METODE EXPLORATORII MULTIDIMENSIONALE

Cornel Lepădatu

cornel_lepadatu@biblacad.ro

Academia Română Bucureşti

Biblioteca Academiei Române

Rezumat: Explorarea datelor este un ansamblu de metode destinate descrierii şi analizei datelor multidimensionale şi utilizate în orice domeniu, atunci când datele sunt mult prea multe pentru a mai putea fi înţelese de o minte omenească. Unele dintre metode, ajută la evidenţierea relaţiilor care pot exista între diferite date şi elaborează informaţii statistice care permit o descriere mai succintă a informaţiei conţinute în aceste date. Altele, permit regrupări ale datelor în scopul de a face să apară clar ceea ce le face omogene şi astfel de a le înţelege şi de a le defini mai bine.

Metodele exploratorii multidimensionale sunt metode descriptive, în cea mai mare parte geometrice, al căror instrument matematic major este algebra matricială şi care se exprimă fără să presupună à priori un model probabilist. Aceste metode permit, în special, prelucrarea şi sinteza informaţiei din tabelele de date de mari dimensiuni pe baza estimării corelaţiilor dintre variabilele studiate, instrumentele statistice utilizate fiind matricea corelaţiilor sau matricea de varianţă-covarianţă.

Un demers exploratoriu îi permite prospectorului de date să abordeze unul dintre principalele obiective ale „data mining" şi anume explorarea multidimensională a datelor sau reducerea de dimensiune: reprezentarea grafică, deducerea unei submulţimi de variabile reprezentative sau a unei mulţimi de componente prealabile pentru alte metode.

Din anii 1980 capacitatea de a stoca informaţii s-a dublat aproximativ la fiecare 40 de luni [10]. Începând cu 2012 au fost create [11], în fiecare zi, 2.5 quintilioane ($2{,}5 \times 10^{18}$) octeţi de date, iar limitarea la ordinul exabyţilor, privind dimensiunile seturilor de date procesabile într-un timp rezonabil [7, 16], constituie deja un subiect de preocupare sistematică a oamenilor de ştiinţă pentru domenii precum meteorologia, genomica, connectomica, simularea fenomenelor fizice complexe, cercetările biologice şi de mediu şi chiar căutarea pe internet, finanţele şi informatica decizională.

Cuvinte cheie: analiza canonică, analiza corespondenţelor multiple, analiza corespondenţelor simple, analiza factorială discriminantă, analiza în componente principale.

Abstract: Data exploring is a set of methods for describing and analyzing multidimensional data used in any area where data are too numerous to be comprehended by a human mind. Some of the methods are helpful in revealing relationships that may exist between different data and in developing statistical information to enable a succinct description of the information contained. Others allow data regrouping to disclose their homogenous part, thus permitting their better understanding and defining.

Multidimensional exploratory methods are descriptive, mostly geometric, based on a major mathematical tool, the matrix algebra, expressing, without assuming a priori, a probabilistic model. These methods allow mainly information processing and a synthesis of large tables of data by estimating the correlations between the variables studied, the statistical tools used being the correlation matrix or the variance-covariance matrix.

An exploratory approach allows data prospector to address one of the main objectives of data mining, that is exploring multidimensional data and dimension reduction: graphical representation, deduction of representative subsets of variables or a set of components preceding other methods.

The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s [10]; as of 2012, every day 2.5 quintillion (2.5×10^{18}) bytes of data were created [11]. As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data [7,16]. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics.

Keywords: Canonical Correlation Analysis, Multiple Correspondence Analysis, Correspondence Analysis, Canonical Discriminant Analysis, Principal Component Analysis.

1. Introducere

Explorarea datelor este un ansamblu de metode care se ocupă cu descrierea şi analiza datelor multidimensionale. Unele dintre metode, ajută la evidenţierea relaţiilor care pot exista între diferite date şi elaborează informaţii statistice care permit o descriere mai succintă a informaţiei conţinute în aceste date. Altele, permit regrupări ale datelor în scopul de a face să apară clar ceea ce le face omogene şi astfel de a le înţelege şi de a le defini mai bine.

Explorarea datelor permite prelucrarea unui număr mare de date şi identificarea celor mai

interesante aspecte ale structurii acestora, computerele fiind acelea care au făcut aceste metode operaţionale şi care le-au permis o utilizare foarte extinsă. Succesul din ultimii ani al acestora se datorează în mare măsură reprezentărilor grafice oferite. Aceste reprezentări pot evidenţia relaţii dificil de sesizat de o analiză directă a datelor dar, mai important şi în contrast cu metodele statistice clasice, aceste reprezentări nu sunt legate de nicio ipoteză privind legile fenomenelor analizate.

Explorarea datelor se bazează pe un set de metode descriptive, în cea mai mare parte geometrice, al căror instrument matematic major este algebra matricială şi care se exprimă fără să presupună à priori un model probabilist. Aceste metode permit, în special, prelucrarea şi sinteza informaţiei din tabelele de date de mari dimensiuni pe baza estimării corelaţiilor dintre variabilele studiate, instrumentele statistice utilizate fiind matricea corelaţiilor sau matricea de varianţă-covarianţă.

Fundamentele matematice ale explorării datelor au început să se dezvolte la începutul secolului al XX-lea dar tehnici de bază privind analiza datelor erau deja cunoscute cu mult înainte. Tabelele de contingenţă, de exemplu, sunt prezente [4] încă din 1588, când Alvarez Paz Salas descrie „Invincibila Armada" sub forma unui tabel în care rândurile reprezintă flote de nave, iar coloanele diverse caracteristici ale navelor cum ar fi tonajul, numărul de soldaţi, etc. sau din 1696, când Nicolas Lamoignon Basville, intendent al regelui Ludovic al XIV-lea, enumeră şi caracterizează mânăstiri şi biserici din regiunea Languedoc. Printre fondatorii metodelor moderne de analiză a datelor se regăsesc Jean-Paul Benzécri, Louis Guttman, Chikio Hayashi, Douglas Carroll şi R.N. Shepard [2].

Într-un proces de explorare a datelor şi descoperire a cunoştinţelor („data mining") un prim demers, inevitabil, constă în efectuarea unei explorări a acestor date: alura distribuţiilor, prezenţa datelor atipice, corelaţii şi coerenţă, transformări eventuale ale datelor. Demersul descriptiv şi exploratoriu permite realizarea de rezumate şi grafice mai mult sau mai puţin elaborate, descrierea mulţimilor de date şi stabilirea de relaţii între variabile, fără a acorda un rol privilegiat vreunei variabile şi care, folosite în mod adecvat, se pot dovedi extrem de utile pentru numeroase probleme şi situaţii din domeniul decizional [5, 6, 12]. Concluziile obţinute privesc doar datele studiate, fără a fi generalizate la o populaţie mai largă. Demersul exploratoriu se sprijină, în mod esenţial, pe noţiuni elementare (medie şi dispersie), pe reprezentări grafice şi pe tehnici descriptive multidimensionale. Metodele exploratorii determină subspaţii de reprezentare (sau factoriale), de dimensiuni mici, care aproximează cel mai bine norii de puncte-indivizi sau de puncte-variabile, astfel încât vecinătăţile măsurate în aceste spaţii să reflecte cât mai exact proximităţile reale.

Demesul exploratoriu îi permite deci prospectorului de date să abordeze unul dintre principalele obiective ale „data mining" şi anume explorarea multidimensională a datelor sau reducerea de dimensiune: reprezentarea grafică, deducerea unei submulţimi de variabile reprezentative sau a unei mulţimi de componente prealabile pentru alte metode. Cele mai frecvent utilizate metode, în funcţie de tipurile variabilelor, sunt [1, 3, 8, 13, 17]: *analiza în componente principale* (*ACP*), *analiza factorială discriminantă* (*AFD*), *analiza corespondenţelor simple* (*ACS*), *analiza corespondenţelor multiple* (*ACM*) şi *analiza canonică* (*AC*).

2. Elemente preliminare

Fie p variabile observate pe n indivizi. Mulţimii de observaţii disponibile i se asociază matricea de valori $\mathbf{X} = \{ (x_{ij}) \mid i = 1 \div n, j = 1 \div p \} \in \mathcal{M}_{n \times p}(\Re)$, x_{ij} reprezentând valoarea variabilei j măsurată pe individul i. Fiecărui individ i se atribuie o pondere ρ_i, $i = 1 \div n$, ($\rho_i > 0$, $\Sigma^n_{i=1} \rho_i = 1$).

Matricea diagonală $\mathbf{D} = \text{diag}(\rho_1, \ldots, \rho_n) \in \mathcal{M}_{n \times n}(\Re)$, se numeşte *matrice de ponderi*, pentru cazul indivizilor echiponderaţi $\mathbf{D} = (1/n)\, \mathbf{I}_n$ unde $\mathbf{I}_n$ este matricea identitate.

O *variabilă* X^j, ($j = 1 \div p$) este identificată prin vectorul-coloană j al matricii $\mathbf{X}$, $\mathbf{x}_j \in \Re^n$, iar un *individ* i, ($i = 1 \div n$) prin vectorul-linie i al matricii $\mathbf{X}$, $\mathbf{x}_i \in \Re^p$. Vectorii-coloană ai matricii $\mathbf{X}$ definesc un nor de p puncte-variabile în $\Re^n$ iar vectorii-linie definesc un nor de n puncte-indivizi în $\Re^p$.

Media de selecţie a unei variabile j este definită prin $\bar{x_j} = \Sigma^n_{i=1}\ \rho_i\, x_{ij}$, iar *dispersia de selecţie* prin $s^2_j = \Sigma^n_{i=1}\ \rho_i(x_{ij} - \bar{x_j})^2$. Vectorul $\mathbf{g}' = (\bar{x_1}, \ldots, \bar{x_p})$ se numeşte *punct mediu* (sau *centru de greutate*) al norului de puncte-indivizi, $\mathbf{g} = \mathbf{X'D1_n}$, unde $\mathbf{1'_n} = (1, \ldots, 1) \in \Re^n$.

Matricea de varianţă-covarianţă $\mathbf{V} \in \mathcal{M}_{p\times p}(\Re)$, asociată matricii **X,** este:

$\mathbf{V} = \{ (v_{jk}) \mid v_{jk} \equiv \mathrm{cov}(\mathbf{x}_j, \mathbf{x}_k) = \sum^n_{i=1} \rho_i(x_{ij} - \bar{x_j})^2 (x_{ij} - \bar{x_{jk}})^2, j = 1 \div p, k = 1 \div p \}$

Matricea de corelaţie $\mathbf{R} \in \mathcal{M}_{p\times p}(\Re)$, asociată matricii **X,** este:

$\mathbf{R} = \{ (r_{jk}) \mid r_{jk} \equiv \mathrm{cor}(\mathbf{x}_j, \mathbf{x}_k) = v_{jk} / s_j s_k, j = 1 \div p, k = 1 \div p \}$

Se numeşte *tabel centrat*, asociat matricii **X**, matricea $\mathbf{Y} \in \mathcal{M}_{n\times p}(\Re)$:

$\mathbf{Y} = \{ (y_{ij}) \mid y_{ij} = (x_{ij} - \bar{x_j}), i = 1 \div n, j = 1 \div p \}$; $\mathbf{Y} = \mathbf{X} - \mathbf{l}_n\mathbf{g}' = (\mathbf{I}_n - \mathbf{l}_n\mathbf{l}_n'\mathbf{D})\mathbf{X}$.

Se numeşte *tabel centrat-redus*, asociat matricii **X**, matricea $\mathbf{Z} \in \mathcal{M}_{n\times p}(\Re)$:

$\mathbf{Z} = \{ (z_{ij}) \mid z_{ij} = y_{ij}/s_j, i = 1 \div n ; j = 1 \div p \}$; $\mathbf{Z} = \mathbf{YD}_{1/s}$, cu $\mathbf{D}_{1/s} = \mathrm{diag}\,(1/s_1, \ldots, 1/s_p)$

Avem: $\mathbf{V} = \mathbf{X'DX} - \mathbf{gg}' = \mathbf{Y'DY}$ şi $\mathbf{R} = \mathbf{D}_{1/s}\mathbf{V}\mathbf{D}_{1/s} = \mathbf{Z'DZ} = \sum^n_{i=1} \rho_i\, \mathbf{x}_i\, \mathbf{x}'_i$.

Fiecare individ, $\mathbf{x}_i$, definit de p coordonate corespunzând valorilor celor p variabile măsurate pe acest individ, este un element dintr-un spaţiu vectorial $\mathcal{E} \subset \Re^p$, având baza canonică $\mathcal{B} = (\mathbf{e}_1, \ldots, \mathbf{e}_p)$, numit *spaţiul indivizilor*.

