Analiza în componente principale

Analiza în componente principale (ACP) este cea mai utilizată metodă de analiză a datelor. A fost propusă de Hotteling în 1938, dar necesitând numeroase și laborioase calcule s-a impus în practică abia începând cu anii 70 odată cu apariția calculatoarelor. ACP sintetizează informația conținută în tabelele de date cantitative de mari dimensiuni, cu un mare număr de instanțe și de variabile. O colectivitate statistică descrisă printr-un număr mare de variabile este greu de studiat. ACP determină un număr mai mic de variabile noi, numite componente principale, care concentrează informația, variabilitatea existentă la nivelul colectivității studiate. Componentele principale sunt construite sub formă de combinație liniară de variabile inițiale, care concentrează o cât mai mare parte din varianță. Astfel, prima componentă principală preia maximul din varianța variabilelor originale, a doua componentă preia maximul de varianță rămasă după eliminarea primei componente șamd

ACP se utilizează ca instrument de analiză statistică în multe domenii de activitate, inclusiv în recunoașterea formelor sau în scalarea multidimensională din grafică. În recunoașterea de forme, prin ACM se asigură selectarea caracteristicilor esențiale, semnificative ale formelor analizate, caracteristici care asigură puterea cea mai mare de discriminare. În grafică se asigură reprezentarea în 2D sau 3D a unor obiecte aflate in spații multidimensionale.

Date prelucrate

Datele analizate apar sub forma unui tabel de observații cu *n* linii și *m* coloane:

$$X = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ \dots & & \\ x_{n1} & .. & x_{nm} \end{bmatrix}$$

unde x_{ij} este valoarea luata de variabila *j* la individul *i*. Variabilele tabelului *X* mai sunt numite și variabile observate sau *variabile cauzale*. Acestea sunt ***standardizate*** (medie 0 și varianță 1).

Se notează cu X_j vectorul coloană format din valorile variabilei *j* pentru cei *n* indivizi.

Obiectivul propus este concentrarea varianței conținute de tabelul *X* într-un număr redus de variabile noi, numite componente principale, absolut necorelate între ele, notate astfel: C_1, C_2, ..., C_s.

Determinarea acestor variabile se face succesiv, astfel:

Etapa 1. Se determină variabila sintetică C_1, prima componentă principală, ca o combinație liniară de variabile X_j:

$$C_1 = a_{11}X_1 + \dots + a_{j1}X_j + \dots + a_{m1}X_m .$$

Valoarea înregistrată de o instanță oarecare, *i*, pentru componenta principală C_1 este:

$$c_{i1} = a_{11}x_{i1} + \dots + a_{j1}x_{ij} + \dots + a_{m1}x_{im} .$$

Notăm cu $a_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \dots \\ a_{m1} \end{bmatrix}$ vectorul care conține coeficienții legăturii liniare dintre variabilele observate și componenta C_1.

Etapa *k*. Se determină variabila sintetică C_k, combinație liniară de variabile *X*:

$$C_k = a_{1k}X_1 + \dots + a_{jk}X_j + \dots + a_{mk}X_m ,$$

unde $a_k = \begin{bmatrix} a_{1k} \\ a_{2k} \\ \dots \\ a_{mk} \end{bmatrix}$ este vectorul coeficienților legăturii liniare cu variabilele observate.

Analiza în componente principale poate fi abordată atât din perspectiva instanțelor cât și din perspectiva variabilelor analizate. Atunci când efectuăm analiza la nivelul instanțelor, componentele principale vor fi determinate astfel încât varianța lor sa fie maximă, deci suma pătratelor valorilor înregistrate de instanțe pentru componentele principale trebuie să fie cât mai mare. Când analiza se efectuează la nivelul variabilelor, deducerea componentelor principale se face astfel încât acestea sa fie maxim corelate cu variabilele observate și absolut necorelate între ele. Ambele modalități de deducere a componentelor principale conduc la aceleași rezultate după cum se va vedea în continuare.

Deducerea componentelor principale în spațiul instanțelor. Abordarea geometrică a modelului

Instanțele formează un nor de *n* puncte într-un spațiu *m*-dimensional, în care cele *m* variabile sunt proiecțiile instanțelor pe *m* axe de reprezentare.

