

METODE EXPLORATORII MULTIDIMENSIONALE

Cornel Lepădatu

cornel_lepadatu@biblacad.ro

Academia Română București

Biblioteca Academiei Române

Rezumat: Explorarea datelor este un ansamblu de metode destinate descrierii și analizei datelor multidimensionale și utilizate în orice domeniu, atunci când datele sunt mult prea multe pentru a mai putea fi înțelese de o minte omenească. Unele dintre metode, ajută la evidențierea relațiilor care pot exista între diferite date și elaborează informații statistice care permit o descriere mai succintă a informației conținute în aceste date. Altele, permit regrupări ale datelor în scopul de a face să apară clar ceea ce le face omogene și astfel de a le înțelege și 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 și care se exprimă fără să presupună a priori un model probabilist. Aceste metode permit, în special, prelucrarea și sinteza informației din tabelele de date de mari dimensiuni pe baza estimării corelațiilor dintre variabilele studiate, instrumentele statistice utilizate fiind matricea corelațiilor sau matricea de varianță-covarianță.

Un demers exploratoriu îi permite prospectului de date să abordeze unul dintre principalele obiective ale „data mining” și anume explorarea multidimensională a datelor sau reducerea de dimensiune: reprezentarea grafică, deducerea unei submulțimi de variabile reprezentative sau a unei mulțimi de componente prealabile pentru alte metode.

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

Cuvinte cheie: analiza canonică, analiza corespondențelor multiple, analiza corespondențelor 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 și analiza datelor multidimensionale. Unele dintre metode, ajută la evidențierea relațiilor care pot exista între diferite date și elaborează informații statistice care permit o descriere mai succintă a informației conținute în aceste date. Altele, permit regrupări ale datelor în scopul de a face să apară clar ceea ce le face omogene și astfel de a le înțelege și de a le defini mai bine.

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

interesante aspecte ale structurii acestora, computerele fiind acelea care au făcut aceste metode operaționale și 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 evidenția relații dificil de sesizat de o analiză directă a datelor dar, mai important și 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 și care se exprimă fără să presupună a priori un model probabilist. Aceste metode permit, în special, prelucrarea și sinteza informației din tabelele de date de mari dimensiuni pe baza estimării corelațiilor dintre variabilele studiate, instrumentele statistice utilizate fiind matricea corelațiilor sau matricea de varianță-covarianță.

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 contingență, 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 soldați, etc. sau din 1696, când Nicolas Lamoignon Basville, intendent al regelui Ludovic al XIV-lea, enumeră și caracterizează mânăstiri și biserici din regiunea Languedoc. Printre fondatorii metodelor moderne de analiză a datelor se regăsesc Jean-Paul Benzécri, Louis Guttman, Chikio Hayashi, Douglas Carroll și R.N. Shepard [2].

Într-un proces de explorare a datelor și descoperire a cunoștințelor („data mining”) un prim demers, inevitabil, constă în efectuarea unei explorări a acestor date: alura distribuțiilor, prezența datelor atipice, corelații și coerență, transformări eventuale ale datelor. Demersul descriptiv și exploratoriu permite realizarea de rezumate și grafice mai mult sau mai puțin elaborate, descrierea mulțimilor de date și stabilirea de relații între variabile, fără a acorda un rol privilegiat vreunei variabile și care, folosite în mod adecvat, se pot dovedi extrem de utile pentru numeroase probleme și situații din domeniul decizional [5, 6, 12]. Concluziile obținute privesc doar datele studiate, fără a fi generalizate la o populație mai largă. Demersul exploratoriu se sprijină, în mod esențial, pe noțiuni elementare (medie și dispersie), pe reprezentări grafice și pe tehnici descriptive multidimensionale. Metodele exploratorii determină subspații de reprezentare (sau factoriale), de dimensiuni mici, care aproximează cel mai bine norii de puncte-indivizi sau de puncte-variabile, astfel încât vecinătățile măsurate în aceste spații să reflecte cât mai exact proximitățile reale.

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

2. Elemente preliminare

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

Matricea diagonală $\mathbf{D} = \text{diag}(\rho_1, \dots, \rho_n) \in \mathcal{M}_{n \times n}(\mathfrak{R})$, se numește *matrice de ponderi*, pentru cazul indivizilor echiponderați $\mathbf{D} = (1/n) \mathbf{I}_n$ 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} , $\mathbf{x}_j \in \mathfrak{R}^n$, iar un *individ* i , ($i = 1 \div n$) prin vectorul-linie i al matricii \mathbf{X} , $\mathbf{x}_i \in \mathfrak{R}^p$. Vectorii-coloană ai matricii \mathbf{X} definesc un nor de p puncte-variabile în \mathfrak{R}^n iar vectorii-linie definesc un nor de n puncte-indivizi în \mathfrak{R}^p .

Media de selecție a unei variabile j este definită prin $\bar{x}_j = \sum_{i=1}^n \rho_i x_{ij}$, iar dispersia de selecție prin $s_j^2 = \sum_{i=1}^n \rho_i (x_{ij} - \bar{x}_j)^2$. Vectorul $\mathbf{g}' = (x_1, \dots, x_p)$ se numește *punct mediu* (sau *centru de greutate*) al norului de puncte-indivizi, $\mathbf{g} = \mathbf{X}'\mathbf{D}\mathbf{1}_n$, unde $\mathbf{1}'_n = (1, \dots, 1) \in \mathfrak{R}^n$.

Matricea de varianță-covarianță $\mathbf{V} \in \mathcal{M}_{p \times p}(\mathfrak{R})$, asociată matricii \mathbf{X} , este:

$$\mathbf{V} = \{ (v_{jk}) \mid v_{jk} \equiv \text{cov}(\mathbf{x}_j, \mathbf{x}_k) = \sum_{i=1}^n \rho_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), j = 1 \div p, k = 1 \div p \}$$

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

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

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

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

Se numește *tabel centrat-reduc*, asociat matricii \mathbf{X} , matricea $\mathbf{Z} \in \mathcal{M}_{n \times p}(\mathfrak{R})$:

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

$$\text{Avem: } \mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{X} - \mathbf{g}\mathbf{g}' = \mathbf{Y}'\mathbf{D}\mathbf{Y} \text{ și } \mathbf{R} = \mathbf{D}_{1/s}\mathbf{V}\mathbf{D}_{1/s} = \mathbf{Z}'\mathbf{D}\mathbf{Z} = \sum_{i=1}^n \rho_i \mathbf{x}_i \mathbf{x}_i'.$$

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

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

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

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

Metricile cele mai uzitate, în spațiul \mathcal{E} al indivizilor, sunt: \mathbf{I}_p , ce induce produsul scalar uzual și distanța euclidiană și \mathbf{D}_{1/s^2} , care conduce la adimensionalizarea variabilelor deoarece fiecare valoare este împărțită cu abaterea standard de selecție a variabilei corespunzătoare (x_{ij} / s_j). Metrica \mathbf{I}_p dă fiecărei variabile aceeași importanță independent de dispersia sa, utilizarea ei va privilegia variabilele cu dispersie mare pentru care diferențele între indivizi sunt mari și va neglija diferențele între celelalte variabile, în schimb metrica \mathbf{D}_{1/s^2} echilibrează influența variabilelor transformându-le în variabile cu dispersia de selecție unu. Utilizarea metricii \mathbf{D}_{1/s^2} pentru tabelul centrat \mathbf{Y} revine la folosirea metricii \mathbf{I}_p pentru tabelul centrat-reduc \mathbf{Z} . Matricea \mathbf{W} a produsului scalar între indivizi poate fi întotdeauna exprimată în funcție de metrica \mathbf{I}_p adică $(\exists) \mathbf{T} : \mathbf{W} = (\mathbf{X}\mathbf{T}')\mathbf{I}_p(\mathbf{T}\mathbf{X}')$ și atunci \mathbf{W} este matricea produsului scalar al tabelului $\mathbf{X}\mathbf{T}'$ față de metrica \mathbf{I}_p . Dacă $\mathbf{M} = \text{diag}(m_1, \dots, m_p)$, atunci $d^2(\mathbf{x}_i, \mathbf{x}_\ell) = \sum_{j=1}^p m_j (x_{ij} - x_{\ell j})^2$ iar coeficienții $\{\sqrt{m_j}\}_{j=1}^p$ pot fi considerați ca ponderi ale variabilelor x_j în distanța dintre indivizi.

