf{g}_i^{(1)}\right) = p P \sum_{i} \sum_{j \in E_{c_i}} d^2\left(j, \mathbf{g}_i^{(1)}\right),$$

iar $\{E_{g_i}\}$ este partiția având punctele fiecărei clase grupate cât mai aproape de $\{\mathbf{g}_i^{(1)}\}$ și cu centrele de greutate $\{\mathbf{g}_i^{(2)}\}$, deci

$$I_W^{(2)} = \sum_{i=1}^{k} P \sum_{j \in E_{g_i}} p\, d^2\left(j, \mathbf{g}_i^{(2)}\right) = pP \sum_i \sum_{j \in E_{g_i}} d^2\left(j, \mathbf{g}_i^{(2)}\right).$$

Conform relației lui Huyghens

$$\sum_i \sum_{j \in E_{g_i}} d^2\left(j, \mathbf{g}_i^{(1)}\right) = I_W^{(2)} + \sum_i d^2\left(\mathbf{g}_i^{(1)}, \mathbf{g}_i^{(2)}\right)$$ căci $\{\mathbf{g}_i^{(1)}\}_i$ nu sunt centrele de greutate ale lui $\{E_{g_i}\}$.

Rezultă

$$I_W^{(2)} \le \sum_i \sum_{j \in E_{g_i}} d^2\left(j, \mathbf{g}_i^{(1)}\right)$$

cu inegalitate strictă dacă $\mathbf{g}_i^{(1)} \neq \mathbf{g}_i^{(2)}$; $(\forall) i = \overline{1,k}$.

Dar

$$\sum_{j \in E_{g_i}} d^2\left(j, \mathbf{g}_i^{(1)}\right) \le \sum_{j \in E_{c_i}} d^2\left(j, \mathbf{g}_i^{(1)}\right)$$ prin construcția celor două partiții, căci $\{E_{g_i}\}$ este partiția în care fiecare clasă E_{g_i} păstrează punctele cele mai apropiate de $\mathbf{g}_i^{(1)}$, deci $\sum_{j \in E_{g_i}} d^2\left(j, \mathbf{g}_i^{(1)}\right)$ este minimă. Egalitatea are loc doar dacă $\{E_{g_i}\} \equiv \{E_{c_i}\}$.

Cu acestea

$$I_W^{(2)} \le \sum_i \sum_{j \in E_{g_i}} d^2\left(j, \mathbf{g}_i^{(1)}\right) \le I_W^{(1)}.$$

Cum $\text{card}(E) = n < \infty$ rezultă $P_{n,k} < \infty$ ceeace implică $j < \infty$.

□

Experiența arată că viteza de convergență este rapidă.

Trebuie remarcat și faptul că, la fiecare pas nefiind necesar decât calculul a *nk* distanțe (distanțele dintrecei *n* indivizi și cele *k* centre de greutate) nu este necesar menținerea în memorie a tabelului cu cele $\frac{n(n-1)}{2}$ distanțe dintre indivizi.

Inconvenientele metodei sunt:

- *k* trebuie cunoscut à priori;
- optimul este dependent de alegerea inițială a punctelor $\{\mathbf{c}_i\}$

În metoda precedentă se așteaptă ca toți indivizii să fie afectați unei clase pentru a calcula centrul de greutate.

Metoda k-mediilor (k-means) a lui MacQueen, 1967 recalculează centrele de greutate după fiecare afectare.

Pentru a înlătura dependența metodei de punctele inițiale se utilizează metoda *norilor dinamici* a lui E. Diday, 1971 care este o generalizare a metodei centrelor mobile în sensul că fiecare clasă nu mai este reprezentată de centrul său de greutate ci de un nucleu de *g*-puncte (cele mai centrale, de exemplu), de o axă principală, de un plan principal.

2.4 CLASIFICARE IERARHICĂ

Principiile generale comune diverselor tehnici de clasificare ascendente ierarhice sunt simple. Aceste principii țin mai mult de bunul simț decât de o teorie formalizată de aceea este dificil să li se găsească o paternitate. Expunerile cele mai sistematice și cele mai vechi sunt poate cele ale lui Sokal&Sneath (1963) apoi cele ale lui Lance&Williams (1967).

Algoritmul constă în crearea, la fiecare etapă, a unei partiții obținută prin agregarea celor mai apropiate două elemente. Se va desemna prin *element* în același timp indivizii sau obiectele de clasat cât și grupările de indivizi generate de algoritm. Există diferite criterii de agregare de unde și un număr important de variante ale acestei tehnici.

Algoritmul nu furnizează o partiție în q clase a unei mulțimi de n obiecte ci o *ierarhie de partiții*. Această ierarhie se prezintă sub forma unui *arbore* numit și *dendogramă* și conține n-1 partiții. Interesul pentru acest arbore este dat de faptul că acesta poate furniza o idee despre numărul de clase ce există efectiv în populație. Fiecare „tăiere" a dendogramei furnizează o partiție având cu atât mai puține clase și clase cu atât mai puțin omogene cu cât tăierea se face mai sus.

2.4.1 ASPECTE FORMALE

Definiția 2.4-1 Fie E o mulțime finită. $\mathcal{H}$ mulțime de mulțimi din $\mathcal{P}(E)$ se numește *ierarhie* dacă și numai dacă

a) E și părțile lui E formate dintr-un element aparțin lui $\mathcal{H}$;

b) $(\forall)\ A,B \in \mathcal{H}$, $A \cap B \in \{A,B,\varnothing\}$.

Definiția 2.4-2 Elementele din $\mathcal{H}$ se numesc *partiții* ale mulțimii E.

Definiția 2.4-3 Elementele unei partiții a lui E se numesc *clase*.

Observații

- Fiecărei ierarhii îi corespunde un arbore de clasificare.
- Fiecare clasă dintr-o ierarhie este reuniunea claselor incluse în ea.

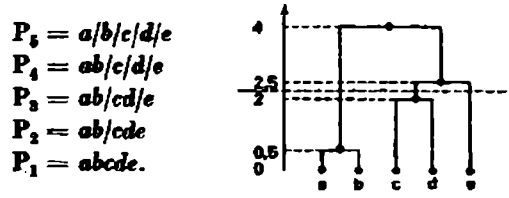
Dacă $\text{card}(E) = n < \infty$, atunci $\text{card}(\mathcal{H}) = n$ căci, datorită condiției b) din definiție, o partiție cu k clase se formează prin regruparea a două clase ale partiției cu $k+1$ clase.

Cum partiția P_n, cu n clase, este formată din elementele mulțimii E, câte un element în fiecare clasă, iar partiția P_1, cu o clasă, este formată din mulțimea E (ambele partiții aparțin prin definiție, condiția a), ierarhiei $\mathcal{H}$) $\mathcal{H}$ conține practic $n-2$ partiții netriviale ale lui E.

Definiția 2.4-4 Se numește *indice al ierarhiei* $\mathcal{H}$ o aplicație $i:\mathcal{H}\to\mathbb{R}_+$ crescătoare $((\forall)A,B\in\mathcal{H}$ cu $A\subset B\Rightarrow i(A)\subseteq i(B))$ și $i(A)=0\ \ (\forall)A\in P_n$.

Definiția 2.4-5 Indicile i al ierarhiei $\mathcal{H}$, dacă există, se mai numește și *nivel de agregare* iar ierarhia $\mathcal{H}$ dotată cu un astfel de indice se numește *ierarhie indexată*.

Exemplu: Fie $E=\{a,b,c,d,e\}$, atunci $n=5=\text{card}(E)$



cu $\quad i(\{a\})=i(\{b\})=i(\{c\})=i(\{d\})=i(\{e\})=0$

$i(\{f\})\equiv i(\{a,b\})=0{,}5 \qquad i(\{h\})\equiv i(\{c,d,e\})=2{,}5$

$i(\{g\})\equiv i(\{c,d\})=2 \qquad i(\{j\})\equiv i(\{a,b,c,d,e\})=4$

Observație:

a) În exemplul de mai sus indicile indică nivelul la care două clase s-au grupat (motivație pentru utilizarea denumirii de nivel de agregare); cu cât indicile este mai mare cu atât mulțimea este mai eterogenă.

b) Cunoscând arborele de clasificare este facil să se obțină o partiție cu un număr mai mic sau mai mare de clase; este suficient pentru aceasta să se taie arborele la un nivel dat și să se considere clasele date de ramurile care cad. Astfel, în exemplul de mai sus, se obține o partiție în 3 clase dacă se taie arborele de-a lungul liniei punctate, adică $\{\{a,b\},\{c,d\},\{e\}\}$.

Propoziția 2.4-1 *Fie E o mulțime, $\delta:E\times E\to\mathbb{R}_+$ o disimilaritate strictă pe E. Atunci:*

$$i(A)=\begin{cases}0 & \text{dacă} \quad A=\{i\},\ i\in E\\ \min\delta(i,j) & \text{dacă} \quad A=A_1\cup A_2,\ A_1\cap A_2=\varnothing \quad i\in A_1,\ j\in A_2\end{cases}$$

induce pe E o ierarhie indexată cu nivelul de agregare i.

Demonstrație Din definiție i este o funcție pozitivă și simetrică.

Trebuie demonstrate două afirmații:

a) că i induce pe E o ierarhie $\mathcal{H}$;

b) că i este indicile acelei ierarhii, adică i este o funcție crescătoare de partiții din $\mathcal{H}$.

a) Fie P_n -pariția formată din n clase a mulțimii E. Din definiția funcției de disimilaritate $i(A)=0\ (\forall)A\in P_n$. Se formează partiția P_{n-1} agregând elementele i, j din P_n pentru care $\delta(i,j)$ este minim. Cum δ este o disimilaritate strictă perechea (i,j) este unică. Din construcție $(\forall)A,B\in P_{n-1}\quad A\cap B=\{A,B,\varnothing\}$.

Se formează partiția P_{n-2} agregând elementele i, j din P_{n-1} pentru care $\delta(i,j)$ este minim și așa mai departe până la obținerea partiției P_1.

Fie $\mathcal{H}\overset{def}{=}\{P_1,P_2,...,P_n\}$. Din construcție $\mathcal{H}$ verifică cele două condiții din definiția ierarhiei.

b) Fie $A,B\in\mathcal{H}$ cu $A\subset B$. Rezultă $B=A\cup C$ și $A\cap C=\varnothing$. Din definiție $\delta(i,j)<\delta(i,k)\ (\forall)i,j\in A,\ k\in C$ căci δ este strictă și dacă $(\exists)k_o$ astfel încât $\delta(i,j)<\delta(i,k_o)$ atunci din agregarea lui A $k_o\in A$ și nu lui C.

Pentru un $j\in A$ fixat pentru moment, dar altfel oarecare, $\min\limits_{i\in A}\delta(i,j)<\min\limits_{i\in A}\left(\min\limits_{k\in C}\delta(i,k)\right)=i(B)$ (din definiție).

Din construcție $i(A)\in\left\{\min\limits_{i\in A}\delta(i,j)/\,j\in A\right\}$. Cum inegalitatea de mai sus este valabilă oricare ar fi $j\in A$, rezultă $i(A)<i(B)$.

□

2.4.2 STRATEGII DE AGREGARE:

Funcție de natura spațiului în care se găsesc îndivizii de agregat, vom distinge între:

- metoda Ward, dacă indivizii formează un nor într-un spațiu euclidian (de exemplu $\mathbb{R}^p$), deci între ei se poate calcula o distanță euclidiană;
- strategii de agregare pe disimilarități, dacă între indivizi se poate calcula o disimilaritate strictă.

2.4.2.1 Metoda Ward

Pe baza distanței euclidiene se poate evalua inerția și astfel se poate utilza principiul de agregare care reunește acele clase pentru care inerția interclase descrește cel mai puțin. (datorită principiului lui Huyghens inerția globală este suma inerțiilor interclase și intraclase. Cu cât clasele sunt mai omogene cu atât inerția intraclase este mai mică, deci inerția interclase mai mare. Clase omogene înseamnă clase cu indivizi cât mai puțini, deci partiții cât mai bogate; firesc ca prin fuzionarea a două clase inerția intraclase să

crească, deci inerţia interclase să scadă. Se va alege deci acea fuzionare pentru care inerţia interclase scade cel mai puţin, adică sunt grupate clasele cele mai asemănătoare, adică cele mai apropiate).

Lema 2.4-1 *Pierderea de inerţie interclase este dată de formula*

$$\delta(A,B)=\frac{P_A P_B}{P_A+P_B}d^2(\mathbf{g}_A,\mathbf{g}_B),$$

unde A, B sunt două clase, cu ponderile P_A *şi* P_B *şi centrele de greutate* $\mathbf{g}_A$ *şi* $\mathbf{g}_B$.

Demonstraţie Inerţia interclase este $I_B=\sum_{j=1}^{k}P_j\cdot d^2(\mathbf{g}_j,\mathbf{g})$; suma va conţine deci şi termenii $P_A\cdot d^2(\mathbf{g}_A,\mathbf{g})+P_B\cdot d^2(\mathbf{g}_B,\mathbf{g})$.

După fuziunea celor două clase, dacă se notează cu $\mathbf{g}_{AB}$ centrul de greutate al noii clase atunci cei doi termeni vor fi înlocuiţi de $(P_A+P_B)\cdot d^2(\mathbf{g}_{AB},\mathbf{g})$

Deci pierderea de inerţie interclase este dată de diferenţa

$$P_A d^2(\mathbf{g}_A,\mathbf{g})+P_B d^2(\mathbf{g}_B,\mathbf{g})-(P_A+P_B)\cdot d^2(\mathbf{g}_{AB},\mathbf{g}) \qquad (1)$$

Din construcţie $\mathbf{g}_{AB}=\frac{P_A\mathbf{g}_A+P_B\mathbf{g}_B}{P_A+P_B}$ adică centrul de greutate al noii clase este pe segmentul $\overline{\mathbf{g}_A\mathbf{g}_B}$.

În $\Delta\mathbf{g}\mathbf{g}_A\mathbf{g}_B$ utilizând o generalizare a teoremei medianei $\left(m_c^2=\frac{1}{2}a^2+\frac{1}{2}b^2-\frac{1}{4}c^2\right)$ rezultă

$$d^2(\mathbf{g},\mathbf{g}_{AB})=\frac{P_A}{P_A+P_B}d^2(\mathbf{g}_A,\mathbf{g})+\frac{P_B}{P_A+P_B}d^2(\mathbf{g}_B,\mathbf{g})-\frac{P_A P_B}{(P_A+P_B)^2}d^2(\mathbf{g}_A,\mathbf{g}_B) \qquad (2)$$

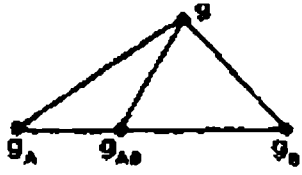


Figura 2.4-1 Teorema medianei generalizate aplicată în $\Delta\mathbf{g}\mathbf{g}_A\mathbf{g}_B$

Introducând rezultatul din formula (2) în formula (1) se obţine rezultatul din enunţul lemei.

□

Lema 2.4-2 *Într-o ierarhie indexată agregată pe baza unei distanțe euclidiene suma indicilor de agregare este egală cu inerția totală.*

Demonstrație Conform principiului lui Huyghens $I = I_W + I_B$ cu I_B -inerția interclase și I_W -inerția intraclase.

La momentul inițial, când E este împărțită în n clase

$$I_W = (P_n) = 0 \Rightarrow I_B = (P_n) = I$$

La momentul final când E are o simplă clasă

$$I_B(P_1) = 0 \Rightarrow I_W = (P_1) = I$$

Cum pierderea de inerție interclase, adică $I_B(P_S) - I_B(P_{S-1})$ este egală tocmai cu indicile de agregare rezultă $\sum_{s=2}^{n} i(P_s) = \sum_{s=2}^{n} \left[I_B(P_S) - I_B(P_{S-1}) \right] = I_B(P_n) - I_B(P_1) = I$.

□

Lema 2.4-3 *(generalizarea formulei Lance-Williams)*

$$\delta(C;(A,B)) = \frac{(P_A + P_C)\delta(A;C) + (P_B + P_C)\delta(B;C) - P_C\delta(A;B)}{P_A + P_B + P_C}.$$

Observație Lema 2.4-3 permite calculul disimilarității dintre două clase fără a fi necesară folosirea distanțelor euclidiene între centrele de greutate al acestor clase. În plus, nici centrele de greutate nu mai trebuiesc calculate.

Aşadar, odată calculate disimilaritățile dintre indivizi se poate lucra numai pe matrici de disimilarități prin aplicarea succesivă a formulei Lance-Williams.

Demonstrație Conform Lema 2.4-1

$$\delta(C;(A,B)) = \frac{P_C \cdot P_{AB}}{P_C + P_{AB}} d^2(\mathbf{g}_C, \mathbf{g}_{AB})$$

unde $P_{AB} = P_A + P_B$ (conform teoremei medianei generalizate). Cum

$$d^2(\mathbf{g}_C;\mathbf{g}_{AB}) = \frac{P_A}{P_A + P_B} d^2(\mathbf{g}_C;\mathbf{g}_A) + \frac{P_B}{P_A + P_B} d^2(\mathbf{g}_C;\mathbf{g}_B) - \frac{P_A P_B}{(P_A + P_B)^2} d^2(\mathbf{g}_A;\mathbf{g}_B)$$

iar, pe de altă parte, tot din Lema 2.4-1

$$\frac{P_C P_A}{P_C + P_A} d^2(\mathbf{g}_C;\mathbf{g}_A) = \delta(A,C); \quad \frac{P_C P_B}{P_C + P_B} d^2(\mathbf{g}_C;\mathbf{g}_B) = \delta(B,C)$$

și $\frac{P_A P_B}{P_A + P_B} d^2(\mathbf{g}_A;\mathbf{g}_B) = \delta(A,B)$

$$\Rightarrow \quad d^2(\mathbf{g}_C;\mathbf{g}_{AB}) = \frac{1}{P_C P_{AB}} \left[(P_C + P_A)\delta(A,C) + (P_C + P_B)\delta(B,C) - P_C\delta(A,B) \right]$$

$$\Rightarrow \quad \delta\big(C;(A,B)\big)=\frac{(P_A+P_C)\delta(A;C)+(P_B+P_C)\delta(B;C)-P_C\delta(A;B)}{P_A+P_B+P_C}$$

□

Rezultatul lemei permite enunțarea următorului algoritm:

Pasul 1 Se înlocuiește matricea **D** a distanțelor dintre indivizi cu matricea $\Delta_n=\left(\delta_{ij}\right)_{i=\overline{1,n}}^{j>i}$ cu $\delta_{ij}=\frac{P_iP_j}{P_i+P_j}d^2\left(\mathbf{e}_i,\mathbf{e}_j\right)$.

Pasul 2 În matricea Δ_n se caută $\min i,j\ \ \delta_{ij}$, se elimină linia și coloana j, iar linia și coloana lui i se notează cu $\widehat{ij}$, formându-se matricea Δ_{n-1}. Indicile de agregare al clasei $\widehat{ij}$ este δ_{ij}.

Pasul 3 Se calculează elementele matricii Δ_{n-1} astfel:

- se copiază coloanele matricii Δ_n;
- coloana $\widehat{ij}$ se calculează după formula $\delta\left(k;\widehat{ij}\right)=\frac{(P_i+P_k)\delta_{ik}+\left(P_j+P_k\right)\delta_{jk}-P_k\delta_{ij}}{P_i+P_j+P_k}$; (formula generalizată a lui Lance-Williams).

Pasul 4 Se pune $n=n-1$;

$\Delta_n=\Delta_{n-1}$;

Dacă $n=1$ atunci STOP; altfel salt la Pasul 2.

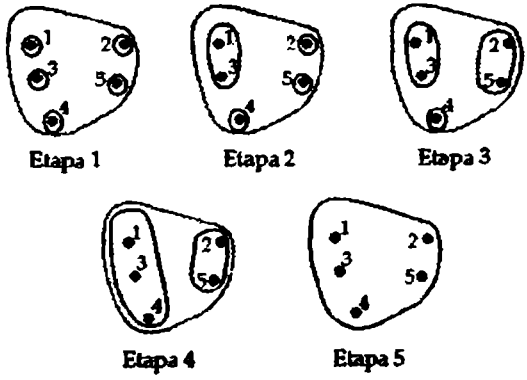


Figura 2.4-2 Aglomerarea progresivă a 5 puncte

Observație. La etapa inițială, inerția intraclase este nulă și inerția interclase este egală cu inerția totală a norului deoarece fiecare element terminal constituie la acest

nivel o clasă. În etapa finală, inerția interclase devine nulă iar inerția intraclase este echivalentă cu inerția totală pentru că la acest nivel există o partiție cu o singură clasă (vezi Figura 2.4-2).

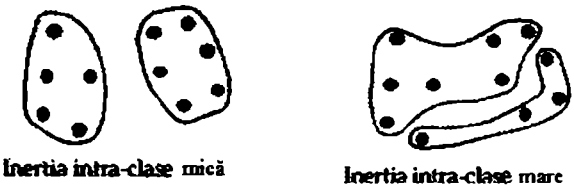


Figura 2.4-3 Calitatea globală a unei partiții

2.4.2.2 Strategii de agregare pe disimilarități

Dacă între indivizi este dată o matrice de disimilaritate strictă atunci se pot imagina mai multe soluții, mai mult sau mai puțin arbitrare. Cele mai utilizate sunt:

- distanța *saltului minimal* (simple linkage)

$$d(A,B) = \min \delta(e_i, e_j) \quad e_i \in A, e_j \in B$$

care favorizează mulțimile cu puncte apropiate;

- distanța *diametrului*

$$d(A,B) = \max \delta(e_i, e_j) \quad e_i \in A, e_j \in B$$

repară limitele primei distanțe dar punctele trebuiesc să fie apropiate;

- distanța *mediei*

$$d(A,B) = \frac{P_x \delta(x,z) + P_y \delta(y,z)}{P_x + P_y} \text{ cu } A = \{x,y\} \quad B\{z\}$$

Observație. Ierarhiile induse de diferitele distanțe sunt în general diferite. Se recomandă aşadar utilizarea mai multor tipuri de clasificări: acestea nu trebuie să difere prea mult când se priveşte partea superioară a arborelui de clasificare. Dacă totuşi acest lucru se întâmplă se poate conchide că mulțimea indivizilor se pretează prost la orice clasificare.