Fie $\mathbf{M} \in \mathcal{M}_{pp}(\Re)$, o matrice simetrică, pozitiv definită, de dimensiune p, cu coeficienţi reali. Se numeşte *matrice a produsului scalar între indivizi* matricea $\mathbf{W} = \mathbf{XMX}' \in \mathcal{M}_{p\times p}(\Re)$:

$\mathbf{W} = \{(w_{i\ell}) \mid w_{i\ell} = \mathbf{x}'_i\, \mathbf{M}\mathbf{x}_\ell = \langle \mathbf{x}_i, \mathbf{x}_\ell \rangle_\mathbf{M} ; i, \ell = 1 \div p \}$,

unde $\langle \mathbf{x}_i, \mathbf{x}_\ell \rangle_\mathbf{M}$ este produsul scalar pe spaţiul $\mathcal{E}$ definit de metrica **M**. *Distanţa* dintre doi indivizi, $\mathbf{x}_i$ şi $\mathbf{x}_\ell$ din spaţiul $\mathcal{E}$, este: $d^2(\mathbf{x}_i, \mathbf{x}_\ell) = \langle \mathbf{x}_i - \mathbf{x}_\ell, \mathbf{x}_i - \mathbf{x}_\ell \rangle_\mathbf{M} = \| \mathbf{x}_i - \mathbf{x}_\ell \|^2_\mathbf{M}$.

Metricile cele mai uzitate, în spaţiul $\mathcal{E}$ al indivizilor, sunt: $\mathbf{I}_p$, ce induce produsul scalar uzual şi distanţa euclidiană şi $\mathbf{D}_{1/s2}$, care conduce la adimensionalizarea variabilelor deoarece fiecare valoare este împărţită cu abaterea standard de selecţie a variabilei corespunzătoare (x_{ij} / s_j). Metrica $\mathbf{I}_p$ 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, în schimb metrica $\mathbf{D}_{1/s2}$ echilibrează influenţa variabilelor transformându-le în variabile cu dispersia de selecţie unu. Utilizarea metricii $\mathbf{D}_{1/s2}$ pentru tabelul centrat **Y** revine la folosirea metricii $\mathbf{I}_p$ pentru tabelul centrat-redus **Z**. Matricea **W** a produsului scalar între indivizi poate fi întotdeauna exprimată în funcţie de metrica $\mathbf{I}_p$ adică $(\exists)\mathbf{T} : \mathbf{W} = (\mathbf{XT}')\mathbf{I}_p(\mathbf{TX}')$ şi atunci **W** este matricea produsului scalar al tabelului **XT'** faţă de metrica $\mathbf{I}_p$. Dacă $\mathbf{M} = \mathrm{diag}(m_1, \ldots, m_p)$, atunci $d^2(\mathbf{x}_i, \mathbf{x}_\ell) = \Sigma^p_{j=1}\, m_j(x_{ij} - x_{\ell j})^2$ iar coeficienţii $\{\sqrt{m_j}\}^p_{j=1}$ pot fi consideraţi ca ponderi ale variabilelor $\mathbf{x}_j$ în distanţa dintre indivizi.

Ipoteza fundamentală a unui demers exploratoriu [3, 9, 14, 15] este aceea că întreaga informaţie este conţinută în distanţele dintre punctele-indivizi ale unui nor, respectiv dispersia punctelor din nor. Se numeşte *inerţie totală* (*globală*) a norului de puncte-indivizi media ponderată a pătratelor distanţelor de la punctele-indivizi la centrul de greutate **g** al norului, adică:

$I_\mathbf{g} = \Sigma^n_{i=1}\, \rho_i(\mathbf{x}_i - \mathbf{g})'\mathbf{M}(\mathbf{x}_i - \mathbf{g}) = \Sigma^n_{i=1}\, \rho_i \| \mathbf{x}_i - \mathbf{g} \|^2_\mathbf{M}$

Prin analogie, *inerţia într-un punct oarecare,* $\mathbf{a} \in \Re^p$, este: $I_\mathbf{a} = \Sigma^n_{i=1}\, \rho_i \| \mathbf{x}_i - \mathbf{a} \|^2_\mathrm{M}$ şi conform formulei lui Huygens: $I_\mathbf{a} = I_\mathbf{g} + (\mathbf{g} - \mathbf{a})'\mathbf{M}(\mathbf{g} - \mathbf{a}) = I_\mathbf{g} + \| \mathbf{g} - \mathbf{a} \|^2_\mathbf{M}$.

Pentru un nor de puncte-indivizi dat, centrul de greutate **g** al norului minimizează inerţia totală. Inerţia totală este media pătratelor distanţelor dintre punctele-indivizi $2I_\mathbf{g} = \Sigma^n_{i=1}\, \Sigma^n_{\ell=1}\, \rho_i\rho_\ell \| \mathbf{x}_i - \mathbf{x}_\ell \|^2_\mathbf{M}$.

Notând cu $\mathrm{tr}(\mathbf{A}) = \Sigma^{n}_{i=1}\, a_{ii}$ urma matricii $\mathbf{A}$, avem $I_{\mathbf{g}} = \mathrm{tr}(\mathbf{MV}) = \mathrm{tr}(\mathbf{VM})$ şi pentru cazul $\mathbf{g} = \mathbf{0}$ avem $I_{\mathbf{g}} = \mathrm{tr}(\mathbf{WD}) = \mathrm{tr}(\mathbf{DW})$.

Dacă $\mathbf{M} = \mathbf{I}_p$, atunci $I_{\mathbf{g}} = \Sigma^{p}_{j=1}\, s^2(\mathbf{x}_j)$, adică inerţia totală este egală cu suma dispersiilor de selecţie a celor p variabile.

Dacă $\mathbf{M} = \mathbf{D}_{1/s2}$, atunci $I_{\mathbf{g}} = \mathrm{tr}(\mathbf{D}_{1/s2}\mathbf{V}) = \mathrm{tr}(\mathbf{D}_{1/s2}\mathbf{V}\mathbf{D}_{1/s2}) = \mathrm{tr}(\mathbf{R}) = \Sigma^{p}_{j=1}\, r_{jj} = \Sigma^{p}_{j=1}\, 1 = p$, adică inerţia totală este egală cu numărul variabilelor şi nu depinde de valorile acestora.

Fiecare variabilă, $\mathbf{x}_j$, definită de n coordonate corespunzând celor n valori ale variabilei j măsurată pe cei n indivizi, este un element dintr-un spaţiu vectorial $\mathcal{F} \subset \Re^n$ cu baza canonică $\mathbb{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_p)$, numit *spaţiul variabilelor*. Metrica utilizată în spaţiul $\mathcal{F}$, al variabilelor, este matricea diagonală a ponderilor indivizilor, $\mathbf{D} = \mathrm{diag}(\rho_1, \ldots, \rho_n) \in \mathcal{M}_{n\times n}(\Re)$.

Pentru variabilele centrate (matricea **Y**):

- produsul scalar dintre două variabile indus de metrica $\mathbf{D}$ este egal cu covarianţa de selecţie dintre cele două variabile necentrate: $\langle \mathbf{y}_j, \mathbf{y}_k \rangle_{\mathbf{D}} = \mathbf{y}_j'\mathbf{D}\mathbf{y}_k = \mathrm{cov}(\mathbf{x}_j, \mathbf{x}_k)$;
- norma („lungimea") unei variabile centrate este egală cu abaterea standard de selecţie a variabilei necentrate: $\|\mathbf{y}_j\|^2_{\mathbf{D}} = s^2(\mathbf{x}_j)$;
- cosinusul unghiului dintre două variabile este egal cu coeficientul de corelaţie de selecţie al variabilelor necentrate: $\cos(\theta_{jk}) = \langle \mathbf{y}_j, \mathbf{y}_k \rangle_{\mathbf{D}} \,/\, \|\mathbf{y}_j\|_{\mathbf{D}}\, \|\mathbf{y}_k\|_{\mathbf{D}} = \mathrm{cor}(\mathbf{x}_j, \mathbf{x}_k)$;
- $(\forall) j, k \in [1, p]$: $\overline{\mathbf{y}_j} = 0$; $s^2(\mathbf{y}_j) = s^2(\mathbf{x}_j)$; $\mathrm{cor}(\mathbf{y}_j, \mathbf{y}_k) = \mathrm{cor}(\mathbf{x}_j, \mathbf{x}_k)$.
- Pentru variabilele centrat-reduse (matricea **Z**):
- $(\forall) j, k \in [1, p]$: $\overline{\mathbf{z}_j} = 0$; $s^2(\mathbf{z}_j) = 1$; $\mathrm{cor}(\mathbf{z}_j, \mathbf{z}_k) = \mathrm{cor}(\mathbf{x}_j, \mathbf{x}_k)$; $d^2(\mathbf{z}_j, \mathbf{z}_k) = 2(1 - r_{jk})$.

Operaţia de centrare a tabelului $\mathbf{X}$ are în spaţiile $\Re^p$ şi $\Re^n$ interpretări geometrice diferite.

În $\Re^p$ această transformare echivalează cu o translaţie a originii axelor în centrul de greutate (punctul mediu) al norului.

În $\Re^n$ această transformare este o proiecţie pe hiperplanul care trece prin originea axelor şi este ortogonal pe dreapta ce trece prin originea axelor având ca parametri directori $\{\rho_i \mid i = 1 \div n\}$. Matricea $\mathbf{P} = \mathbf{I}_n - \mathbf{l}_n\mathbf{l}'_n\mathbf{D}$, asociată acestei transformări, este matricea proiecţiei $\mathbf{M}$-ortogonale pe subspaţiul generat de vectorii coloană liniar-independenţi ai matricii $\mathbf{Y}$. Coordonatele acestor vectori satisfac relaţia $\sum^{n}_{i=1} \rho_i y_{ij} = 0$, $(\forall) j = 1 \div p$ reprezentând ecuaţia unui hiperplan în $\Re^n$ care trece prin originea axelor şi are ca normală în punctul $\mathbf{0}_n$ dreapta de parametri directori $\{\rho_i \mid i = 1 \div n\}$. Dacă $\mathbf{D} = (1/n)\mathbf{I}_n$ atunci hiperplanul este ortogonal pe prima bisectoare.

Toate punctele-variabilă se află pe hipersfera de rază 1, centrată în originea axelor numită *sfera de corelaţie*. Planurile în care vor fi proiectate variabilele intersectează sfera după cercuri diametrale, numite *cercuri de corelaţie*, de rază 1 şi în interiorul cărora se află proiecţiile punctelor-variabile.

Dacă în spaţiul indivizilor interesează distanţa dintre puncte, în spaţiul variabilelor interesează unghiurile dintre ele. Proximitatea între punctele-variabile se interpretează în termeni de corelaţii. Sistemul de proximităţi dintre două puncte-variabile, indus de relaţia $d^2(\mathbf{z}_j, \mathbf{z}_k) = 2(1 - r_{jk})$, evidenţiază că:

- două variabile puternic corelate sunt sau foarte apropiate una de cealaltă (deoarece $r_{jk} \approx 1$ implică $d^2(\mathbf{z}_j, \mathbf{z}_k) \approx 0$) sau foarte depărtate ($r_{jk} \approx -1$ implică $d^2(\mathbf{z}_j, \mathbf{z}_k) \approx 4$);
- două variabile necorelate, deci ortogonale, sunt la distanţă medie (deoarece $r_{jk} \approx 0$ implică $d^2(\mathbf{z}_j, \mathbf{z}_k) \approx 2$).

3. Analiza în componente principale

În funcţie de proveniență variabilele care pot face obiectul unei *ACP* pot lua valori cantitative obţinute în urma unor măsurători, pot lua valori calitative obţinute în urma unor notaţii dar sunt asimilabile cu variabilele cantitative sau pot lua valori calitative ordinale obţinute în urma unor clasamente dar pot fi transformate în variabile continue.

Obiectivele urmărite de *ACP* sunt:

- reprezentarea grafică „optimală" a indivizilor (liniilor), minimizând deformările norului de puncte, într-un subspaţiu $\mathcal{E}_q$ de dimensiune q ($q < p$);
- reprezentarea grafică a variabilelor intr-un subspaţiu $\mathcal{F}_q$ explicitând „cel mai bine" legăturile iniţiale între aceste variabile;
- reducerea dimensiunii (compresia), sau aproximarea matricii **X** printr-o matrice de rang $q < p$.

Poziţia punctelor într-un 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. O modalitate de a reda vizual forma unui nor este aceea de a-1 proiecta pe o dreaptă sau pe un plan minimizând deformările pe care această proiecţie le implică.