Ipozeza fundamentală a unui demers exploratoriu [3, 9, 14, 15] este aceea că întreaga informație este conținută în distanțele dintre punctele-indivizi ale unui nor, respectiv dispersia punctelor din nor. Se numește *inerție totală (globală)* a norului de puncte-indivizi media ponderată a pătratelor distanțelor de la punctele-indivizi la centrul de greutate \mathbf{g} al norului, adică:

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

Prin analogie, *inerția într-un punct oarecare*, $\mathbf{a} \in \mathfrak{R}^p$, este: $I_a = \sum_{i=1}^n \rho_i \|\mathbf{x}_i - \mathbf{a}\|_{\mathbf{M}}^2$ și conform formulei lui Huygens: $I_a = I_g + (\mathbf{g} - \mathbf{a})' \mathbf{M} (\mathbf{g} - \mathbf{a}) = I_g + \|\mathbf{g} - \mathbf{a}\|_{\mathbf{M}}^2$.

Pentru un nor de puncte-indivizi dat, centrul de greutate \mathbf{g} al norului minimizează inerția totală. Inerția totală este media pătratelor distanțelor dintre punctele-indivizi $2I_g = \sum_{i=1}^n \sum_{\ell=1}^n \rho_i \rho_\ell \|\mathbf{x}_i - \mathbf{x}_\ell\|_{\mathbf{M}}^2$.

Notând cu $\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii}$ urma matricii \mathbf{A} , avem $I_g = \text{tr}(\mathbf{M}\mathbf{V}) = \text{tr}(\mathbf{V}\mathbf{M})$ și pentru cazul $\mathbf{g} = \mathbf{0}$ avem $I_g = \text{tr}(\mathbf{W}\mathbf{D}) = \text{tr}(\mathbf{D}\mathbf{W})$.

Dacă $\mathbf{M} = \mathbf{I}_p$, atunci $I_g = \sum_{j=1}^p s^2(\mathbf{x}_j)$, adică inerția totală este egală cu suma dispersiilor de selecție a celor p variabile.

Dacă $\mathbf{M} = \mathbf{D}_{1/s_2}$, atunci $I_g = \text{tr}(\mathbf{D}_{1/s_2}\mathbf{V}) = \text{tr}(\mathbf{D}_{1/s_2}\mathbf{V}\mathbf{D}_{1/s_2}) = \text{tr}(\mathbf{R}) = \sum_{j=1}^p r_{jj} = \sum_{j=1}^p 1 = p$, adică inerția totală este egală cu numărul variabilelor și 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 spațiu vectorial $\mathcal{F} \subset \mathfrak{R}^n$ cu baza canonică $\mathcal{F} = (\mathbf{f}_1, \dots, \mathbf{f}_p)$, numit *spațiul variabilelor*. Metrica utilizată în spațiul \mathcal{F} , al variabilelor, este matricea diagonală a ponderilor indivizilor, $\mathbf{D} = \text{diag}(\rho_1, \dots, \rho_n) \in \mathcal{M}_{n \times n}(\mathfrak{R})$.

Pentru variabilele centrate (matricea \mathbf{Y}):

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

Operația de centrare a tabelului \mathbf{X} are în spațiile \mathfrak{R}^p și \mathfrak{R}^n interpretări geometrice diferite.

În \mathfrak{R}^p această transformare echivalează cu o translație a originii axelor în centrul de greutate (punctul mediu) al norului.

În \mathfrak{R}^n această transformare este o proiecție pe hiperplanul care trece prin originea axelor și este ortogonal pe dreapta ce trece prin originea axelor având ca parametri directori $\{\rho_i \mid i = 1 \div n\}$. Matricea $\mathbf{P} = \mathbf{I}_n - \mathbf{I}_n \mathbf{1}'_n \mathbf{D}$, asociată acestei transformări, este matricea proiecției \mathbf{M} -ortogonale pe subspațiul generat de vectorii coloană liniar-independenți ai matricii \mathbf{Y} . Coordonatele acestor vectori satisfac relația $\sum_{i=1}^n \rho_i v_{ij} = 0$, $(\forall) j = 1 \div p$ reprezentând ecuația unui hiperplan în \mathfrak{R}^n care trece prin originea axelor și 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 corelație*. Planurile în care vor fi proiectate variabilele intersectează sfera după cercuri diametrale, numite *cercuri de corelație*, de rază 1 și în interiorul cărora se află proiecțiile punctelor-variabile.

Dacă în spațiul indivizilor interesează distanța dintre puncte, în spațiul variabilelor interesează unghiurile dintre ele. Proximitatea între punctele-variabile se interpretează în termeni de corelații. Sistemul de proximități dintre două puncte-variabile, indus de relația $d^2(\mathbf{z}_j, \mathbf{z}_k) = 2(1 - r_{jk})$, evidențiază 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 distanță medie (deoarece $r_{jk} \approx 0$ implică $d^2(\mathbf{z}_j, \mathbf{z}_k) \approx 2$).

3. Analiza în componente principale

În funcție de proveniență variabilele care pot face obiectul unei *ACP* pot lua valori cantitative obținute în urma unor măsurători, pot lua valori calitative obținute în urma unor notații dar sunt asimilabile cu variabilele cantitative sau pot lua valori calitative ordinale obținute 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 subspațiu \mathcal{E}_q de dimensiune q ($q < p$);
- reprezentarea grafică a variabilelor într-un subspațiu \mathcal{F}_q explicitând „cel mai bine” legăturile inițiale între aceste variabile;
- reducerea dimensiunii (compresia), sau aproximarea matricii \mathbf{X} printr-o matrice de rang $q < p$.

Poziția punctelor într-un nor este dată de mulțimea distanțelor între toate punctele și determină *forma norului*. Forma norului este cea care caracterizează natura și intensitatea relațiilor între indivizi (liniile) și între variabile (coloanele) și relevă structurile de informații conținute în date. O modalitate de a reda vizual forma unui nor este aceea de a-l proiecta pe o dreaptă sau pe un plan minimizând deformările pe care această proiecție le implică.

Matricea $\mathbf{W} = \mathbf{YMY}' \in \mathcal{M}_{n \times n}(\mathcal{R})$ 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 și ℓ . Se numește *image euclidiană a indivizilor*, asociată produselor scalare $w_{i\ell}$, un nor compus din n puncte S_1, \dots, S_n și 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 $\mathbf{V} = \mathbf{Y}'\mathbf{D}\mathbf{Y} \in \mathcal{M}_{n \times n}(\mathcal{R})$ (de varianță-covarianță 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 și k . Se numește *image euclidiană a variabilelor* asociată produselor scalare v_{jk} , un nor compus din p puncte T_1, \dots, T_p și 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 aceluiași nor de puncte. Două imagini euclidiene sunt *echivalente* dacă ele reconstituie aceleași produse scalare.