Exemplu

Fie matricea de disimilaritate (căci $\delta(c.e) > \delta(c,d) - \delta(d,e)$ pentru că $6 > 2 + \frac{1}{2}$) dintre indivizii $\{a,b,c,d,e\}$

	a	b	c	d	e
a	0	3	7	3	4
b		0	4	4	1
c			0	2	6
d				0	½
e					0

Să aplicăm algoritmul de clasificare ierarhică ascendentă folosind pe rând disimilaritățile enumerate mai sus.

- Astfel, pentru disimilaritatea saltului minimal (Inf) se obțin următoarele etape

1° $f=\{d,e\}; i(f)=\frac{1}{2}$

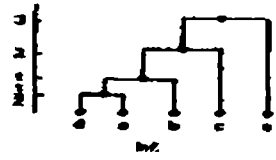
	a	b	c	f
a	0	3	7	3
b		0	4	1
c			0	2
f				0

2° $g=\{f,b\}; i(g)=1$

	a	b	g
a	0	7	3
b		0	2
g			0

3° $h=\{c,g\}; i(h)=2$

	a	h
a	0	3
h		0



4° $i=\{a,h\}; i(i)=3$

- Pentru disimilaritatea diametrului (Sup) se obțin următoarele etape

1° $f=\{d,e\}; i(f)=\frac{1}{2}$

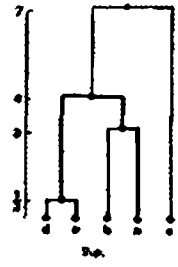
	a	b	c	f
a	0	3	7	4
b		0	4	4
c			0	6
f				0

2° $g=\{a,b\}; i(g)=3$

	c	f	g
c	0	6	7
f		0	4
g			0

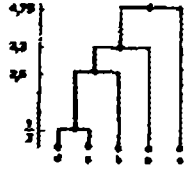
3° $h=\{f,g\}; i(h)=4$

	c	h
c	0	7
h		0



4° $i=\{h,c\}; i(c)=7$

- Analog pentru disimilaritatea medie se obține următoarea dendogramă



În pofida faptului că fiecare arbore începe cu agregarea lui *d* şi a lui *e* într-o singură clasă *f* urmează imediat diferențe importante atunci când se calculează distanțele de la *f* la ceilalți indivizi:

$$d_{\inf}(b,f)=\inf\left(d(b,d);d(b,e)\right)=1;$$

$$d_{\sup}(b,f)=\sup\left(d(b,d);d(b,e)\right)=4;$$

$$d_{\text{med}}(b,f)=2{,}5.$$

Să notăm însă că una din principalele dificultăți în clasificare constă în definirea unei distanțe sau disimilarități între indivizi, mai ales când aceştia sunt descrişi prin caractere calitative.

2.5 CLASIFICARE MIXTĂ ŞI DESCRIEREA STATISTICĂ A CLASELOR

Algoritmii de clasificare sunt mai mult sau mai puțin adaptați pentru volume mari de date. Astfel:

- metode de agregare în jurul centrilor mobili pot manipula volume mari cu prețuri mici dar au dezavantajul că produc partiții dependente de numărul ales de clase şi de centrii inițiali;
- metode de agregare ierarhice sunt „deterministe" (în sensul că dau întotdeauna acelaşi rezultat dacă datele inițiale sunt aceleaşi), dau indicații privind numărul de clase ce trebuie reținut dar sunt prost adaptate la volume de date mari.

Combinarea celor două metode a dat naştere unui algoritm mixt (*hybrid clustering*, Wong, 1982).

Algoritmul de clasificare mixtă procedează în trei etape: mulțimea elementelor de clasificat este partiționată (centrii mobili) în câteva zeci, eventual sute de partiții omogene; se procedează apoi la agregarea ierarhică a acestor grupe cu scopul ca dendrograma obținută să sugereze numărul de clase finale ce trebuiesc reținute; în fine, se optimizează (folosind iarăşi tehnica centrilor mobili) partiția obținută prin tăierea arborelui.

Figura 2.5-1 schematizează etapele algoritmului de clasificare mixtă.

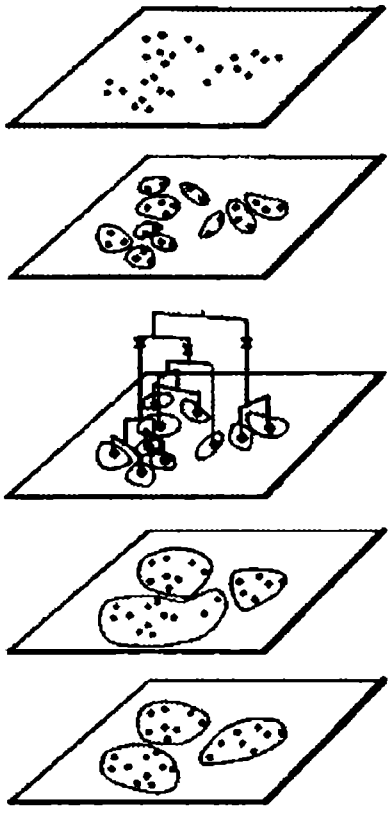


Figura 2.5-1 Schema clasificării mixte

Etapele algoritmului sunt:

1. *Partiționarea inițială*. Această etapă vizează obținerea rapidă şi cu un preţ scăzut a unei partiții de n obiecte în k clase omogene, unde k este mult mai mare decât s numărul de clase dorit dar mult mai mic decât n. În acest scop este utilizat algoritmul centrilor mobili. Optimalitatea nu este desigur atinsă dar partiția obținută poate fi ameliorată pornind de la grupările stabile (grupuri de indivizi sau elemente care apar mereu în aceleaşi clase). Aceste grupări vor fi elementele de bază în etapa următoare.
2. *Agregarea ierarhică a claselor obținute*. Această etapă constă în efectuarea unei clasificări ierarhice ascendente în care elementele terminale ale arborelui sunt cele k clase ale partiției inițiale. Scopul acestei etape este de a reconstitui clasele care au fost fragmentate şi de a agrega elementele aparent dispersate în jurul centrelor de origine. Arborele este construit după strategia Ward care ține seamă de mase în momentul alegerii elementelor de agregat.

3. *Partiția finală.* Partiția finală a populației este dată prin tăierea arborelui obținut în etapa precedentă. Omogenitatea claselor obținute poate fi optimizată prin reafectare.

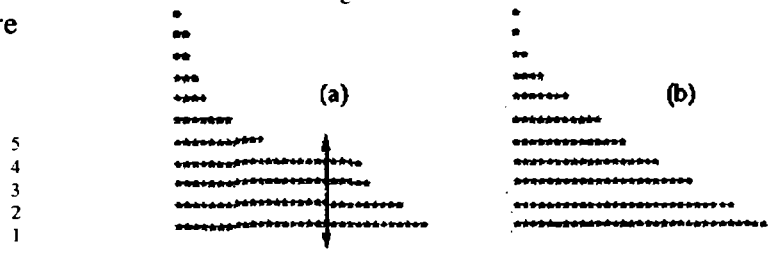
2.5.1 ALEGEREA CLASELOR PRIN „TĂIEREA" ARBORELUI

Alegerea nivelului de tăiere şi astfel al numărului de clase ale partiției poate fi facilitată de inspecția vizuală a arborelui; tăierea trebuie să se facă în intervalul dintre indici de valori mici, corespunzând unor clase omogene şi indici de valori mari ce disociază clase bine conturate.

Într-o manieră generală, cu cât se grupează mai mulți indivizi, altfel spus cu cât ne apropiem de vârful arborelui, cu atât mai mare va fi distanța între două clase vecine iar indicile de agregare va fi mai mare. Tăind arborele la nivelul unui salt important al acestui indice se poate spera în obținerea unei partiții de bună calitate în sensul că indivizii grupați sub nivelul de tăiere erau apropiați şi cei grupați deasupra nivelului de tăiere sunt necesarmente depărtați (ceeace corespunde definiției unei bune partiții).

În practică situația nu este însă atât de clar definită; ca şi în cazul analizei factoriale se utilizează criterii empirice –histograma indicilor de agregare.

Tabelul 2.5-1 Histogramele indicilor de nivel



Cazul favorabil
când apare un palier evident între al 4-lea şi al 5-lea indice (de jos în sus) –sugerând o bună poziție în 5 clase

Cazul nefavorabil

2.5.2 CARACTERIZAREA STATISTICĂ A CLASELOR

Elementele unei aceleiaşi clase se aseamănă din punct de vedere al criteriilor alese pentru a le descrie. Rămâne de precizat care sunt criteriile care se află la originea grupărilor obținute. Se procedează la descrierea automată a claselor ceeace constituie în practică o etapă indispensabilă oricărei proceduri de clasificare.

Descrierea automată a claselor este, în general, bazată pe compararea mediilor sau a procentelor din interiorul claselor cu mediile sau procentele obținute pe întreaga populație. Pentru a selecționa variabilele continue sau modalitățile variabilelor nominale caracteristice fiecărei clase se măsoară ecartul dintre valorile specifice clasei și valorile globale. Aceste statistici pot fi convertite într-un criteriu numit *valoare-test* care permite să operăm o selecție asupra variabilelor, desemnând astfel variabilele cele mai reprezentative (conform Morineau, 1984).

2.5.2.1 Valori test pentru variabile continue

Pentru a caracteriza o clasă prin variabile continue, se compară $\overline{x_k}$ media variabilei x în clasa k, cu media $\bar{x}$ în întreg norul. Valoarea-test este aici

$$t_k = \frac{\overline{x_k} - \bar{x}}{s_k(x)}$$

cu $s_k^2(x) = \frac{n - n_k}{n-1} \cdot \frac{s^2(x)}{n_k}$ estimatorul dispersiei lui x în clasa k ($s^2(x)$ este dispersia empirică a lui x în întreg norul). Se recunoaște aici în ${s_k}^2(x)$ dispersia unei medii, în cazul extragerii fără revenire a k elemente.

În ipoteza nulă a unei extrageri aleatoare, fără revenire, a n_k indivizi din clasa k, variabila $\overline{x_k}$ reprezentând media empirică în acea clasă are ca medie și dispersie empirică globală pe $\bar{x}$ respectiv ${s_k}^2(x)$.

Valoarea test $t_k(x)$ urmează, aproximativ, o distribuție Gauss-Laplace centrat–redusă (teorema limită centrală). Ea măsoară distanța între media clasei și media generală în ecarturi tip.

E de la sine înțeles că această interpretare nu are sens decât pentru o variabilă x suplimentară care nu a participat la construcția claselor (nu se poate stipula o independență între claselele unei partiții și variabilele care au participat la definirea partiției). Se calculează apoi probabilitatea ca variabila să depășească valoarea absolută a diferenței observate. Cu cât valoarea test este mai mare (cu atât probabilitatea este mai mică) cu atât ipoteza de a avea n_k valori ale variabilei x extrase la întâmplare dintre valorile posibile este discutabilă. În acest caz, media în clasă diferă de media generală și variabila este caracteristică clasei. Ordonarea variabilelor în funcție de probabilitățile crescătoare de a depăși media generală este echivalentă cu ordonarea în funcție de valorile-test descrescătoare.

Dacă interpretarea probabilistică a valorilor-test pentru variabilele active nu este licită este totuși posibil să fie folosite pentru a obține un clasament al acestora în vederea caracterizării fiecărei clase. Modulul acestor valori-test reprezintă atunci simple măsuri ale similarității între variabile și clasă.

2.5.2.2 Valori test pentru valori nominale

O modalitate (sau categorie) a unei variabile nominale este considerată caracteristică pentru clasă dacă abundența în clasă este apreciată ca semnificativ superioară față de restul populației. Notând cu n_{kj} numărul de indivizi având modalitatea j din cei n_k indivizi ai clasei k, n_j numărul total de indivizi având modalitatea j dintr-un total de n, abundența modalității j este definită, comparând procentul ei în clasă, adică $\frac{n_{jk}}{n_k}$, cu procentul în toată populația adică $\frac{n_j}{n}$.

În ipoteza nulă, unde cei n_k indivizi ai clasei k sunt extraşi aleator fără revenire, din populația de n indivizi, procentajul indivizilor clasei k având modalitatea j de o parte, şi procentajul indivizilor având modalitatea j în întreaga populație, pe de altă parte, ar trebui să coincidă, modulo o fluctuație aleatoare, adică:

$$\frac{n_{jk}}{n_k} \approx \frac{n_j}{n}.$$

În ipoteza de independență cei N indivizi ai clasei k care au modalitatea j este o variabilă aleatoare care urmează o lege hiper-geometrică de parametrii $\mathrm{Hg}\left(n_k, n, \frac{n_j}{n}\right)$ (n_k -numărul de succese dintr-un total de n cu probabilitatea de succes de $\frac{n_j}{n}$).

Suntem deci interesați de calculul lui

$$p_k(j) = \mathrm{Prob}\left(N \geq n_{kj}\right) = \sum_{x=n_{kj}}^{x=n_k} \mathrm{Prob}(N = x) = \sum \frac{C_{n_j}^{x} C_{n-n_j}^{n_k - x}}{C_n^{n_k}}.$$

Cu cât această probabilitate este mai mică, cu atât ipoteza unei extrageri aleatoare este mai dificil de acceptat. Vom folosi această probabilitate pentru a ordona modalitățile caracteristice clasei (cea mai caracteristică corespunzând celei mai mici probabilități).

Această probabilitate este adesea foarte mică; este comod să i se substituie valoarea $t_k(N)$ a variabilei Dauss-Laplace corespunzând aceleiaşi probabilități. Ea măsoară distanța între proporția în clasă şi proporția generală în număr de abateri standard ai legii normale. Cum

$$E(N) = n_k \frac{n_j}{n} \text{ şi } s_k^2(N) = n_k \frac{n - n_k}{n-1} \cdot \frac{n_j}{n}\left(1 - \frac{n_j}{n}\right) \Rightarrow t_k(N) = \frac{N - E(N)}{s_k(N)}.$$

Aceasta este valoarea-test pentru o modalitate a unei variabile nominale. Acesta este un criteriu statistic doar pentru variabilele ilustrative.

3. METODE EXPLICATIVE UZUALE

Acest capitol face legătura între *demersurile exploratorii* prezentate în capitolele 1 şi 2, şi *demersul inferenţial* şi *confirmatoriu* care constituie partea clasică cea mai amplă a statisticii matematice.

Să recapitulăm, pe scurt, caracteristicile celor două familii de metode cărora le corespund demersuri complementare.

Statistica descriptivă şi exploratorie permite rezumate şi grafice mai mult sau mai puţin elaborate, descrierea mulţimilor de date statistice şi stabilirea de relaţii între variabile, fără a acorda un rol privilegiat vreunei variabile. Concluziile obţinute în această etapă privesc doar datele studiate fără a fi generalizate la o populaţie mai largă. Analiza exploratorie se sprijină în mod esenţial pe noţiuni elementare care sunt noţiunile de medie şi dispersie, pe reprezentări grafice şi pe tehnice descriptive multidimensionale de tipul celor abordate în primele două capitole.

Statsitica inferenţială şi confirmatorie permite validarea sau infirmarea, pornind de la teste statistice sau modele probabiliste ipotezelor formulate a priori (sau urmarea unui demers exploratoriu) şi extrapolarea acestora de la nivelul eşantionului la cel al unei populaţii mai mari. Statistica confirmatorie face apel, în special, la metodele numite explicative8 şi previzionale destinate, aşa cum le indică numele, să explice apoi să prevadă, urmând reguli de decizie, o variabilă privilegiată cu ajutorul uneia sau mai multor variabile explicative.

Demersurile sunt complementare, explorarea şi descrierea trebuind, în general, să preceadă etapele explicative şi predictive. Într-adevăr, o explorare preliminară este adesea utilă pentru a avea o primă indee despre natura legăturilor între variabile şi pentru a trata cu prudenţa variabilele corelate şi deci redundante care riscă să încarce inutil modelul.

Metodele explicative prezentate în secţiunile 3.1-3.3 acoperă utilizările cele mai curente.

Analiza discriminantă (secţiunile 3.1 şi 3.2) este schematic vorbind, analogă cu regresia multiplă când variabila endogenă **y** este normală. În acest caz variabila de explicat defineşte clasele unei partiţii a priori a populaţiei. Scopul analizei îl constituie

---

8 Statistica nu explică nimic, dar furnizează elemente potenţiale de explicaţii. De altfel, termenii de variabilă explicativă sau variabilă de explicat nu sunt cei mai judicioşi. Se mai spune independent şi dependent sau exogen şi endogen. Ultimii doi termeni sunt poate cei mai adecvaţi dar nu sunt destul de evaocatori. Adjectivul independent este, în schimb, sursă de confuzie.

studierea legaturilor între variabilele explicative şi clasele partiţiei (secţiunea 3.1). Se definesc astfel funcţii discriminante care vor permite, într-o etapă decizională afectarea de noi indivizi la aceste calse (secţiunea 3.2).

Tehnicile de *segmentare prin arbore binar* (secţiunea 3.3) sunt prezentate în cadrul acestui capitol din diferite motive. Pe de o parte ele se aplică la toate variabelele oricare ar fi statutul sau natura lor, şi pe alta parte ele integrează simultan faza explicativă şi cea decizională. Ele constituie astfel o metodă de previziune foarte accesibilă, a căror rezultate sunt uşor de interpretat.

3.1 ANALIZĂ DISCRIMINANTĂ

Desemnăm sub numele de analiză discriminantă o familie de tehnici destinate să claseze (să afecteze la clase preexistente) indivizi caracterizaţi print-un număr de variabile continue sau discrete.

Originea metodei se găseşte în lucrările lui Fisher (1936) sau într-o manieră mai puţin directă în cele ale lui Mahalanobis (1930) .

Analiza discriminantă este una din tehnicile de analiză multidimensională cele mai folosite în practică (diagnostic automat, controlul calităţii, previziunea riscului, recunoaşterea formelor).

3.1.1 NOTAŢII ŞI FORMULAREA PROBLEMEI

Dispunem de n observaţii (sau indivzi) asupra a p variabile $(x_1, x_2, ..., x_p)$, observaţii repartizate în q clase definite à priori de variabila y –nominală cu q modalităţi (în cele ce urmează vom nota cu **y** vectorul n –dimensional cu componente numere naturale $\{1,...,q\}$ reprezentând numărul clasei din care face parte observaţia / individul i şi cu **Y** matricea disjunctivă $n \times q$ corespunzătoare).

Analiza discriminantă îşi propune, într-o primă etapă, să separe cât se poate de bine cele q clase cu ajutorul celor p variable explicative, iar apoi, într-o a doua etapă, să rezolve problema afectării unui individ nou, caracterizat prin cele p variabile la una din clasele deja identificate pe baza eşantionului de n indivizi (numit eşantion de învăţare).

Se disting, în consecinţă, două demersuri:

- primul descriptiv, ce constă în căutarea funcţiilor de discriminare liniare pe eşantionul de volum n (adică găsirea combinaţiilor liniare de variabile explicative $x_1, x_2, ..., x_p$ a căror valori separă cel mai bine cele q clase);
- al doilea decizional, ce constă în aflarea claselor de afectare a celor n' indivizi noi descrişi prin variabilele explicative $(x_1, x_2, ..., x_p)$ (numit eşantion de test).

  Este vorba aici de o problemă de clasare în clase preexistente, în opoziţie cu

problemele de clasificare (capitolul 2) care constau în construirea de clase cât mai omogen posibil într-un eşantion dat.

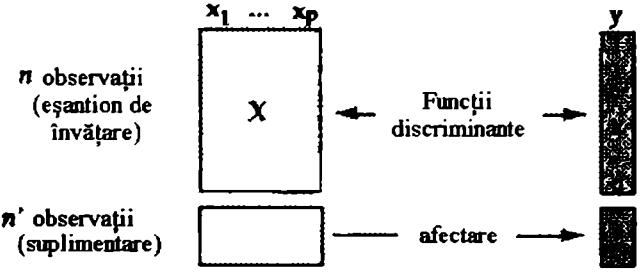


Figura 3.1-1 Principiul analizei discriminante

3.1.2 ANALIZA FACTORIALĂ DISCRIMINANTĂ

Fie tabelul observaţiilor $\mathbf{X} \in \mathcal{M}_{n\times p}(\mathbb{R})$ cu $\mathbf{X} = \left(x_{ij}\right)_{i=\overline{1,n}}^{j=\overline{1,p}}$

Cei n indivizi sunt împărţiţi în q clase (se cunoaşte afectarea fiecărui individ la o clasă, clase presupuse disjuncte).

Fiecare clasă k caracterizează un subnor I_k de n_k indivizi, unde

$$\sum_{k=1}^{q} n_k = n$$

Se notează cu $\mathbf{g}_k$ centrul de greutate al clasei k si cu $\mathbf{g}$ centrul de greutate al norului, adică

$$\mathbf{g}_k = \left(\bar{x}_{kj}\right)_{j=\overline{1,p}} \text{ unde } \bar{x}_{kj} = \frac{1}{n_k}\sum_{i\in I_k} x_{ij},$$

respectiv

$$\mathbf{g} = \left(\bar{x}_j\right)_{j=\overline{1,p}} \text{ cu } \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \sum_{k=1}^{q} \frac{n_k}{n}\bar{x}_{kj}.$$

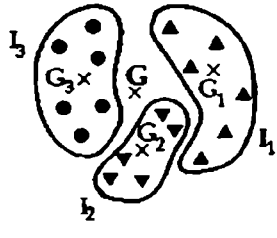


Figura 3.1-2 Reprezentarea norului de indivizi discriminaţi

Pentru precizarea ideilor să considerăm o mulțime **X** de date dintr-un spațiu bidimensional. Valorilor caracteristicilor x_1, și x_2 ale datelor sunt date de proiecțiile norului **X** pe axele de coordonate Ox, și Oy. Structura de clase a lui **X** se poate în acest caz detecta prin simpla inspecție vizuală, în unele situații se poate constata că nu există în **X** o structură de clase bine definită. Diferiți observatori pot indica diferite moduri de grupare a datelor în clase. Aceasta relevă faptul că puterea de discriminare a caracteristicilor (axelor) este slabă pentru datele considerate. Există două posibilități: fie că nu s-au ales cele mai bune caracteristici ale datelor, fie că datele sunt, prin natura lor, foarte asemănătoare. Este uneori posibil să determinăm un nou sistem de coordonate față de care structura de clase a norului **X** să fie mai evidentă decât în sistemul inițial. Axele noului sistem au deci o putere de discriminare a claselor din **X** superioară celei a axelor inițiale, în unele situații este suficient să determinăm o singură axă discriminantă, astfel încât proiecțiile norului **X** pe această axă să conste din clase compacte și bine separate. În Figura 3.1-3 axa 1 are o bună putere discriminantă în timp ce axa 2 (care este axa principală uzuală) nu permite o separare a proiecțiilor celor două grupe.