Matricea $\mathbf{W} = \mathbf{YMY}' \in \mathcal{M}_{n\times n}(\Re)$ este o matrice simetrică, de dimensiune *n,* al cărui termen general, $w_{i\ell} = \mathbf{y}'_i\mathbf{M}\mathbf{y}_\ell$, este un produs scalar între indivizii i şi ℓ. Se numeşte *imagine euclidiană a indivizilor,* asociată produselor scalare $w_{i\ell}$, un nor compus din n puncte $S_1,...,S_n$ şi dintr-un punct O din $\mathcal{E}$ astfel încât aceste puncte să reconstituie produsele scalare $w_{i\ell}$, adică $\langle OS_i, OS_\ell \rangle = w_{i\ell}$ $(\forall) i, \ell = 1 \div n$ unde produsul scalar $\langle \circ, \circ \rangle$ este definit de metrica euclidiană $\mathbf{M} = \mathbf{I}_p$.

Matricea $\mathbf{V} = \mathbf{Y'DY} \in \mathcal{M}_{n\times n}(\Re)$ (de varianţă-covarianţă a variabilelor centrate) este o matrice simetrică, de dimensiune *p,* al cărui termen general, $v_{jk} = \mathbf{y}'_j\mathbf{D}\mathbf{y}_k$, este un produs scalar între variabilele j şi k. Se numeşte *imagine euclidiană a variabilelor* asociată produselor scalare v_{jk}, un nor compus din p puncte $T_1,...,T_p$ şi dintr-un punct O din $\mathcal{F}$ astfel încât aceste puncte să reconstituie produsele scalare v_{jk} , adică $\langle OT_j, OT_k \rangle = v_{jk}$ $(\forall) j, k = 1 \div p$ unde produsul scalar $\langle \circ, \circ \rangle$ este definit de metrica euclidiană $\mathbf{D} = \mathbf{I}_n$.

Există o infinitate de imagini euclidiene ale aceluiaşi nor de puncte. Două imagini euclidiene sunt *echivalente* dacă ele reconstituie aceleaşi produse scalare.

Dacă dimensiunea spaţiului vectorial în care se lucrează este mai mică sau egală cu 3 atunci imaginea euclidiană a unui nor de puncte poate fi vizualizată, dacă nu atunci trebuie căutată o imagine euclidiană aproximativă. Mai precis, pornindu-se de la o imagine euclidiană dintr-un spaţiu afin de dimensiune d se doreşte obţinerea unei imagini euclidiene într-un spaţiu afin de dimensiune mult mai mică $q << d$.

Reprezentarea indivizilor. În spaţiul $\mathcal{E} \subset \Re^p$ al indivizilor, **Y** (tabelul centrat asociat lui **X)** poate fi reprezentat ca un nor de n puncte-indivizi centrate în punctul mediu al norului şi ale căror p coordonate reprezintă liniile lui **Y**.

Dacă rang(**Y**) = q atunci problema aproximării este practic rezolvată. Este suficient să se determine o bază a subspaţiului vectorial de dimensiune q din $\Re^p$ ce conţine norul de puncte-indivizi şi să se calculeze coordonatele punctelor în noua bază.

Dacă rang(**Y**) > q, demersul de mai sus se realizează prin proiecţia punctelor-indivizi pe un subspaţiu $\mathcal{E}_q$ de dimensiune q, obţinut astfel încât media pătratelor distanţelor între proiecţii să fie maximă sau, inerţia norului proiectat pe $\mathcal{E}_q$ să fie maximă sau, în fine, deformarea distanţelor prin

proiecţie să fie minimă. Astfel problema ce trebuie rezolvată capătă următorul enunţ: s*ă se găsească* $\mathcal{H} \equiv \mathcal{E}_q$ *astfel încât* $\max \sum^{n}_{i=1} d^2(\mathbf{y}_i, \mathbf{0})$, iar soluţia este dată de următoarea teoremă: *subspaţiul de dimensiune q pe care se proiectează optim, în sensul celor mai mici pătrate, cele n puncte din* $\Re^p$ *este generat de primii q vectori proprii ai matricii* $\mathbf{A} = \mathbf{VM} \in M_{pp}(\Re)$ *corespunzători valorilor proprii* $\lambda_1 > \lambda_2 > ... > \lambda_q$, *unde* **V** *este matricea de variantă-covarianţă asociată tabelului* **X**, *iar* **M** *este metrica spaţiului indivizilor.*

Valorile proprii ale matricii **A** sunt reale şi pozitive, **A** fiind **M**-simetrică, pozitiv definită şi cu coeficienţi reali. Vectorii proprii ai matricii **A** sunt **M**-ortonormaţi. Matricea **A** se numeşte *matricea inerţiei* şi $I_{\mathbf{g}} = \text{tr}(\mathbf{A}) = \Sigma^{p}_{j=1} \lambda_j$.

Imaginea euclidiană a norului de puncte-indivizi obţinută prin proiecţia pe subspaţiul $\mathcal{H}$ se numeşte *imaginea euclidiană a punctelor-indivizi* asociată aproximaţiei de ordinul *q* a produselor scalare. Se numesc *axe principale de inerţie* vectorii proprii, **M**-normaţi, $\mathbf{a}_j$, ai matricii de inerţie **A**. Se numeşte *factor principal* asociat axei principale $\mathbf{a}_j$ şi se notează cu $\mathbf{u}_j$ forma liniară din $\Re^p$ definită de relaţia $\mathbf{u}_j = \mathbf{M}\mathbf{a}_j$. Factorii principali $\{\mathbf{u}_j \mid j = 1 \div p\}$ sunt vectorii proprii ai matricii **MV** asociaţi valorilor proprii $\{\lambda_j \quad j = 1 \div p\}$ ale matricii $\mathbf{A} = \mathbf{VM}$. Se numeşte *plan factorial principal* subspaţiul $\mathcal{E}_2$, generat de vectorii $\{\mathbf{u}_1, \mathbf{u}_2\}$. Se numeşte *componentă principală* asociată factorului principal $\mathbf{u}_j$ şi se notează cu $\mathbf{c}_j$ forma liniară din $\Re^n$ definită de relaţia $\mathbf{c}_j = \mathbf{Y}\mathbf{u}_j$, $\mathbf{c}_j$ este proiecţia M-ortogonală a indivizilor pe axa principală $\mathbf{a}_j$.

Componentele principale $\{\mathbf{c}_j \mid j = 1 \div p\}$, sunt vectorii proprii ai matricii **WD** asociaţi valorilor proprii $\{\lambda_j \mid j = 1 \div p\}$ ale matricii **A** şi sunt **D**-ortogonale, deci necorelate. Mediile de selecţie ale componentelor principale sunt nule (pe datele centrate şi centrat reduse). Dispersia de selecţie a componentei principale $\mathbf{c}_j$ este λ_j valoarea proprie a matricii inerţiei, **A**, pentru $(\forall) j = 1 \div p$.

Componentele principale sunt combinaţii liniare de variabilele iniţiale, de dispersie maximă şi care satisfac restricţiile $\mathbf{u}'_j\mathbf{M}^{-1}\mathbf{u}_j = \mathbf{1}$. În cazul *ACP* normate $(\mathbf{Z}, \mathbf{I}_p)$, componentele principale $\{\mathbf{c}_j \quad j = 1 \div p\}$ asociate valorilor proprii $\{\lambda_j \quad j = 1 \div p\}$ ale matricii **A** sunt variabilele cele mai „legate" de variabilele iniţiale, $\mathbf{z}_1, ..., \mathbf{z}_p$, în sensul că suma pătratelor coeficienţilor de corelaţie, $\Sigma^{p}_{k=1} \text{cor}^2(\mathbf{c}_j, \mathbf{z}_k)$, este maximă pentru $(\forall) j = 1 \div p$.

Reprezentarea variabilelor. În spaţiul $\mathcal{F} \subset \Re^n$ al variabilelor, **Y** (tabelul centrat asociat lui **X)** poate fi reprezentat ca un nor de *p* puncte-variabilă ale căror *n* coordonate sunt coloanele lui **Y**. La fel ca şi în cazul norului de puncte-indivizi, se doreşte găsirea axelor principale şi a subspaţiului afin *q*-dimensional, $\mathcal{F}_q \subset \Re^n$, generat de aceste axe, care aproximează optim norul de puncte-variabilă. Aceasta înseamnă să fie maximizată media pătratelor distanţelor dintre cele *p* proiecţii pe $\mathcal{F}_q$, adică de rezolvat problema de programare pătratică cu restricţii liniare: $\max_{(\mathbf{b})} \mathbf{b'DYMY'Db} \mid \mathbf{b'Db} = \mathbf{1}$ a cărei soluţie, **b,** este vectorul propriu al matricii $\mathbf{B} = \mathbf{YMY'D}$ (**D**-simetrică, reală), corespunzând celei mai mari valori proprii μ.

Ecuaţia axei factoriale **b** din $\Re^n$ este: $\mathbf{YMY'Db} = \mu\mathbf{b} \mid \mathbf{b'Db} = \mathbf{1}$; ecuaţia factorului principal **v** din $(\Re^n)$ este: $\mathbf{v} = \mathbf{Db}$; ecuaţia componentei principale **d** din $\Re^n$ este: $\mathbf{d} = \mathbf{Y'v}$ sau $\mathbf{d} = \mathbf{Z'v}$.

Se numeşte *cerc de corelaţie principal* subspaţiul $\mathcal{F}_2$ generat de vectorii $\{\mathbf{v}_1, \mathbf{v}_2\}$.

Analog ca în cazul norului de puncte-indivizi:

- Factorii principali $\mathbf{v}_i \in (\Re^n)$, $i=1\div n$, sunt $\mathbf{D}^{-1}$-ortonormaţi şi satisfac relaţiile $\mathbf{DYMY'v}_i = \mu_i\mathbf{v}_i$.
- Componentele principale $\mathbf{d}_i \in \Re^p$, $i=1\div n$ sunt **M**-ortogonale, au dispersia de selecţie egală cu μ şi satisfac relaţiile $\mathbf{X'DXMd}_i = \mu_i\mathbf{d}_i$.

În cazul *ACP* normate norul de puncte-variabile se află pe hipersfera de corelaţie deci planul factorial va intersecta această hipersferă după un cerc diametral.

Relaţii de tranziţie între cele două spaţii. Din punct de vedere numeric, o analiză în componente principale se reduce la calculul primelor *q* valori proprii şi al vectorilor proprii asociaţi

pentru matricile $\mathbf{VM} = \mathbf{Y'DYM} \in \mathcal{M}_{p,p}(\Re)$ şi $\mathbf{WD} = \mathbf{YMY'D} \in \mathcal{M}_{n,n}(\Re)$.

O întrebare naturală este dacă există o relaţie între elementele principale dintr-o *ACP* pe spaţiul variabilelor $(\mathcal{F}, \mathbf{M}) \subset \Re^p$ şi elementele principale dintr-o *ACP* pe spaţiul indivizilor $(\mathcal{E}, \mathbf{D}) \subset \Re^n$ iar răspunsul, privind relaţiile de tranziţie între cele două spaţii, este dat de următoarea teoremă: *toate valorile proprii nenule ale matricilor* **Y'DYM** *şi* **YMY'D** *sunt egale având, eventual, acelaşi ordin de multiplicitate şi pentru* $\lambda_j \neq 0$ *sunt adevărate următoarele relaţii de tranziţie*:

- $\mathbf{b}_j = (1/\sqrt{\lambda_j})\ \mathbf{YMa}_j = (1/\sqrt{\lambda_j})\ \mathbf{Yu}_j = (1/\sqrt{\lambda_j})\ \mathbf{c}_j$ *şi*
- $\mathbf{a}_j = (1/\sqrt{\lambda_j})\ \mathbf{Y'Db}_j = (1/\sqrt{\lambda_j})\ \mathbf{Y'v}_j = (1/\sqrt{\lambda_j})\ \mathbf{d}_j$ *unde* $j = 1 \div \text{rang}(\mathbf{Y'Y})$.

Cum, în general, $p < n$ este suficientă *ACP* pe norul de puncte-indivizi, elementele principale pentru norul de puncte-variabile obţinându-se prin relaţiile de tranziţie.

Coordonalele punctelor pe o axă factorială în $\Re^p$ sunt proporţionale cu componentele axei factoriale din $\Re^n$ corespunzătoare aceleiaşi valori proprii şi reciproc, deoarece $\mathbf{c} = \mathbf{Xu}$ şi $\mathbf{d} = \mathbf{X'v}$ implică $\mathbf{c} = (\sqrt{\lambda})\mathbf{b}$ şi $\mathbf{d} = (\sqrt{\lambda})\mathbf{a}$).

Orientarea axelor factoriale este arbitrară deoarece vectorii proprii sunt determinaţi modulo semnul lor. Acest lucru nu impietează asupra formei norului, adică a distanţelor între puncte.