Dacă dimensiunea spațiului 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 spațiu afin de dimensiune d se dorește obținerea unei imagini euclidiene într-un spațiu afin de dimensiune mult mai mică $q \ll d$.

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

Dacă $\text{rang}(\mathbf{Y}) = q$ atunci problema aproximării este practic rezolvată. Este suficient să se determine o bază a subspațiului vectorial de dimensiune q din \mathcal{R}^p ce conține norul de puncte-indivizi și să se calculeze coordonatele punctelor în noua bază.

Dacă $\text{rang}(\mathbf{Y}) > q$, demersul de mai sus se realizează prin proiecția punctelor-indivizi pe un subspațiu \mathcal{E}_q de dimensiune q , obținut astfel încât media pătratelor distanțelor între proiecții să fie maximă sau, inerția norului proiectat pe \mathcal{E}_q să fie maximă sau, în fine, deformarea distanțelor prin proiecție să fie minimă. Astfel problema ce trebuie rezolvată capătă următorul enunț: *să se*

găsească $\mathcal{H} \equiv \mathcal{E}_q$ astfel încât $\max \sum_{i=1}^n d^2(\mathbf{y}_i, \mathbf{0})$, iar soluția este dată de următoarea teoremă: subspațiul de dimensiune q pe care se proiectează optim, în sensul celor mai mici pătrate, cele n puncte din \mathfrak{R}^p este generat de primii q vectori proprii ai matricii $\mathbf{A} = \mathbf{VM} \in M_{pp}(\mathfrak{R})$ corespunzătorilor valorilor proprii $\lambda_1 > \lambda_2 > \dots > \lambda_q$, unde \mathbf{V} este matricea de varianță-covarianță asociată tabelului \mathbf{X} , iar \mathbf{M} este metrica spațiului indivizilor.

Valorile proprii ale matricii \mathbf{A} sunt reale și pozitive, \mathbf{A} fiind \mathbf{M} -simetrică, pozitiv definită și cu coeficienți reali. Vectorii proprii ai matricii \mathbf{A} sunt \mathbf{M} -ortonormați. Matricea \mathbf{A} se numește matricea inerției și $I_g = \text{tr}(\mathbf{A}) = \sum_{j=1}^p \lambda_j$.

Imaginea euclidiană a norului de puncte-indivizi obținută prin proiecția pe subspațiul \mathcal{H} se numește imaginea euclidiană a punctelor-indivizi asociată aproximației de ordinul q a produselor scalare. Se numesc axe principale de inerție vectorii proprii, \mathbf{M} -normați, \mathbf{a}_j , ai matricii de inerție \mathbf{A} . Se numește factor principal asociat axei principale \mathbf{a}_j și se notează cu \mathbf{u}_j forma liniară din \mathfrak{R}^p definită de relația $\mathbf{u}_j = \mathbf{M}\mathbf{a}_j$. Factorii principali $\{\mathbf{u}_j | j = 1 \div p\}$ sunt vectorii proprii ai matricii \mathbf{MV} asociați valorilor proprii $\{\lambda_j | j = 1 \div p\}$ ale matricii $\mathbf{A} = \mathbf{VM}$. Se numește plan factorial principal subspațiul \mathcal{E}_2 , generat de vectorii $\{\mathbf{u}_1, \mathbf{u}_2\}$. Se numește componentă principală asociată factorului principal \mathbf{u}_j și se notează cu \mathbf{c}_j forma liniară din \mathfrak{R}^n definită de relația $\mathbf{c}_j = \mathbf{Y}\mathbf{u}_j$, \mathbf{c}_j este proiecția \mathbf{M} -ortogonală a indivizilor pe axa principală \mathbf{a}_j .

Componentele principale $\{\mathbf{c}_j | j = 1 \div p\}$, sunt vectorii proprii ai matricii \mathbf{WD} asociați valorilor proprii $\{\lambda_j | j = 1 \div p\}$ ale matricii \mathbf{A} și sunt \mathbf{D} -ortogonale, deci necorelate. Mediile de selecție ale componentelor principale sunt nule (pe datele centrate și centrat reduce). Dispersia de selecție a componentei principale \mathbf{c}_j este λ_j valoarea proprie a matricii inerției, \mathbf{A} , pentru $(\forall)j = 1 \div p$.

Componentele principale sunt combinații liniare de variabilele inițiale, de dispersie maximă și care satisfac restricțiile $\mathbf{u}_j \mathbf{M}^{-1} \mathbf{u}_j = \mathbf{1}$. În cazul ACP normate (\mathbf{Z}, \mathbf{I}_p), componentele principale $\{\mathbf{c}_j | j = 1 \div p\}$ asociate valorilor proprii $\{\lambda_j | j = 1 \div p\}$ ale matricii \mathbf{A} sunt variabilele cele mai „legate” de variabilele inițiale, $\mathbf{z}_1, \dots, \mathbf{z}_p$, în sensul că suma pătratelor coeficienților de corelație, $\sum_{k=1}^p \text{cor}^2(\mathbf{c}_j, \mathbf{z}_k)$, este maximă pentru $(\forall)j = 1 \div p$.

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

Ecuția axei factoriale \mathbf{b} din \mathfrak{R}^n este: $\mathbf{Y}\mathbf{M}\mathbf{Y}'\mathbf{D}\mathbf{b} = \mu\mathbf{b} \mid \mathbf{b}'\mathbf{D}\mathbf{b} = \mathbf{1}$; ecuația factorului principal \mathbf{v} din (\mathfrak{R}^n) este: $\mathbf{v} = \mathbf{D}\mathbf{b}$; ecuația componentei principale \mathbf{d} din \mathfrak{R}^n este: $\mathbf{d} = \mathbf{Y}'\mathbf{v}$ sau $\mathbf{d} = \mathbf{Z}'\mathbf{v}$.

Se numește cerc de corelație principal subspațiul \mathcal{F}_2 generat de vectorii $\{\mathbf{v}_1, \mathbf{v}_2\}$.

Analog ca în cazul norului de puncte-indivizi:

- Factorii principali $\mathbf{v}_i \in (\mathfrak{R}^n)$, $i=1 \div n$, sunt \mathbf{D}^{-1} -ortonormați și satisfac relațiile $\mathbf{D}\mathbf{Y}\mathbf{M}\mathbf{Y}'\mathbf{v}_i = \mu_i \mathbf{v}_i$.
- Componentele principale $\mathbf{d}_i \in \mathfrak{R}^p$, $i=1 \div n$ sunt \mathbf{M} -ortogonale, au dispersia de selecție egală cu μ și satisfac relațiile $\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}\mathbf{d}_i = \mu_i \mathbf{d}_i$.

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

Relații de tranziție între cele două spații. Din punct de vedere numeric, o analiză în componente principale se reduce la calculul primelor q valori proprii și al vectorilor proprii asociați

pentru matricile $\mathbf{VM} = \mathbf{Y}'\mathbf{D}\mathbf{Y}\mathbf{M} \in \mathcal{M}_{p,p}(\mathfrak{R})$ și $\mathbf{WD} = \mathbf{Y}\mathbf{M}\mathbf{Y}'\mathbf{D} \in \mathcal{M}_{n,n}(\mathfrak{R})$.