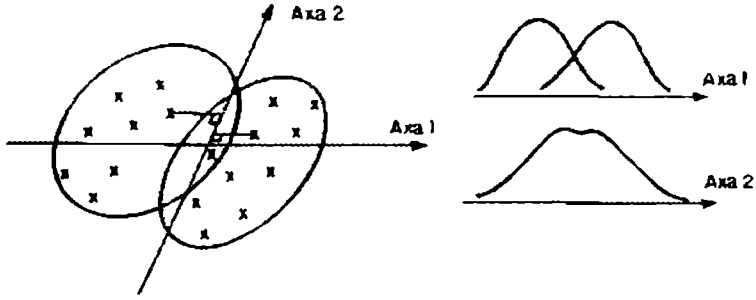


Figura 3.1-3 Axe cu proprietăți de discriminare diferite

Mărirea puterii discriminante a axelor poate fi așadar reclamată de datele problemei, pentru a putea „vedea" o anumită structură în date. Determinarea axelor discriminante poate servi și ca o tehnică de reducere a dimensiunii spațiului variabilelor. Prin această tehnică sunt selectate cele mai relevante caracteristici. Reducerea dimensiunii poate fi impusă și de necesitatea vizualizării claselor prin proiectarea datelor într-un spațiu cu una sau două dimensiuni, în acest caz cerința fundamentală este ca prin proiectarea datelor într-un spațiu de dimensiune redusă la clase compacte și bine separate din spațiul inițial să corespundă clase compacte și bine separate din noul spațiu.

Fie combinația liniară, pentru individul i, formată cu cele p variabile

$$a(i)=\sum_{j=1}^{p} a_j\left(x_{ij}-\bar{x}_j\right)^2,\ i=\overline{1,n}.$$

Atunci, dispersia empirică a lui $\mathbf{a}=\left(a(i)\right)_{i=1}^{n}$ este

$$D^2[\mathbf{a}]=\frac{1}{n}\sum_{i=1}^{n}a^2(i)=\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p}a_j\left(x_{ij}-\bar{x}_j\right)^2\right]$$

$$=\frac{1}{n}\sum_{i}\sum_{j}\sum_{j'=1}^{p}a_j a_{j'}\left(x_{ij}-\bar{x}_j\right)\left(x_{ij'}-\bar{x}_{j'}\right)$$

căci componentele sunt centrate. Inversând ordinea de sumare şi notând

$$t_{jj'}=\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)\left(x_{ij'}-\bar{x}_{j'}\right)=\operatorname{cov}\left(x_j,x_{j'}\right)$$

dispersia empirică a variabilei $\mathbf{a}$ se poate scrie

$$D^2(\mathbf{a})=\sum_{j=1}^{p}\sum_{j'=1}^{p}a_j a_{j'}\operatorname{cov}\left(x_j,x_{j'}\right)=\mathbf{a}'\mathbf{T}\mathbf{a}\,.$$

Ca şi în analiza dispersională (vezi, de exemplu, Văduva (1970)) să descompunem matricea de covarianţă într-o componentă intra-clase (în interiorul claselor) şi o componentă inter-clase (între clase) obţinând formula de descompunere a lui Huyghens (sau *ecuaţia analizei dispersionale*).

Să pornim de la identitatea

$$x_{ij}-\bar{x}_j=\left(x_{ij}-\bar{x}_{kj}\right)+\left(x_{kj}-\bar{x}_j\right)$$

şi să observăm că, din definiţia lui $\bar{x}_{kj}$,

$$\sum_{i\in I_k}\left(x_{ij}-\bar{x}_{kj}\right)\left(x_{kj'}-\bar{x}_{j'}\right)=\left(x_{kj'}-\bar{x}_{j'}\right)\sum_{i\in I_k}\left(x_{ij}-\bar{x}_{kj}\right)=0$$

şi în mod analog

$$\sum_{i\in I_k}\left(x_{kj}-\bar{x}_j\right)\left(x_{ij'}-\bar{x}_{kj'}\right)=0$$

Aşadar

$$t_{jj'}=\frac{1}{n}\sum_{k=1}^{q}\left[\sum_{i\in I_k}\left(x_{ij}-\bar{x}_j\right)\left(x_{ij'}-\bar{x}_{j'}\right)\right]$$

$$=\frac{1}{n}\sum_{k=1}^{q}\left[\sum_{i\in I_k}\left[\left(x_{ij}-\bar{x}_{kj}\right)+\left(\bar{x}_{kj}-\bar{x}_j\right)\right]\cdot\left[\left(x_{ij'}-\bar{x}_{kj'}\right)+\left(\bar{x}_{kj'}-\bar{x}_{j'}\right)\right]\right].$$

$$=\frac{1}{n}\sum_{k=1}^{q}\sum_{i\in I_k}\left(x_{ij}-\bar{x}_{kj}\right)\left(x_{ij'}-\bar{x}_{kj'}\right)+\frac{1}{n}\sum_{k=1}^{q}\sum_{i\in I_k}\left(\bar{x}_{kj}-\bar{x}_j\right)\left(\bar{x}_{kj'}-\bar{x}_{j'}\right)$$

Notând cu $d_{jj'}=\frac{1}{n}\sum_{k=1}^{q}\sum_{i\in I_k}\left(x_{ij}-\bar{x}_{kj}\right)\left(x_{ij'}-\bar{x}_{kj'}\right)$ şi cu $e_{jj'}=\sum_{k=1}^{q}\frac{n_k}{n}\left(\bar{x}_{kj}-\bar{x}_j\right)\left(\bar{x}_{kj'}-\bar{x}_{j'}\right)$

relaţia de mai sus se scrie matricial

$$\mathbf{T} = \mathbf{D} + \mathbf{E} \tag{1}$$

Astfel, dispersia unei combinații liniare de variabile **a** se descompune în

$$\mathbf{a}'\mathbf{T}\mathbf{a} = \mathbf{a}'\mathbf{D}\mathbf{a} + \mathbf{a}'\mathbf{E}\mathbf{a} \tag{2}$$

Reamintim că, dintre toate combinațiile liniare de variabile, sunt căutate cele care au o dispersie intra-clase minimă și o dispersie inter-clase maximă. În proiecție pe axa discriminantă **a** fiecare subnor trebuie să fie, în măsura posibilului, în același timp bine grupat și bine separat de ceilalți subnori.

Trebuie găsit **a** astfel încât

$\frac{\mathbf{a}'\mathbf{E}\mathbf{a}}{\mathbf{a}'\mathbf{D}\mathbf{a}}$ să fie maximă (sau echivalentul $\frac{\mathbf{a}'\mathbf{D}\mathbf{a}}{\mathbf{a}'\mathbf{E}\mathbf{a}}$ minimă)

sau, conform (2), să se maximizeze $f(\mathbf{a}) = \frac{\mathbf{a}'\mathbf{E}\mathbf{a}}{\mathbf{a}'\mathbf{T}\mathbf{a}}$ (raportul dintre dispersia inter-clase și dispersia totală).

Așadar, un punct staționar a lui $f(\mathbf{a})$ se află rezolvând ecuația

$$f'(\mathbf{a}) = 0 \Rightarrow \frac{(\mathbf{a}'\mathbf{T}\mathbf{a})(2\mathbf{E}\mathbf{a}) - (\mathbf{a}'\mathbf{E}\mathbf{a})(2\mathbf{T}\mathbf{a})}{(\mathbf{a}'\mathbf{T}\mathbf{a})^2} = 0$$

căci $\frac{d}{da}(\mathbf{a}'\mathbf{E}\mathbf{a}) = 2\mathbf{E}\mathbf{a}$ dacă **E** este simetrică (și este căci **E** și **T** sunt matrici de covarianță, în plus **T** este inversabilă). Rezultă

$$(\mathbf{a}'\mathbf{T}\mathbf{a})\mathbf{E}\mathbf{a} = (\mathbf{a}'\mathbf{E}\mathbf{a})\mathbf{T}\mathbf{a}$$

$$\mathbf{E}\mathbf{a} = \left(\frac{\mathbf{a}'\mathbf{E}\mathbf{a}}{\mathbf{a}'\mathbf{T}\mathbf{a}}\right)\mathbf{T}\mathbf{a} \quad \left| \times \mathbf{T}^{-1} \right. \tag{3}$$

$$\mathbf{T}^{-1}\mathbf{E}\mathbf{a} = \left(\frac{\mathbf{a}'\mathbf{E}\mathbf{a}}{\mathbf{a}'\mathbf{T}\mathbf{a}}\right)\mathbf{a}$$

Așadar $f(\mathbf{a})$ este maximă dacă este egală cu λ valoarea proprie maximă a matricii $\mathbf{T}^{-1}\mathbf{E}$ iar **a** este vector propriu corespunzător lui λ maxim.

Observații

1. $\mathbf{T}^{-1}\mathbf{E}$ este o matrice $p \times p$, în general nesimetrică. Din punct de vedere al calculului numeric, având în vedere că $q << p$, este interesant de aflat vectorii și valorile proprii ale unei matrici simetrice de dimensiune $q \times q$.
2. Să observăm că **E** este produsul unei matrci $\mathbf{C} \in \mathcal{M}_{p,q}$ (având coeficienții $c_{jk} = \sqrt{\frac{n_k}{n}}\left(\bar{x}_{kj} - \bar{x}_j\right)$) cu transpusa sa deci revenind în (3) $\mathbf{T}^{-1}\mathbf{C}'\mathbf{C}\mathbf{a} = \lambda\mathbf{a}$ sau $\mathbf{C}\mathbf{C}'\mathbf{a} = \lambda\mathbf{T}\mathbf{a}$ și punând $\mathbf{a} = \mathbf{T}^{-1}\mathbf{C}\mathbf{w}$ rezultă

$$\mathbf{C}\mathbf{C}'\mathbf{T}^{-1}\mathbf{C}\mathbf{w} = \lambda\mathbf{C}\mathbf{w} \tag{4}$$

Dacă **w** este vector propriu corespunzător lui λ al matricii $\mathbf{C'T^{-1}C}$ atunci el verifică relația și **a** și λ verifică relația (3). Cum $\mathbf{C'T^{-1}C} \in \mathcal{M}_{p\times q}(\mathbb{R})$ și este simetrică, în practică se diagonalizează această matrice iar apoi se află $\mathbf{a} = \mathbf{T^{-1}Cw}$.

3. λ_{max} se numește *putere discriminantă* și din (1) este mai mică sau egală cu unu.

Dacă:

- $\lambda_{max} = 1$ corespunde cazului A) din Figura 3.1-4. În proiecție pe axa **a** dispersiile intraclase sunt nule. Cei k nori sunt fiecare într-un hiperplan ortogonal pe **a**. Discriminarea pe această axă este perfectă dacă centrele de greutate se proiectează în puncte diferite.
- $\lambda_{max} = 0$ corespunde cazului în care cea mai bună axă discriminantă nu poate să separe centrele de greutate **g**, pentru că acestea sunt confundate. Norii sunt deci concentrici și neliniari separabili (cazul B) din Figura 3.1-4). Este posibil să existe posibilitatea unei suprafețe de decizie neliniară; în cazul de față este vorba de o funcție pătratică.

Valoarea proprie λ este o măsură pesimistă a puterii de discriminare a unei axe. Cazul C) din Figura 3.1-4 arată că cele două clase sunt liniar separabile pe axa considerată în pofida faptului că $\lambda < 1$.

Numărul dre valori proprii nenule, deci a axelor discriminante, este egal cu $q-1$ în cazul obișnuit, unde $n > p > q$ și variabilele nu sunt legate prin relații liniare.

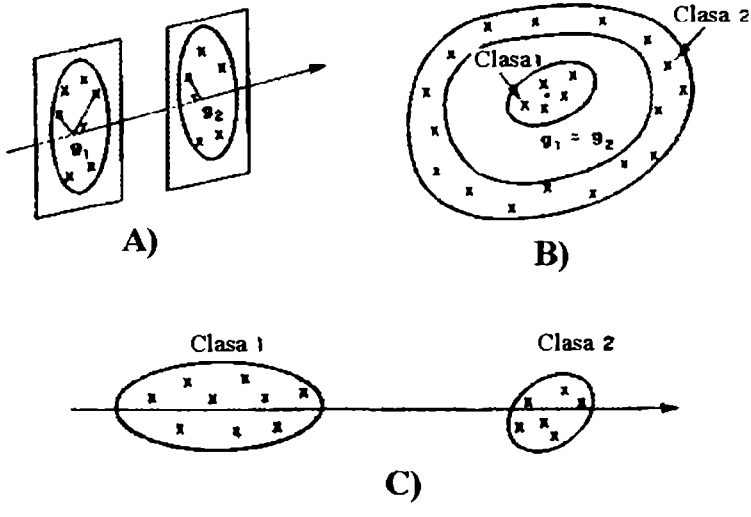


Figura 3.1-4 Exemplificarea diferitelor puteri de discriminare a unei axe

Odată găsite axele cu puterea de discriminare cea mai bună pasul următor constă în găsirea suprafețelor de dicizie.

3.1.3 METODE GEOMETRICE

Metodele geometrice de analiză discriminantă, esențialmente descriptive, se bazează pe noțiunea de distanță și nu utilizează nici o noțiune probabilistă. Pentru detalii privind această secțiune pot fi consultate monografiile Anderberg (1973), Duda&Hart (1973)

3.1.3.1 Suprafețe de decizie

În context gometric, discriminarea poate fi interpretată ca o împărțire a spațiului variabilelor în regiuni, numite *regiuni de decizie*, fiecare regiune fiind asociată cu o clasă de obiecte. Regiunile de decizie și implicit clasele corespunzătoare, se zic *separabile* dacă ele pot fi separate prin suprafețe din spațiul variabilelor.

Suprafețele de separare ale regiunilor de decizie se numesc și *suprafețe de decizie*. Dacă suprafețele de decizie sunt hiperplane, clasele de zic *liniar separabile*.

Sprafețele de decizie pot fi descrise cu ajutorul unei mulțimi de *funcții de discriminare* sau *funcții de decizie*.

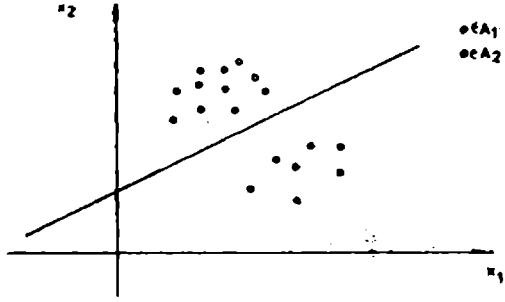


Figura 3.1-5 Două clase liniar separabile din $\mathbb{R}^2$, notate A_1 și A_2

Clasele ce apar în multe probleme concrete nu pot fi, în general, precis definite, deoarece apartenența unor elemente la una sau alta din clase poate fi incertă. Aceste clase fără margini precise, în care tranziția de la apartenență la neapartenență este mai degrabă graduală, pot fi descrise prin *mulțimi nuanțate* (*fuzzy* sau cu *apartenență divizată*. Vezi, de exemplu, Dumitrescu (1999)).

Vom considera cazul claselor separabile. Funcția de discriminare atașează fiecare obiect unei regiuni R din spațiul variabilelor, regiune delimitată prin intermediul unei mulțimi de suprafețe de decizie. O *funcție de discriminare instruibilă* (cu *învățare*) tinde să reducă numărul caracteristicilor incorecte, făcând acest număr cât mai mic posibil, eventual nul. Acest lucru se realizează prin ajustarea mulțimii $\mathcal{R}$ a regiunilor de decizie

ca răspuns la observațiile făcute asupra unei mulțimi de obiecte de instruire. Mulțimea obiectelor de instruire se numeşte *mulțime de instruire*. Ajustarea regiunilor de decizie ca rezultat al observațiilor asupra mulțimii de instruire reprezintă *faza de învățare* sau instruire a funcției de discriminare.

Dacă se cunoaşte dinainte numărul claselor şi pentru fiecare obiect din mulțimea de instruire ştim clasa căruia acesta aparține, învățarea se zice *supervizată* sau *cu profesor*. Dacă structura de instruire nu este cunoscută, adică pentru nici un obiect din această mulțime nu cunoaştem dinainte clasa, instruirea se zice *nesupervizată* sau *fără profesor*.

Procedura prin care regiunile de decizie sunt ajustate ca răspuns la observațiile privind clasarea vectorilor din mulțimea de instruire, constituie procedura de instruire. După ce clasele şi suprafețele de decizie sunt stabilite prin faza de instruire (funcția de discriminare este instruită), funcției de discriminare i se prezintă date ale căror clase nu se cunosc. Această fază în care obiecte noi sunt asociate uneia sau alteia din clasele stabilite, se numeşte *fază de lucru*, sau *decizională* sau încă *de afectare*. Uneori faza de instruire şi cea de lucru pot să coincidă sau să se suprapună parțial. Este ceea ce se întâmplă în cazul clasificării nesupervizate.

Să considerăm că în mulțimea datelor sunt prezente q clase, notate $A_1,\ldots,A_q$. Distingem următoarele trei cazuri de separabilitate:

Cazul 1. Fiecare clasă A_i este separabilă de toate celelalte printr-o singură suprafață de decizie. Există q funcții de decizie. Notăm cu $g_i:\mathbb{R}^p\to\mathbb{R}$ funcția de decizie corespunzătoare clasei A_i. Ecuația suprafeței de decizie ce separă clasa A_i de toate celelalte clase este $g_i(\mathbf{x})=0$.

Pentru fiecare clasă regula de afectare este

dacă $\mathbf{x}\in A_i$, *atunci* $g_i(\mathbf{x})>0$.

Dacă pentru un punct $\mathbf{x}$ nou considerat avem

$$g_i(\mathbf{x})>0 \text{ şi } g_j(\mathbf{x})<0,\quad j=1,\ldots,q,\ j\neq i.$$

atunci $\mathbf{x}$ este ataşat clasei A_i.

Regiunea de decizie R corespunzătoare clasei A_i va fi aşadar

$$R_i=\left\{\mathbf{x}\in\mathbb{R}^p \mid g_i(\mathbf{x})>0 \text{ si } g_j(\mathbf{x})<0,\quad j=1,\ldots,q,\ j\neq i\right\}.$$

Punctele ce nu aparțin nici unei regiuni de decizie formează o regiune de nedeterminare (RN). Suprafețele de decizie aparțin regiunii de nedeterminare. Este posibil ca regiunea de nedeterminare RN să conțină şi alte puncte decât cele aparținând suprafețelor de decizie.

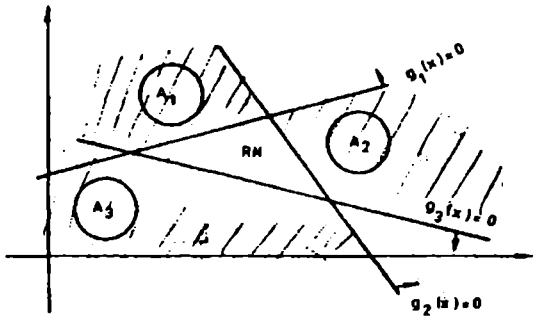


Figura 3.1-6 Cazul 1 de separabilitate

Cazul 2. Fiecare clasă este separată de oricare alta printr-o suprafață de decizie distinctă. Clasele sunt aşadar două câte două separabile. Există $q(q-1)/2$ suprafețe de decizie generate de funcțiile $g_{ij}:\mathbf{R}^p \to \mathbf{R}$. Suprafața de decizie corespunzătoare claselor A_i şi A_j are ecuația $g_{ij}(\mathbf{x})=0$. Funcțiile de decizie satisfac condiția $g_{ij}(\mathbf{x})=-g_{ij}(\mathbf{x}), \forall \mathbf{x}$.

Punctele clasei A_i se află de partea pozitivă a suprafeței $g_{ij}(\mathbf{x})=0$. Regula de decizie este:

$$\mathbf{x} \in A_i \Leftrightarrow g_{ij}(\mathbf{x}) > 0, \ \forall\, j \neq i.$$

Regiunea de decizie R corespunzătoare clasei A_i este

$$R_i = \left\{ \mathbf{x} \in \mathbf{R}^p \mid g_{ij}(\mathbf{x}) > 0 \ \ j \neq i \right\}.$$

La fel ca şi în condițiile cazului 1 de separabilitate, este posibil să existe o regiune de nedeterminare, neaparținând nici unei regiuni de decizie.

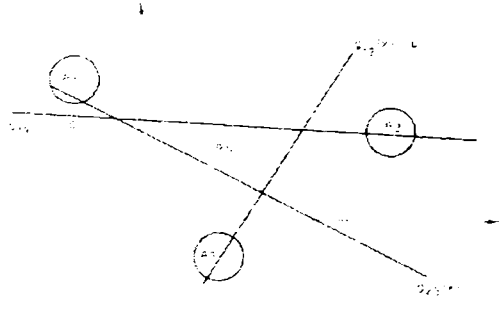


Figura 3.1-7 Cazul 2 de separabilitate

Cazul 3. Există k funcții de. Regula de decizie se formulează astfel:

$\mathbf{x} \in A_i$ *dacă şi numai dacă* $g_i(\mathbf{x}) > g_j(\mathbf{x}), \quad \forall j \neq i$.

Regiunea de decizie R corespunzătoare clasei A_i va fi aşadar

$$R_i = \left\{ \mathbf{x} \in \mathbb{R}^p \mid g_i(\mathbf{x}) > g_j(\mathbf{x}) \ \ \forall j \neq i \right\}.$$

Suprafaţa de decizie dintre clasele A_i şi A_j are ecuaţia

$$g_i(\mathbf{x}) = g_j(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^p, j \neq i.$$

Obiectele clasei A_i se află de partea pozitivă a suprafeţei de separare.

Observaţie. Separabilitatea de tip 3 implică separabilitatea de tip 2. Într-adevăr, să punem

$$g_{ij}(\mathbf{x}) = g_i(\mathbf{x}) - g_j(\mathbf{x})$$

şi să admitem separabilitatea, în condiţiile cazului 3, a claselor $A_1, \ldots, A_q$. Dacă $\mathbf{x}$ aparţine regiunii clasei A_i, atunci $g_i(\mathbf{x}) > g_j(\mathbf{x}), \quad \forall j \neq i$. Avem că $g_{ij}(\mathbf{x}) > 0, \quad \forall j \neq i$. Rezultă aşadar că dacă clasele sunt separabile faţă de condiţiile cazului 3, ele sunt separabile şi faţă de cazul 2. Reciproca nu este în general valabilă.