ACP 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 între ele poate exista o relaţie de ordin superior lui *1* (relaţie neliniară).

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 cazul datelor centrat-reduse, $\Sigma^p_{j=1}\ \text{cor}^2\ (\mathbf{z}_k, \mathbf{v}_j) = \mathbf{a}'_k\mathbf{Ma}_k = 1$.

Reconstituirea datelor iniţiale. Pornind de la relaţia $\mathbf{c} = \mathbf{Yu}$ se obţine relaţia $\mathbf{Y} = \Sigma^p_{j=1}\ \mathbf{c}_j\mathbf{u}'_j\mathbf{M}^{-1}$ numită *formula de reconstituire* a tabelului de date **Y** pornind de la componentele şi factorii principali. Analog, pornind de la relaţia $\mathbf{c} = \mathbf{Xu}$ se poate reconstitui tabelul **X** precum şi $\mathbf{MV} = \sum^p_{j=1} \lambda_j\mathbf{u}_j\mathbf{u}'_j\mathbf{M}^{-1}$ şi $\mathbf{VM} = \sum^p_{j=1} \lambda_j\mathbf{a}_j\mathbf{a}'_j\mathbf{M}$

Dacă $\mathbf{M} = \mathbf{I}$, adică în cazul metricii euclidiene, axele principale coincid cu factorii principali şi, conform formulelor de tranziţie, se obţine formula de reconstituire $\mathbf{Y} = \sum^p_{j=1}\ \mathbf{c}_j\mathbf{u}'_j = \sum^p_{j=1}\ (\sqrt{\lambda_j})\mathbf{v}_j\mathbf{u}'_j$ cu $\mathbf{v}_j$ vectori proprii normaţi ai matricii **YY'** şi $\mathbf{u}_j$ vectori proprii normaţi ai matricii **Y'Y**. Dacă în formula de mai sus sumarea se face doar după primii $q < p$ termeni (valorile proprii sunt ordonate descrescător), atunci se obţine cea mai bună aproximare, în sensul celor mai mici pătrate, a lui **Y** printr-o matrice de rang *q*. Privite doar din acest punct de vedere, metodele de analiză factorială se reduc la *metode de compresie a datelor.*

Reprezentarea simultană. Analiza norului de variabile este dedusă din analiza norului de indivizi, reprezentarea variabilelor pe axele factoriale în $\Re^n$ ajută la interpretarea axelor factoriale în $\Re^p$ şi reciproc. Trebuie totuşi evitată interpretarea distanţei dintre un punct-individ şi un punct-variabilă deoarece aceste puncte nu fac parte nici din acelaşi nor, nici din acelaşi spaţiu şi nici nu sunt reprezentate în acelaşi reper. Dacă însă se consideră în loc de puncte-variabile direcţiile variabilelor în $\Re^p$, atunci se pot reprezenta simultan, în acest spaţiu, atât punctele-indivizi, cât şi vectorii reprezentând variabilele.

În spaţiul $\Re^p$ al celor *n* puncte-indivizi, după transformarea tabelului de date, există două sisteme de axe: vechile axe unitare $\{\mathbf{e}_1, ..., \mathbf{e}_p\}$ şi noile axe unitare $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$, formate din axele factoriale. Posibilitatea unei reprezentări simultane rezidă în acest context în proiecţia, ca individ suplimentar, a vechii axe $\mathbf{e}_j$ pe noua axă $\mathbf{u}_k$. Coordonata proiecţiei lui $\mathbf{e}_j$ pe $\mathbf{u}_k$ este $\mathbf{e}'_j\ \mathbf{u}_k = u_{kj}$. Este, astfel, posibil să se reprezinte în $\Re^p$ direcţiile date de variabilele iniţiale pe planul factorial al norului de indivizi. Această reprezentare a variabilelor este diferită de reprezentarea norului de variabile.

Se numeşte *reprezentare simultană* proiectarea reperului ortonormat al axelor de origine în

planul factorial al norului de indivizi. În $\Re^n$, în metrica euclidiană, coordonata variabilei *j* pe axa *k* este egală cu coeficientul de corelaţie între variabilă şi factor: $d_{kj} = \sqrt{\lambda_k}\, u_{kj}$. 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 $\Re^n$, 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 $\Re^n$ (norul proiectat al extremităţilor vectorilor unitari din $\Re^p$ şi norul extremităţilor vectorilor variabile în $\Re^n$ au în general forme asemănătoare, vectorii proprii fiind totuşi comparabili, deci dilatările fiind puţin deformante).

Ţinând cont de aceste consideraţii, are totuşi sens să se compare, în reprezentarea simultană, poziţia a doi indivizi faţă de ansamblul variabilelor, sau poziţia a două variabile faţă de ansamblul indivizilor. La intersecţia axelor se găsesc valorile medii ale tuturor variabilelor. 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. Pe direcţia unei variabile prezintă interes distanţele între indivizi.

Interpretarea rezultatelor. 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 aspecte:

- axele factoriale permit obţinerea celei mai bune vizualizări aproximative, în sensul celor mai mici pătrate, ale distanţelor dintre indivizi, respectiv dintre variabile şi în acest sens, primul demers care se impune este legat de măsurarea calităţii acestei aproximări;
- metoda naturală de a da o semnificaţie unei componente principale **c** este de a o corela cu variabilele iniţiale $\mathbf{x}_j$, în acest sens sunt calculaţi coeficienţii de corelaţie liniară cor(**c**, $\mathbf{x}_j$) şi sunt puşi în evidenţă coeficienţii cu valori absolute mari;
- practica frecvent utilizată este de a împărţi în două mulţimea variabilelor: o parte din variabile, numite *variabile active,* urmând să fie utilizate pentru determinarea axelor principale iar cealaltă parte, numite *variabile pasive* (*suplimentare* sau *ilustrative*), să fie corelate, à posteriori, cu componentele principale;
- într-un mod asemănător se procedează şi în cazul mulţimii indivizilor, distingându-se între *indivizi activi* şi *indivizi suplimentari*, care nu sunt luaţi în considerare la calculul matricilor de covarianţă / corelaţie.

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

4. Analiza factorială discriminantă

Se dispune de observaţii privind *p* variabile cantitative X^1, ..., X^p , jucând rolul de variabile explicative şi o variabilă calitativă *T* cu *q* modalităţi $\{\tau_1, ..., \tau_q\}$, jucând rolul de variabilă de explicat. Cele *p* variabile explicative au fost observate pe un eşantion de *n* indivizi, variabila nominală *T* generează o partiţie a celor *n* indivizi în *q* clase I_k, $k = 1 \div q$.

În anumite situaţii se poate constata că puterea de discriminare a caracteristicilor (axelor) este slabă pentru datele considerate, fie că nu s-au ales cele mai bune caracteristici ale datelor, fie că datele sunt prin natura lor foarte asemănătoare. Pentru astfel de situaţii este uneori posibilă determinarea unui nou sistem de coordonate faţă de care structura de clase este mai evidentă decât în sistemul iniţial, axele noului sistem având o putere de discriminare a claselor superioară celei a axelor iniţiale.

Fie $\mathbf{X} = \{ x_{ij} \mid i = 1 \div n,\ j = 1 \div p \} \in \mathcal{M}_{n \times p}(\Re)$ matricea observaţiilor. Fiecare clasă *k* caracterizează un subnor I_k de n_k indivizi, unde $\Sigma^q_{k=1}\, n_k = n$.

Se notează cu $\mathbf{g}_k$ centrul de greutate al clasei k, adică $\mathbf{g}_k = (\overline{x_j^k})_{j=\overline{1\div p}}$, unde $\overline{x_j^k} = (1/n_k)\Sigma_{i\in Ik}\, x_{ij}$ şi respectiv cu $\mathbf{g}$ centrul de greutate al norului, adică $\mathbf{g} = (\overline{x_j})_{j=\overline{1\div p}}$, cu $\overline{x_j} = (1/n)\Sigma^n_{i=1}\, x_{ij} = \Sigma^q_{k=1}\, (n_k/n)\, \overline{x_j^k}$

Variabila $\mathbf{a} = \{\, a(i) \mid a(i) = \Sigma^p_{j=1}\ a_j\,(x_{ij} - \overline{x_j}),\ i = \overline{1\div n}\}$, combinaţie liniară a celor p variabile, are media empirică 0 (este centrată) şi dispersia empirică: $D^2(\mathbf{a}) = \Sigma^p_{j=1}\Sigma^p_{k=1}\, a_j\, a_k\, \mathrm{cov}(\mathbf{x}_j, \mathbf{x}_k) = \mathbf{a'Va}$.

Conform formulei lui Huygens, matricea de covarianţă $\mathbf{V}$ se descompune într-o componentă intraclase (sau reziduală) $\mathbf{V}_r$ şi o componentă interclase (sau explicată) $\mathbf{V}_e$, $\mathbf{V} = \mathbf{V}_r + \mathbf{V}_e$, astfel încât dispersia combinaţiei liniare $\mathbf{a}$ de variabile devine $D^2(\mathbf{a}) = \mathbf{a'Va} = \mathbf{a'V_ra} + \mathbf{a'V_ea}$. Dintre toate combinaţiile liniare de variabile, sunt căutate cele care au o dispersie intraclase $\mathbf{V_r}$ minimă şi o dispersie interclase $\mathbf{V_e}$ maximă pentru ca în proiecţie pe axa discriminantă $\mathbf{a}$, fiecare subnor să fie, în măsura posibilului, în acelaşi timp bine grupat şi bine separat de ceilalţi subnori. Cu alte cuvinte, trebuie găsit $\mathbf{a}$ astfel încât raportul $\mathbf{a'V_ea} / \mathbf{a'V_ra}$ să fie maxim (sau $\mathbf{a'V_ra} / \mathbf{a'V_ea}$ să fie minim) sau, conform cu $D^2(\mathbf{a}) = \mathbf{a'V_ra} + \mathbf{a'V_ea}$, să se maximizeze $f(\mathbf{a}) = \mathbf{a'V_ea} / \mathbf{a'Va}$ adică raportul dintre dispersia interclase $\mathbf{V_e}$ şi dispersia totală $\mathbf{V}$. Un punct staţionar al lui $f(\mathbf{a})$ se află rezolvând ecuaţia: $\partial f(\mathbf{a})/\partial \mathbf{a} = 0$ adică ecuaţia $[\ (\mathbf{a'Va})\ (2\ \mathbf{V_ea}) - (\mathbf{a'}\ \mathbf{V_ea})\ (2\mathbf{Va})\ ]\ /\ (\mathbf{a'Va})^2 = 0$ deoarece $\partial(\mathbf{a'V_ea}) / \partial \mathbf{a} = 2\mathbf{V_ea}$ dacă $\mathbf{V_e}$ este simetrică şi este deoarece atât $\mathbf{V_e}$ cât şi $\mathbf{V}$ sunt matrici de covarianţă, în plus $\mathbf{V}$ este inversabilă. Deci $(\mathbf{a'Va})\ (\mathbf{V_ea}) = (\mathbf{a'}\ \mathbf{V_ea})\ (\mathbf{Va})$ sau $\mathbf{V^{-1}V_ea} = (\ \mathbf{a'V_ea} / \mathbf{a'Va}\ )\mathbf{a}$ adică $\mathbf{V^{-1}V_ea} = f(\mathbf{a})\mathbf{a}$. $f(\mathbf{a})$ este maximă dacă este egală cu λ_{max}, valoarea proprie maximă a matricii $\mathbf{V^{-1}V_e}$ iar $\mathbf{a}$ este vectorul propriu corespunzător lui λ_{max}.

Matricea $\mathbf{V^{-1}V_e} \in \mathcal{M}_{p\times p}$ este, în general, o matrice nesimetrică. Din punct de vedere al calculului numeric, având în vedere că $q \ll p$, este mai uşor a afla vectorii şi valorile proprii ale unei matrici simetrice de dimensiune $q \times q$ şi a găsi o exprimare a lui $\mathbf{a}$ în funcţie de aceste elemente. $\mathbf{V_e}$ este produsul matricii $\mathbf{C} = \{\ c_{jk} \mid c_{jk} = \sqrt{[(n_k / n)\ (\overline{x_j^k} - \overline{x_j})]},\ j = \overline{1\div p},\ k = \overline{1\div q})\ \} \in \mathcal{M}_{p\times q}(\Re)$ cu transpusa sa, $\mathbf{V_e} = \mathbf{CC'}$ deci $\mathbf{V^{-1}CC'a} = \lambda\mathbf{a}$ sau $\mathbf{CC'a} = \lambda\mathbf{Va}$. Luând $\mathbf{a} = \mathbf{V^{-1}Cw}$ avem relaţia $\mathbf{CC'V^{-1}Cw} = \lambda\mathbf{Cw}$, dacă $\mathbf{w}$ este vector propriu al matricii $\mathbf{C'V^{-1}C}$, corespunzător lui λ, atunci el verifică această relaţie, iar $\mathbf{a}$ şi λ verifică relaţia $\mathbf{CC'a} = \lambda\mathbf{Va}$. Deoarece $\mathbf{C'V^{-1}C} \in \mathcal{M}_{q\times q}(\Re)$ este simetrică, se diagonalizează această matrice şi apoi se află $\mathbf{a} = \mathbf{V^{-1}Cw}$.