O întrebare naturală este dacă există o relație între elementele principale dintr-o *ACP* pe spațiul variabilelor $(\mathcal{F}, \mathbf{M}) \subset \mathfrak{R}^p$ și elementele principale dintr-o *ACP* pe spațiul indivizilor $(\mathcal{E}, \mathbf{D}) \subset \mathfrak{R}^n$ iar răspunsul, privind relațiile de tranziție între cele două spații, este dat de următoarea teoremă: *toate valorile proprii nenule ale matricilor $\mathbf{Y}'\mathbf{D}\mathbf{Y}\mathbf{M}$ și $\mathbf{Y}\mathbf{M}\mathbf{Y}'\mathbf{D}$ sunt egale având, eventual, același ordin de multiplicitate și pentru $\lambda_j \neq 0$ sunt adevărate următoarele relații de tranziție:*

- $\mathbf{b}_j = (1/\sqrt{\lambda_j}) \mathbf{Y}\mathbf{M}\mathbf{a}_j = (1/\sqrt{\lambda_j}) \mathbf{Y}\mathbf{u}_j = (1/\sqrt{\lambda_j}) \mathbf{c}_j$ și
- $\mathbf{a}_j = (1/\sqrt{\lambda_j}) \mathbf{Y}'\mathbf{D}\mathbf{b}_j = (1/\sqrt{\lambda_j}) \mathbf{Y}'\mathbf{v}_j = (1/\sqrt{\lambda_j}) \mathbf{d}_j$ unde $j = 1 \div \text{rang}(\mathbf{Y}'\mathbf{Y})$.

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

Coordonatele punctelor pe o axă factorială în \mathfrak{R}^p sunt proporționale cu componentele axei factoriale din \mathfrak{R}^n corespunzătoare aceleiași valori proprii și reciproc, deoarece $\mathbf{c} = \mathbf{X}\mathbf{u}$ și $\mathbf{d} = \mathbf{X}'\mathbf{v}$ implică $\mathbf{c} = (\sqrt{\lambda})\mathbf{b}$ și $\mathbf{d} = (\sqrt{\lambda})\mathbf{a}$.

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

ACP nu pune în evidență decât legăturile liniare între variabile. Un coeficient de corelație slab între două variabile semnifică doar că acestea sunt independente liniar, în timp ce între ele poate exista o relație de ordin superior lui 1 (relație 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 corelație al variabilei cu factorul \mathbf{v}_j considerat ca o variabilă artificială ale cărei coordonate sunt date de cele n proiecții ale indivizilor pe această axă, conform relațiilor de tranziție. În cazul datelor centrat-reduce, $\sum_{j=1}^p \text{cor}^2(\mathbf{z}_k, \mathbf{v}_j) = \mathbf{a}'_k \mathbf{M} \mathbf{a}_k = 1$.

Reconstituirea datelor inițiale. Pornind de la relația $\mathbf{c} = \mathbf{Y}\mathbf{u}$ se obține relația $\mathbf{Y} = \sum_{j=1}^p \mathbf{c}_j \mathbf{u}'_j \mathbf{M}^{-1}$ numită *formula de reconstituire* a tabelului de date \mathbf{Y} pornind de la componentele și factorii principali. Analog, pornind de la relația $\mathbf{c} = \mathbf{X}\mathbf{u}$ se poate reconstitui tabelul \mathbf{X} precum și $\mathbf{M}\mathbf{V} = \sum_{j=1}^p \lambda_j \mathbf{u}_j \mathbf{u}'_j \mathbf{M}^{-1}$ și $\mathbf{V}\mathbf{M} = \sum_{j=1}^p \lambda_j \mathbf{a}_j \mathbf{a}'_j \mathbf{M}$

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 $\mathbf{Y}\mathbf{Y}'$ 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 (valorile proprii sunt ordonate descrescător), atunci se obține cea mai bună aproximare, în sensul celor mai mici pătrate, a lui \mathbf{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 \mathfrak{R}^n ajută la interpretarea axelor factoriale în \mathfrak{R}^p și reciproc. Trebuie totuși evitată interpretarea distanței 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 direcțiile variabilelor în \mathfrak{R}^p , atunci se pot reprezenta simultan, în acest spațiu, atât punctele-indivizi, cât și vectorii reprezentând variabilele.

În spațiul \mathfrak{R}^p al celor n puncte-indivizi, după transformarea tabelului de date, există două sisteme de axe: vechile axe unitare $\{\mathbf{e}_1, \dots, \mathbf{e}_p\}$ și noile axe unitare $\{\mathbf{u}_1, \dots, \mathbf{u}_p\}$, 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 \mathfrak{R}^p direcțiile date de variabilele inițiale pe planul factorial al norului de indivizi. Această reprezentare a variabilelor este diferită de reprezentarea norului de variabile.

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

planul factorial al norului de indivizi. În \mathfrak{R}^n , în metrica euclidiană, coordonata variabilei j pe axa k este egală cu coeficientul de corelație între variabilă și 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 dilatație $\sqrt{\lambda_k}$.

În cazul reprezentării simultane, care este de fapt o reprezentare în \mathfrak{R}^n , distanța dintre două variabile nu se interpretează în termeni de corelație deoarece este vorba de extremitățile unor vectori ortonormați (distanță egală cu $\sqrt{2}$ în spațiul complet). Interpretarea distanței între două variabile, în termeni de corelație, nu se poate face decât în \mathfrak{R}^n (norul proiectat al extremităților vectorilor unitari din \mathfrak{R}^p și norul extremităților vectorilor variabile în \mathfrak{R}^n au în general forme asemănătoare, vectorii proprii fiind totuși comparabili, deci dilatările fiind puțin deformante).

Ținând cont de aceste considerații, are totuși sens să se compare, în reprezentarea simultană, poziția a doi indivizi față de ansamblul variabilelor, sau poziția a două variabile față de ansamblul indivizilor. La intersecția axelor se găsesc valorile medii ale tuturor variabilelor. Direcția unei variabile definește zone pentru indivizi: de o parte indivizii ce iau valori mari pentru această variabilă și în partea opusă, indivizii care iau valori mici. Pe direcția unei variabile prezintă interes distanțele între indivizi.

Interpretarea rezultatelor. ACP construiește variabile noi, artificiale și reprezentări grafice ce permit vizualizarea relațiilor între variabile și a eventualelor grupe de indivizi și de variabile. Interpretarea rezultatelor este o fază delicată ce trebuie întreprinsă respectând următoarele aspecte:

- axele factoriale permit obținerea celei mai bune vizualizări aproximative, în sensul celor mai mici pătrate, ale distanțelor dintre indivizi, respectiv dintre variabile și în acest sens, primul demers care se impune este legat de măsurarea calității acestei aproximări;
- metoda naturală de a da o semnificație unei componente principale c este de a o corela cu variabilele inițiale x_j , în acest sens sunt calculați coeficienții de corelație liniară $\text{cor}(c, x_j)$ și sunt puși în evidență coeficienții cu valori absolute mari;
- practica frecvent utilizată este de a împărți în două mulțimea 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, a posteriori, cu componentele principale;
- într-un mod asemănător se procedează și în cazul mulțimii indivizilor, distingându-se între *indivizi activi* și *indivizi suplimentari*, care nu sunt luați în considerare la calculul matricilor de covarianță / corelație.

În funcție 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-redușă) este cea mai utilizată.

4. Analiza factorială discriminantă

Se dispune de observații privind p variabile cantitative X^1, \dots, X^p , jucând rolul de variabile explicative și o variabilă calitativă T cu q modalități $\{\tau_1, \dots, \tau_q\}$, jucând rolul de variabilă de explicat. Cele p variabile explicative au fost observate pe un eșantion de n indivizi, variabila nominală T generează o partiție a celor n indivizi în q clase $I_k, k = 1 \div q$.

În anumite situații 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 situații este uneori posibilă determinarea unui nou sistem de coordonate față de care structura de clase este mai evidentă decât în sistemul inițial, axele noului sistem având o putere de discriminare a claselor superioară celei a axelor inițiale.