In condiţiile cazului 3 de separabilitate nu există alte regiuni de nedeterminare decât suprafeţele de separare (vezi Figura 3.1-8).

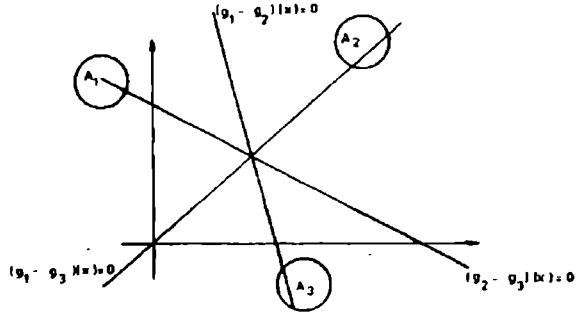


Figura 3.1-8 Cazul 3 de separabilitate

În cele ce urmează, prin separabilitatea a două clase vom înţelege, în absenţa altei precizări, separabilitatea sub condiţiile cazului 3.

3.1.3.2 Funcţii de decizie afine şi liniare

De o mare importanţă practică este cazul când clasele sunt liniar separabile. În această situaţie funcţiile de decizie sunt funcţii afine.

O *funcţie afină* de decizie g este o *aplicaţie afină* $g : \mathbb{R}^p \to \mathbb{R}$, adică g se poate scrie sub forma

$$g(\mathbf{x}) = \mathbf{w}'\mathbf{x} + w_{p+1}, \quad \mathbf{x} \in \mathbb{R}^p.$$

unde vectorii caracteristică se consideră ca fiind vectori coloană şi $\mathbf{w} = (w_1, \ldots, w_p)'$. $w_j \in \mathbb{R}$ se numeşte *vector pondere* sau *vector parametru*.

O convenţie uzuală este să se adauge w_{p+1} ca ultimă componentă a vectorului $\mathbf{w}$. Se defineşte astfel vectorul pondere extins $\mathbf{v} = (w_1, \ldots, w_p, w_p)'$ şi, respectiv, vectorul caracteristică caracteristică extins $\mathbf{y} = (x_1, \ldots, x_p, 1)'$. Vectorii $\mathbf{y}$ vor fi elemente ale spaţiului extins al caracteristicilor, spaţiu notat cu $\mathcal{Y}$. Prin această mărire a dimensiunii spaţiului caracteristicilor, proprietăţile geometrice ale claselor nu sunt alterate. Cu noile notaţii introduse funcţia afină de decizie se transformă într-o funcţie liniară de decizie $g: \mathcal{Y} \to \mathbb{R}$, $\mathcal{Y} \subset \mathbb{R}^{p+1}$ dată de expresia

$$g(\mathbf{y}) = \mathbf{v}'\mathbf{y}, \quad \mathbf{y} \in \mathcal{Y}.$$

Dacă g. este funcţia de decizie liniară corespunzând clasei A_i atunci, în conformitate cu cazul 3 de separabilitate, un obiect $\mathbf{y}$ este ataşat clasei A_i dacă

$$g_i(\mathbf{y}) = g_j(\mathbf{y}), \quad j \neq i.$$

Considerăm o funcţie $r: \mathcal{Y} \to \{1, 2, \ldots, q\}$. Funcţia r a fiecărui vector caracteristică $\mathbf{y}$ face să-i corespundă indicele unei clase. Regula de decizie se reformulează astfel:

$r(\mathbf{y}) = i$ *dacă şi numai dacă* $g_i(\mathbf{x}) > g_j(\mathbf{x}), \quad \forall j \neq i$.

În cazul când există doar două clase, putem considera o singură funcţie de decizie $g: \mathcal{Y} \to \mathbb{R}$ dată de relaţia

$$g(\mathbf{y}) = g_1(\mathbf{y}) - g_2(\mathbf{y}).$$

Dacă $g(\mathbf{y}) > 0$, atunci $\mathbf{y}$ aparţine clasei A_1, iar dacă $g(\mathbf{y}) < 0$, atunci $\mathbf{y}$ aparţine clasei A_2.

3.1.3.3 Ecuaţia unui hiperplan

Reamintim că ecuaţia unui hiperplan $\mathcal{H}$ ce trece printr-un punct $\mathbf{x}_0$ şi este normal pe un vector unitar $\mathbf{u}$ se poate scrie sub forma

$$(\mathbf{u}, \mathbf{x} - \mathbf{x}_0) = \mathbf{u}'(\mathbf{x} - \mathbf{x}_0) = 0$$

cu produsul scalar uzual.

Ecuaţia dreptei Δ ce trece printr-un punct $\mathbf{z}_0$ şi este ortogonală pe hiperplanul $\mathcal{H}$ de ecuaţie se scrie

$$\mathbf{x} - \mathbf{z}_0 = t\mathbf{u}, \quad t \in \mathbb{R}$$

adică

$$\mathbf{x} = \mathbf{z}_0 + t\mathbf{u}, \quad t \in \mathbb{R}$$

Pentru a găsi intersecţia lui $\mathcal{H}$ cu Δ înlocuim ecuaţia dreptei în ecuaţia hiperplanului. Obţinem

$$\mathbf{u}^T(\mathbf{z}_0 + t\mathbf{u} - \mathbf{x}_0) = 0,$$

şi deci

$$t\mathbf{u}^T\mathbf{u} = \mathbf{u}^T(\mathbf{x}_0 - \mathbf{z}_0),$$

de unde, ţinând cont că $\|\mathbf{u}\| = 1$, găsim

$$t = \frac{\mathbf{u}^T(\mathbf{x}_0 - \mathbf{z}_0)}{\|\mathbf{u}\|^2} = \mathbf{u}^T(\mathbf{x}_0 - \mathbf{z}_0).$$

Punctul de intersecţie al dreptei cu hiperplanul $\mathcal{H}$ va fi aşadar

$$\mathbf{x}' = \mathbf{z}_0 + \mathbf{u}^T(\mathbf{x}_0 - \mathbf{z}_0)\mathbf{u}.$$

Distanţa de la punctul $\mathbf{z}_0$ la hiperplan este deci

$$\begin{aligned} d(\mathcal{H}, \mathbf{z}_0) &= \|\mathbf{x}' - \mathbf{z}_0\| \\ &= |\mathbf{u}'(\mathbf{x}_0 - \mathbf{z}_0)| \cdot \|\mathbf{u}\|. \\ &= |\mathbf{u}'(\mathbf{x}_0 - \mathbf{z}_0)| \end{aligned}$$

Distanţa de la originea spaţiului la hiperplan se obţine punând în relaţia de mai sus $\mathbf{z}_0 = 0$ şi deci

$$D = d(\mathcal{H}, 0) = |\mathbf{u}'\mathbf{x}_0|.$$

3.1.3.4 Hiperplane de separare

Într-un clasificator liniar regiunile de decizie sunt mărginite de hiperplane sau de porţiuni de hiperplane. Dacă regiunile R_i, şi R_j au o frontieră comună, suprafaţa de decizie ce le separă este hiperplanul de ecuaţie

$$g_i(\mathbf{y}) - g_j(\mathbf{y}) = (\mathbf{v}_i^T - \mathbf{v}_j^T)\mathbf{y} = 0.$$

Observăm că în spaţiul extins al caracteristicilor toate hiperplanele de separare trec prin originea spaţiului.

În spaţiul caracteristicilor ecuaţia suprafeţei de decizie este

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

deci avem

$$\mathbf{w}_i'\mathbf{x} + \mathbf{w}_{i,p+1} = \mathbf{w}_j'\mathbf{x} + \mathbf{w}_{j,p+1} \text{ sau } \mathbf{w}'\mathbf{x} - \mathbf{w}_{p+1} = 0$$

unde am notat

$$\mathbf{w} = \mathbf{w}_i - \mathbf{w}_j$$

$$\mathbf{w}_{p+1} = \mathbf{w}_{i,p+1} - \mathbf{w}_{j,p+1}$$

Din relaţia de mai sus rezultă că ecuaţia hiperplanului de separare în spaţiul caracteristicilor se mai poate scrie sub forma

$$\frac{\mathbf{w}'}{\|\mathbf{w}\|}\mathbf{x} + \frac{\mathbf{w}_{p+1}}{\|\mathbf{w}\|} = 0$$

Comparând această ecuaţie cu ecuaţia generală

$$\mathbf{u}^T\mathbf{x} - \mathbf{u}^T\mathbf{x}_0 = 0$$

a hiperplanului ce trece prin punctul $\mathbf{x}_0$, obținem că vectorul unitar normal pe hiperplan este

$$\mathbf{u} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

şi

$$\mathbf{u}'\mathbf{x}_0 = \frac{\mathbf{w}_{p+1}}{\|\mathbf{w}\|}.$$

Rezultă că distanța de la origine la hiperplanul de separare se poate scrie

$$D = |\mathbf{u}'\mathbf{x}_0| = \frac{|\mathbf{w}_{p+1}|}{\|\mathbf{w}\|}.$$

Distanța de la punctul $\mathbf{z}_0$ la hiperplan va fi

$$\begin{aligned} d(\mathcal{H}, \mathbf{z}_0) &= |\mathbf{u}'(\mathbf{x}_0 - \mathbf{z}_0)| \\ &= \left| -\frac{\mathbf{w}_{p+1}}{\|\mathbf{w}\|} - \frac{\mathbf{w}'}{\|\mathbf{w}\|}\mathbf{z}_0 \right| \\ &= \frac{1}{\|\mathbf{w}\|}|\mathbf{w}'\mathbf{z}_0 + \mathbf{w}_{p+1}|. \end{aligned}$$

Formulele stabilite ne vor fi utile în studiul geometriei funcțiilor discriminante liniare.

3.1.4 FUNCȚII DISCRIMINANTE CU DISTANȚĂ MINIMĂ

În această secțiune ne propunem să arătăm cum clasarea prin minimizarea unei funcții criteriu ne conduce la o clasă de funcții discriminante liniare. Funcția criteriu considerată aici este distanța de la vectorii caracteristică la prototipurile claselor.

Pătratul distanței euclidiene de la un vector $\mathbf{x}$ din $\mathcal{X}$ la prototipul $\mathbf{L}_i$ al clasei A_i, se scrie

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{L}_i) &= \|\mathbf{x} - \mathbf{L}_i\|^2 = (\mathbf{x} - \mathbf{L}_i)'(\mathbf{x} - \mathbf{L}_i) \\ &= \mathbf{x}'\mathbf{x} - 2\mathbf{x}'\mathbf{L}_i + \mathbf{L}_i'\mathbf{L}_i \end{aligned}$$

Un vector $\mathbf{x}$ este ataşat acelei clase A_i de al cărei prototip $\mathbf{x}$ este mai aproape, adică $\mathbf{x} \in A_i$ dacă $d(\mathbf{x}, \mathbf{L}_i) = \min_j d(\mathbf{x}, \mathbf{L}_j)$.

Distanțele fiind întotdeauna pozitive, a minimiza d este echivalent cu a minimiza d^2. Deoarece $\mathbf{x}'\mathbf{x}$ nu depinde de clasa i distanța lui $\mathbf{x}$ la prototipul $\mathbf{L}_i$ se mai scrie

$$d^2(\mathbf{x}, \mathbf{L}_i) = \mathbf{x}'\mathbf{x} - 2\left(\mathbf{x}'\mathbf{L}_i - \tfrac{1}{2}\mathbf{L}_i'\mathbf{L}_i\right).$$

O clasificare echivalentă cu regula de asignare de mai sus se obține considerând funcția $g_i : \mathbb{R}^p \to \mathbb{R}$ dată de

$$g_i(\mathbf{x}) = \mathbf{x}'\mathbf{L}_i - \frac{1}{2}\mathbf{L}_i'\mathbf{L}_i .$$

Regula de decizie devine

$\mathbf{x} \in A_i$ dacă $g_i(\mathbf{x}) = \max_j g_j(\mathbf{x})$.

Am obținut că g_i este o funcție afină de decizie. Notând

$\mathbf{c}_i = \mathbf{L}_i$ și $c_{i,p+1} = \frac{1}{2}\mathbf{L}_i'\mathbf{L}_i$.

putem scrie g_i sub forma standard

$$g_i(\mathbf{x}) = \mathbf{c}_i'\mathbf{x} + c_{i,p+1} .$$

Suprafața de decizie ce separă clasele A_i și A_j are ecuația

$g_i(\mathbf{x}) = g_j(\mathbf{x})$,

adică, ținând cont de forma lui g_i avem că

$$(\mathbf{L}_i - \mathbf{L}_j)^T\mathbf{x} + \frac{1}{2}(\mathbf{L}_j^T\mathbf{L}_j - \mathbf{L}_i^T\mathbf{L}_i) = 0,$$

ceea ce se mai poate scrie sub forma

$$(\mathbf{L}_i - \mathbf{L}_j)^T(\mathbf{x} - \frac{1}{2}(\mathbf{L}_i + \mathbf{L}_j)) = 0.$$

Notând

$\mathbf{c} = \mathbf{L}_i - \mathbf{L}_j$ și $\mathbf{x}_0 = \frac{1}{2}(\mathbf{L}_i + \mathbf{L}_j)$

ecuația suprafeței de decizie devine:

$\mathbf{c}^T(\mathbf{x} - \mathbf{x}_0) = 0$

Suprafața de separare este deci un hiperplan ce trece prin punctul $\mathbf{x}_0$ și este ortogonal pe vectorul $\mathbf{c}$. Cu alte cuvinte, hiperplanul de separare este ortogonal pe dreapta ce unește prototipurile claselor, pe care o intersectează într-un punct $\mathbf{x}_0$ situat la jumătatea distanței dintre prototipuri.

Funcția discriminantă cu distanță minimă este adecvat pentru cazurile când punctele unei clase tind să se aglomereze în vecinătatea unui punct prototip, formând un nor (*cluster*) de puncte.

3.2 METODE PROBABILISTE DE DISCRIMINARE

Această secțiune este dedicată aspectului inferențial al analizei discriminante prin abordarea probabilistă a metodelor de discriminare. Principalul instrument folosit este teoria bayesiană a deciziilor. Se vor considera diferite metode de estimare a parametrilor necunoscuți din densitatea de probabilitate atașată mulțimii datelor.

3.2.1 PRELIMINARII

Definiția 3.2-1 Fie $(\Omega, \mathcal{K}, P)$ un câmp de probabilitate și $A, B \in \mathcal{K}$ cu $P(B) > 0$

$$P_B : \mathcal{K} \to \mathbb{R} \text{ cu } P_B(A) \equiv P(A|B) = \frac{P(A \cap B)}{P(B)}$$

se numeşte *probabilitatea condiţionată* a evenimentului A relativ la evenimentul B.

Lema 3.2-1 *Fie* $(\Omega, \mathcal{K}, P)$ *un câmp de probabilitate şi* $\{A_i\}_{i \in I}$ *un sistem complet de evenimente. Are loc următoarea egalitate (formula lui Bayes a probabilităţii cauzelor)*

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)\dfrac{P(B \cap A_i)}{P(A_i)}}{P(B)}$$

$$= \frac{P(A_i)P(B|A_i)}{\sum_i P(A_i)P(B|A_i)}$$

cu $\{P(A_i)\}$ *-probabilităţi à priorice şi* $\{P(B|A_i)\}$ *-probabilităţi à posteriori.*

Definiţia 3.2-2 Fie $(\Omega, \mathcal{K}, P)$ un câmp de probabilitate, X variabilă aleatoare şi $A \in \mathcal{K}$ cu $P(A) > 0$, $F_A : \mathbb{R} \to [0,1]$

$$F_A(x) \equiv F(x|A) = P(X < x|A) \ (\forall)\ x \in \mathbb{R}$$

se numeşte *funcţie de repartiţie a variabilei aleatoare X condiţionată de evenimentul A.*

Analog $f(\cdot|A) : \mathbb{R} \to \mathbb{R}$ se numeşte *densitate de repartiţie condiţionată,* unde $F(x|A) = \int_{-\infty}^{x} f(t|A)dt$.

Observaţie. $f(x|A) = F'(x|A)$ aproape peste tot.

Lema 3.2-2 $P(A|X = x) = \dfrac{P(A)f(x|A)}{f(x)}$.

Fie (X,Y) variabilă aleatoare bidimensională, cu densitatea de probabilitate h şi funcţia de repartiţie F, adică

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} h(t,s)dt\,ds$$

Funcţia de repartiţie a lui X este

$$F_X(x) = P(X < x) = P(X < x, Y < \infty) = F(x,\infty) = \int_{-\infty}^{x}\int_{\mathbb{R}} h(t,s)dt\,ds$$

şi densitatea de probabilitate a lui X este $f(x) = F_X'(x) = \int_{\mathbb{R}} h(x,s)ds$.

Analog, densitatea de probabilitate a lui Y este $g(y) = F_Y'(y) = \int_{\mathbb{R}} h(t,y)dt$.

Lema 3.2-3 *Dacă h este densitatea de probabilitate a variabilei aleatoare* (X,Y) *f este densitatea de probabilitate a variabilei aleatoare X şi g este densitatea de probabilitate a variabilei aleatoare Y, atunci*

1) $f(x)=\int_{\mathbb{R}} h(x,y)dy$;

2) $g(y)=\int_{\mathbb{R}} h(x,y)dx$;

3) $f(x|y)=\frac{h(x,y)}{g(y)}$ *dacă* $g(y)>0$ *altfel arbitrar;*

4) $g(y|x)=\frac{h(x,y)}{f(x)}$ *dacă* $f(x)>0$ *altfel arbitrar;*

5) $f(x)=\int_{\mathbb{R}} f(x|y)g(y)dy$;

6) $g(y)=\int_{\mathbb{R}} g(y|x)f(x)dx$;

7) $g(y|x)=\frac{f(x|y)g(y)}{f(x)}=\frac{f(x|y)g(y)}{\int_{\mathbb{R}} f(x|t)g(t)dt}$ *(formula lui Bayes pentru densităţi de probabilitate).*

3.2.2 FORMULAREA BAYESIANĂ A PROBLEMEI DE DISCRIMINARE

Problema de discriminare (sau clasare. *atenţie! nu de clasificare)* formulată în termenii teoriei statistice a deciziei este următoarea:

Dându-se:

- *K grupe (populaţii)* $\Pi_1,\Pi_2,.....\Pi_K$ *specificate prin distribuţiile lor de probabilitate* $P_i(\mathbf{x})=P(X=\mathbf{x}\mid \mathbf{x}\in\Pi_i)$ *cu* $i=\overline{1,K}$;
- q_i, $i=\overline{1,K}$, *probabilităţi a priori, ca un individ (observaţie) să provină din populaţiile* Π_i, $i=\overline{1,K}$ $(\{q_i\}_{i=1}^{K}$ *formează un sistem complet de probabilităţi, adică* $\sum_i q_i=1$*)*:
- $\mathcal{X}$ *spaţiul observaţiilor asupra a p variabile aleatoare* $\xi_1,.....\xi_p$ *(predictori);*
- $\{C(j|i)\}_{i,j=1}^{K}$ *costurile erorii de clasare (costul clasării unei observaţii provenind din populaţia* Π_i *în populaţia* Π_j, $i\neq j$*);*
- *să se găsească o partiţie* $\mathcal{R}=\{R_i\}_{i=1}^{K}$ *a spaţiului* $\mathcal{X}$ *(adică*

$\mathcal{X}=\bigcup_{1}^{K}R_i,\quad R_i\cap R_j=\varnothing,\quad i\neq j,\quad i,j=\overline{1,K}$ *astfel încât*

$$\sum_{i=1}^{K}q_i\left\{\sum_{\substack{j=1\\ j\neq i}}^{K}\mathsf{C}(j|i)P(j|i,\mathcal{R})\right\}$$

să fie minimă.

În cele de mai sus am notat cu: $P(j|i,\mathcal{R})=\int_{R_j}P_i(\mathbf{x})d\mathbf{x}\qquad i\neq j,\qquad i,j=\overline{1,K}$

probabilitățile de eroare pentru o partiție $\mathcal{R}$ dată.

3.2.2.1 Regula Bayes pentru distribuții cunoscute

În această secțiune presupunem $\{q_i\}_i$ şi $\{P_i\}_i$ cunoscute. Aceasta va permite să se construiască procedura de clasare cu proprietăți de optimalitate, dar cu aplicabilitate practică directă redusă, deoarece în realitate, cel puțin distribuțiile $\{P_i\}_i$ sunt necunoscute.

Fie $Y=\{1,\ldots,K\}$ spațiul etichetelor claselor şi fie distribuția de probabilitate pe Y $P_Y(x)=\sum_{i=1}^{K}q_i\delta_i(x)$, unde s-a notat cu $\delta_i(x)$ funcția Dirac (adică $\delta_i(x)=1$ dacă $x=i$ şi 0 în rest).

Definiția 3.2-3 O funcție $\mathsf{c}:\mathcal{X}\to Y$ ce estimează clasa $\mathsf{c}(\mathbf{x})=y\in Y$ a lui $\mathbf{x}$, după ce $\mathbf{x}\in\mathcal{X}$ a fost observat se numeşte *plasator*.

Pentru a aprecia calitatea plasatorului este natural să se studieze probabilitatea de misclasare pentru clasa k :

$$pmc(k)=P\left[\left\{\mathsf{c}(\mathbf{x})\neq k\mid\{\mathbf{x}\in\Pi_k\}\right\}\right].$$

Se consideră funcția de pierdere discretă $\ell(\mathsf{c}(\mathbf{x}),j)$ pentru plasatorul c față de clasa j şi riscul funcțional al plasatorului c

$$\mathsf{R}(\mathsf{c})=\mathsf{M}_{\mu}\left[\ell(\mathsf{c}(\mathbf{x}),j)\right]=\sum_{i=1}^{K}q_i\,pmc(i)=\sum_{i=1}^{K}q_i\sum_{\substack{j=1\\ j\neq i}}^{K}\int_{R_j}P_i(\mathbf{x})d\mathbf{x}$$

căci în acest caz particular, distribuția de probabilitate pe $\mathcal{X}\times Y$ este din construcție, $\mu(\mathbf{x},i)=q_iP_{\mathbf{e}(\mathbf{x})}(\mathbf{x})$, cu $\mathbf{e}(\mathbf{x})\in Y$ notație pentru clasa lui $\mathbf{x}$.