Valoarea $\lambda_{max} \in [0, 1]$ şi se numeşte *putere discriminantă*.

- Cazul $\lambda_{max} = 1$. În proiecţia pe axa $\mathbf{a}$ dispersiile intraclase sunt nule. Cei k nori sunt fiecare într-un hiperplan ortogonal pe $\mathbf{a}$. Discriminarea pe această axă este perfectă dacă centrele de greutate se proiectează în puncte diferite.
- Cazul $\lambda_{max} = 0$ corespunde cazului în care cea mai bună axă discriminantă nu poate să separe centrele de greutate $\mathbf{g}_k$ pentru că acestea sunt confundate. Subnorii sunt, deci, concentrici şi neliniari separabili. Este posibilă existenţa unei suprafeţe de decizie neliniare.

Valoarea proprie este o *măsură pesimistă* a puterii de discriminare a unei axe, clasele pot fi liniar separabile pe axa considerată în pofida faptului că $\lambda < 1$. Numărul de valori proprii nenule, deci al axelor discriminante, este egal cu $q - 1$ în cazul obişnuit unde $n > p > q$ şi variabilele nu sunt legate prin relaţii liniare.

Odată găsite axele cu puterea de discriminare cea mai bună, pasul următor constă în găsirea suprafeţelor de decizie. Metodele geometrice de analiză discriminantă, esenţialmente descriptive, se bazează pe noţiunea de distanţă şi nu utilizează nicio noţiune probabilistă. În context geometric, discriminarea poate fi interpretată ca o împărţire a spaţiului variabilelor în regiuni, numite *regiuni de decizie,* fiecare regiune fiind asociată cu o clasă de indivizi. Regiunile de decizie şi implicit clasele corespunzătoare, se zic *separabile* dacă pot fi separate prin suprafeţe din spaţiul variabilelor. Suprafeţele de separare ale regiunilor de decizie se numesc şi *suprafeţe de decizie.* Suprafeţele de decizie pot fi descrise cu ajutorul unei mulţimi de *funcţii de discriminare* (sau *de decizie*). Funcţia de discriminare ataşează fiecare vector-individ unei regiuni din spaţiul variabilelor, regiune delimitată prin intermediul unei mulţimi de suprafeţe de decizie.

O funcţie de discriminare *instruibilă* tinde să reducă numărul indivizilor clasaţi incorect făcând

acest număr cât mai mic posibil, eventual nul. Acest lucru se realizează prin ajustarea mulţimii regiunilor de decizie ca răspuns la observaţiile făcute asupra unei mulţimi de indivizi *de instruire.*

După ce clasele şi suprafeţele de decizie sunt stabilite (prin o fază de instruire), respectiv funcţia de discriminare este instruită, funcţiei de discriminare i se prezintă date ale căror clase nu se cunosc. Această fază, în care indivizi noi sunt asociaţi uneia sau alteia dintre clasele stabilite, se numeşte *fază de lucru* (sau *decizională* sau *de afectare*). Uneori faza de instruire şi cea de lucru pot să coincidă sau să se suprapună parţial.

Intro *AFD* se disting, în consecinţă, două demersuri:

- primul, *descriptiv,* ce constă în căutarea funcţiilor de discriminare liniare pe eşantionul de volum n respectiv găsirea combinaţiilor liniare de variabile explicative ale căror valori separă cel mai bine cele q clase;
- al doilea, *decizional,* ce constă în aflarea claselor de afectare a n' indivizi noi, descrişi prin variabilele explicative $(X^1, ..., X^p)$.

5. Analiza corespondenţelor simple

Se dispune de observaţii privind două variabile calitative (nominale sau categoriale), X cu n modalităţi $\{x_1, ..., x_n\}$ şi respectiv Y cu p modalităţi $\{y_1, ..., y_p\}$. Variabilele nominale X şi Y au fost observate simultan pe un eşantion de k indivizi şi generează fiecare câte o partiţie a celor k indivizi.

Un tabel ale cărui linii, respectiv coloane, desemnează două partiţii ale aceleiaşi mulţimi, partiţii date de modalităţile a două variabile nominale, se numeşte *tabel de contingenţă* (*de dependenţă* sau *încrucişat*). De exemplu, într-un scrutin electoral cu mai mulţi candidaţi, dacă pentru un eşantion de alegători se cunosc circumscripţiile electorale şi opţiunile acestora atunci este convenabil să se grupeze datele într-un tabel de contingenţă **K** ale cărui elemente k_{ij} reprezintă numărul de persoane din circumscripţia i care optează pentru candidatul j.

Analiza corespondenţelor simple (*ACS*) se poate aplica unor tabele de contingenţă cu toate valorile nenegative şi tratează în mod echivalent atât liniile cât şi coloanele. Abordările curente constau în a defini *ACS* ca fiind rezultatul a două *ACP*, pentru profiluri-linii şi pentru profiluri-coloane, utilizând metrica χ^2.

Fie $\mathbf{K} = \{ k_{ij} \mid i = 1 \div n, j = 1 \div p\} \in \mathcal{M}_{n \times p}(\Re)$ tabelul de contingenţă cu n linii, p coloane şi elementele k_{ij}, unde k_{ij} este numărul de indivizi având simultan modalitatea i a variabilei X şi modalitatea j a variabilei Y.

Se numesc *efective marginale* (sau *marje*) cantităţile $k_{i\bullet} = \sum_{j=1}^{p} k_{ij}$ şi $k_{\bullet j} = \sum_{i=1}^{n} k_{ij}$, $(\forall) i = 1 \div n$ şi $(\forall) j = 1 \div p$ îndeplinind condiţiile $\sum_{i=1}^{n} k_{i\bullet} = \sum_{j=1}^{p} k_{\bullet j} = \sum_{i=1}^{n} \sum_{j=1}^{p} k_{ij} = k$. Se numesc *frecvenţe relative* cantităţile $f_{ij} = k_{ij} / k$, $(\forall) i = 1 \div n$ şi $(\forall) j = 1 \div p$. Se numesc *frecvenţe marginale* (sau *marje*) cantităţile $f_{i\bullet} = \sum_{j=1}^{p} f_{ij}$; $(\forall) i = 1 \div n$ şi $f_{\bullet j} = \sum_{i=1}^{n} f_{ij}$, $(\forall) j = 1 \div p$ îndeplinind condiţiile $\sum_{i=1}^{n} f_{i\bullet} = \sum_{j=1}^{p} f_{\bullet j} = \sum_{i=1}^{n} \sum_{j=1}^{p} f_{ij} = f = 1$.

Fie $\mathbf{F} = \{ f_{ij} \mid i = 1 \div n, \ j = 1 \div p\} \in \mathcal{M}_{n \times p}(\Re)$ matricea frecvenţelor relative. După cum este considerată privilegiată una sau alta dintre variabilele X sau Y sunt posibile două lecturi: pe linii, cu frecvenţele $\{ f_{ij} / f_{i\bullet} \}$, *profilurile-linie*, şi respectiv pe coloane, cu frecvenţele $\{ f_{ij} / f_{\bullet j} \}$, *profilurile-coloană*. Distanţele euclidiene între profilurile-linie, $d^2(i, \ell) = \sum_{j=1}^{p} (f_{ij} / f_{i\bullet} - f_{\ell j} / f_{\ell\bullet})^2$ şi respectiv între profilurile-coloană, $d^2(j, k) = \sum_{i=1}^{n} (f_{ij} / f_{\bullet j} - f_{ik} / f_{\bullet k})^2$, favorizează coloanele (respectiv liniile) care au o masă $f_{\bullet j}$ (respectiv $f_{i\bullet}$) importantă, adică modalităţile j (respectiv i) care sunt bine reprezentate în populaţia studiată. Pentru a remedia acest lucru cât şi din alte considerente, se ponderează fiecare diferenţă cu inversa masei coloanei, obţinîndu-se distanţa χ^2, $d^2_\chi(i, \ell) = \sum_{j=1}^{p} (1 / f_{\bullet j}) (f_{ij} / f_{i\bullet} - f_{\ell j} / f_{\ell\bullet})^2$ şi respectiv, $d^2_\chi(j, k) = \sum_{i=1}^{n} (1 / f_{i\bullet}) (f_{ij} / f_{\bullet j} - f_{ik} / f_{\bullet k})^2$.

Distanţa χ^2 este invariantă la agregarea liniilor, respectiv a coloanelor, cu acelaşi profil. Această proprietate poartă numele de *principiul echivalenţei distribuţiilor*. Echivalenţa distribuţională

permite agregarea a două modalităţi (ale aceleiaşi variabile) cu profiluri identice (în $\Re^p$ ele se confundă) într-o nouă modalitate cu o pondere sumată fără însă a afecta prin aceasta nici distanţele între modalităţile variabilei nou formate, nici distanţele între modalităţile celeilalte variabile. Din punct de vedere practic, această proprietate este fundamentală deoarece garantează o oarecare invarianţă a rezultatelor faţă de nomenclatura aleasă pentru construcţia modalităţilor unei variabile, cu condiţia regrupării modalităţilor asemănătoare. Nu se pierde astfel informaţia prin agregarea unor clase şi nu se câştigă informaţie prin divizarea claselor omogene.

ACS pe tablelul centrat este echivalentă cu *ACS* pe tabelul necentrat. Este o particularitate a *ACS*, în comparaţie cu *ACP*, echivalenţa dintre analiza realizată pe tabloul necentrat (adică cu originea în *O*) şi cea realizată pe tabloul centrat (adică cu originea în *G*) cu condiţia ignorării, în primul caz, a axei factoriale care uneşte pe *O* cu *G* (această axă este asociată valorii proprii egală cu unu, numită valoare proprie trivială).

Profilurile-linie şi profilurile-coloană au mase: $\{ f_{i\bullet} \mid i = 1 \div n \}$ şi respectiv $\{f_{\bullet j} \mid j = 1 \div p\}$ şi atunci matricile de pondere respective sunt $\mathbf{D}_n = \text{diag}\,(f_{i\bullet}) \in \mathcal{M}_{n\times n}(\Re)$, cu marjele liniilor pe diagonala principală şi $\mathbf{D}_p = \text{diag}\,(f_{\bullet j}) \in \mathcal{M}_{p\times p}(\Re)$, cu marjele coloanelor pe diagonala principală.

Metrica spaţiului $\Re^p$ este $\mathbf{M} = \mathbf{D}_p^{-1}$, metrica spaţiului $\Re^n$ este $\mathbf{M} = \mathbf{D}_n^{-1}$. Centrul de greutate al profilurilor-linie este $\mathbf{x}_{G\ell} = (f_{\bullet 1}, \ldots, f_{\bullet p})'$, centrul de greutate al profilelor-coloană este $\mathbf{x}_{Gc} = (f_{1\bullet}, \ldots, f_{n\bullet})$.

Reprezentările grafice ale proximităţilor între profiluri se fac, pe rînd, în cele două spaţii, în centrul de greutate al norului corespunzător.

Problemele de optimizat şi matricile de diagonalizat sunt:

- în $\Re^p$, spaţiul profilurilor-linie: $\max_{\mathbf{u}} \{ \sum_{i=1}^{n} f_{i\bullet}\, d^2(i, 0) \} \mid \mathbf{u}'\mathbf{D}_p^{-1}\mathbf{u} = 1$. Soluţia $\mathbf{u}$ este vectorul propriu al matricii $\mathbf{S} = \mathbf{F}'\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1}$, asociat celei mai mari valori proprii $\lambda \neq 1$.
- în $\Re^n$, spaţiul profilurilor-coloană: $\max_{\mathbf{v}} \{ \sum_{j=1}^{p} f_{\bullet j}\, d^2(j, 0) \} \mid \mathbf{v}'\mathbf{D}_n^{-1}\mathbf{v} = 1$. Soluţia $\mathbf{v}$ este vector propriu al matricii $\mathbf{T} = \mathbf{F}\mathbf{D}_p^{-1}\mathbf{F}'\mathbf{D}_n^{-1}$, asociat celei mai mari valori proprii $\lambda \neq 1$.