Fie $\mathbf{X} = \{x_{ij} \mid i = 1 \div n, j = 1 \div p\} \in \mathcal{M}_{n \times p}(\mathfrak{R})$ matricea observațiilor. 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 , adică $\mathbf{g}_k = (x_j^k)_{j=1 \div p}$, unde $x_j^k = (1/n_k) \sum_{i \in I_k} x_{ij}$ și respectiv cu \mathbf{g} centrul de greutate al norului, adică $\mathbf{g} = (x_j)_{j=1 \div p}$, cu $x_j = (1/n) \sum_{i=1}^n x_{ij} = \sum_{k=1}^q (n_k/n) x_j^k$

Variabila $\mathbf{a} = \{a(i) \mid a(i) = \sum_{j=1}^p a_j (x_{ij} - \bar{x}_j), i = 1 \div n\}$, combinație liniară a celor p variabile, are media empirică 0 (este centrată) și dispersia empirică: $D^2(\mathbf{a}) = \sum_{j=1}^p \sum_{k=1}^q a_j a_k \text{cov}(\mathbf{x}_j, \mathbf{x}_k) = \mathbf{a}'\mathbf{V}\mathbf{a}$.

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

Matricea $\mathbf{V}^{-1}\mathbf{V}_e \in M_{p \times p}$ este, în general, o matrice nesimetrică. Din punct de vedere al calculului numeric, având în vedere că $q \ll p$, este mai ușor a afla vectorii și valorile proprii ale unei matrici simetrice de dimensiune $q \times q$ și a găsi o exprimare a lui \mathbf{a} în funcție de aceste elemente. \mathbf{V}_e este produsul matricii $\mathbf{C} = \{c_{jk} \mid c_{jk} = \sqrt{[(n_k/n)(x_j^k - \bar{x}_j)]}, j = 1 \div p, k = 1 \div q\} \in M_{p \times q}(\mathbb{R})$ cu transpusa sa, $\mathbf{V}_e = \mathbf{C}\mathbf{C}'$ deci $\mathbf{V}^{-1}\mathbf{C}\mathbf{C}'\mathbf{a} = \lambda\mathbf{a}$ sau $\mathbf{C}\mathbf{C}'\mathbf{a} = \lambda\mathbf{V}\mathbf{a}$. Luând $\mathbf{a} = \mathbf{V}^{-1}\mathbf{C}\mathbf{w}$ avem relația $\mathbf{C}\mathbf{C}'\mathbf{V}^{-1}\mathbf{C}\mathbf{w} = \lambda\mathbf{C}\mathbf{w}$, dacă \mathbf{w} este vector propriu al matricii $\mathbf{C}'\mathbf{V}^{-1}\mathbf{C}$, corespunzător lui λ , atunci el verifică această relație, iar \mathbf{a} și λ verifică relația $\mathbf{C}\mathbf{C}'\mathbf{a} = \lambda\mathbf{V}\mathbf{a}$. Deoarece $\mathbf{C}'\mathbf{V}^{-1}\mathbf{C} \in M_{q \times q}(\mathbb{R})$ este simetrică, se diagonalizează această matrice și apoi se află $\mathbf{a} = \mathbf{V}^{-1}\mathbf{C}\mathbf{w}$.

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

- Cazul $\lambda_{max} = 1$. În proiecția 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, concentrice și neliniari separabili. Este posibilă existența unei suprafețe 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 obișnuit unde $n > p > q$ și variabilele nu sunt legate prin relații liniare.

Odată găsite axele cu puterea de discriminare cea mai bună, pasul următor constă în găsirea suprafețelor de decizie. Metodele geometrice de analiză discriminantă, esențialmente descriptive, se bazează pe noțiunea de distanță și nu utilizează nicio noțiune probabilistă. În context geometric, discriminarea poate fi interpretată ca o împărțire a spațiului variabilelor în regiuni, numite *regiuni de decizie*, fiecare regiune fiind asociată cu o clasă de indivizi. Regiunile de decizie și implicit clasele corespunzătoare, se zic *separabile* dacă pot fi separate prin suprafețe din spațiul variabilelor. Suprafețele de separare ale regiunilor de decizie se numesc și *suprafețe de decizie*. Suprafețele de decizie pot fi descrise cu ajutorul unei mulțimi de *funcții de discriminare* (sau *de decizie*). Funcția de discriminare atașează fiecare vector-individ unei regiuni din spațiul variabilelor, regiune delimitată prin intermediul unei mulțimi de suprafețe de decizie.

O funcție de discriminare *instruibilă* tinde să reducă numărul indivizilor clasați incorect făcând

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

După ce clasele și suprafețele de decizie sunt stabilite (prin o fază de instruire), respectiv funcția de discriminare este instruită, funcției de discriminare i se prezintă date ale căror clase nu se cunosc. Această fază, în care indivizi noi sunt asociați uneia sau alteia dintre clasele stabilite, se numește *fază de lucru* (sau *decizională* sau *de afectare*). Uneori faza de instruire și cea de lucru pot să coincidă sau să se suprapună parțial.

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

- primul, *descriptiv*, ce constă în căutarea funcțiilor de discriminare liniare pe eșantionul de volum n respectiv găsirea combinațiilor 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, descriși prin variabilele explicative (X^1, \dots, X^p) .

5. Analiza corespondențelor simple

Se dispune de observații privind două variabile calitative (nominale sau categoriale), X cu n modalități $\{x_1, \dots, x_n\}$ și respectiv Y cu p modalități $\{y_1, \dots, y_p\}$. Variabilele nominale X și Y au fost observate simultan pe un eșantion de k indivizi și generează fiecare câte o partiție a celor k indivizi.

Un tabel ale cărui linii, respectiv coloane, desemnează două partiții ale aceleiași mulțimi, partiții date de modalitățile a două variabile nominale, se numește *tabel de contingență* (*de dependență* sau *încruciat*). De exemplu, într-un scrutin electoral cu mai mulți candidați, dacă pentru un eșantion de alegători se cunosc circumscripțiile electorale și opțiunile acestora atunci este convenabil să se grupeze datele într-un tabel de contingență \mathbf{K} ale cărui elemente k_{ij} reprezintă numărul de persoane din circumscripția i care optează pentru candidatul j .

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

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

Se numesc *efective marginale* (sau *marje*) cantitățile $k_{i\cdot} = \sum_{j=1}^p k_{ij}$ și $k_{\cdot j} = \sum_{i=1}^n k_{ij}$, $(\forall) i = 1 \div n$ și $(\forall) j = 1 \div p$ îndeplinind condițiile $\sum_{i=1}^n k_{i\cdot} = \sum_{j=1}^p k_{\cdot j} = \sum_{i=1}^n \sum_{j=1}^p k_{ij} = k$. Se numesc *frecvențe relative* cantitățile $f_{ij} = k_{ij} / k$, $(\forall) i = 1 \div n$ și $(\forall) j = 1 \div p$. Se numesc *frecvențe marginale* (sau *marje*) cantitățile $f_{i\cdot} = \sum_{j=1}^p f_{ij}$, $(\forall) i = 1 \div n$ și $f_{\cdot j} = \sum_{i=1}^n f_{ij}$, $(\forall) j = 1 \div p$ îndeplinind condițiile $\sum_{i=1}^n f_{i\cdot} = \sum_{j=1}^p f_{\cdot j} = \sum_{i=1}^n \sum_{j=1}^p f_{ij} = f = 1$.