În abordarea geometrică determinăm un sistem de axe ortonormat (axe ortogonale și de norma 1) în care vor fi reprezentate cele *n* puncte. Fiecare axă corespunde unei componente principale, iar vectorii a_k vor fi vectori unitari ai axelor (versori), deci $\sum_{j=1}^{m} a_{kj}{}^2 = 1$, $k = \overline{1,s}$, unde *s* este numărul maxim de axe.

Etapa 1. Se determină axa 1, corespunzătoare primei componente principale , astfel încât indivizii să fie cat mai bine reprezentați pe această axă (varianța componentei sa fie maximă). Se notează cu O centrul de greutate al norului de puncte.

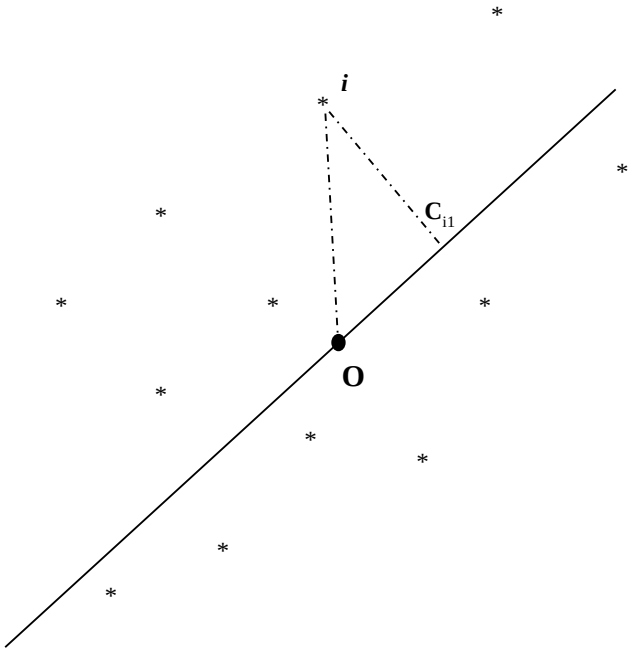

Figura 2.1

Distanța de la un punct/individ oarecare *i* la axa D_1 corespunzătoare primei componente principale se notează cu $d(i,D_1)$, iar distanţa de la punctul *i* la originea **O** cu $d(i,\mathbf{O})$. Există următoarea relație între distanțe:

$$d(i,\mathbf{O})^2 = d(i,D_1)^2 + c_{i1}{}^2,$$

unde c_{i1} este proiecția punctului *i* pe axa D_1.

Pentru toate punctele:

$$\frac{1}{n}\sum_{i=1}^{n} d(i,O)^2 = \frac{1}{n}\sum_{i=1}^{n} d(i,D_1)^2 + \frac{1}{n}\sum_{i=1}^{n} c_{i1}^2 .$$

Suma distanțelor către centrul de greutate este constantă (nu depinde de alegerea axei). Varianța explicată prin axa 1 este $\frac{1}{n}\sum_{i=1}^{n} c_{i1}^2$ care matriceal se poate scrie $\frac{1}{n}(C_1)^t C_1 = \frac{1}{n}(a_1)^t X^t X a_1$. Problema se poate pune complementar în două moduri cu același rezultat:

1. Se maximizează varianța explicată prin axa 1,
2. Se minimizează suma distanțelor punctelor la axa 1.

Problema de optim care se rezolvă este următoarea:

$$\begin{cases} \underset{a^1}{Max} \frac{1}{n}(a_1)^t X^t X a_1 \\ (a_1)^t a_1 = 1 \end{cases}.$$

Folosind metoda multiplicatorilor Langrange pentru rezolvarea acestei probleme de extrem cu restricții, funcția Lagrangean asociată acestei probleme este: $L(a_1,\lambda) = \frac{1}{n}(a_1)^t X^t X a_1 - \lambda((a_1)^t a_1 - 1)$.

Prin anularea derivatelor parțiale se obține:

$$\frac{\partial L}{\partial a_1} = 2\frac{1}{n} X^t X a_1 - 2\lambda a_1 = 0,$$

$$\frac{\partial L}{\partial \lambda} = (a_1)^t a_1 - 1 = 0.$$

Din prima relație rezultă $\frac{1}{n} X^t X a_1 = \lambda a_1$. Prin urmare a_1 este vector propriu al matricei $\frac{1}{n} X^t X$ corespunzător valori proprii λ. Înmulțind această ultimă relație la stânga cu $(a_1)^t$ rezultă: $\frac{1}{n}(a_1)^t X^t X a_1 = \lambda$. Deoarece cantitatea $\frac{1}{n}(a_1)^t X^t X a_1$ este tocmai cea care se maximizează, rezultă că λ este cea mai mare valoare proprie iar a_1 este vectorul propriu corespunzător ei. Vom nota λ cu α_1.