Dacă se consideră costurile misclasării $\{\mathsf{C}(j|i)\}_{i,j=1}^{K}$ egale cu unitatea (ipoteză naturală în multe situații practice, excepție făcând situațiile din medicină (când costul erorii de a considera un bolnav sănătos, poate fi dramatic, față de costul erorii considerării unui om sănătos ca bolnav) atunci, un plasator va fi optim dacă

minimizează riscul funcțional $R(c)$ (adică, exact funcționala din enunțul problemei de clasare).

Să mai notăm că *probabilitatea a posteriori* a unei clase i, dându-se $X = \mathbf{x}$ este

$$P(i|\mathbf{x}) = \frac{q_i P_i(\mathbf{x})}{\sum_{j=1}^{K} q_j P_j(\mathbf{x})}.$$

Cu acestea se pot enunța următoarele rezultate:

Teorema 3.2-1 (a „regiunilor") (Anderson, 1958) *Partiția $\mathcal{R}$ a lui X care minimizează riscul funcțional este $\mathcal{R} = \{R_i\}_{i=1}^{K}$ cu*

$$R_i = \left\{ \mathbf{x} \in X \mid \sum_{\substack{j=1 \\ j \neq i}}^{K} q_j P_j(\mathbf{x}) \leq \sum_{\substack{j=1 \\ j \neq k}}^{K} q_j P_j(\mathbf{x}), \quad k \neq i,\ k = \overline{1,K} \right\}, \quad i = \overline{1,K}.$$

Demonstație: Pentru simplificarea demonstrației să presupunem $k = 2$ (doar două populații) și $C(1|2) = C(2|1) = 1$. Atunci media costului misclasificării este

$$q_1 \int_{R_1} P_1(\mathbf{x}) d\mathbf{x} + q_2 \int_{R_2} P_2(\mathbf{x}) d\mathbf{x} \tag{1}$$

Pentru a minimiza pe (1), un $\mathbf{x}$ dat va fi asignat populației cu probabilitate à posteriori maximă. Astfel, dacă

$$\frac{q_1 P_1(\mathbf{x})}{q_1 P_1(\mathbf{x}) + q_2 P_2(\mathbf{x})} \geq \frac{q_2 P_2(\mathbf{x})}{q_1 P_1(\mathbf{x}) + q_2 P_2(\mathbf{x})} \tag{2}$$

atunci $\mathbf{x}$ va fi asigurat lui Π_1, altfel lui Π_2.

Cum minimizăm probabilitatea de misclasare în fiecare punct, minimizăm costul misclasării pe tot spațiul.

Așadar regiunile de decizie sunt:

$$\begin{aligned} R_1 &: \mathbf{x} \in X \; q_1 P_1(\mathbf{x}) \geq q_2 P_2(\mathbf{x}) \\ R_2 &: \mathbf{x} \in X \; q_1 P_1(\mathbf{x}) < q_2 P_2(\mathbf{x}) \end{aligned} \tag{3}$$

Dacă $q_1 P(\mathbf{x}) = q_2 P_2(\mathbf{x})$ punctul poate fi clasificat fie în Π_1 fie în Π_2 (arbitrar în (3) a fost asignat lui Π_1).

Dacă $q_1 P(\mathbf{x}) + q_2 P_2(\mathbf{x}) = 0$ la fel punctul poate fi asignat oricărei regiuni.

Să arătăm acum că (3) este cea mai bună procedură. Pentru orice partiție $\mathfrak{R}^* = (R_1^*, R_2^*)$ a lui X probabilitatea de misclasare este

$$q_1 \int_{R_2^*} P_1(\mathbf{x}) d\mathbf{x} + q_2 \int_{R_1^*} P_2(\mathbf{x}) d\mathbf{x} = \int_{R_2^*} \left(q_1 P_1(\mathbf{x}) - q_2 P_2(\mathbf{x}) \right) d\mathbf{x} + q_2 \int_{R_1^*} P_2(\mathbf{x}) d\mathbf{x} \tag{4}$$

Dar $q_2\int_{R_2^*} P_2(\mathbf{x})d\mathbf{x} + q_2\int_{R_1^*} P_2(\mathbf{x})d\mathbf{x} = q_2\int_{\mathcal{X}} P_2(\mathbf{x})d\mathbf{x}$ $(= q_2$ dacă $\operatorname{supp} P_2 \subseteq \mathcal{X}$ sau constantă în caz contrar).

Relația (4) va fi minimă dacă R_2^* va include punctele $\mathbf{x}$ pentru care $q_1P_1(\mathbf{x}) - q_2P_2(\mathbf{x}) < 0$ şi va exclude punctele pentru care $q_1P_1(\mathbf{x}) - q_2P_2(\mathbf{x}) > 0$; adică $R_2^* = R_2 \Rightarrow R_1^* = R_1$ (ca partiții ale aceluiaşi spațiu).

Dacă, în plus $P\left(\frac{P_1(\mathbf{x})}{P_2(\mathbf{x})} = \frac{q_2}{q_1} \mid \Pi_i\right) = 0 \quad i = 1,2$ atunci procedura Bayes este unică excepție o mulțime de probabilitate zero.

Dacă $C(1|2) \neq C(2|1) \neq 1$ atunci regiunile de decizie se scriu

$$\begin{aligned} R_1 : \mathbf{x} \in \mathcal{X} \quad & \frac{P_1(\mathbf{x})}{P_2(\mathbf{x})} \geq \frac{C(1|2)q_2}{C(2|1)q_1} \\ R_2 : \mathbf{x} \in \mathcal{X} \quad & \frac{P_1(\mathbf{x})}{P_2(\mathbf{x})} < \frac{C(1|2)q_2}{C(2|1)q_1} \end{aligned} \tag{5}$$

□

Observație. Regiunile de decizie Bayes se înscriu în cazul 3 de separabilitate.

Corolar 3.2-1 (Ripley, 1996) *Plasatorul care minimizează riscul funcțional, este* $c_B(\mathbf{x}) = j$. *dacă* $P(j \mid \mathbf{x}) = \max_{1 \leq i \leq K} P(i \mid \mathbf{x})$.

Dacă maximul din enunțul de mai sus este atins pentru $k << K$ clase, atunci lui $c_B(\mathbf{x})$ i se va asigna una din cele k valori, selectată aleator.

Dacă probabilitatea ca maximul să fie atins pentru mai mult de un i, pentru $\mathbf{x}$ dat, este zero, atunci plasatorul şi deci şi partiția R sunt unice, modulo o mulțime de măsură nulă.

Nu există nici o restricție pentru tipul de densități $P_1,\ldots,P_K$: în particular, acestea nu trebuie să fie densități față de măsura Lebesque.

Definiția 3.2-4 Plasatorul $c_B(\mathbf{x})$ se numeşte *plasator Bayes*, riscul funcțional pe care acesta îl minimizează se numeşte *risc Bayes* sau *eroare Bayes*, iar partiția $\mathcal{R}$ care determină şi este determinată de plasatorul Bayes, se numeşte *procedură de discriminare (clasare) bayesiană*.

Dată fiind importanța conceptului, vom prezenta şi alte proprietăți ale procedurilor de discriminare bayesiană.

Fie $r(i,j,\mathcal{R}) = C(j|i)P(j|i,\mathcal{R})$ costul misclasării unei observații din populația Π_i în populația Π_j prin procedura de clasare dată de partiția $\mathcal{R}$ a spațiului $\mathcal{X}$ (numită în cele ce urmează *procedură de clasare*).

Definiția 3.2-5 Procedura $\mathcal{R}$ este *mai bună* decât procedura $\mathcal{R}^* \Leftrightarrow$

$$r(i,j,\mathcal{R}) \le r(i,j,\mathcal{R}^*) \quad (\forall)\, i \ne j, \quad i,j = \overline{1,K}$$

și cel puțin una dintre inegalități este strictă.

Definiția 3.2-6 Procedura $\mathcal{R}$ este *admisibilă* dacă și numai dacă nu există o procedură $\mathcal{R}^*$ mai bună decât ea.

Definiția 3.2-7 O clasă de proceduri este *completă* dacă pentru orice procedură ce nu aparține clasei, există întotdeauna o procedură în clasă care este mai bună decât ea.

Definiția 3.2-8 O clasă de proceduri este *minimală și completă* dacă nici una din submulțimile sale nevide nu formează o clasă completă.

Propoziția 3.2-1 (Anderson, 1958) *Dacă* $P\left(P_j(\mathbf{x}) = 0 | \mathbf{x} \in \Pi_i\right) = 0 \quad (\forall) i \ne j$, $i,j = \overline{1,K}$ *atunci orice procedură bayesiană este admisibilă.*

Cu alte cuvinte, Propoziția 3.2-1 afirmă că o condiție necesară pentru ca o procedură să fie admisibilă (să nu existe o procedură de clasare mai bună decât ea) este ca suporturile tuturor distribuțiilor de probabilitate $\{P_i\}_{i=1}^K$ să difere între ele doar pe o mulțime de probabilitate nulă.

Demonstarție: Fie $\mathfrak{R} = (R_1, R_2)$. Prin reducere la absurd presupunem că $\mathfrak{R}$ procedura Bayes nu este admisibilă, atunci

$$P(1|2,\mathfrak{R}^*) \le P(1|2,\mathfrak{R})$$

și $P(2|1,\mathfrak{R}^*) \le P(2|1,\mathfrak{R})$ cu cel puțin una din inegalități stricte.

Dar $\mathfrak{R}$ este procedură Bayes (adică minimizează media costului / probabilității de misclasare), deci

$$\begin{aligned} & q_1 P(2|1,\mathfrak{R}) + q_2 P(1|2,\mathfrak{R}) \le q_1 P(2|1,\mathfrak{R}^*) + q_2 P(1|2,\mathfrak{R}^*) \\ \Rightarrow\ & q_1\left[P(2|1,\mathfrak{R}) - P(2|1,\mathfrak{R}^*)\right] \le q_2\left[P(1|2,\mathfrak{R}^*) - P(1|2,\mathfrak{R})\right] \end{aligned} \qquad (1)$$

Dacă $q_1 > 0$ și $P(1|2,\mathfrak{R}^*) \le P(1|2,\mathfrak{R}) \Rightarrow$ membrul stâng al inegalității (1) este nepozitiv $\Rightarrow P(2|1,\mathfrak{R}) \le P(2|1,\mathfrak{R}^*)$. Contradicție, $\mathfrak{R}^*$ nu este admisibilă.

Dacă $q_2 > 0$, similar $\Rightarrow P(1|2,\mathfrak{R}) \le P(1|2,\mathfrak{R}^*)$ deci iarăși contradicție.

Dacă $q_1 = 0$ atunci

$$0 \le P\left(1|2,\Re^*\right) - P\left(1|2,\Re\right) \tag{2}$$

şi regiunea $R_1 : \mathbf{x} \in \mathcal{X}\ q_1 P_1(\mathbf{x}) \ge q_2 P_2(\mathbf{x})$ a oricărei proceduri Bayes va conţine doar punctele $\mathbf{x}$ pentru care $P_2(\mathbf{x}) = 0 \Rightarrow P\left(1|2,\Re\right) = 0$ (căci $P\left(1|2,\Re\right) = \int_{R_1} P_2(\mathbf{x}) d\mathbf{x}$) şi din inegalitatea de mai sus $\Rightarrow P\left(1|2,\Re^*\right) = 0$.

Din ipoteza $P\left(P_2(\mathbf{x}) = 0 | \Pi_1\right) = 0$ rezultă, ca evenimente complementare, $P\left(P_2(\mathbf{x}) > 0 | \Pi_1\right) = 1$

Să observăm că

$$P\left(2|1,R\right) = P\left(P_2(\mathbf{x}) > 0 | \Pi_1\right) = 1$$

şi cum $\Re^*$ este admisibilă trebuie ca şi

$$P\left(2|1,R^*\right) = 1 \tag{3}$$

Din (2) şi (3) rezultă că nici una din inegalităţile de definiţie a admisibilităţii lui $\Re^*$ nu sunt verificate. Contradicţie.

Dacă $q_2 = 0$ atunci $P\left(2|1,R\right) \le P\left(2|1,R^*\right)$ contradicţie cu ipoteza de admisibilitate a lui $\Re^*$.

□

Propoziţia 3.2-2 (Anderson, 1958) *Dacă* $P\left(\frac{P_i(\mathbf{x})}{P_j(\mathbf{x})} = b \mid \mathbf{x} \in \Pi_k\right) = 0$, $(\forall) i \ne j$, $i,j,k = \overline{1,K}$ *şi* $0 \le b < \infty$, *atunci fiecare procedură admisibilă este o procedură bayesiană.*

Cu alte cuvinte Propoziţia 3.2-2 afirmă că o condiţie suficientă pentru ca o procedură bayesiană să fie admisibilă este ca oricare două distribuţii de probabilitate P_i şi P_j $i,j = \overline{1,K}$ să fie proporţionale între ele cel mult pe o mulţime de probabilitate nulă.

Demonstraţie În condiţia $P\left(\frac{P_1(\mathbf{x})}{P_2(\mathbf{x})} = b | \Pi_i\right) = 0 \quad i = \overline{1,2} \quad 0 \le b < \infty$

$\frac{P_1(\mathbf{x})}{P_2(\mathbf{x})} = \infty$ înseamnă că $P_2(\mathbf{x}) = 0$.

Atunci oricare ar fi q_1 procedura Bayes este unică. În plus funcţia de repartiţie a lui $P_1(\mathbf{x}) / P_2(\mathbf{x})$ este continuă.

Fie $\mathfrak{R}$ o procedură admisibilă. Atunci, există b astfel încât $P(2|1,\mathfrak{R}) = P\left(\frac{P_1(\mathbf{x})}{P_2(\mathbf{x})} \le b|\Pi_1\right) = P(2|1,\mathfrak{R}^*)$ unde $\mathfrak{R}^*$ este procedura Bayes corespunzând lui $q_2/q_1 = b$ căci $R_1^*: \frac{P_1}{P_2} > \frac{q_2}{q_1}$; şi $R_2^*: \frac{P_1}{P_2} \le \frac{q_2}{q_1}$.

Cum $\mathfrak{R}$ este admisibilă $P(1|2,\mathfrak{R}) \le P(1|2,\mathfrak{R}^*)$ (1).

Din propoziţia de mai sus $\mathfrak{R}^*$ Bayes este admisibilă (căci sunt verificate ipotezele propoziţiei din cazurile particulare $b=0$, $b=\infty$) deci

$$P(1|2,\mathfrak{R}) \ge P(1|2,\mathfrak{R}^*) \quad (2).$$

Din (1) şi din (2) $\Rightarrow P(1|2,\mathfrak{R}) = P(1|2,\mathfrak{R}^*)$ deci $\mathfrak{R}$ este o procedură Bayes; din unicitatea procedurii Bayes $\mathfrak{R}$ este aceeaşi cu $\mathfrak{R}^*$.

□

Cu acestea, rezultatul cheie al analizei discriminante clasice este:

Teorema 3.2-2 (Anderson, 1958) *Dacă* $P\left(\frac{P_i(\mathbf{x})}{P_j(\mathbf{x})} = b \mid \mathbf{x} \in \Pi_k\right) = 0 \quad (\forall) i \ne j,$

$i,j,k = \overline{1,K}$ *şi* $0 \le b < \infty$, *atunci clasa procedurilor bayesiene este minimală şi completă.*

Acest rezultat justifică de ce, atunci când ipotezele din Propoziţia 3.2-1, Propoziţia 3.2-2 şi cele de la începutul acestui paragraf sunt îndeplinite, întreaga cercetare se reduce la a construi o procedură admisibilă sau a aproxima, într-un anumit sens, o astfel de procedură.

3.2.2.2 Clasificarea Bayes în cazul a două populaţii normale multidimensionale cu parametrii cunoscuţi

Fie $k=2$ populaţii caracterizate de densităţile de probabilitate

$$P_i(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{1}{2}p}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]$$

adică $X \in \Pi_i \Rightarrow X \sim N(\boldsymbol{\mu}_i, \mathbf{\Sigma})$ cu $\boldsymbol{\mu}_i \in \mathcal{M}_{p\times 1}(\mathbb{R})$ vectorul medie şi $\mathbf{\Sigma} \in \mathcal{M}_{p\times p}$ matricea de varianţă-covarianţă.

Raportul densităţilor este

$$\frac{P_1(\mathbf{x})}{P_1(\mathbf{x})} = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right]}{\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right]}$$

$$= \exp\left\{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)-(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right]\right\}$$

Conform teoremei de mai sus, regiunea de clasificare în Π_1, R_1, este mulțimea punctelor $\mathbf{x} \in \mathbb{R}^p$ pentru care raportul densităților este $\geq c$ (cu c ales convenabil). Cum funcția logaritmică este monoton crescătoare condiția de definire a lui R_1 poate fi rescrisă ca:

$$-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)-(\mathbf{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right] \geq \log c$$

Termenul stâng al inegalității de mai sus, după desfacerea parantezelor și efectuarea reducerilor, devine:

$$\mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

Observație. Primul termen al formulei de mai sus este binecunoscuta *funcție discriminantă a lui Fisher.*

Corolar 3.2-2 *(al teoremei „regiunilor") Dacă* Π_i, $i=1.2$ *sunt populații multidimensionale normal distribuite de medie* $\boldsymbol{\mu}_i$ *și matricea de varianță-covarianță comună* $\boldsymbol{\Sigma}$, *atunci cele mai bune regiuni de clasificare sunt date de:*

$$R_1 : \mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \geq \ln c$$

$$R_2 : \mathbf{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) < \ln c$$

Dacă probabilitățile à priorice q_1 *și* q_2 *sunt cunoscute, atunci c este dat de*

$$c = \frac{q_2 C(1|2)}{q_1 C(2|1)}.$$

Observație. Cazul particular când $q_1 = q_2$ și $C(1|2) = C(2|1) \Rightarrow c = 1$ și $\ln c = 0$.

Dacă notăm cu $\mathbf{L}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$ prototipul populației Π_i atunci suprafața de separare a celor două regiuni este hiperplanul

$$(\mathbf{L}_1 - \mathbf{L}_2)'\left[\mathbf{x} - \frac{1}{2}(\mathbf{L}_1 + \mathbf{L}_2)\right] = 0$$

iar clasificatorul obținut este un clasificator cu distanță minimă.

Dacă probabilitățile à priorice nu sunt cunoscute atunci $C = \ln c$ va fi ales astfel încât costurile misclasării să fie egale. Mai riguros:

Teorema 3.2-3_(a egalității costurilor misclasării): *Dacă* $\Pi_i \sim N(\mu_i, \Sigma), i = 1,2$ *regiunile Bayes sunt date de relațiile din corolarul* Corolar **3.2-2** *cu* $C = \ln c$ *ales astfel încât*

$$C(1|2)\left[1 - \Phi\left(\frac{C + \frac{1}{2}\alpha}{\sqrt{2}}\right)\right] = C(1|2)\Phi\left[1 - \left(\frac{C - \frac{1}{2}\alpha}{\sqrt{2}}\right)\right]$$

unde $C(i|j)$ *sunt cele două costuri ale misclasării,* $\alpha = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ *este distanța Mahalanobis dintre cele două populații, iar* $\Phi(x)$ *este funcția de repartiție a variabilei aleatoare Gauss-Laplace (adică* $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$*).*

Demonstrație:

Fie $U = \mathbf{X}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

Regiunile Bayes sunt, conform Corolar 3.2-2

$R_1 : U \geq C$ şi $R_2 : U < C$,

iar costurile misclasării sunt

$$C(2|1)\int_{R_2} f(U|X \in \Pi_1)du = C(2|1)\int_{-\infty}^{c} f_{1,U}(t)dt$$

pentru U construit pe baza unei observații $\mathbf{X} \in \Pi_1$

şi

$$C(1|2)\int_{R_1} f(U|X \in \Pi_2)du = C(1|2)\int_{c}^{\infty} f_{2,U}(t)dt$$

pentru U construit pe baza unei observații $\mathbf{X} \in \Pi_2$.

Soluția minimax de alegere a lui C impune ca $C(2|1)\int_{-\infty}^{c} f_{1,U}(t)dt = C(1|2)\int_{c}^{\infty} f_{2,U}(t)dt$.

Pentru a finaliza demonstrația mai rămâne de evaluat distribuțiile condiționate ale lui $U\ f(U|X\in\Pi_i)$.

Fie $\mathbf{X}\in\Pi_1\Rightarrow\mathbf{X}\sim N(\boldsymbol{\mu}_1,\Sigma)$ atunci

$$U=\mathbf{X}'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

este distribuită normal (căci combinații liniare de normale este tot o normală), de medie

$$M[U]=\boldsymbol{\mu}_1'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

şi dispersie

$$\begin{aligned}D^2(U)&=D^2\left[\mathbf{X}'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]\\&=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}D^2[\mathbf{X}]\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\&=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\\&=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\end{aligned}$$

Notând „distanța" dintre cele două populații cu α, $\Rightarrow U\sim N\left(\frac{1}{2}\alpha,\alpha\right)$.

Dacă $X\sim N(\boldsymbol{\mu}_2,\Sigma)$ atunci $U\sim N\left(-\frac{1}{2}\alpha,\alpha\right)$.

În concluzie

$$f_{1,U}(t)=\frac{1}{\sqrt{2\pi\alpha}}e^{\frac{1}{2}\left(t-\frac{1}{2}\alpha\right)^2\Big/\alpha}\quad\text{şi}\quad f_{2,U}(t)=\frac{1}{\sqrt{2\pi\alpha}}e^{\frac{1}{2}\left(t+\frac{1}{2}\alpha\right)^2\Big/\alpha}$$

Cu acestea egalitatea costurilor misclasificării se scrie

$$C(2|1)\int_{-\infty}^{c}\frac{1}{\sqrt{2\pi\alpha}}e^{\frac{\left(t-\frac{1}{2}\alpha\right)^2}{2\alpha}}dt=C(1|2)\int_{c}^{\infty}\frac{1}{\sqrt{2\pi\alpha}}e^{\frac{\left(t-\frac{1}{2}\alpha\right)^2}{2\alpha}}dt$$

În membrul stâng al egalității se face transformarea $z=\dfrac{t-\frac{1}{2}\alpha}{\sqrt{2}}$, iar în membrul drept al egalității se face transformarea $z=\dfrac{t+\frac{1}{2}\alpha}{\sqrt{2}}$

Cu jacobianul transformării (același pentru ambele transformări) egal cu $\frac{1}{\sqrt{a}}$, se obține în final

$$C(2|1)\int_{-\infty}^{c-\frac{1}{2}\alpha}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz = C(1|2)\int_{c-\frac{1}{2}\alpha}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz$$

şi ţinând cont că $\int_x^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy = 1-\Phi(x)$ se obţine egalitatea din enunţul teoremei.