Axele factoriale:

- Matricile $\mathbf{S}$ şi $\mathbf{T}$ au aceleaşi valori proprii nenule: $\mathbf{S}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$, $\mathbf{u} \in \Re^p$ şi $\mathbf{T}\mathbf{v}_\alpha = \lambda_\alpha \mathbf{v}_\alpha$, $\mathbf{v} \in \Re^n$.
- Valorile proprii λ_α sunt subunitare ($\lambda_\alpha \leq 1$, $(\forall)\alpha$).
- Între vectorii proprii normaţi $\mathbf{u}_\alpha$ ai lui $\mathbf{S}$ asociaţi lui λ_α şi vectorii proprii normaţi $\mathbf{v}_\alpha$ ai lui $\mathbf{T}$ asociaţi aceleiaşi valori proprii există relaţiile: $\mathbf{v}_\alpha = (1/\sqrt{\lambda_\alpha})\mathbf{F}\mathbf{D}_p^{-1}\mathbf{u}_\alpha$ şi $\mathbf{u}_\alpha = (1/\sqrt{\lambda_\alpha})\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{v}_\alpha$.

De asemenea:

- $\sum_{j=1}^{p} f_{ij} / f_{i\bullet} = 1$, $(\forall) i = 1 \div n \Rightarrow$ în *ACS* punctele sunt conţinute în hiperplanul $\mathcal{H}$ de dimensiune $p - 1$ (pentru $\Re^p$).
- $\sum_{j=1}^{p} \mathbf{x}^j_{G\ell} = \sum_{j=1}^{p} f_{\bullet j} = 1 \Rightarrow G_\ell \in \mathcal{H}$.
- $\mathbf{x}'_{G\ell}\mathbf{M}\mathbf{x}_{G\ell} = 1 \Rightarrow G_\ell$ se află la distanţa *1* de origine.
- $\langle OG_\ell, \mathbf{x}_{G\ell} \rangle = 0 \Rightarrow \text{OG}_\ell \perp \mathcal{H}$

În analiza în raport cu originea, prima direcţie $\mathbf{u}_1$ este axa ce leagă originea de centrul de greutate al norului şi este ortonormală pe $\mathcal{H}$. Inerţia proiectată pe această axă este 1, egală cu distanţa dintre *O* şi G_ℓ deoarece toate punctele norului se proiectează pe această axă în acelaşi punct

G_ℓ. Următoarele $p-1$ axe ($\mathbf{u}_2$, ..., $\mathbf{u}_p$) conţinute în $\mathcal{H}$ constituie o bază, definind direcţii de inerţie maximă ale norului. Ele coincid cu primele $p-1$ axe ale *ACS* în raport cu G_ℓ şi ($\mathbf{u}^\ell_1$, $\mathbf{u}^\ell_2$, ..., $\mathbf{u}^\ell_p$), a p-a axă corespunde lui $\mathbf{u}_1 = OG_\ell$ şi nu indică nicio direcţie în $\mathcal{H}$ deoarece nu este conţinută în $\mathcal{H}$. Inerţia sa (valoarea proprie asociată) este nulă.

Coordonatele pe axele factoriale:

- În $\Re^p$: $\Psi_\alpha = D_n^{-1} F D_p^{-1} u_\alpha$ cu $\psi_{\alpha i} = \sum_{j=1}^{p} (f_{ij} / f_{i\bullet} f_{\bullet j}) u_{\alpha j}$.
- În $\Re^n$: $\Phi_\alpha = D_p^{-1} F' D_n^{-1} v_\alpha$ cu $\varphi_{\alpha j} = \sum_{i=1}^{n} (f_{ij} / f_{i\bullet} f_{\bullet j}) v_{\alpha i}$.

Coordonata modalităţii i a unei variabile reprezintă media modalităţilor j ale celeilalte variabile, ponderate de frecvenţele condiţionate ale profilului i. Analog, coordonata modalităţii j reprezintă media mulţimii modalităţilor i ponderate de frecvenţele condiţionate ale profilului j.

Relaţiile de tranziţie între spaţii (formulele quasi-baricentrice):

- $\mathbf{\Psi}_\alpha = (1 / \sqrt{\lambda_\alpha})\, \mathbf{D}_n^{-1} \mathbf{F}\, \mathbf{\Phi}_\alpha$ cu $\psi_{\alpha i} = (1/\sqrt{\lambda_\alpha}) \sum_{j=1}^{p} \varphi_{\alpha j} f_{ij} / f_{i\bullet}$
- $\mathbf{\Phi}_\alpha = (1 / \sqrt{\lambda_\alpha})\, \mathbf{D}_p^{-1} \mathbf{F}\, \mathbf{\Psi}_\alpha$ cu $\varphi_{\alpha j} = (1/\sqrt{\lambda_\alpha}) \sum_{i=1}^{n} \psi_{\alpha i} f_{ij} / f_{\bullet j}$

Astfel, modulo coeficientul de dilataţie ($1/\sqrt{\lambda_\alpha}$), proiecţiile punctelor unui nor sunt, pe o axă, coordonatele baricentrice ale proiecţiilor punctelor celuilalt nor. Relaţiile quasi-baricentrice justifică reprezentarea simultană a liniilor şi a coloanelor.

Rămâne în continuare valabilă observaţia de la *ACP* legată de faptul că distanţa dintre un punct-linie şi un punct-coloană este lipsită de sens deoarece acestea se situează în spaţii diferite. *ACS* oferă totuşi posibilitatea de a poziţiona şi interpreta un punct dintr-un nor în raport cu punctele din celălalt nor.

6. Analiza corespondenţelor multiple

Se dispune de observaţii privind s variabile calitative X^q ($q = 1 \div s$, $s > 2$), având respectiv modalităţile $\{ (1, ..., p_q) \}$. Modalităţile fiecărei variabile se exclud reciproc, fiecare modalitate este observată cel puţin o dată. Variabilele au fost observate simultan pe un eşantion de n indivizi, fiecare individ alege una şi numai una dintre modalităţile fiecărei variabile.

Analiza corespondenţelor multiple (ACM) este o tehnică de descriere a datelor calitative, folosită în special în anchetele unde întrebările sunt cu răspunsuri multiple.

Fie $p = \sum_{q=1}^{s} p_q$ numărul total de modalităţi ale celor s variabile nominale şi fie r_{iq} ($r_{iq} \leq p_q$) numărul modalităţii alese de individul i, dintre cele p_q modalităţi ale variabilei X^q. Se numeşte *tabel de date condensat* matricea $\mathbf{R} = \{r_{iq} \mid i = 1 \div n, q = 1 \div s\} \in \mathcal{M}_{n \times s}(\Re)$. Tabelul $\mathbf{R}$ care descrie cele s modalităţi alese de cei n indivizi nu este exploatabil, sumele pe linii sau pe coloane nu au sens, fiind necesar un alt mod de descriere a informaţiilor respective.

Pentru variabila nominală X^q, ($q = 1 \div s$) se numeşte *variabilă auxiliară a modalităţii* j ($j = 1 \div p_q$}) variabila $z_{ij,\,q}$ definită astfel: $z_{ij,\,q} = (z_{ij,\,q} = 0) \wedge [(r_{iq} \neq 0) \Rightarrow (z_{ij,\,q} = 1)]$; $(\forall) i \in [1, n]$.

Matricea $\mathbf{Z}_q = \{z_{ij,\,q} \mid i = 1 \div n, j = 1 \div p_q\}$; $(\forall) q \in [1, s]\} \in \mathcal{M}_{n \times pq}(\Re)$, în care fiecare linie conţine $p_q - 1$ zerouri şi un singur unu, se numeşte *matrice auxiliară a modalităţilor* variabilei nominale X^q. Matricea $\mathbf{Z} = [\mathbf{Z}_1, ..., \mathbf{Z}_q, ..., \mathbf{Z}_s] \in \mathcal{M}_{n \times p}(\Re)$, obţinută prin concatenarea matricilor $\mathbf{Z}_q$, se numeşte *tabel disjunctiv complet*. Avem: $z_{i\bullet} = \sum_{j=1}^{p} z_{ij,\,q} = s$; $z_{\bullet j} = \sum_{i=1}^{n} z_{ij,\,q}$ = numărul de indivizi care au ales modalitatea j a întrebării q; $n = \sum_{j=1}^{pq} z_{\bullet j} = z_q$; $z = \sum_{i=1}^{n} z_{i\bullet} = \sum_{q=1}^{s} z_q = \sum_{i=1}^{n} \sum_{j=1}^{p} z_{ij} = ns$ = efectivul total.

Matricea $\mathbf{B} = \mathbf{Z}'\mathbf{Z} \in \mathcal{M}_{p \times p}(\Re)$, se numeşte *tabel de contingenţă Burt* asociat tabelului disjunctiv complet $\mathbf{Z}$, având termenul general $b_{jj'} = \sum_{i=1}^{n} z_{ij} z_{ij'}$, marjele $b_j = \sum_{j'=1}^{p} b_{jj'} = s z_{\bullet j}$, efectivul total $b =$

$\sum^{p}_{j=1} b_j = s^2 n$ iar termenii de pe diagonală sunt efectivele $\{z_{\bullet j}\}$ ale modalităţilor întrebării q. Se notează cu $\mathbf{D} \in \mathcal{M}_{p\times p}(\Re)$ matricea diagonală definită de relaţiile: $d_{jj} = b_{jj} = z_{\bullet j}$ şi $d_{jj'} = 0$, $(\forall) j, j' \in [1, p], j \neq j'$.

Analiza corespondenţelor multiple (*ACM*) este analiza corespondenţelor simple (*ACS*) aplicată unui tabel disjunctiv complet. În consecinţă se aplică aceleaşi transformări tabelului de date pentru obţinerea profilurilor-linie sau profilurilor-coloană, aceleaşi ponderi ale punctelor funcţie de profilurile marginale, aceeaşi distanţă, distanţa χ^2. Indivizii sunt toţi afectaţi de o pondere identică $m_i = z_{i\bullet} / ns = 1/n$, $(i = 1 \div n)$, fiecare modalitate j este ponderată de frecvenţa sa, $m_j = z_{\bullet j} / ns$. Pe un tabel disjunctiv:

- în $\Re^n$ distanţa χ^2 între modalităţi, se scrie: $d^2(j, j') = \sum^{n}_{i=1} n(z_{ij} / z_{\bullet j} - z_{ij'} / z_{\bullet j'})^2$ şi este nulă dacă modalităţile j şi j' sunt alese de aceiaşi indivizi. Modalităţile de efectiv scăzut, adică cele alese de puţini indivizi, sunt depărtate faţă de celelalte modalităţi.
- în $\Re^p$ distanţa χ^2 între indivizi, se scrie $d^2(i, i') = (1/s)\sum^{p}_{j=1} (n / z_{\bullet j})(z_{ij} - z_{i'j})^2$ şi este nulă dacă indivizii i şi i' au ales aceleaşi modalităţi . Ei sunt cu atât mai depărtaţi cu cât au răspuns mai diferit. O modalitate j intervine în distanţa dintre indivizi cu atât mai mult cu cât masa ei este mai mică.

Reluând rezultatele analizei corespondenţelor simple şi notaţiile adoptate rezultă: $\mathbf{F} = (1/ns)\mathbf{Z}$, cu termenul general $f_{ij} = z_{ij}/ns$, $\mathbf{D}_p = (1/ns)\mathbf{D}$, cu termenul general $f_{\bullet j} = \delta_{ij}(z_{\bullet j}/ns)$ şi $\mathbf{D}_n = (1/n)\mathbf{I}_n$, cu termenul general $f_{i\bullet} = \delta_{ij}/n$).

Pentru a găsi axele factoriale $\mathbf{u}_\alpha$ se diagonalizează matricea $\mathbf{S} = \mathbf{F'D}^{-1}_{n}\mathbf{FD}^{-1}_{p} = (1/s)\mathbf{Z'ZD}^{-1}$ cu termenul general $s_{jj'} = (1/s \cdot z_{\bullet j'}) \sum^{n}_{i=1} z_{ij} z_{ij'}$:

- în $\Re^p$, ecuaţia celei de-a α-a axe factoriale $\mathbf{u}_\alpha$ este $(1/s)\mathbf{Z'ZD}^{-1}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$ şi ecuaţia celui de-al α-lea factor $\mathbf{\Phi}_\alpha = \mathbf{D}^{-1}\mathbf{u}_\alpha$ este $(1/s)\mathbf{D}^{-1}\mathbf{Z'Z}\mathbf{\Phi}_\alpha = \lambda_\alpha \mathbf{\Phi}_\alpha$;
- în $\Re^n$, ecuaţia celui de-al α-lea factor $\mathbf{\Psi}_\alpha$ este: $(1/s)\ \mathbf{ZD}^{-1}\mathbf{Z'}\mathbf{\Psi}_\alpha = \lambda_\alpha \mathbf{\Psi}_\alpha$.