Etapa 2. Se determină axa 2 de vector a_2 astfel încât aceasta să fie ortogonală în raport cu axa 1 și să maximizeze varianța explicată (punctele reprezentând indivizii să fie cât mai răsfirați pe axă). Optimizarea aplicată este:

$$\begin{cases} \underset{a_2}{Max} \frac{1}{n}(a_2)^t X^t X a_2 \\ (a_2)^t a_2 = 1 \\ (a_2)^t a_1 = 0 \end{cases}$$

Funcția Lagrangean asociată este:

$L(a_2,\lambda_1, \lambda_2) = \frac{1}{n}(a_2)^t X^t X a_2 - \lambda_1((a_2)^t a_2 - 1) - \lambda_2 (a_2)^t a_1$.

Anularea derivatei parțiale în funcție de a_2:

$$\frac{\partial L}{\partial a_2} = 2\frac{1}{n} X^t X a_2 - 2\lambda_1 a_2 - \lambda_2 a_1 = 0.$$

Dacă înmulțim această relație la stânga cu $(a_1)^t$ obținem:

$$2\frac{1}{n}(a_1)^t X^t X a_2 - 2\lambda_1 (a_1)^t a_2 - \lambda_2 (a_1)^t a_1 = 0.$$

Avem $(a_1)^t a_2 = 0$. Deoarece $\frac{1}{n} X^t X a_1 = \alpha_1 a_1$, prin transpunere rezultă că și $(a_1)^t \frac{1}{n} X^t X = \alpha_1 (a_1)^t$ deoarece matricea X^tX este simetrică.

Atunci: $2\frac{1}{n}(a_1)^t X^t X a_2 = 2\frac{1}{n}\alpha_1 (a_1)^t a_2 = 0$. Prin urmare $\lambda_2 = 0.$

Înlocuind în derivată, obținem $\frac{1}{n} X^t X a_2 = \lambda_1 a_2$, deci a_2 este vector propriu corespunzător valorii proprii λ_1, iar această valoare proprie este maximă conform relației, $\frac{1}{n}(a_2)^t X^t X a_2 = \lambda_1$, deoarece cantitatea $\frac{1}{n}(a_2)^t X^t X a_2$ este cea care se maximizează la această etapă. Vom nota această valoare proprie cu α_2.

Etapa *k*. Se determină axa *k* de vector a_k astfel încât aceasta să fie ortogonală în raport cu axele anterioare și sa maximizeze, de asemenea, varianța explicată.
Problema de optim care se rezolvă este următoarea:

$$\begin{cases} \underset{a^k}{Max} \frac{1}{n}(a_k)^t X^t X a_k \\ (a_k)^t a_k = 1 \\ (a_k)^t a_j = 0,\ j = 1, k-1 \end{cases}.$$

Funcția Lagrangean asociată este:

$L(a_k,\lambda_1, \lambda_2,..., \lambda_k) = \frac{1}{n}(a_k) X^t X a_k - \lambda_1((a_k)^t a_k - 1) - \lambda_2 (a_k)^t a_1 - ... - \lambda_k (a_k)^t a_{k-1}$.

Anulând derivata parțială în a_k obținem:

$$\frac{\partial L}{\partial a_k} = 2\frac{1}{n} X^t X a_k - 2\lambda_1 a_k - \lambda_2 a_1 - ... - \lambda_k a_{k-1} = 0 .$$

Folosim procedeul de la etapa 2: înmulțim prima relație, succesiv, cu $(a_1)^t$, $(a_2)^t$,..., $(a_{k-1})^t$, și obținem $\lambda_2 = 0$, $\lambda_3 = 0$, ..., $\lambda_k = 0$. Revenind cu aceste rezultate în prima derivată parțială, obținem $\frac{1}{n} X^t X a_k = \lambda_1 a_k$, ceea ce ne duce la concluzia că a_k este vector propriu al matricei $\frac{1}{n} X^t X$, corespunzător valorii proprii λ_1, și mai mult, deoarece cantitatea $\frac{1}{n}(a_k)^t X^t X a_k$ este cea care se maximizează la acest pas, λ_1 este valoarea proprie de ordin *k*. Notăm valoarea proprie λ_1 cu α_k.