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

Distanța χ^2 este invariantă la agregarea liniilor, respectiv a coloanelor, cu același profil. Această proprietate poartă numele de *principiul echivalenței distribuțiilor*. Echivalența distribuțională permite agregarea a două modalități (ale aceleiași variabile) cu profiluri identice (în \mathbb{R}^p ele se

confundă) într-o nouă modalitate cu o pondere sumată fără însă a afecta prin aceasta nici distanțele între modalitățile variabilei nou formate, nici distanțele între modalitățile celeilalte variabile. Din punct de vedere practic, această 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.

ACS pe tabelul centrat este echivalentă cu ACS pe tabelul necentrat. Este o particularitate a ACS, în comparație cu ACP, echivalența dintre analiza realizată pe tabloul necentrat (adică cu originea în O) și cea realizată pe tabloul centrat (adică cu originea în G) cu condiția ignorării, în primul caz, a axei factoriale care unește pe O cu G (această axă este asociată valorii proprii egală cu unu, numită valoare proprie trivială).

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

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

Reprezentările grafice ale proximităților între profiluri se fac, pe rând, în cele două spații, în centrul de greutate al norului corespunzător.

Problemele de optimizat și matricile de diagonalizat sunt:

- în \mathcal{R}^p , spațiul profilurilor-linie: $\max_{\mathbf{u}} \{ \sum_{i=1}^n f_{i\cdot} d^2(i, 0) \} \mid \mathbf{u}'\mathbf{D}_p^{-1}\mathbf{u} = 1$. Soluția \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 \mathcal{R}^n , spațiul profilurilor-coloană: $\max_{\mathbf{v}} \{ \sum_{j=1}^p f_{\cdot j} d^2(j, 0) \} \mid \mathbf{v}'\mathbf{D}_n^{-1}\mathbf{v} = 1$. Soluția \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} și \mathbf{T} au aceleași valori proprii nenule: $\mathbf{S}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$, $\mathbf{u} \in \mathcal{R}^p$ și $\mathbf{T}\mathbf{v}_\alpha = \lambda_\alpha \mathbf{v}_\alpha$, $\mathbf{v} \in \mathcal{R}^n$.
- valorile proprii λ_α sunt subunitare ($\lambda_\alpha \leq 1$, $(\forall)\alpha$).
- între vectorii proprii normați \mathbf{u}_α ai lui \mathbf{S} asociați lui λ_α și vectorii proprii normați \mathbf{v}_α ai lui \mathbf{T} asociați aceleiași valori proprii există relațiile: $\mathbf{v}_\alpha = (1/\sqrt{\lambda_\alpha})\mathbf{F}\mathbf{D}_p^{-1}\mathbf{u}_\alpha$ și $\mathbf{u}_\alpha = (1/\sqrt{\lambda_\alpha})\mathbf{F}'\mathbf{D}_n^{-1}\mathbf{v}_\alpha$

De asemenea:

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

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 ($\mathbf{u}_2, \dots, \mathbf{u}_p$) 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 ($\mathbf{u}_1^t, \mathbf{u}_2^t, \dots, \mathbf{u}_p^t$), a p -a axă corespunde lui $\mathbf{u}_1 = OG_\ell$ și nu indică nicio direcție în \mathcal{H} deoarece nu este conținută în \mathcal{H} . Inerția sa (valoarea proprie asociată) este nulă.

Coordonatele pe axele factoriale:

- în \mathfrak{R}^p : $\Psi_\alpha = D_n^{-1} F D_p^{-1} u_\alpha$ cu $\psi_{ai} = \sum_{j=1}^p (f_{ij}/f_i f_j) u_{aj}$.
- în \mathfrak{R}^n : $\Phi_\alpha = D_p^{-1} F' D_n^{-1} v_\alpha$ cu $\varphi_{aj} = \sum_{i=1}^n (f_{ij}/f_i f_j) v_{ai}$.

Coordonata modalității i a unei variabile reprezintă media modalităților j ale celeilalte variabile, ponderate de frecvențele condiționate ale profilului i . Analog, coordonata modalității j reprezintă media mulțimii modalităților i ponderate de frecvențele condiționate ale profilului j .

Relațiile de tranziție între spații (formulele quasi-baricentrice):

- $\Psi_\alpha = (1/\sqrt{\lambda_\alpha}) D_n^{-1} F \Phi_\alpha$ cu $\psi_{ai} = (1/\sqrt{\lambda_\alpha}) \sum_{j=1}^p \varphi_{aj} f_{ij}/f_i$.
- $\Phi_\alpha = (1/\sqrt{\lambda_\alpha}) D_p^{-1} F' \Psi_\alpha$ cu $\varphi_{aj} = (1/\sqrt{\lambda_\alpha}) \sum_{i=1}^n \psi_{ai} f_{ij}/f_j$.

Astfel, modulo coeficientul de dilatație ($1/\sqrt{\lambda_\alpha}$), proiecțiile punctelor unui nor sunt, pe o axă, coordonatele baricentrice ale proiecțiilor punctelor celuilalt nor. Relațiile quasi-baricentrice justifică reprezentarea simultană a liniilor și a coloanelor.

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

6. Analiza corespondențelor multiple

Se dispune de observații privind s variabile calitative X^q ($q = 1 \div s$, $s > 2$), având respectiv modalitățile $\{ (1, \dots, p_q) \}$. Modalitățile fiecărei variabile se exclud reciproc, fiecare modalitate este observată cel puțin o dată. Variabilele au fost observate simultan pe un eșantion de n indivizi, fiecare individ alege una și numai una dintre modalitățile fiecărei variabile.

Analiza corespondențelor 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 modalități ale celor s variabile nominale și fie r_{iq} ($r_{iq} \leq p_q$) numărul modalității alese de individul i , dintre cele p_q modalități ale variabilei X^q . Se numește *tabel de date condensat* matricea $\mathbf{R} = \{r_{iq} \mid i = 1 \div n, q = 1 \div s\} \in \mathcal{M}_{n \times s}(\mathfrak{R})$. Tabelul \mathbf{R} care descrie cele s modalități 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 informațiilor respective.

Pentru variabila nominală X^q , ($q = 1 \div s$) se numește *variabilă auxiliară a modalității* 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 $\mathbf{Z}_q = \{z_{ij,q} \mid i = 1 \div n, j = 1 \div p_q\}$; $(\forall) q \in [1, s] \in \mathcal{M}_{n \times p_q}(\mathfrak{R})$, în care fiecare linie conține $p_q - 1$ zerouri și un singur unu, se numește *matrice auxiliară a modalităților* variabilei nominale X^q . Matricea $\mathbf{Z} = [\mathbf{Z}_1, \dots, \mathbf{Z}_q, \dots, \mathbf{Z}_s] \in \mathcal{M}_{n \times p}(\mathfrak{R})$, obținută prin concatenarea matricilor \mathbf{Z}_q , se numește *tabel disjunctiv complet*. Avem: $z_{i\cdot} = \sum_{j=1}^{p_q} z_{ij,q} = s$; $z_{\cdot j} = \sum_{i=1}^n z_{ij,q} =$ numărul de indivizi care au ales modalitatea j a întrebării q ; $n = \sum_{j=1}^{p_q} z_{\cdot j} = z_{\cdot q}$; $z = \sum_{i=1}^n z_{i\cdot} = \sum_{q=1}^s z_{\cdot q} = \sum_{i=1}^n \sum_{j=1}^{p_q} z_{ij} = ns =$ efectivul total.