Factorii $\mathbf{\Phi}_\alpha$ şi $\mathbf{\Psi}_\alpha$ (de normă λ_α) reprezintă coordonatele punctelor linie şi ale punctelor coloană pe axa factorială α.

Relaţiile de tranziţie între factorii $\mathbf{\Phi}_\alpha$ şi $\mathbf{\Psi}_\alpha$ sunt: $\mathbf{\Phi}_\alpha = (1 / \sqrt{\lambda_\alpha})\ \mathbf{D}^{-1}\mathbf{Z'}\mathbf{\Psi}_\alpha$; $\mathbf{\Psi}_\alpha = (1 / s\sqrt{\lambda_\alpha})\ \mathbf{Z}\mathbf{\Phi}_\alpha$.

Coordonatele factoriale ale individului $\boldsymbol{i}$ pe axa α sunt date de: $\psi_{\alpha, i} = (1/\sqrt{\lambda_\alpha})\Sigma^{p}_{j=1}(z_{ij}/z_{i\bullet})\varphi_{\alpha,j} = (1/s\sqrt{\lambda_\alpha})\Sigma_{j\in p(i)}\varphi_{\alpha,j}$ unde $p(i)$ desemnează mulţimea modalităţilor alese de individul i. Modulo coeficientul $1/\sqrt{\lambda_\alpha}$ individul i se găseşte proiectat în planul factorial principal în centrul de greutate (punctul de coordonate media aritmetică) al modalităţilor pe care le-a ales.

Coordonatele factoriale ale modalităţii $\boldsymbol{j}$ pe axa α sunt date de: $\varphi_{\alpha, j} = (1/\sqrt{\lambda_\alpha})\Sigma^{n}_{i=1}(z_{ij}/ z_{\bullet j})\ \psi_{\alpha, i} = (1/z_{\bullet j}\sqrt{\lambda_\alpha})\Sigma_{i\in n(j)}\ \psi_{\alpha, i}$ unde $n(j)$ desemnează mulţimea indivizilor care au ales modalitatea j.

În formulele de mai sus, modalităţile/indivizii nu sunt ponderaţi; coordonatele sunt simple medii aritmetice. Norul modalităţilor din $\Re^n$ poate fi descompus în s submulţimi, a $(q - a)$-a submulţime (subnor) corespunzând mulţimii p_q a modalităţilor variabilei q. Centrele de greutate ale celor s submulţimi ale norului modalităţilor din $\Re^n$ coincid cu centrul de greutate al norului global. Dacă tabelul $\mathbf{Z}$ nu este complet disjunctiv, adică dacă pentru cel puţin un individ nicio modalitate a unei întrebări nu a fost aleasă, modalităţile acelei variabile nu mai sunt centrate în centrul de greutate al norului global.

Coordonatele modalităţilor în $\Re^n$ sunt coloanele matricii $\mathbf{ZD}^{-1}$. Acestea generează un subspaţiu a cărui dimensiune este rangul lui $\mathbf{ZD}^{-1}$, adică $p - s + 1$. Rangul maxim al matricii $\mathbf{D}^{-1}\mathbf{Z'Z}$ de diagonalizat va fi deci $p - s + 1$. Dar, în analiza norului în raport cu originea O, prima bisectoare este vectorul propriu corespunzînd valorii proprii 1. În analiza în raport cu centrul de greutate G vor fi găsite $p - s$ valori proprii nenule. Alegând o bază în suportul norului, revine la a căuta valorile proprii ale unei matrici de ordin $p - s$.

Distanţa de la o modalitate $\boldsymbol{j}$ la centrul de greutate G este $d^2(\boldsymbol{j}, G) = (\boldsymbol{j} - G)'\mathbf{D}_n^{-1}(\boldsymbol{j} - G) = n / z_{\bullet j} - 1$.

Inerţia unei modalităţi $\boldsymbol{j}$ este, prin definiţie, $I(\boldsymbol{j}) = m_j d^2(\boldsymbol{j}, G)$ cu $m_j = z_{\cdot j} / ns$ sau, efectuând substituţiile, $I(\boldsymbol{j}) = (1/s)(1 - z_{\cdot j} / n)$. Inerţia unei modalităţi este cu atât mai mare cu cât efectivul $z_{\cdot j}$ al acestei modalităţi, adică numărul de indivizi care au ales-o, este mai mic. Maximul $1/s$ va fi atins pentru modalităţile de efectiv nul. În consecinţă, în momentul codificării, se va evita introducerea unor modalităţi susceptibile de a fi alese de puţini indivizi pentru a nu introduce perturbaţii în primele axe factoriale.

Inerţia unei întrebări q este, prin definiţie: $I(q) = \Sigma^{pq}_{j=1} I(j) = (1/s)(p_q - 1)$. Inerţia unei întrebări este cu atât mai mare cu cât numărul de modalităţi asociat, p_q, este mai mare. Minimul $1/s$ este atins de întrebările cu doar două modalităţi de răspuns. In consecinţă, dacă se doreşte ca întrebările să joace un rol aproximativ egal, se va echilibra sistemul de întrebări (variabilele vor fi "decupate " într-un număr egal de modalităţi).

Inerţia totală este $I = \Sigma^{s}_{q=1} I(q) = \Sigma^{p}_{j=1} (z_{\cdot j} / ns) d^2(\boldsymbol{j}, G) = p/s - 1$, ($\Sigma^{s}_{q=1} p_q = p$). În particular $I = 1$ dacă toate întrebările au două modaliăţi de răspuns, adică $p = 2s$. În consecinţă depinzînd exclusiv de numărul de întrebări şi de modalităţile asociate acestora, inerţia globală nu are, în cazul *ACM* (ca şi în cazul *ACP* normat, de altfel), nicio semnficaţie statistică, deoarece nu depinde de legătura între variabile.

A spune că există afinităţi între răspunsuri este acelaşi lucru cu a spune că există indivizi care au profile asemănătoare din punct de vedere al atributelor alese spre a-i descrie. Ţinînd cont de distanţele între elementele tabelului disjunctiv complet şi de relaţiile baricentrice se poate interpreta că:

- proximitatea între indivizi semnifică faptul că au ales global aceleaşi modalităţi ca răspuns la întrebările puse;
- proximitatea între modalităţi ale unor întrebări diferite semnifică faptul că ele au fost alese ca răspuns de grupe de indivizi asemănători (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 (modalităţile unei aceleiaşi variabile se exclud).

Regulile de interpretare a rezultatelor privind elementele active ale unei *ACM* sunt asemănătoare cu cele corespunzătoare unei *ACS*.

7. Analiza canonică

Obiectivul general al analizei canonice este explorarea relaţiilor ce pot exista între două grupe de variabile cantitative observate pe aceeaşi mulţime de indivizi. De exemplu, analizele medicale efectuate, pe acelaşi eşantion de indivizi, de către două laboratoare diferite. Scopul analizei canonice este de a compara cele două grupuri de variabile pentru a vedea dacă acestea descriu acelaşi fenomen, caz în care prospectorul de date ar putea renunţa la unul din cele două grupe de variabile.

Fie n numărul indivizilor, p numărul variabilelor din primul grup şi q numărul variabilelor din cel de al doilea grup, fie $\mathbf{X} \in \mathcal{M}_{n\times p}(\Re)$ matricea conţinând observaţiile relative la primul grup de variabile şi fie $\mathbf{Y} \in \mathcal{M}_{n\times q}(\Re)$ matricea conţinând observaţiile relative la cel de al doilea grup de variabile.

Coloana j a matricii $\mathbf{X}$, ($j = 1 \div p$), conţine observaţiile x_{ij} asupra variabilei X^j din primul grup iar coloana k a matricii $\mathbf{Y}$, ($k = 1 \div q$), conţine observaţiile y_{ik} asupra variabilei Y^j din al doilea grup, pentru $i = 1 \div n$. Din motive de comoditate matricile $\mathbf{X}$ şi $\mathbf{Y}$ se presupun centrate şi reduse şi de asemenea se presupune $p \le q \le n$, rang($\mathbf{X}$) = p şi rang($\mathbf{Y}$) = q.

În spaţiul $\mathcal{F}$ al variabilelor, respectiv $\Re^n$ înzestrat cu o bază canonică $\mathtt{F}$ şi cu o metrică $\mathbf{M}$, se pot defini două subspaţii vectoriale:

- $\mathcal{F}_{\mathbf{X}}$ generat de vectorii $\mathbf{x}_j$ ($j = 1 \div p$), în general de dimensiune p şi

- $\mathcal{F}_{\mathbf{Y}}$ generat de vectorii $\mathbf{y}_k$ ($k = 1 \div q$), în general de dimensiune q.

În general $\mathbf{M} = \mathbf{I}$, uneori $\mathbf{M} = \text{diag}(\rho_i)$ sau $\mathbf{M} = (1/n)\mathbf{I}$ dacă ponderile sunt egale.

Pentru indivizi pot fi luate în consideraţie două spaţii vectoriale:

- $\mathcal{E}_1 = (\Re^p, \mathtt{E}, \mathbf{M})$, generat de vectorii $\mathbf{x}_i$ ($i = 1 \div n$) şi
- $\mathcal{E}_2 = (\Re^q, \mathtt{E}, \mathbf{M})$, generat de vectorii $\mathbf{y}_i$ ($i = 1 \div n$)

Se caută, pentru început, un cuplu de variabile (V^1, W^1), V^1 combinaţie liniară a variabilelor X^j (deci un element din $\mathcal{F}_{\mathbf{X}}$), normată şi W^1 combinaţie liniară a variabilelor Y^k (deci un element din $\mathcal{F}_{\mathbf{Y}}$), normată astfel încât V^1 şi W^1 să fie cele mai corelate posibil. Se caută, apoi, cuplul normat (V^2, W^2), V^2 combinaţie liniară a variabilelor X^j, necorelată cu V^1 şi W^2 combinaţie liniară a variabilelor Y^k, necorelată cu V^1, astfel încât V^2 şi W^2 să fie cele mai corelate posibil. Se continuă în acelaşi mod rezultând, în final, un număr de p cupluri de variabile (V^s, W^s), $s = 1 \div p$.

Variabilele V^s constituie o bază ortonormată în $\mathcal{F}_{\mathbf{X}}$ (fiind combinaţii liniare de variabile centrate, sunt centrate; fiind necorelate sunt ortogonale în metrica identitate). Variabilele W^s constituie un sistem ortonormat al lui $\mathcal{F}_{\mathbf{Y}}$ (ele nu formează o bază decât dacă $q = p$). Cuplurile (V^s, W^s), şi în special primele dintre ele, ţin cont de legăturile liniare dintre cele două grupe de variabile iniţiale.

Variabilele V^s şi W^s se numesc *variabile canonice*. Corelaţiile lor succesive se numesc *coeficienţi de corelaţie canonică* (sau *corelaţii canonice*) şi se notează cu δ_s ($1 \geq \delta_1 \geq \delta_2 \geq ... \geq \delta_p \geq 0$). Orice variabilă canonică $V^{s'}$ este, prin construcţie, necorelată (deci ortogonală) cu celelalte variabile canonice V^s, $s \neq s'$. Deasemenea $V^{s'}$ este necorelată cu W^s, dacă $s \neq s'$ şi bineînţeles proprietatea este adevărată pentru orice variabilă $W^{s'}$ în raport cu variabilele V^s, $s \neq s'$. Sistemul de variabile W^s ($s = 1 \div p$) poate fi completat pentru a se obţine o bază ortonormată în $\mathcal{F}_{\mathbf{Y}}$, în care ultimele variabile W^s ($s = p+1 \div q$) sunt asociate cu coeficienţi de corelaţie canonică nuli ($\delta_s = 0$, $s = p+1 \div q$).

Fie $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ şi $\mathbf{P}_{\mathbf{Y}} = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'$ matricile proiecţiilor ortogonale pe subspaţiile $\mathcal{F}_{\mathbf{X}}$ şi $\mathcal{F}_{\mathbf{Y}}$, ale lui $\mathcal{F}$ înzestrat cu metrica $\mathbf{I}$. Atunci:

- vectorii V^s sunt vectorii proprii normaţi ai matricii $\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{Y}}$ corespunzători valorilor proprii λ_s ordonate descrescător, $\lambda_s \in [0, 1]$;
- vectorii W^s sunt vectorii proprii normaţi ai matricii $\mathbf{P}_{\mathbf{Y}}\mathbf{P}_{\mathbf{X}}$ corespunzători aceloraşi valori proprii λ_s;
- coeficienţii de corelaţie canonică sunt $\delta_s = \sqrt{\lambda_s}$, $s = 1 \div p$.