□

Observaţii

a) Reprezentarea grafică a problemei este dată în graficul

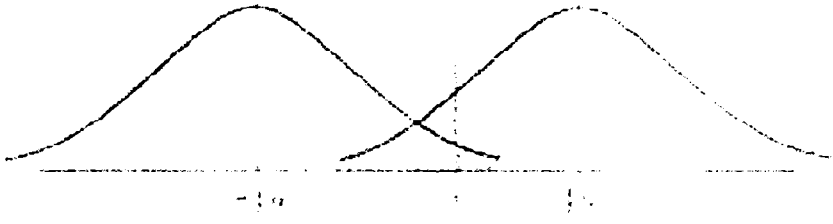


Figura 3.2-1 Zona de misclasare în cazul a două populaţii normale unidimensionale

Zona haşurată este zona de misclasare.

Să notăm că cele două condiţii pentru ca procedura de clasificare să fie minimală şi completă, anume $P(P_1(\mathbf{x})=0|\Pi_2)=0$ şi $P(P_2(\mathbf{x})=0|\Pi_1)=0$ cât şi $P\left(\frac{P_1(\mathbf{x})}{P_2(\mathbf{x})}=k|\Pi_1\right)=0$ şi $P\left(\frac{P_1(\mathbf{x})}{P_2(\mathbf{x})}=k|\Pi_2\right)=0$ sunt îndeplinite.

b) Dacă $C(1|2)=C(2|1)$ atunci egalitatea probabilităţilor de misclasare implică $C=0$ şi deci probabilitatea misclasării este $\int_{\frac{\sqrt{\alpha}}{2}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy = 1-\Phi\left(\frac{\sqrt{\alpha}}{2}\right)$.

c) Determinarea lui C care satisface cu o precizie suficientă condiţia din enunţul teoremei se află rezolvând numeric, pe baza tabelelor existente, ecuaţia $k\Phi(x)+\Phi\left(x+\sqrt{\alpha}\right)=1$, unde $k=\frac{C(2|1)}{C(1|2)}$, iar $C=\sqrt{\alpha}\left(x+\frac{1}{2}\sqrt{\alpha}\right)$.

d) În condiţiile de definire a regiunilor (R_1,R_2) apare termenul $\boldsymbol{\delta}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$. Este interesant de notat că $\mathbf{x}'\boldsymbol{\delta}$ este funcţie liniară care maximizează

$$\frac{\left[M(\mathbf{x}'\mathbf{d}|\mathbf{X}\in\Pi_1)-M(\mathbf{x}'\mathbf{d}|\mathbf{X}\in\Pi_2)\right]^2}{D^2(\mathbf{x}'\mathbf{d})}$$

(nu importă de unde „vine" $\mathbf{x}$ căci cele două populații au aceeași matrice de varianță-covarianță Σ).

Acesta este demersul folosit de Fisher pentru obținerea funcției de discriminare liniară ce-i poartă numele.

Numărătorul câtului de mai sus este

$$\left[\boldsymbol{\mu}_1'\mathbf{d}-\boldsymbol{\mu}_2'\mathbf{d}\right]^2=\mathbf{d}'\left[(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\right]\mathbf{d}$$

iar numitorul este

$$\mathbf{d}'M\left[(\mathbf{X}-M(\mathbf{X}))(\mathbf{X}-M(\mathbf{X}))'\right]\mathbf{d}=\mathbf{d}'\Sigma\mathbf{d}.$$

Problema s-a redus la următoarea problemă de optimizare pătratică cu restricții

$$\begin{cases}\max\limits_{\mathbf{d}\in\mathbb{R}^r}\dfrac{\mathbf{d}'\left[(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\right]\mathbf{d}}{\mathbf{d}'\Sigma\mathbf{d}}\\ \mathbf{d}'\Sigma\mathbf{d}=1\end{cases}$$

care se rezolvă folosind tehnica multiplicatorilor lui Lagrange.

Fie deci lagrangeanul

$$\mathbb{L}=\mathbf{d}'\left[(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\right]\mathbf{d}-\lambda(\mathbf{d}'\Sigma\mathbf{d}-1)\text{ cu }\lambda\text{ multiplicatorul lui Lagrange.}$$

$$\frac{\partial\mathbb{L}}{\partial\lambda}0\Rightarrow 2\left[(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\right]\mathbf{d}=2\lambda\Sigma\mathbf{d}\text{ căci }\Sigma\text{ este simetrică.}$$

Cum $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\mathbf{d}=s$ este un scalar, ecuația de mai sus se rescrie

$$\boldsymbol{\mu}_1-\boldsymbol{\mu}_2=\frac{\lambda}{s}\Sigma\mathbf{d}\Rightarrow\mathbf{d}=\frac{s}{\lambda}\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

căci Σ este pozitiv definită, deci inversabilă.

$\mathbf{d}$ este proporțional cu $\boldsymbol{\delta}$. Pentru determinarea lui $\frac{s}{\lambda}$ se utilizează Σ-normarea lui $\mathbf{d}$, adică

$$\mathbf{d}'\Sigma\mathbf{d}=1\Rightarrow\left(\frac{s}{\lambda}\right)^2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}\Sigma(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=1\Rightarrow\frac{s}{\lambda}=\frac{1}{\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|}.$$

Așadar $\mathbf{x}'\boldsymbol{\delta}$ este funcția liniară care are cea mai mare dispersie între clase (dispersia interclase) relativ la dispersia în clase (dispersia intraclase).

Atunci când populațiile sunt cunoscute, criteriul folosit este optim din punct de vedere al minimizării erorii de clasare; când probabilitățile à priori nu sunt cunoscute procedura generează o clasă de proceduri admisibile. Ce se poate spune despre cazul estimațiilor?

3.2.2.3 Clasificarea Bayes în cazul a două populații normale multidimensionale cu parametrii necunoscuți

- **Estimatori de resubstituție (*plug-in*)**

Fie $\mathbf{x}_1^{(i)},\ldots,\mathbf{x}_{n_i}^{(i)} \in N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$, $i=1,2$ două selecții bernoulliene.

Se cunosc rezultatele următoare

$$\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_j^{(i)} \quad i=1,2$$

$$(n_1+n_2-2)\mathbf{S} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\mathbf{x}_j^{(i)}-\bar{\mathbf{x}}_i\right)\left(\mathbf{x}_j^{(i)}-\bar{\mathbf{x}}_i\right)'$$

sunt estimatori nedeplasați, de verosimilitate maximă, ai lui $\boldsymbol{\mu}_i$, $i=1,2$ şi $\boldsymbol{\Sigma}$.

Fie

$$\mathbf{Z}_{12} = \mathbf{X} - \frac{1}{2}\left(\bar{\mathbf{x}}^{(1)}+\bar{\mathbf{x}}^{(2)}\right)$$

$$\mathbf{Y}_{12} = \bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}$$

atunci

$$\mathbf{V}_{12} = \mathbf{X}'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\right) - \frac{1}{2}\left(\bar{\mathbf{x}}^{(1)}+\bar{\mathbf{x}}^{(2)}\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\right)$$

$$= \left[\mathbf{X}-\frac{1}{2}\left(\bar{\mathbf{x}}^{(1)}+\bar{\mathbf{x}}^{(2)}\right)\right]'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\right) = \mathbf{Z}_{12}\mathbf{S}^{-1}\mathbf{Y}_{12}$$

Din construcție

$$\mathbf{Y}_{12} \sim N\left(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)},\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\boldsymbol{\Sigma}\right)$$

iar

$$\mathbf{Z}_{12} \sim N\left[\frac{1}{2}\left(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)}\right),\left(1+\frac{1}{4n_1}+\frac{1}{4n_2}\right)\boldsymbol{\Sigma}\right] \text{ dacă } \mathbf{X} \sim N\left(\boldsymbol{\mu}^{(1)},\boldsymbol{\Sigma}\right)$$

$$\mathbf{Z}_{12} \sim N\left[\frac{1}{2}\left(\boldsymbol{\mu}^{(2)}-\boldsymbol{\mu}^{(1)}\right),\left(1+\frac{1}{4n_1}+\frac{1}{4n_2}\right)\boldsymbol{\Sigma}\right] \text{ dacă } \mathbf{X} \sim N\left(\boldsymbol{\mu}^{(2)},\boldsymbol{\Sigma}\right)$$

$$\operatorname{cov}\left(\mathbf{Z}_{12},\mathbf{Y}_{12}\right) = -\left(\frac{1}{2n_1}-\frac{1}{2n_2}\right)\boldsymbol{\Sigma}.$$

Dacă $N_1 = N_2$ atunci $\text{cov}(\mathbf{Z}, \mathbf{Y}) = 0$. În acest caz distribuția lui V când $\mathbf{X} \in \Pi_1$ este aceeași cu a lui V când $\mathbf{X} \in \Pi_2$. Atunci, dacă $R_1 = \{x \in \mathcal{X} / V(x) \geq 0\}$, probabilitățile de misclasare sunt egale.

Asimptotic, cum

$$\bar{\mathbf{x}}^{(1)} \underset{n_1 \to \infty}{\overset{p}{\to}} \boldsymbol{\mu}^{(1)}; \; \bar{\mathbf{x}}^{(2)} \underset{n_2 \to \infty}{\overset{p}{\to}} \boldsymbol{\mu}^{(2)} \text{ și}$$

$$\mathbf{S} \underset{n_1, n_2 \to \infty}{\overset{p}{\to}} \boldsymbol{\Sigma}$$

rezultă

$$\mathbf{S}^{-1}\left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right) \underset{n_1, n_2 \to \infty}{\overset{p}{\to}} \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}\right)$$

$$\left(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}\right)' \mathbf{S}^{-1}\left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right) \underset{n_1, n_2 \to \infty}{\overset{p}{\to}} \left(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)}\right)' \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}\right)$$

adică distribuția asimptotică a lui V este $\mathbf{U}_{12}$.

Concluzie: Pentru selecții suficient de mari folosirea estimațiilor în locul valorilor exacte implică erori mici.

Urmându-l pe Anderson (1958) vom substitui parametrii estimați în relațiile de definiție ale regiunilor de decizie obținând

$$R_1 : \mathbf{x}'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) - \frac{1}{2}\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) \geq \ln k$$

$$R_2 : \mathbf{x}'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) - \frac{1}{2}\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) < \ln k$$

Anderson argumentează în favoarea acestui criteriu că minimizează costurile misclasării dacă parametrii populațiior sunt cunoscuți și continuă „it seems intuitively reasonable that the above relations should give good results".

Dacă se dorește clasificarea selecțiior reunite ca un tot atunci se utilizează următorii estimatori, respectiv criteriu

$$n = n_1 + n_2$$

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j$$

cu

$$\mathbf{x}_j \in \Pi_1 / \in \Pi_2,$$

$$(n_1 + n_1 + n - 3)\bar{\mathbf{S}} = \mathbf{S} + \sum_{j=1}^{n}\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)'$$

respectiv

$$R_1 : \left[\bar{\mathbf{x}} - \frac{1}{2}\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right)\right] S^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right) \geq c .$$

Se poate arăta că $N \to \infty \Rightarrow P(1|2), P(2|1) \to 0$

Regăsirea rezultatului din Teorema 3.2-1.

a) Cazul $k=2$ (două clase). Particularizând regiunile de decizie de mai sus se obține

$$R_1 = \left\{x \in \mathcal{X} \mid q_2 P_2(x) \leq q_1 P_1(x)\right\}$$

$$= \left\{x \in \mathcal{X} \mid \frac{P_1(x)}{P_2(x)} \geq \frac{q_2}{q_1}\right\}$$

Punând $P_i = p_i$, $q_1 = q_2$ și $\boldsymbol{\mu}^{(i)}$ și $\boldsymbol{\Sigma}$ estimați rezultă

$R_1 = \left\{x \in \mathcal{X} \mid V_{12} \geq 0\right\}$ și $R_2 = \mathcal{X} - R_1$.

Când $p=1$ atunci $V_{12}(x) = 0 \Rightarrow x = \dfrac{\bar{x}^{(1)} + \bar{x}^{(2)}}{2}$ „suprafața" de decizie este un punct;

$p=2$ atunci $V_{12}(x) = 0 \Rightarrow$ „suprafața" de decizie este o dreaptă;

$p=3$ atunci $V_{12}(x) = 0 \Rightarrow$ suprafața de decizie este un plan;

$p \geq 4$ atunci $V_{12}(x) = 0 \Rightarrow$ suprafața de decizie este un hiperplan.

b) Cazul $k=3$ (trei clase). Particularizând, se obțin următoarele regiuni de decizie:

$$R_1 = \left\{\begin{array}{c} x \in \mathcal{X} \mid q_2 P_2 + q_3 P_3 \leq q_1 P_1 + q_3 P_3, \\ q_2 P_2 + q_3 P_3 \leq q_1 P_1 + q_2 P_2 \end{array}\right\} = \left\{x \in \mathcal{X} \mid \frac{P_1}{P_2} \geq \frac{q_2}{q_1}, \frac{P_1}{P_3} \geq \frac{q_3}{q_1}\right\},$$

$$R_2 = \left\{\begin{array}{c} x \in \mathcal{X} \mid q_1 P_1 + q_3 P_3 \leq q_1 P_1 + q_2 P_2, \\ q_1 P_1 + q_3 P_3 \leq q_2 P_2 + q_3 P_3 \end{array}\right\} = \left\{x \in \mathcal{X} \mid \frac{P_2}{P_3} \geq \frac{q_3}{q_2}, \frac{P_2}{P_1} \geq \frac{q_1}{q_2}\right\},$$

$$R_3 = \left\{\begin{array}{c} x \in \mathcal{X} \mid q_1 P_1 + q_2 P_2 \leq q_1 P_1 + q_3 P_3, \\ q_1 P_1 + q_2 P_2 \leq q_2 P_2 + q_3 P_3 \end{array}\right\} = \left\{x \in \mathcal{X} \mid \frac{P_3}{P_2} \geq \frac{q_2}{q_3}, \frac{P_3}{P_1} \geq \frac{q_1}{q_3}\right\}$$

și punând $P_i = p_i$, $q_1 = q_2 = q_3$ și $\boldsymbol{\mu}^{(i)}$ și $\boldsymbol{\Sigma}$ estimați rezultă

$$R_1 = \left\{x \in \mathcal{X} \mid V_{12} > 0, V_{13} > 0\right\}$$

$$R_2 = \left\{x \in \mathcal{X} \mid V_{21} > 0, V_{23} > 0\right\} = \left\{x \in \mathcal{X} \mid V_{12} < 0, V_{13} > V_{12}\right\}$$

căci $V_{ij} = -V_{ji}$ și $V_{23} = V_{13} - V_{12}$ și

$$R_3 = \left\{x \in \mathcal{X} \mid V_{32} > 0, V_{31} > 0\right\} = \left\{x \in \mathcal{X} \mid V_{13} < 0, V_{12} > V_{13}\right\}.$$

Dacă $p=1$ (o singură caracteristică) și presupunând $\bar{x}^{(1)} < \bar{x}^{(2)} < \bar{x}^{(3)}$ atunci regiunile de decizie devin semidrepte și segment de dreaptă, adică:

$$R_1 : x \in \Re \text{ cu } x < \frac{\bar{x}^{(1)} + \bar{x}^{(2)}}{2},$$

$$R_2 : x \in \Re \text{ cu } \frac{\bar{x}^{(1)} + \bar{x}^{(2)}}{2} \le x \le \frac{\bar{x}^{(2)} + \bar{x}^{(3)}}{2},$$

$$R_2 : x \in \Re \text{ cu } \frac{\bar{x}^{(2)} + \bar{x}^{(3)}}{2} < x.$$

Când $p = 2$ regiunile de decizie devin semiplane (vezi Figura 3.2-2).

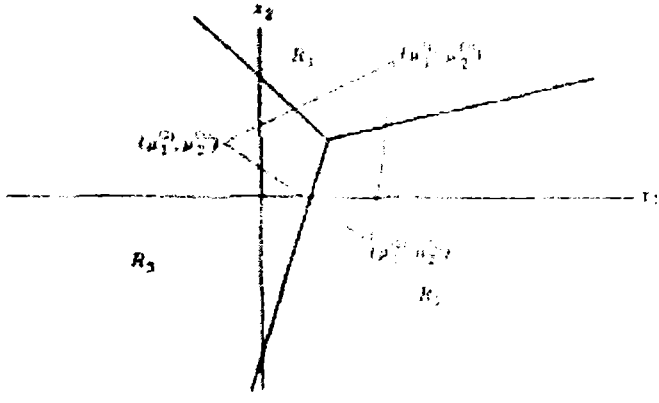


Figura 3.2-2 Exemplu de regiuni de decizie în cazul normalei bidimensionale

- **Estimatori de verosimilității maxime**

Fie ipoteză compozită

$$H_O : \mathbf{x}, \mathbf{x}_1^{(1)}, ..., \mathbf{x}_{n_1}^{(1)} \in N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$\mathbf{x}_1^{(2)}, ..., \mathbf{x}_{n_2}^{(2)} \in N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

$$H_A : \mathbf{x}_1^{(1)}, ..., \mathbf{x}_{n_1}^{(1)} \in N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$\mathbf{x}, \mathbf{x}_1^{(2)}, ..., \mathbf{x}_{n_2}^{(2)} \in N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

În ipoteza H_O estimatorii de verosimilitate maximă sunt

$$\hat{\boldsymbol{\mu}}_1^{(0)} = \left(n_1 \bar{\mathbf{x}}_1 + \mathbf{x}\right) / \left(n_1 + 1\right)$$

$$\hat{\boldsymbol{\mu}}_2^{(0)} = \bar{\mathbf{x}}_2$$

$$\hat{\Sigma}^{(0)} = \frac{1}{n_1 + n_1 + 1}\left[\begin{array}{l}\sum_{j=1}^{n_1}\left(\mathbf{x}_j^{(1)} - \hat{\boldsymbol{\mu}}_1^{(0)}\right)\left(\mathbf{x}_j^{(1)} - \hat{\boldsymbol{\mu}}_1^{(0)}\right)' \\ +\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1^{(0)}\right)\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1^{(0)}\right)' \\ +\sum_{j=1}^{n_2}\left(\mathbf{x}_j^{(2)} - \hat{\boldsymbol{\mu}}_2^{(0)}\right)\left(\mathbf{x}_j^{(2)} - \hat{\boldsymbol{\mu}}_2^{(0)}\right)'\end{array}\right].$$

Deoarece

$$\sum_{j=1}^{n_1}\left(\mathbf{x}_j^{(1)} - \hat{\boldsymbol{\mu}}_1^{(0)}\right)\left(\mathbf{x}_j^{(1)} - \hat{\boldsymbol{\mu}}_1^{(0)}\right)'$$

$$+\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1^{(0)}\right)\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1^{(0)}\right)' =$$

$$= \sum_{j=1}^{n_1}\left(\mathbf{x}_j^{(1)} - \bar{\mathbf{x}}_1\right)\left(\mathbf{x}_j^{(1)} - \bar{\mathbf{x}}_1\right)' + n\left(\bar{\mathbf{x}}_1 - \hat{\boldsymbol{\mu}}_1^{(0)}\right)\left(\bar{\mathbf{x}}_1 - \hat{\boldsymbol{\mu}}_1^{(0)}\right)'$$

$$+\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1^{(0)}\right)\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_1^{(0)}\right)'$$

$$= \sum_{j=1}^{n_1}\left(\mathbf{x}_j^{(1)} - \bar{\mathbf{x}}_1\right)'\left(\mathbf{x}_j^{(1)} - \bar{\mathbf{x}}_1\right)' + \frac{n_1}{n_1 + 1}\left(\mathbf{x} - \bar{\mathbf{x}}_1\right)\left(\mathbf{x} - \bar{\mathbf{x}}_1\right)'$$

Rezultă $\hat{\Sigma}^{(0)} = \frac{1}{n_1 + n_1 + 1}\left[C + \frac{n_1}{n_1 + 1}\left(\mathbf{x} - \bar{\mathbf{x}}_1\right)\left(\mathbf{x} - \bar{\mathbf{x}}_1\right)'\right]$

cu $C = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}_i\right)\left(\mathbf{x}_j^{(i)} - \bar{\mathbf{x}}_i\right)'$.

Analog, sub H_A estimatorii de verosimilitate maximă sunt:

$$\hat{\boldsymbol{\mu}}_1^{(A)} = \bar{\mathbf{x}}_1$$

$$\hat{\boldsymbol{\mu}}_2^{(A)} = \left(n_2\bar{\mathbf{x}}_2 + \mathbf{x}\right)/\left(n_2 + 1\right)$$

$$\hat{\Sigma}^{(A)} = \frac{1}{n_1 + n_2 + 1}\left[C + \frac{n_2}{n_2 + 1}\left(\mathbf{x} - \bar{\mathbf{x}}_2\right)\left(\mathbf{x} - \bar{\mathbf{x}}_2\right)'\right]$$

Raportul de verosimilitate devine aşadar

$$\Lambda = \frac{1+\frac{n_2}{n_2+1}\left(\mathbf{x}-\bar{\mathbf{x}}_2\right)'\mathbf{C}^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}_2\right)}{1+\frac{n_1}{n_1+1}\left(\mathbf{x}-\bar{\mathbf{x}}_1\right)'\mathbf{C}^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}_1\right)} \text{ iar}$$

R_1 : $\mathbf{x}$ cu $\Lambda \geq \mathbf{C}$ (acele puncte $\mathbf{x}$ care maximizează pe Λ).

- **Estimare bayesiană**

Natura discuției din acest paragraf este conceptual foarte diferită de abordarea anterioară. Anterior am prezentat o metodologie pornind de la un punct de vedere frecventist; am presupus o selecție aleatoare $x_1,...,x_n$ dintr-o populație având densitatea de probabilitate (funcția de repartiție $f(x;\theta)$) cu $x \in \mathfrak{X}$ și $\theta \in \Theta$. Parametrul necunoscut θ este presupus fixat. O procedură de inferența frecventistă depinde de funcția de verosimilitate $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ unde θ este necunoscut dar fix.

În demersul bayesian experimentatorul presupune/crede înainte de a „vedea datele" (à priori adică) că parametrul necunoscut θ este o variabilă aleatoare având o distribuție de probabilitate proprie pe Θ (spațiul parametrilor), notată $h(\theta)$ și numită *distribuția à priorică* (prior distribution) a lui θ; $f(x;\theta)$ devine în acest context $f(x|\theta)$.

Distribuția à priorică $h(\theta)$ reflectă adesea intuiția subiectivă a statisticianului privitoare la ce valori ale lui θ sunt mai puțin probabile când se consideră întreg spațiul parametrilor Θ.