Matricea $\mathbf{B} = \mathbf{Z}'\mathbf{Z} \in \mathcal{M}_{p \times p}(\mathfrak{R})$, se numește *tabel de contingență Burt* asociat tabelului disjunctiv complet \mathbf{Z} , având termenul general $b_{jj'} = \sum_{i=1}^n z_{ij} z_{ij'}$, marjele $b_j = \sum_{j'=1}^p b_{jj'} = sz_{\cdot j}$, efectivul total $b = \sum_{j=1}^p b_j = s^2 n$ iar termenii de pe diagonală sunt efectivele $\{z_{\cdot j}\}$ ale modalităților întrebării q . Se notează cu $\mathbf{D} \in \mathcal{M}_{p \times p}(\mathfrak{R})$ matricea diagonală definită de relațiile: $d_{jj} = b_{jj} = z_{\cdot j}$ și $d_{jj'} = 0$, $(\forall) j, j' \in [1, p], j \neq j'$.

Analiza corespondențelor multiple (ACM) este analiza corespondențelor simple (ACS) aplicată unui tabel disjunctiv complet. În consecință se aplică aceleași transformări tabelului de date pentru obținerea profilurilor-linie sau profilurilor-coloană, aceleași ponderi ale punctelor funcție de

profilurile marginale, aceeași distanță, distanța χ^2 . Indivizii sunt toți afectați de o pondere identică $m_i = z_i / ns = 1/n$, ($i = 1 \div n$), fiecare modalitate j este ponderată de frecvența sa, $m_j = z_j / ns$. Pe un tabel disjunctiv:

- în \mathfrak{R}^n distanța χ^2 între modalități, se scrie: $d^2(j, j') = \sum_{i=1}^n n(z_{ij} / z_j - z_{ij'} / z_{j'})^2$ și este nulă dacă modalitățile j și j' sunt alese de aceiași indivizi. Modalitățile de efectiv scăzut, adică cele alese de puțini indivizi, sunt depărtate față de celelalte modalități.
- în \mathfrak{R}^p distanța χ^2 între indivizi, se scrie $d^2(i, i') = (1/s) \sum_{j=1}^p (n / z_j)(z_{ij} - z_{i'j})^2$ și este nulă dacă indivizii i și i' au ales aceleași modalități. Ei sunt cu atât mai depărtați cu cât au răspuns mai diferit. O modalitate j intervine în distanța dintre indivizi cu atât mai mult cu cât masa ei este mai mică.

Reluând rezultatele analizei corespondențelor simple și notațiile 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 $f_{.j} = \delta_{ij}(z_j/ns)$ și $\mathbf{D}_n = (1/n)\mathbf{I}_n$, cu termenul general $f_{.i} = \delta_{ij}/n$.

Pentru a găsi axele factoriale \mathbf{u}_α se diagonalizează matricea $\mathbf{S} = \mathbf{F}'\mathbf{D}^{-1}\mathbf{F}\mathbf{D}_p^{-1} = (1/s)\mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1}$ cu termenul general $s_{jj'} = (1/s \cdot z_j \cdot z_{j'}) \sum_{i=1}^n z_{ij}z_{ij'}$:

- în \mathfrak{R}^p , ecuația celei de-a α -a axe factoriale \mathbf{u}_α este $(1/s)\mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$ și ecuația celui de-al α -lea factor $\Phi_\alpha = \mathbf{D}^{-1}\mathbf{u}_\alpha$ este $(1/s)\mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z}\Phi_\alpha = \lambda_\alpha \Phi_\alpha$;
- în \mathfrak{R}^n , ecuația celui de-al α -lea factor Ψ_α este: $(1/s)\mathbf{Z}\mathbf{D}^{-1}\mathbf{Z}'\Psi_\alpha = \lambda_\alpha \Psi_\alpha$.

Factorii Φ_α și Ψ_α (de normă λ_α) reprezintă coordonatele punctelor linie și ale punctelor coloană pe axa factorială α .

Relațiile de tranziție între factorii Φ_α și Ψ_α sunt: $\Phi_\alpha = (1 / \sqrt{\lambda_\alpha}) \mathbf{D}^{-1}\mathbf{Z}'\Psi_\alpha$; $\Psi_\alpha = (1 / s\sqrt{\lambda_\alpha}) \mathbf{Z}\Phi_\alpha$.

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

Coordonatele factoriale ale modalității j pe axa α sunt date de: $\varphi_{\alpha, j} = (1/\sqrt{\lambda_\alpha}) \sum_{i=1}^n (z_{ij}/z_j) \psi_{\alpha, i} = (1/z_j\sqrt{\lambda_\alpha}) \sum_{i \in n(j)} \psi_{\alpha, i}$ unde $n(j)$ desemnează mulțimea indivizilor care au ales modalitatea j .

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

Coordonatele modalităților în \mathfrak{R}^n sunt coloanele matricii $\mathbf{Z}\mathbf{D}^{-1}$. Acestea generează un subspațiu a cărui dimensiune este rangul lui $\mathbf{Z}\mathbf{D}^{-1}$, adică $p - s + 1$. Rangul maxim al matricii $\mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z}$ de diagonalizat va fi deci $p - s + 1$. Dar, în 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 $p - s$ valori proprii nenule. Alegând o bază în suportul norului, revine la a căuta valorile proprii ale unei matrici de ordin $p - s$.

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

Inerția unei modalități j este, prin definiție, $I(\mathbf{j}) = m_j d^2(\mathbf{j}, G)$ cu $m_j = z_j / ns$ sau, efectuând substituțiile, $I(\mathbf{j}) = (1/s)(1 - z_j / n)$. Inerția unei modalități este cu atât mai mare cu cât efectivul z_j al acestei modalități, adică numărul de indivizi care au ales-o, este mai mic. Maximul $1/s$ va fi atins pentru modalitățile de efectiv nul. În consecință, în momentul codificării, se va evita introducerea unor modalități susceptibile de a fi alese de puțini indivizi pentru a nu introduce perturbații în primele axe factoriale.

Inerția unei întrebări q este, prin definiție: $I(q) = \sum_{j=1}^{p_q} I(j) = (1/s)(p_q - I)$. Inerția unei întrebări este cu atât mai mare cu cât numărul de modalități asociat, p_q , este mai mare. Minimul $1/s$ este atins de întrebările cu doar două modalități de răspuns. În consecință, dacă se dorește ca întrebările să joace un rol aproximativ egal, se va echilibra sistemul de întrebări (variabilele vor fi "decupate" într-un număr egal de modalități).

Inerția totală este $I = \sum_{q=1}^s I(q) = \sum_{j=1}^p (z_j / ns) d^2(j, G) = p/s - I$, ($\sum_{q=1}^s p_q = p$). În particular $I = 1$ dacă toate întrebările au două modalități de răspuns, adică $p = 2s$. În consecință depinzînd exclusiv de numărul de întrebări și de modalitățile asociate acestora, inerția globală nu are, în cazul ACM (ca și în cazul ACP normat, de altfel), nicio semnificație statistică, deoarece nu depinde de legătura între variabile.

A spune că există afinități între răspunsuri este același lucru cu a spune că există indivizi care au profile asemănătoare din punct de vedere al atributelor alese spre a-i descrie. Ținînd cont de distanțele între elementele tabelului disjunctiv complet și de relațiile baricentrice se poate interpreta că:

- proximitatea între indivizi semnifică faptul că au ales global aceleași modalități ca răspuns la întrebările puse;
- proximitatea între modalități 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 modalitățile aceleiași întrebări semnifică faptul că grupele de indivizi care le-au ales sunt asemănătoare (modalitățile unei aceleiași 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 relațiilor ce pot exista între două grupe de variabile cantitative observate pe aceeași mulțime de indivizi. De exemplu, analizele medicale efectuate, pe același eșantion 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 același fenomen, caz în care prospectorul de date ar putea renunța la unul din cele două grupe de variabile.