Noile axe formează un nou spațiu numit *spațiul principal*. Semnificația informațională a axelor este dată de cantitatea de varianță explicată de fiecare axă.

Deducerea componentelor principale în spațiul variabilelor

Analiza în componente principale realizată în spațiul variabilelor își propune identificarea directă a componentelor principale astfel încât acestea să fie maxim corelate cu variabilele inițiale și absolut necorelate între ele.

Etapa 1. Se determină prima variabilă sintetică C_1 astfel încât aceasta să fie maxim corelată cu variabilele inițiale:

$\sum_{j=1}^{m} R^2(C_1, X_j)$ să fie maximă.

$$R^2(C_1, X_j) = \frac{Cov(C_1, X_j)^2}{Var(C_1)Var(X_j)} = \frac{1}{n}\frac{(C_1)^t X_j (X_j)^t C_1}{(C_1)^t C_1}$$

$$\sum_{j=1}^{m} R^2(C_1, X_j) = \frac{1}{n}\sum_{j=1}^{m}\frac{(C_1)^t X_j (X_j)^t C_1}{(C_1)^t C_1} = \frac{1}{n}\frac{(C_1)^t X X^t C_1}{(C_1)^t C_1}$$

Problema care se rezolvă:

$$\underset{C_1}{Maxim}\frac{1}{n}\frac{(C_1)^t X X^t C_1}{(C_1)^t C_1}.$$

Dacă notăm cu C'1 = $\frac{C_1}{\sqrt{C_1^t C_1}}$, vectorul C_1 normat ($\sqrt{C_1^t C_1}$ este norma lui C_1) , problema de optim va deveni:

$$\begin{cases} \underset{C^2}{Maxim}\frac{1}{n}(C_1')^t X X^t C'_1 \\ (C'_1)^t C'_1 = 1 \end{cases}.$$

Această problemă se va rezolva în același fel ca problema de optim pentru prima etapă la abordarea geometrică a modelului (capitolul precedent).

Soluția, C'_1 o constituie vectorul propriu al matricei $\frac{1}{n} XX^t$, corespunzător celei mai mari valori proprii β_1. Cum C_1 și C'_1 diferă prin normă, rezultă că și C_1 este vector propriu al matricei $\frac{1}{n} XX^t$ corespunzător aceleiași valori proprii, β_1.

Etapa 2. Se determină a doua componentă principală C_2, maxim corelată cu variabilele inițiale și absolut necorelată cu prima componentă principală C_1.

Problema de optim este:

$$\begin{cases} \underset{C^2}{Maxim}\frac{1}{n}\frac{(C_2)^t X X^t C_2}{(C_2)^t C_2} \\ R(C_1, C_2) = 0 \end{cases}.$$

Făcând substituirea prin normă, vom avea:

$$\begin{cases} \underset{C^2}{Maxim} \frac{1}{n}(C_2')^t XX^t C'_2 \\ (C'_2)^t C'_2 = 1 \\ (C'_2)^t C'_1 = 0 \end{cases},$$

deoarece $R(C_1, C_2) = 0$ implică $(C'_2)^t C'_1 = 0$. Soluția, C'_2 și implicit, C_2, sunt vectori proprii ai matricei $\frac{1}{n} XX^t$ corespunzători celei de-a doua valori proprii β_2: $\frac{1}{n} XX^t \cdot C_2 = \beta_2 \cdot C_2$

Etapa *k*. În mod identic se determină variabila C_k maxim corelată cu variabilele inițiale dar absolut necorelată cu celelalte variabile noi, C_i, i=1,k-1.

Problema care se rezolvă:

$$\begin{cases} \underset{C^1}{Maxim} \frac{1}{n} \frac{(C_k)^t XX^t C_k}{(C_k)^t C_k} \\ R(C_k, C_i) = 0, i = 1, k-1 \end{cases} \Leftrightarrow \begin{cases} \underset{C^2}{Maxim} \frac{1}{n}(C_k')^t XX^t C'_k \\ (C'_k)^t C'_k = 1 \\ (C'_k)^t C'_j = 0, \quad j = 1, k-1 \end{cases}$$

Soluția o constituie vectorul propriu al matricei $\frac{1}{n} XX^t$ corespunzător valorii proprii β_k: $\frac{1}{n} XX^t \cdot C_k = \beta_k \cdot C_k$.