Distribuția à priorică este, în cazul ideal, dată/fixată înainte de începerea experimentului (a culegerii selecției bernoulliene).

Paradigma bayesiană implică combinarea informațiilor à priorice cu cele date de funcția de verosimilitate și obținerea a ceea ce este numit *distribuție à posteriori*, via teorema Bayes.

Ca fapt istoric este de reținut opoziția vehementă a lui R.A.Fisher la tot ce era bayesian.

Se cunosc următoarele fapte

- distribuția comună a lui x și θ este dată de
  $f(x|\theta)h(\theta)$ $(\forall)\, x \in \mathfrak{X}$ și $\theta \in \Theta$
- distribuția marginală a lui x este atunci
  $m(x) = \int_{\Theta} f(x|\theta)h(\theta)d\theta$

$\Rightarrow$ distribuția lui θ condiționată de evenimentul $X = x$ este, conform teoremei lui Bayes

$$h(\theta|x)=h(\theta|X=x)=\frac{f(x|\theta)h(\theta)}{m(x)} \quad x\in\mathfrak{X} \text{ şi } \theta\in\Theta \text{ cu } m(x)>0.$$

Definiția 3.2-9 $h(\theta|x)$ se numeşte *distribuția à posteriori* a lui θ.

Definiția 3.2-10 Fie $h(\theta)\in\mathfrak{D}$ (≡ familie de distribuții particulare). $h(\theta)$ se numeşte *distribuția à priorică conjugată* $\Leftrightarrow h(\theta|x)\in\mathfrak{D}$.

Propoziția 3.2-3 *Dacă* $\boldsymbol{\theta}\sim N(\mathbf{m},S)$ şi $\mathbf{x}\sim N(\theta,\Sigma)$ *atunci* $h(\theta|x)$*este densitatea de probabilitate a unei* $N(\boldsymbol{\mu},\mathbf{C})$ *cu* $\boldsymbol{\mu}=\mathbf{S}(\mathbf{S}+\boldsymbol{\Sigma})^{-1}\mathbf{x}+\boldsymbol{\Sigma}(\mathbf{S}+\boldsymbol{\Sigma})^{-1}\mathbf{m}$ *şi* $\mathbf{C}=\boldsymbol{\Sigma}(\mathbf{S}+\boldsymbol{\Sigma})^{-1}\mathbf{S}$.

Demonstrație. După observarea lui $\mathbf{x}$ densitatea condiționată

$$h(\theta|\mathbf{x})=\frac{h(\theta)f(\mathbf{x}|\theta)}{\int_{\mathbf{R}}h(\theta)f(\mathbf{x}|\theta)\,d\theta}$$
$$=C\,h(\theta)f(\mathbf{x}|\theta)$$

cu C factor ce depinde de $\mathbf{x}$ dar nu şi de θ.

Din ipotezele propoziției rezultă

$$h(\theta|\mathbf{x})=c_1\exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\mathbf{m})'\mathbf{S}^{-1}(\boldsymbol{\theta}-\mathbf{m})\right]\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\mathbf{x})\right]$$
$$=c_1\exp\left[-\frac{1}{2}\mathbf{m}'\mathbf{S}^{-1}\mathbf{m}-\frac{1}{2}(\boldsymbol{\theta}'\mathbf{S}^{-1}\boldsymbol{\theta}-2\boldsymbol{\theta}'\mathbf{S}^{-1}\mathbf{m})\right]\times$$
$$\exp\left[-\frac{1}{2}(\boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}-2\boldsymbol{\theta}'\mathbf{S}^{-1}\mathbf{x}+\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})\right]$$

În final se obține

$$h(\theta|\mathbf{x})=c_2\exp\left\{-\frac{1}{2}\left[\boldsymbol{\theta}'(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})\boldsymbol{\theta}-2\boldsymbol{\theta}'(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\mathbf{m})\right]\right\} \qquad (1)$$

unde factorii care nu depind de $\boldsymbol{\theta}$ au fost absorbiți în c_1 şi c_2.

Deoarece paranteza dreaptă din exponentul egalității (1) este o formă pătratică, rezultă că densitatea de probabilitate $h(\theta|\mathbf{x})$ este o densitate a unei variabile aleatoare normale. Pentru a determina parametrii acestei legi scriem pe $h(\theta|\mathbf{x})$ sub forma

$$h(\theta|\mathbf{x})=c_3\exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})'\mathbf{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]=c_4\exp\left[-\frac{1}{2}(\boldsymbol{\theta}'\mathbf{C}^{-1}\boldsymbol{\theta}-2\boldsymbol{\theta}\mathbf{C}^{-1}\boldsymbol{\mu})\right] \qquad (2)$$

Comparând (1) cu (2) se obține

$\mathbf{C}^{-1}=\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1}$ şi $\mathbf{C}^{-1}\boldsymbol{\mu}=\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\mathbf{m}\Rightarrow\boldsymbol{\mu}=\mathbf{C}\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{C}\mathbf{S}^{-1}\mathbf{m}$.

Se observă că, dacă $\mathbf{C}^{-1}=\mathbf{\Sigma}^{-1}+\mathbf{S}^{-1}$ atunci $\mathbf{C}=\mathbf{\Sigma}(\mathbf{S}+\mathbf{\Sigma})^{-1}\mathbf{S}=\mathbf{S}(\mathbf{S}+\mathbf{\Sigma})^{-1}\mathbf{\Sigma}$.

Într-adevăr

$$\begin{aligned}\mathbf{C}^{-1}&=\left[\mathbf{\Sigma}(\mathbf{S}+\mathbf{\Sigma})^{-1}\mathbf{S}\right]^{-1}=\mathbf{S}^{-1}\left[\mathbf{\Sigma}(\mathbf{S}+\mathbf{\Sigma})^{-1}\right]^{-1}\\&=\mathbf{S}^{-1}(\mathbf{S}+\mathbf{\Sigma})\mathbf{\Sigma}^{-1}=\mathbf{\Sigma}^{-1}+\mathbf{S}^{-1}\\&=\mathbf{C}^{-1}\end{aligned}$$

$$\begin{aligned}\mathbf{C}^{-1}&=\left[\mathbf{S}(\mathbf{S}+\mathbf{\Sigma})^{-1}\mathbf{\Sigma}\right]^{-1}=\mathbf{\Sigma}^{-1}\left[\mathbf{S}(\mathbf{S}+\mathbf{\Sigma})^{-1}\right]^{-1}\\&=\mathbf{\Sigma}^{-1}(\mathbf{S}+\mathbf{\Sigma})\mathbf{S}^{-1}=\mathbf{\Sigma}^{-1}+\mathbf{S}^{-1}\\&=\mathbf{C}^{-1}\end{aligned}$$

Înlocuind în expresia lui $\boldsymbol{\mu}$ rezultă formula din enunț.

□

Corolar 3.2-3 *Dacă* $\theta\sim N(\tau,\sigma_0^2)$ *și* $x\sim N(\theta,\sigma_1^2)$ *atunci densitatea à posteriori a lui* θ *este* $N(\mu,\sigma^2)$ *cu* $\mu=\left(\frac{x}{\sigma_1^2}+\frac{\tau}{\sigma_0^2}\right)\left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma_1^2}\right)^{-1}$ *și* $\sigma^2=\frac{\sigma_0^2\sigma_1^2}{\sigma_0^2+\sigma_1^2}=\left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma_1^2}\right)^{-1}$

Definiția 3.2-11 Fie X variabila aleatoare $:\Omega\to\mathbb{R}$ cu densitatea de probabilitate $f(\mathbf{x};\theta)$ depinzând de θ. O funcție $T:\Omega\to\mathbb{R}$ se numește *statistică suficientă* pentru $\theta\Leftrightarrow$

$$f(\mathbf{x}|T(\mathbf{x})=t,\theta)=f(\mathbf{x}|T(\mathbf{x})=t)\quad(\forall)\,t\in\mathcal{T}\subseteq\mathbb{R},$$

densitatea de probabilitate condiționată a lui X este independentă de θ.

Fie $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ o selecție bernoulliană asupra unei variabile aleatoare ce depinde de un parametru θ.

Fie $\delta\equiv\delta(T)$ un estimator a lui θ și fie *funcția de pierdere* ce se obține estimând pe θ prin $\delta(T)$:

$$L^*(\theta,\delta)\equiv L^*(\theta,\delta(T))=\left[\delta(T)-\theta\right]^2$$

Riscul funcțional este atunci

$$R^*(\theta,\delta)\equiv M\left[L^*(\theta,\delta)\right]=\int_{\mathcal{T}}L^*(\theta,\delta(t))f(t|\theta)\,dt$$

Definiția 3.2-12 Se numește *risc bayesian*

$$r^*(\theta,\delta)=\int_{\Theta}R^*(\theta,\delta)h(\theta)\,d\theta.$$

Definiția 3.2-13 Se numește *estimator bayesian*

$r^*(\theta,\delta^*)=\inf_{\delta\in\mathcal{B}} r^*(\theta,\delta)$, $\delta^*\in\mathcal{B}$ (clasa estimatorilor pentru care riscul bayesian este finit).

Teorema 3.2-4 *În cazul funcției de pierdere „suma pătratelor erorilor", estimatorul bayesian* $\delta^*\equiv\delta^*(t)$ *este media distribuției à posteriori* $h(\theta|t)$ *adică*

$$\delta^*(t)=\int_\Theta \theta h(\theta|t)\,d\theta\equiv M\left[\theta|T(x)=t\right]$$

pentru toate valorile posibile observate $t\in\mathcal{T}$.

Demonstrație Pentru a determina pe $\delta^*(t)$ trebuie minimizat

$$r^*(\theta,\delta)=\int_\Theta\int_{\mathcal{T}} L^*\left(\theta,\delta(t)\right)f(t|\theta)h(\theta)dt\,d\theta$$
$$=\int_{\mathcal{T}}\left[\int_\Theta L^*\left(\theta,\delta(t)\right)f(\theta|t)d\theta\right]m(t)dt$$

unde $m(t)=\dfrac{1}{\int_\Theta h(\theta)f(t|\theta)d\theta}$ (conform teoremei Fubini și a faptului că integranzii sunt nenegativi)

$$\int_\Theta L^*\left(\theta,\delta(t)\right)f(\theta|t)d\theta=\int_\Theta\left[\theta^2-2\theta\delta(t)+\delta^2(t)\right]f(\theta|t)d\theta$$
$$=\alpha(t)-2\delta(t)M\left[\theta|T(x)=t\right]+\delta^2(t)$$

unde s-a notat cu $\alpha(t)=\int_\Theta\theta^2 f(\theta|t)d\theta$ și s-a folosit egalitatea $\int_\Theta f(\theta|t)d\theta=1$.

Considerăm expresia $\alpha(t)-2\delta(t)M\left[\theta|T(x)=t\right]+\delta^2(t)$ ca o funcție de $\delta\equiv\delta(t)$ pe care dorim să o minimizăm. Minimul este atins (căci expresia ca funcție de δ este o parabolă cu coeficientul lui δ^2 pozitiv)

$$\Rightarrow\frac{\partial}{\partial\delta}\left(\alpha(t)-2\delta M\left[\theta|t\right]+\delta^2\right)=0\Rightarrow\delta^*=M\left[\theta|T(x)=t\right].$$

□

Corolar 3.2-4 *Fie* $x_1,\ldots,x_2$ *variabile aleatoare independente și identic repartizate* $N(\theta,\sigma_1^2)$ *cu* θ *necunoscut și* $\sigma_1>0$ *dat. Considerăm statistica* $T=\dfrac{1}{n}\sum_{i=1}^{n}x_i$ *care este suficientă pentru* θ. *Se presupune că distribuția à priori a lui* θ *pe spațiul* $\Theta=\mathbb{R}$ *este* $N(\tau,\sigma_0^2)$ *cu* τ *și* $\sigma_0>0\in\mathbb{R}$ *și dați. Distribuția à posteriori a lui* θ *condiționată de observațiile* $x_1,\ldots,x_2$ *este, conform propoziției anterioare* $N(\mu,\sigma^2)$ *cu*

$$\mu = \frac{n\sigma_0^2}{n\sigma_0^2 + n\sigma_1^2} T(x) + \frac{\sigma_1^2}{n\sigma_0^2 + \sigma_1^2}\tau$$

$$\sigma^2 = \frac{\sigma_0^2\sigma_1^2}{n\sigma_0^2 + \sigma_1^2}.$$

Observație Să observăm că μ este o combinație convexă între $\bar{x}$ ($= T(x)$) și τ deci se află între aceste valori.

Dacă σ_0, dispersia mediei necunoscute θ, este mai mare ca σ_1, atunci $\mu \approx \bar{x}$. În acest caz cunoașterea mediei à priorice τ este de importanță redusă. Dacă, dimpotrivă $\sigma_0 = 0$ atunci $\mu = \tau$ indiferent de observațiile efectuate.

Raportul $a = \frac{\sigma_1^2}{\sigma_0^2}$ măsoară încrederea à priori că τ este o estimare corectă a mediei.

Dacă $a < \infty$ atunci $\lim_{n\to\infty} \mu = \lim_{n\to\infty} \bar{x}$.

În concluzie, dacă dispersia inițială este mică, media estimată tinde să rămână în apropierea mediei inițiale τ chiar dacă media empirică $\bar{x}$ diferă considerabil de aceasta. Dacă raportul a este mic, atunci media și dispersia à priori au doar o influență redusă asupra estimării parametrilor care sunt determinați aproape exclusiv din datele empirice.

În lumina teoremei de mai sus estimatorul Bayes a mediei unei variabile aleatoare $N(\mu, \sigma^2)$ este, dacă $T(\mathbf{x}) = t$

$$\delta(t) \equiv \hat{\theta}_B = \left(\frac{n}{\sigma_1^2}t + \frac{1}{\sigma_0^2}\tau\right)\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma_1^2}\right)^{-1}$$

Analog pentru cazul multidimensional

$$\hat{\boldsymbol{\theta}}_B = \mathbf{S}\left(\mathbf{S} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\mathbf{t} + \frac{1}{n}\boldsymbol{\Sigma}\left(\mathbf{S} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\mathbf{m}$$

Fie $\mathbf{x} = (x_1, \ldots, x_n)$ o selecție bernoulliană din populațiile Π_1 și Π_2.

Dacă $X \in \Pi_i$ atunci densitatea de probabilitate este $f_i(x|\theta)$, $\theta \in \theta_i$ și densitatea à priorică este $h_i(\theta)$, $i = 1,2$. Dându-se probabilitățile à priorice ale populațiilor $\{\Pi_1, \Pi_2\}$, fie acestea $q_i, i = 1,n$, teorema Bayes calculează probabilitățile à posteriori

$$P(\Pi_i|\mathbf{x}) = \frac{m_i(\mathbf{x})q_i}{m_1(\mathbf{x})q_1 + m_2(\mathbf{x})q_2} \quad i = 1,2$$

unde $m_i(\mathbf{x}) = \int_{\Theta_i} f_i(\mathbf{x}|\theta) h_i(\theta) d\theta$ este densitatea de probabilitate marginală a lui $\mathbf{x}$ condiționat de faptul că provine din Π_i.

Este evident că o procedură bayesiană de discriminare este

- $\mathbf{x} \in \Pi_1$ *dacă* $\frac{P(\Pi_1|\mathbf{x})}{P(\Pi_2|\mathbf{x})} = B_{12}(\mathbf{x}) \frac{q_1}{q_2} \geq 1$;
- $\mathbf{x} \in \Pi_2$ *în caz contrar*,

unde $B_{12}(\mathbf{x}) = m_1(\mathbf{x}) / m_2(\mathbf{x})$ este cunoscut ca *factorul Bayes al populației* Π_1 *versus* Π_2.

3.3 SEGMENTARE

Metodele de segmentare urmăresc să rezolve problemele de discriminare şi de regresie împărțind progresiv eşantionul într-un *arbore de decizie binară*.

Pionieri în acest domeniu sunt considerați a fi Sonquist&Morgan (1964) şi Morgan&Messenger (1973) cu metoda AID (*Automatic Interaction Detection*). Au urmat numeroase contribuții dar lucrările lui Breiman et al. (1984) cu metoda CART (*Classification and Regression Tree*) au îmbogățit domeniul şi au resuscitat interesul pentru segmentare.

Proprietățile metodei de segmentare pot fi sintetizate astfel:

- **avantajele metodei**:
  - lizibilitatea regulilor de afectare, interpretarea rezultatelor fiind directă şi intuitivă;
  - tehnica este neparametrică şi impune puține restricții asupra variabilelor. Se pot utiliza concomitent ca variabile explicative, variabile continue, ordinale şi nominale fără un codaj prealabil. În plus metoda oferă din oficiu selecția variabilelor ținând cont de eventualele interacții;
  - tehnica este robustă față de valorile eronate sau față de valorile aberante şi gestionează valorile lipsă atât la construcția arborelui şi la estimarea erorii sale de misclasare cât şi în cazul unui nou subiect;
  - metoda foloseşte acelaşi principiu, tehnici, algoritm atât pentru a analiza o variabilă discretă (analiza discriminantă) cât şi una continuă (analiza de regresie);
- **dezavantajele metodei**:
  - regulile de afectare, pot apărea uneori "aberante" şi prea sensibile la perturbații uşoare ale datelor ;

- lipsa unei funcții de afectare globală (ce utilizează toate variabilele) ce privează utilizatorul de o reprezentare geometrică.

3.3.1 FORMULAREA PROBLEMEI, PRINCIPIU ŞI VOCABULAR

Ne poziționăm în cadrul analizei discriminante (o variabilă "privilegiată" y discretă – cu k modalități– este "explicată" de variabilele $x_1,\ldots,x_p$).

Metoda de segmentare constă în a calcula mai întâi variabile x_j care explică cel mai bine variabila y. Această variabilă defineşte o primă împărțire a eşantionului în două submulțimi, numite *segmente*. Se reiterează procedeul în interiorul fiecărui segment căutându-se a doua cea mai bună variabilă şi aşa mai departe.

Se construieşte astfel un *arbore de decizie binară* prin împărțirea succesivă a eşantionului în câte două submulțimi. Distingem astfel :

- *segmentele intermediare* sau *nodurile* din care pornesc câte 2 segmente descendente;
- *segmentele terminale* care nu mai sunt împărțite;
- *ramurile* unui segment care conține toate segmentele descendente din t, fără t ;
- *arborele binar complet* notat A_{max} ;
- un *sub-arbore* A obținut din A_{max} , prin elagarea uneia sau mai multor ramuri.

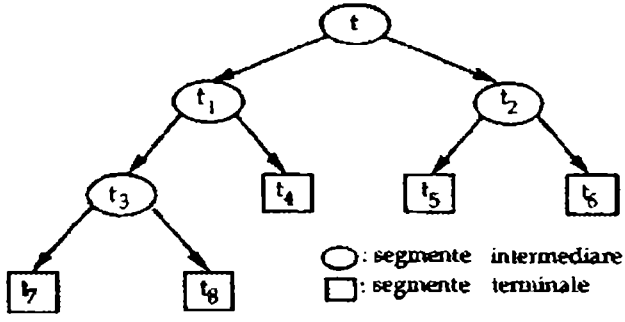


Figura 3.3-1 Arbore de decizie binară

3.3.1.1 Construcția arborelui de decizie binară

Ideea de bază constă în efectuarea diviziunii unui nod astfel încât cele două segmente descendente să fie mai omogene decât nodul părinte iar ele să fie cât mai diferite între ele față de variabila y.

Aşadar fazele de construire ale arborelui sunt :

a) stabilirea pentru fiecare nod a mulțimii diviziunilor admisibile;

b) definirea unui criteriu de selecționare "a celei mai bune" diviziuni a unui nod;
c) definirea unei reguli care să permită declararea unui nod ca terminal sau intermediar;
d) afectarea fiecărui nod terminal unei clase;
e) estimarea riscului de misclasare.

Variabilele explicative pot fi de natură oarecare; să le considerăm pentru moment variabile continue.

1) La început există un singur segment conținând toți indivizii;

2) Sunt examinate secvențial toate variabilele explicative. Pentru o variabilă dată x_j sunt trecute în revistă toate diviziunile posibile $x_j < \alpha$, cu α o valoare oarecare din suportul lui x_j. Fiecare diviziune împarte eşantionul în segmente descendente: segmentul din stânga t_s conține indivizii ce verifică condiția $x_j \leq \alpha$, iar segmentul din dreapta t_d conține indivizii ce verifică condiția $x_j > \alpha$. Din toate diviziunile d_j^m admisibile (se numeşte diviziune admisibilă o diviziune posibilă cu segmentele descendente nevide ale lui x_j, unde m reprezintă a m-diviziune (sau a m valoare ordonată a variabilei x_j din eşantion), procedura selecționează "pe cea mai bună", notată d_j^*, în sensul unui criteriu ce urmează a fi precizat.

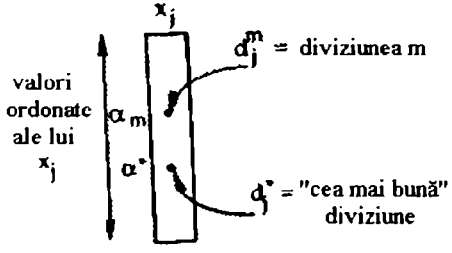


Figura 3.3-2 Diviziuni posibile pentru variabila x_j

Se obține astfel, pentru fiecare din cele p variabile, diviziunea optimă "locală" şi se va reține, în final, din cele p diviziuni, pe cea notată cu d^* care va furniza cele două segmente "cele mai caracteristice" vis-à-vis de y.

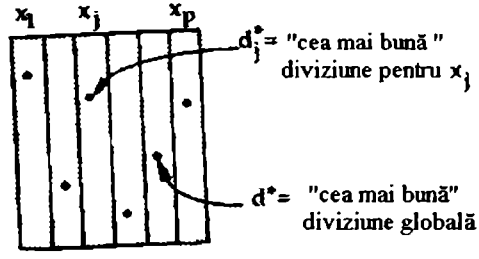


Figura 3.3-3 Cea mai bună diviziune pentru toate variabilele

3) Se aplică iterativ pasul 2 fiecărui segment descendent obținut.
4) Procedeul se opreşte când toate segmentele sunt declarate terminale:
- fie că nu mai necesită vreo diviziune;
- fie pentru că "talia lor" (numărul de indivizi afectați nodului) este inferior unui efectiv fixat (în practică acesta se alege între 1 şi 5).

Afectarea unui individ nou se face prin "coborârea" lui pe ramurile arborelui.