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

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

În spațiul \mathcal{F} al variabilelor, respectiv \mathcal{R}^n înzestrat cu o bază canonică \mathcal{F} și cu o metrică \mathbf{M} , se pot defini două subspații vectoriale:

- $\mathcal{F}_{\mathbf{X}}$ generat de vectorii \mathbf{x}_j ($j = 1 \div p$), în general de dimensiune p și
- $\mathcal{F}_{\mathbf{Y}}$ 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 considerație două spații vectoriale:

- $\mathcal{E}_1 = (\mathcal{R}^p, \mathcal{E}, \mathbf{M})$, generat de vectorii \mathbf{x}_i ($i = 1 \div n$) și

- $\mathcal{E}_2 = (\mathcal{R}^q, \mathcal{E}, \mathbf{M})$, generat de vectorii $\mathbf{y}_i (i = 1 \div n)$

Se caută, pentru început, un cuplu de variabile (V^1, W^1) , V^1 combinație liniară a variabilelor X^j (deci un element din \mathcal{F}_X), normată și W^1 combinație liniară a variabilelor Y^k (deci un element din \mathcal{F}_Y), normată astfel încât V^1 și W^1 să fie cele mai corelate posibil. Se caută, apoi, cuplul normat (V^2, W^2) , V^2 combinație liniară a variabilelor X^j , necorelată cu V^1 și W^2 combinație liniară a variabilelor Y^k , necorelată cu V^1 , astfel încât V^2 și W^2 să fie cele mai corelate posibil. Se continuă în același 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}_X (fiind combinații liniare de variabile centrate, sunt centrate; fiind necorelate sunt ortogonale în metrica identitate). Variabilele W^s constituie un sistem ortonormat al lui \mathcal{F}_Y (ele nu formează o bază decât dacă $q = p$). Cuplurile (V^s, W^s) , și în special primele dintre ele, țin cont de legăturile liniare dintre cele două grupe de variabile inițiale.

Variabilele V^s și W^s se numesc *variabile canonice*. Corelațiile lor succesive se numesc *coeficienți de corelație canonică* (sau *corelații canonice*) și se notează cu $\delta_s (1 \geq \delta_1 \geq \delta_2 \geq \dots \geq \delta_p \geq 0)$. Orice variabilă canonică $V^{s'}$ este, prin construcție, necorelată (deci ortogonală) cu celelalte variabile canonice V^s , $s \neq s'$. Deasemenea $V^{s'}$ este necorelată cu W^s , dacă $s \neq s'$ și bineînțeles 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 obține o bază ortonormată în \mathcal{F}_Y , în care ultimile variabile $W^s (s = p+1 \div q)$ sunt asociate cu coeficienți de corelație canonică nuli ($\delta_s = 0$, $s = p+1 \div q$).

Fie $\mathbf{P}_X = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ și $\mathbf{P}_Y = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'$ matricile proiecțiilor ortogonale pe subspațiile \mathcal{F}_X și \mathcal{F}_Y , ale lui \mathcal{F} înzestrat cu metrica \mathbf{I} . Atunci:

- vectorii V^s sunt vectorii proprii normați ai matricii $\mathbf{P}_X\mathbf{P}_Y$ corespunzători valorilor proprii λ_s , ordonate descrescător, $\lambda_s \in [0, 1]$;
- vectorii W^s sunt vectorii proprii normați ai matricii $\mathbf{P}_Y\mathbf{P}_X$ corespunzători aceluiași valori proprii λ_s ;
- coeficienții de corelație canonică sunt $\delta_s = \sqrt{\lambda_s}$, $s = 1 \div p$.

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

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

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

În măsura în care graficul astfel obținut este „bun” el poate fi utilizat pentru a interpreta relațiile (proximități, opoziții, depărtări) dintre cele două mulțimi de variabile. Prin construcție, acest grafic reprezintă corelațiile dintre variabilele canonice V^s și variabilele inițiale X^j și Y^k , corelații care stau la baza interpretării lui. Interpretarea poate fi facilitată utilizând coeficienții de corelație liniară dintre variabilele X^j , dintre variabilele Y^k și dintre variabilele X^j și Y^k . Toți acești coeficienți sunt furnizați de produsul software.

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

Cele două grafice (în \mathcal{F}_X și în \mathcal{F}_Y) având aceeași calitate și conducând la aceleași interpretări este suficient unul singur pentru a interpreta rezultatele unei analize.

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

8. Concluzii

Explorarea datelor este utilizată în orice domeniu, atunci când datele sunt mult prea multe pentru a mai putea fi înțelese de o minte omenească. În studiile care vizează un număr important de variabile, respectiv indivizi reprezentabili în spații de mari dimensiuni, o dificultate majoră o constituie obținerea unei reprezentări grafice adecvate a cărei vizualizare și interpretare să faciliteze înțelegerea structurii datelor analizate. Analiza în componente principale are un rol esențial fiind metoda care servește drept fundament teoretic și 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 (rețelele neuronale).

Analiza discriminantă este una dintre tehnicile de analiză multidimensională cele mai folosite în practică: diagnostic automat, controlul calității, previziunea riscului, recunoașterea formelor. Scopul analizei discriminante îl constituie studierea legăturilor între variabilele explicative și clasele partiției și definirea funcțiilor 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 și ca o tehnică de reducere a dimensiunii spațiului variabilelor, prin această tehnică fiind selectate cele mai relevante caracteristici. Reducerea dimensiunii poate fi deasemenea impusă și de necesitatea vizualizării claselor prin proiectarea datelor într-un spațiu cu una sau două dimensiuni.

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

Analiza corespondențelor multiple este o generalizare posibilă a analizei corespondențelor simple, are însă proceduri de calcul și reguli de interpretare specifice și se pretează la un număr mare de aplicații. Ea este în mod deosebit adaptată la descrierea tabelelor mari de variabile nominale, specifice anchetelor socio-economice sau medicale, modalitățile 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 și de asemenea poate fi privită ca un caz particular de analiză în componente principale a două pachete de variabile într-un spațiu înzestrat cu o metrică specială. Multă vreme analiza canonică, nefiind ușor aplicabilă, a avut puține aplicații 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 și ulterior, prin apariția datelor de expresie genomică (biochip-uri) combinate cu variabile biologice pentru situații tipice de analiză canonică.

Analiza canonică prezintă anumite analogii atât cu analiza în componente principale, privind construirea și interpretarea graficelor, cât și cu regresia liniară, privind natura datelor. Analiza canonică este apropiată de regresia liniară multiplă (explicarea unei variabile cantitative prin o mulțime 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 regăsește regresia). Deasemenea, când unul din cele două grupuri de variabile este înlocuit de variabilele auxiliare (modalitățile) unei variabile calitative se regăsește analiza factorială discriminantă, iar când fiecare din cele două grupuri este înlocuit cu variabilele auxiliare ale unei variabile calitative se regăsește analiza corespondențelor simple. Mai mult, există anumite generalizări ale analizei canonice la mai mult de două grupuri de variabile cantitative, iar acestea permit atât regăsirea analizei corespondențelor multiple (înlocuind fiecare grup prin variabilele auxiliare ale unei variabile calitative), cât și regăsirea analizei în componente principale (lăsând câte o singură variabilă cantitativă în fiecare grup).

Analiza în componente independente, mai recentă, rezultată din fizica semnalului și cunoscută inițial 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 populații, 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 și aplicații. București: 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, decidenți - metode de bază și instrumente informatice asociate. Ediția a 2-a, rev. București: Editura Tehnică, 2005, 376 p.
6. **FILIP, F. G.:** Sisteme suport pentru decizii. Ediția a 2-a, rev. București: 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.