Reprezentările grafice ale rezultatelor analizei canonice se fac întro dimensiune d redusă, $1 \leq d \leq p$.

Fie $\mathbf{v}^s \in \mathcal{F}_{\mathbf{X}}$ şi $\mathbf{w}^s \in \mathcal{F}_{\mathbf{Y}}$ vectorii asociaţi variabilelor canonice V^s şi respectiv W^s. În $\mathcal{F}_{\mathbf{X}}$ se consideră baza ortonormată ($\mathbf{v}^1$, ..., $\mathbf{v}^p$) restrânsă la ($\mathbf{v}^1$, ..., $\mathbf{v}^d$). Se reprezintă fiecare dintre variabilele iniţiale X^j cu ajutorul coordonatelor acestora pe $\mathbf{v}^s$ obţinute calculând produsele scalare $\langle \mathbf{x}_j, \mathbf{v}^s \rangle$.

Variabilele X^j fiind centrate şi reduse vectorii $\mathbf{x}_j$ sunt centraţi şi normaţi (la fel şi vectorii $\mathbf{v}^s$), astfel încât aceste produse scalare sunt egale cu corelaţiile dintre variabilele iniţiale X^j şi variabilele canonice V^s, cu coeficient n deoarece s-a considerat metrica $\mathbf{I}$. În acelaşi spaţiu pot fi reprezentate şi variabilele Y^k ale celuilalt grup proiectând mai întâi vectorii $\mathbf{y}_k$ în $\mathcal{F}_{\mathbf{X}}$, cu ajutorul lui $\mathbf{P}_{\mathbf{X}}$, apoi luând produsul scalar al acestor proiecţii cu vectorii $\mathbf{v}^s$, $\langle \mathbf{P}_{\mathbf{X}}(\mathbf{y}_k), \mathbf{v}^s \rangle = \langle \mathbf{y}_k, \mathbf{P}_{\mathbf{X}}(\mathbf{v}^s) \rangle = \langle \mathbf{y}_k, \mathbf{v}^s \rangle$ deasemenea egale cu corelaţiile dintre variabilele iniţiale Y^k şi variabilele canonice V^s.

În măsura în care graficul astfel obţinut este „bun” el poate fi utilizat pentru a interpreta relaţiile (proximităţi, opoziţii, depărtări) dintre cele două mulţimi de variabile. Prin construcţie, acest grafic reprezintă corelaţiile dintre variabilele canonice V^s şi variabilele iniţiale X^j şi Y^k, corelaţii care stau la baza interpretării lui. Interpretarea poate fi facilitată utilizând coeficienţii de corelaţie liniară dintre variabilele X^j, dintre variabilele Y^k şi dintre variabilele X^j şiY^k. Toţi aceşti coeficienţi sunt furnizaţi de produsul software.

În mod simetric, în $\mathcal{F}_\mathbf{Y}$ se restrânge sistemul ($\mathbf{w}^1$, ..., $\mathbf{w}^p$) la primele variabile ($\mathbf{w}^1$, ..., $\mathbf{w}^d$), în raport cu care se reprezintă la fel de bine variabilele iniţiale, atât X^j cât şi Y^k , conform aceluiaşi principiu descris mai sus (coordonatele sunt corelaţii).

Cele două grafice (în $\mathcal{F}_\mathbf{X}$ şi în $\mathcal{F}_\mathbf{Y}$) având aceeaşi calitate şi conducând la aceleaşi interpretări este suficient unul singur pentru a interpreta rezultatele unei analize.

În fiecare din spaţiile relative la indivizi ($\mathcal{E}_1$ şi $\mathcal{E}_2$), se poate deasemenea obţine câte o reprezentare grafică a acestor indivizi în dimensiunea d, cele două reprezentări fiind comparabile (cu atât mai comparabile cu cât corelaţiile canonice sunt mai mai ridicate). Coordonatele indivizilor pe axele canonice în aceste două reprezentări sunt date de liniile matricilor $\mathbf{V}_d \in \mathcal{M}_{n\times d}(\Re)$ (în $\mathcal{E}_1$) şi $\mathbf{W}_d \in \mathcal{M}_{n\times d}(\Re)$ (în $\mathcal{E}_2$), ale căror coloane conţin coordonatele primelor d variabile canonice în baza canonică F a spaţiului $\mathcal{F}$.

8. Concluzii

Explorarea datelor este utilizată în orice domeniu, atunci când datele sunt mult prea multe pentru a mai putea fi înţelese de o minte omenească. În studiile care vizează un număr important de variabile, respectiv indivizi reprezentabili în spaţii de mari dimensiuni, o dificultate majoră o constituie obţinerea unei reprezentări grafice adecvate a cărei vizualizare şi interpretare să faciliteze înţelegerea structurii datelor analizate. Analiza în componente principale are un rol esenţial fiind metoda care serveşte drept fundament teoretic şi pentru celelalte metode de explorare multidimensională numite factoriale. Obiectivele realizate de analiza în componente principale permit utilizarea prealabilă a acestei metode în cazul altor metode care preferă, fie variabile ortogonale (regresia liniară), fie un număr redus de intrări (reţelele neuronale).

Analiza discriminantă este una dintre tehnicile de analiză multidimensională cele mai folosite în practică: diagnostic automat, controlul calităţii, previziunea riscului, recunoaşterea formelor. Scopul analizei discriminante îl constituie studierea legăturilor între variabilele explicative şi clasele partiţiei şi definirea funcţiilor discriminante care vor permite, într-o etapă ulterioară, afectarea de noi indivizi la aceste clase. Mărirea puterii discriminante a axelor poate fi reclamată de datele problemei, cu scopul de a putea „vedea” o anumită structură în date. Determinarea axelor discriminante poate servi şi ca o tehnică de reducere a dimensiunii spaţiului variabilelor, prin această tehnică fiind selectate cele mai relevante caracteristici. Reducerea dimensiunii poate fi deasemenea impusă şi de necesitatea vizualizării claselor prin proiectarea datelor într-un spaţiu cu una sau două dimensiuni.

Analiza corespondenţelor simple este o metodă descriptivă ce revine la efectuarea unei analize a unui nor de puncte ponderate într-un spaţiu cu o metrică specială. Analiza corespondenţelor simple este în principal utilizată pentru tabele mari de date, comparabile între ele (dacă este posibil exprimate în aceeaşi unitate de măsură) şi permite analiza şi interpretarea datelor calitative complexe întâlnite în general în domeniul ştiinţelor umane şi sociale, dar nu numai. Analiza corespondenţelor simple este conceptual similară cu analiza în componente principale, se poate aplica unor tabele de contingenţă şi tratează în mod echivalent atât liniile, cât şi coloanele. Abordările curente constau în a defini analiza corespondenţelor simple ca fiind rezultatul a două analize în componente principale (pentru profiluri-linii şi pentru profiluri-coloane) utilizând metrica χ^2.

Analiza corespondenţelor multiple este o generalizare posibilă a analizei corespondenţelor

simple, are însă proceduri de calcul şi reguli de interpretare specifice şi se pretează la un număr mare de aplicaţii. Ea este în mod deosebit adaptată la descrierea tabelelor mari de variabile nominale, specifice anchetelor socio-economice sau medicale, modalităţile acestora fiind de cele mai multe ori răspunsuri la întrebări. Deasemenea este de multe ori utilizată pentru determinarea unor scoruri prealabile unor metode de clasificare (metoda norilor dinamici).

Analiza canonică este considerată, pe plan teoretic, una din metodele descriptive multidimensionale centrale deoarece generalizează diverse alte metode şi de asemenea poate fi privită ca un caz particular de analiză în componente principale a două pachete de variabile într-un spaţiu înzestrat cu o metrică specială. Multă vreme analiza canonică, nefiind uşor aplicabilă, a avut puţine aplicaţii practice, dar lucrurile s-au schimbat mai ales datorită dezvoltării, la mijlocul anilor 1990, a regresiei *PLS* („partial least squares") metodă destul de apropiată cu analiza canonică şi ulterior, prin apariţia datelor de expresie genomică (biochip-uri) combinate cu variabile biologice pentru situaţii tipice de analiză canonică.

Analiza canonică prezintă anumite analogii atât cu analiza în componente principale, privind construirea şi interpretarea graficelor, cât şi cu regresia liniară, privind natura datelor. Analiza canonică este apropiată de regresia liniară multiplă (explicarea unei variabile cantitative prin o mulţime de alte variabile cantitative) metodă pentru care analiza canonică constituie de altfel o generalizare (dacă unul din grupuri se reduce la o singură variabilă se regăseşte regresia). Deasemenea, când unul din cele două grupuri de variabile este înlocuit de variabilele auxiliare (modalităţile) unei variabile calitative se regăseşte analiza factorială discriminantă, iar când fiecare din cele două grupuri este înlocuit cu variabilele auxiliare ale unei variabile calitative se regăseşte analiza corespondenţelor simple. Mai mult, există anumite generalizări ale analizei canonice la mai mult de două grupuri de variabile cantitative, iar acestea permit atât regăsirea analizei corespondenţelor multiple (înlocuind fiecare grup prin variabilele auxiliare ale unei variabile calitative), cât şi regăsirea analizei în componente principale (lăsând câte o singură variabilă cantitativă în fiecare grup).

Analiza în componente independente, mai recentă, rezultată din fizica semnalului şi cunoscută iniţial ca „metodă de separare oarbă a sursei", este mai apropiată, intuitiv, de metodele de clasificare nesupravegheată. Clasificarea automată, analiza factorială discriminantă sau analiza discriminantă permit identificarea, în interiorul unei populaţii, a grupurilor omogene din punctul de vedere al variabilelor studiate.

BIBLIOGRAFIE

1. **BACCINI**, **A.; BESSE**, **P.:** Data mining / Exploration Statistique. Toulouse: INSA, 2010, 111 p.
2. **BENZÉCRI**, **J.-P.:** Histoire et Préhistoire de l'Analyse des données: Partie 5. Les Cahiers de l'analyse des données, vol. 2, no.1, 1977, pp. 9-40.
3. **ENĂCHESCU**, **D.:** Data Mining - metode şi aplicaţii. Bucureşti: Editura Academiei Române, 2009, 277 p.
4. **FALGUEROLLES**, **A.:** L'analyse des données: before and around., Electronic Journal for History of Probability and Statistics, vol. 4, no. 2, dec. 2008.
5. **FILIP**, **F. G.:** Decizie asistată de calculator: decizii, decidenţi - metode de bază şi instrumente informatice asociate. Ediţia a 2-a, rev. Bucureşti: Editura Tehnică, 2005, 376 p.
6. **FILIP**, **F. G.:** Sisteme suport pentru decizii. Ediţia a 2-a, rev. Bucureşti: Editura Tehnică, 2007, 364 p.
7. **FRANCIS**, **M.:** Future telescope array drives development of exabyte processing. 2012 (http://arstechnica.com/science/2012/04/future-telescope-array-drives-development-of-exabyte-processing/ , accesat 2012-12-18).

8. **GORUNESCU**, **F.:** Data Mining, Concepts, Models and Techniques. Springer-Heidelberg, series Intelligent Systems Reference Library, 2011, 372 p.
9. **HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J.:** The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Springer-Verlag, Springer Series in Statistics, 2008, 763 p.
10. **HILBERT**, **M.; LOPEZ**, **P.:** The World's Technological Capacity to Store, Communicate, and Compute Information. Science, Vol. 332, 6025, apr. 2011 p. 60-65.
11. **IBM** – Bringing big data to the enterprise (http://www-01.ibm.com/software/data/bigdata/, accesat 2012-12-18).
12. **MĂRGINEAN**, **N.:** Sisteme inteligente pentru asistarea deciziilor. Editura Risoprint, Cluj-Napoca, 2006, 239 p.
13. **PENG**, **Y.; KOU**, **G.; SHI**, **Y.; CHEN**, **Z.:** A descriptive framework for the field of data mining and knowledge discovery. International Journal of Information Technology & Decision Making, Vol. 7, No. 4, 2008, pp. 639-682.
14. **TAN, P-N.; STEINBACH**, **M.; KUMAR**, **V.:** Introduction to Data Mining. Addison-Wesley, 2006, 769 p.
15. **TUFFERY**, **S.:** Data mining et statistique décisionnelle, 3ème Edition. Editions TECHNIP, 2010, 705 p.
16. **WATTERS**, **A.:** The Age of Exabytes: Tools and Approaches for Managing Big Data. Hewlett-Packard Development Company, 2010 (http://readwrite.com/2012/03/05/big-data , accesat 2012-12-18).
17. **WU**, **X.; KUMAR**, **V. (ed.):** The Top Ten Algorithms in Data Mining. Chapman & Hall / CRC DMKD Series, 2009, 232 p.