Dacă printre variabilele explicative se numără şi variabile discrete atunci diviziunile posibile pot fi:

- una singură, dacă variabila explicativă x_j este binară. În această situație segmentul t_s va conține toate observațiile pentru care $x_j = 1$, iar segmentul t_d toate observațiile pentru care $x_j = 2$ (am presupus că valorile luate de variabila binară sunt 1 şi 2);
- $k-1$, dacă variabila explicativă x_j are k modalități ordonate ($1,2,\ldots,k$, cu $k>2$). Într-adevăr, prima diviziune va dirija toate observațiile pentru care $x_j = 1$ spre segmentul t_s şi toate observațiile pentru care $x_j \in \{2,3,\ldots,k\}$ spre segmentul t_d. A doua diviziune va dirija toate observațiile pentru care $x_j \in \{1,2\}$ spre segmentul t_s şi toate observațiile pentru care $x_j \in \{3,\ldots,k\}$ spre segmentul t_d. A $k-1$ diviziune va dirija toate observațiile pentru care $x_j \in \{1,2,\ldots,k-1\}$ spre segmentul t_s şi toate observațiile pentru care $x_j = k$ spre segmentul t_d;
- $2^{k-1}-1$, dacă variabila explicativă x_j are k modalități neordonate.

Pentru selectarea celei mai bune diviziuni a unui nod se pot utiliza mai multe criterii; Breiman et al. (1984) recomandă utilizarea criteriilor bazate pe noțiunea de impuritate.

Impuritatea unui segment (nod) a, notată $i(a)$, este o funcție nenegativă de $P[1/a],\ldots,P[k/a]$ (probabilitatea condiționată de apartenență la un grup $G_r, r=\overline{1,k}$, a mulțimii observațiilor din nodul a) care verifică condițiile următoare:

i) $i(a)$ este maximă pentru $P[1/a]=\frac{1}{k}$, $(\forall)r=\overline{1,k}$ (impuritatea unui nod a e maximală când, pentru acest nod, probabilitățile de apartenență la diferite grupe sunt egale între ele;

ii) $i(a)$ este nulă pentru $P[r/a]=1$ și $P[s/a]=0$, $(\forall)s\neq r$ și $r,s=\overline{1,k}$ (impuritatea este nulă dacă nodul conține observații aparținând unui singur grup);

iii) $i(a)$ este o funcție simetrică de probabilități $P[r/a]$, $r=\overline{1,k}$.

Funcțiile de impuritate cele mai folosite sunt :

$$i(a)=-\sum_{r=1}^{k}P[r/a]\ln(P[r/a])$$

și

$$i(a)=\sum_{r\neq s}P[r/a]\,P[s/a].$$

Prima funcție e derivată din noțiunea de informație sau de *entropie Shannon*; a doua, numită *indicile de diversitate Gini*, a fost propusă de Goodman & Kruskal (1954).

Fie o diviziune d care împarte nodul a în t_s și t_d cu probabilitățile $p_s\equiv P[t_s/a]=\frac{P(t_s)}{P(a)}$, respectiv $p_d=\frac{P(t_d)}{P(a)}$.

Se definește $\Delta i(d,a)=i(a)-p_s i(t_s)-p_d i(t_d)$ reducerea impurității nodului a datorat diviziunii d.

Lema 3.3-1 *Orice diviziune d a unui nod a duce la o reducere pozitivă sau nulă a impurității, adică:*

$$\Delta i(d,a)\geq 0,$$

egalitatea fiind obținută $\Leftrightarrow P(r/t_s)=P(r/t_d)=P(r/a)$, $(\forall)r=\overline{1,k}$.

Demonstrație.

$$p_s i(t_s)+p_d i(t_d)=p_s f(P[1/t_s],\ldots,P[k/t_s])+p_d f(P[1/t_d],\ldots,P[k/t_d])\leq$$
$$\leq f(p_s P[1/t_s]+p_d P[1/t_d],\ldots,p_s P[k/t_s]+p_d P[k/t_d])$$

(1)

căci $i(a)$ este strict concavă.

Pe de altă parte

$$p_s P[r/t_s] + p_d P[r/t_d] = P[r/a] \quad (\forall) r = \overline{1,k}$$

deci

$$f\left(p_s P[1/t_s] + p_d P[1/t_d], \ldots, p_s P[k/t_s] + p_d P[k/t_d]\right) = f\left(P[1/a], \ldots, P[k/a]\right) = i(a)$$

(2).

Aşadar, din (1) şi (2)

$$\Delta i(d,a) \geq 0$$

(3)

Dacă în (2) $P[r/t_s] = P[r/t_d]$, $(\forall) r = \overline{1,2}$ atunci (1) devine egalitate şi deci şi (3) devine egalitate.

□

Cele două funcţii de impuritate de mai sus sunt strict concave, deci criteriile de diviziune bazate pe cele două funcţii conduc întotdeauna la reducerea pozitivă a impurităţii.

Cea mai bună diviziune d_j^* este aceea pentru care reducerea impurităţii este maximă, adică:

$$d_j^* = \underset{m \in d_j}{\operatorname{argmax}} \Delta i\left(d_j^m, t\right)$$

unde d_j este mulţimea diviziunilor admisibile ale variabilei x_j.

Pe mulţimea p a variabilelor, diviziunea nodului t este efectuată cu ajutorul variabilei care asigură

$$d^* = \max_{1 \leq j \leq p} \left\{d_j^*\right\}.$$

3.3.1.2 Reguli de afectare

La fiecare etapă de construire a lui $A_{\max}$ este posibil să afectăm toate nodurile terminale a ale arborelui curent A uneia din cele k grupe.

Fiecărei erori de clasare i se asociază un preţ $\gamma(s/r)$, $(s = 1,\ldots,k;\ \ r = 1,\ldots,k)$ de misclasare. Costul misclasării este atunci $\sum_r \gamma(s/r) p(r/a)$ şi nodul va fi asignat acelei clase pentru care

$$s^* = \underset{1 \leq s \leq k}{\arg\min} \sum_r \gamma(s/r) p(r/a).$$

Dacă minimul este atins pentru cel puţin două clase atunci nodul este afectat arbitrar uneia din clase.

Următoarea proprietate este foarte utilă în practică:

Lema 3.3-2 *Dacă* $\gamma(s/r)=1$, $(\forall)s\neq r$ *şi* $\gamma(s/s)=0$, $(\forall)s$ *atunci nodul va fi asignat clasei cu cei mai mulţi reprezentanţi în el.*

Demonstraţie. Într-adevăr, fie s_0 acea clasă. Să observăm că

$$p(r/a)=\frac{n_r}{n_a}$$

cu n_r - numărul de indivizi din clasa r aflaţi în nodul a ;

n_a - numărul de indivizi din nodul a.

Conform ipotezei

$$\sum_{\substack{r=1\\ r\neq s_0}}^{k} n_r < \sum_{\substack{r=1\\ r\neq j}}^{k} n_r \qquad \begin{matrix} j=\overline{1,k} \\ j\neq s_0 \end{matrix},$$

adică un sistem de $k-1$ inegalităţi cu acelaşi membru stâng.

Reducând termenii asemenea se obţin $k-1$ inegalităţi de forma

$$n_r < n_{s_0} \qquad \begin{matrix} r=\overline{1,k} \\ r\neq s_0 \end{matrix},$$

adică n_{s_0} este maximal.

□

Costul misc lasării unei observaţii aparţinând nodului a, notat $c(a)$, este egal deci

$$c(a)=\min_s \sum_r \gamma(s/r)\,p(r/a).$$

Costul misclasării datorată nodului a, notat $C(a)$, este egal cu

$$C(a)=c(a)\,p(a)$$

unde $p(a)$ probabilitatea nodului a.

Riscul erorii de afectare datorat arborelui A sau *rata erorii aparente de clasare* datorată arborelui A, notată *TEA* (*taux d'erreur apparent*) este

$$TEA(A)=\sum_{a\in\tilde{A}} C(a)=\sum_s \sum_{a\in\tilde{A}(s)} \sum_r \gamma(s/r)\,p(r/a)\,\pi_r$$

$$=\sum_s \sum_r \gamma(s/r)\frac{n_{sr}}{n_r}\cdot\frac{n_r}{n}=\sum_s \sum_r \gamma(s/r)\frac{n_{sr}}{n}$$

cu $\tilde{A}$ - mulţimea nodurilor terminale ale lui A ;

$\tilde{A}(s)$ - mulţimea nodurilor terminale ale lui A asignate clasei s ;

π_r - probabilitatea a priori ca un nod să provină din clasa r ;

n_{sr} - numărul de indivizi din clasa r clasați în clasa $s\,(s \neq r)$.

3.3.2 ELAGAREA ARBORELUI MAXIMAL

O ramură A^a a arborelui $A_{\max}$, având ca rădăcină nodul intermediar a este constituită din toți descendenții lui a. *Elagarea ramurii* A^a *din arborele* $A_{\max}$ înseamnă îndepărtarea din $A_{\max}$ a tuturor descendenților lui a excepție el însuși. Se notează cu $A_{\max} - A^a$ arborele astfel obținut. Dacă arborele A este obținut din $A_{\max}$ prin elagări succesive atunci A este un subarbore a lui $A_{\max}$.

Prin „cel mai bun" subarbore se înțelege acel arbore care conține minimul de segmente terminale cu *TEA* minimă și furnizând o estimație corectă a erorii teoretice de clasare.

Metoda propusă de Breiman et al. pentru obținerea celui mai bun subarbore se bazează pe utilizarea unui eșantion-test și prezintă un dublu avantaj :

- determină "cel mai bun" subarbore fără să utilizeze teste statistice pentru definirea unei reguli de oprire a diviziunii ;
- determină o estimație precisă a erorii teoretice de clasare.

3.3.2.1 Procedura de selecție a subarborelui optimal

Se împarte eșantionul de bază în două părți –un *eșantion de învățare* (de exemplu 2/3 din eșantionul de bază) și un *eșantion de testare* (restul de 1/3 din eșantionul de bază).

Pornind de la eșantionul de învățare se construiește arborele $A_{\max}$.

Operația de elagare a arborelui $A_{\max}$ constă în construirea unui șir optimal de subarbori incluși $\{A_H, \ldots, A_k, \ldots, A_1\}$ cu $A_H = A_{\max}$, A_h este subarborele cu h segmente terminale, iar A_1 este eșantionul total. Fiecare subarbore A_h din acest șir este optimal în sensul că *eroarea aparentă* (*EA*) a subarborelui este minimală printre toți subarborii având același număr de segmente terminale, adică

$$EA(A_h) = \min_{A \in S_h} EA(A)$$

cu $S_h = \{\text{subarborii lui } A_{\max} \text{ cu } h \text{ segmente terminale}\}$.

Se selectează din șirul de arbori optimali subarborele A^* care prezintă eroarea teoretică (*ET*) minimă, adică

$$ET(A^*) = \min_{1 \le h \le H} ET(A_h).$$

Eroarea teoretică se estimează după formula

$$\widehat{ET}(A) = \sum_{t \in A} \tilde{R}_t$$

cu $\hat{R}_t = \frac{\tilde{n}_t}{\tilde{n}} \times \tilde{s}_t^2$, unde $\tilde{n}$ este volumul eşantionului test, $\tilde{n}_t$ este numărul de indivizi din eşantionul test aparţinând segmentului t, iar $\tilde{s}_t^2$ este dispersia de selecţie a variabilei y în interiorul segmentului t, adică

$$\tilde{s}_t^2 = \frac{1}{\tilde{n}_t} \sum_{i=1}^{\text{card}(t)} \left(y_i - \tilde{\bar{y}}_t\right)^2,$$

unde $\tilde{\bar{y}}_t$ este media de selecţie în interiorul segmentului t.

3.3.2.2 Diviziuni echi-reductive şi echi-divizante

Cea mai bună diviziune d^* a unui nod este cea care asigură cea mai mare reducere a dispersiei reziduale sau a impurităţii prin trecerea de la acel nod la segmentele descendente. Această definiţie este foarte strictă, putând exista diviziuni aproximativ la fel de bune dar foarte importante la nivelul interpretării. Se pot defini astfel alte două tipuri de diviziuni:

- *diviziunile echi-reductive* care asigură după diviziunea d^* cele mai mari reduceri ale impurităţii sau cele mai mici dispersii reziduale. Ele permit alegerea "celei mai bune" variabile explicative;
- *diviziunile echi-divizante* care furnizează repartizările cele mai apropiate de cea mai bună diviziune d^*. Ele permit clasarea indivizilor cu valori lipsă tocmai la variabila(lele) ce defineşte(sc) diviziunea.

Diviziunile echi-reductive se obţin înlocuind variabila x^* ce dă diviziunea optimă d^* cu variabila $x_i \left(x_i \neq x^*\right)$ ce dă diviziunea d_i^* cu reducerea impurităţii cea mai bună după d^*; este, în alţi termeni, a doua cea mai bună diviziune a nodului t. Prin extensie se poate defini a 3-a, a 4-a..., diviziune echi-reductivă.

Diviziunile echi-divizante (numite uneori supleante) permit clasarea unui individ nou ce are ca dată lipsă tocmai măsurătoarea ce defineşte diviziunea. În acest caz se caută variabila care înlocuieşte cel mai bine variabila care divizează nodul în sensul asigurării unei separări a indivizilor cît mai apropiate de separarea realizată de d^*. Analog se poate defini a 2-a, a 3-a,..., diviziune echi-divizantă.

BIBLIOGRAFIE

[1]. ANDERBERG M.R. (1973) *Cluster Analysis for Applications.* Academic Press, New York.

[2]. ANDERSON T.W. (1958) *An Introduction to Multivariate Statistical Analysis.* J. Wiley, New York.

[3]. ANDERSON T.W. (1963) Asymptotic theory for principal component analysis: the non-normal case. *Australian J. of Statist.*, **19**, p.206-212.

[4]. BENZÉCRI J.P. (1973) *L'Analyse des Données.* Tome 1: *La Taxinomie.* Tome 2: *L'Analyse des Correspondances* (2^{de}. éd. 1976). Dunod, Paris.

[5]. BREIMAN L., FRIEDMAN, J.H., OHLSEN R.A., STONE C.J. (1984) *Classification and Regression Trees.* Wadsworth, Belmont

[6]. BURT C. (1950) The factorial analysis of qualitative data. *British J. of Statist. Psychol.* **3**, 3, p.166-185

[7]. COX D. R. (1972) *Analyse des données binaires.* Dunod, Paris.

[8]. DAUDIN J.J., DUBY C., TRÉCOURT P. (1988) Stability of principal components studied by the bootstrap method. *Statistics,* **19**, p 241-258.

[9]. DAZY F., LE BARZIC J.F. (1996) *L'analyse des données évolutives. Méthodes et Applications.* Ed. Technip, Paris.

[10]. DEMIDOVITCH B., MARON I. (1973) *Eléments de calcul numérique.* Ed. Mir, Moscou.

[11]. DEMPSTER A.P. (1971) An overview of multivariate data analysis. *J. Mult. Analysis,* **1**, p 316-346.

[12]. DIDAY E. (1971) La méthode des nuées dynamiques. *Revue Statist. Appl.,* **19**, 2, p 19-34.

[13]. DODGE Y. (ed.) (1987) *Statsitical data Analysis Based on the L_1 -Norm and Related Methodes.* North Holland, Amsterdam.

[14]. DOMENGES D., VOLLE M. (1979) Analyse factorielle sphérique: une exploration. *Annales de l'INSEE*, no **35**.

[15]. DUDA R.O., HART P.E. (1973) *Pattern Classification and Scène Analysis.* J. Wiley, New

[16]. DUMITRESCU D. (1999) *Principiile matematice ale teoriei clasificării.* Ed. Academiei Române, Bucureşti

[17]. ENĂCHESCU C. (1999) *Aplicaţii ale reţelelor neuronale în teoria statistică a învăţării.* Ed. Sigma, Bucureşti

[18]. ENĂCHESCU C., ENĂCHESCU D. (2000) *Some simple rules for interpreting outputs of principal compnents and correspondence analysis.* Analele Univ. Buc., Informatică, **XLIX**, p.3-8

[19]. FALISSARD B. (1995) Déploiement d'une matrice de corrélation sur la sphère unité de $\mathbf{R}^3$. *Revue de Statist. Appl.*, **43**(2), p.35-48.

[20]. FISHER R. A. (1936) The use of multiple measurements in taxonomic problems. *Ann. of Eugenics,* **7**, p 179-188.

[21]. GIFI A. (1990) *Non Linear Multivariate Analysis.* J.Wiley, Chichester.

[22]. GOLDSTEIN M., DILLON W. R. (1978) *Discrète Discriminant Analysis*, J. Wiley, Chichester

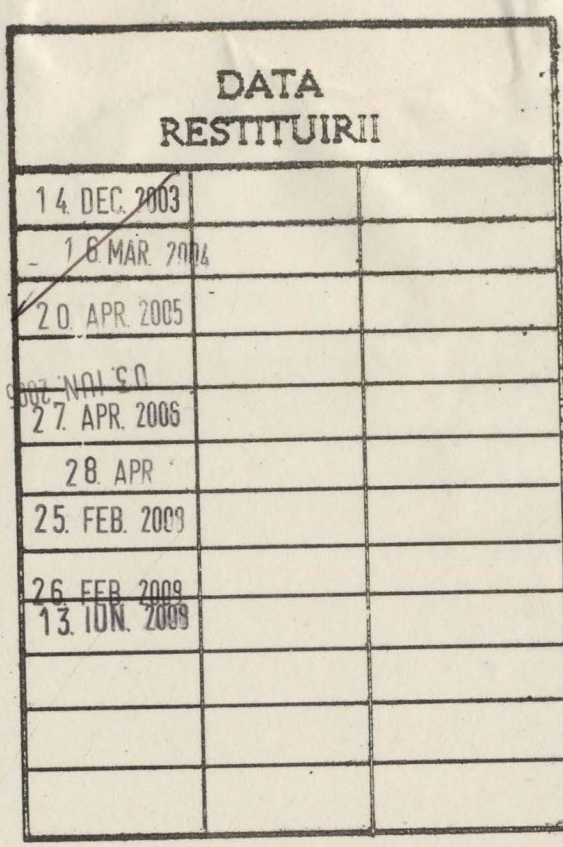
[23]. GOODMAN L.A., KRUSKAL W.H. (1954) Measures of association for cross classification. *J. of Amer. Statist. Assoc.,* **49**, p 732-764.

[24]. GUTTMAN L. (1941) - The quantification of a class of attributes: a theory and method of a scale constructuion. In: *The prédiction of personal adjustment* (Horst P., ed.) p 251 -264, SSCR New York.

[25]. HAND D. J. (1981) *Discrimination and Classification.* J. Wiley, New York

[26]. HARMAN H.H. (1967) *Modern Factor Analysis* (2nd ed.). Chicago University Press, Chicago.

[27]. HAYASHI C.(1956) Theory and examples of quantification. (II) *Proc. of the Institute of Statist. Math.*, **4** (2), p 19-30.
[28]. JAMBU M. (1991) *Exploration statistique et informatique des données.* Dunod, Paris
[29]. KAZMIERCZAK J.B. (1985) Analyse logarithmique: deux examples d'application. *Revue de Satist. Appl.*, **33**(1), p.13-24.
[30]. LANCE G. N., WILLIAMS W. T. (1967) A general theory of classification sorting strategies. *Computer J.*, **9**, p 373-380.
[31]. LEBART L. (1975) L'orientation du dépouillement de certaines enquêtes par l'analyse des correspondances multiples. *Consommation*, **2**, p 73-96. Dunod
[32]. LEBART L., MORINEAU A., PIRON M. (1995) *Statistique exploratoire multidimensionnelle.* Dunod, Paris.
[33]. Macqueen J. B. (1967) - Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability* (5th), Berkeley, **1**, p 281-297, Univ. of Calif. Press, Berkeley
[34]. MAHALANOBIS P.C. (1936) On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India,* **12**, p 49-55.
[35]. MALINVAUD E. (1987) Data analysis in applied socio-économie statistics with spécial considération of correspondence analysis. *Marketing Science Conférence Proceedings,* HEC-ISA, Jouy en Josas.
[36]. MEYER R. (1994) An eigenvector algorithm to fit L_p-distances matrices. In: *New Approches in Classification and Data Analysis*, Diday E. et al. (eds.), Springer Verlag, Berlin, p.502-509.
[37]. MORGAN J. M., MESSENGER R. C. (1973) *THAID : a sequential search program for the analysis of nominal scale dépendent variables.* Institute for Social Research, University of Michigan, Ann Arbor
[38]. MORINEAU A. (1984) Note sur la caractérisation statistique d'une classe et les valeurs-tests, *Bull. Techn. du Centre de Statist. et d'Infor. Appl,* **2**, p 20-27
[39]. NISHISATO S.(1980) *Analysis of Categorical Data. Dual Scaling and its Application.* Univ. of Toronto Press.
[40]. RAO C.R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhya*, serie A, **26**, p.329-357
[41]. RIPLEY B. D. (1994) Neural nerworks and related methods of classification. *J. R. Statist. Soc. B,* **56**, n°3, p 409-456
[42]. SAPORTA G. (1990) *Probabilités, Analyse des Donnés et Statistique.* Ed. Technip, Paris
[43]. SOKAL R. R., SNEATH P. H. A. (1963) *Principles of Numerical Taxonomy.* Freeman and co., San-Francisco.
[44]. SONQUIST J. A. AND MORGAN J. N. (1964) *The Détection of Interaction Effects.* Institute for Social Research, University of Michigan, Ann Arbor.
[45]. TUKEY J. W. (1977) *Exploratory Data Analysis.* Addison Wesley, Reading, Mass
[46]. VĂDUVA I. (1970) *Analiză dispersională.* Ed. Tehnică, Bucureşti
[47]. van RIJCKEVORSEL J. (1987) *The application of fuzzy coding and horseshoes in multiple correspondances analysis.* DSWO Press, Leiden.
[48]. WARD J.H. (1963) Hierarchical grouping to optimize an objective fonction. *J. of Amer.Statist. Assoc.*, **58**, p 236-244.
[49]. WISHART D. (1969) Mode analysis: a generalization of nearest neighbour which reduces chaining effects. *Numerical Taxonomy* (A.J. Cole éd.) p 282-311, Academic Press, London
[50]. WONG M.A. (1982) A hybrid clustering method for identifying high density clusters. *J of Amer. Statist. Assoc.*, **77**, p 841-847.



95.000 lei

ISBN 973-575-814-8

Lei 95000