Legătura dintre cele două abordări. În spațiul indivizilor , la etapa *k* se determină vectorul propriu a_k, care este vectorul unitar al axei *k* corespunzătoare componentei C_k:

$\frac{1}{n} X^t X \cdot a_k = \alpha_k a_k$.

Înmulțind această relație la stânga cu *X* obținem:

$\frac{1}{n} XX^t X a_k = X\alpha_k a_k \;\Rightarrow \frac{1}{n} XX^t C_k = \alpha_k C_k$.

Relația obținută, $\frac{1}{n} XX^t C_k = \alpha_k C_k$, este aceeași cu cea obținută la abordarea problemei în spațiul variabilelor, considerând $\beta_k=\alpha_k$.

Numărul maxim de etape în spațiul indivizilor este poate fi *m* (gradul matricei $\frac{1}{n} X^t X$) iar în spațiul variabilelor numărul maxim de etape poate fi *n* (gradul matricei $\frac{1}{n} X \cdot X^t$). Numărul valorilor proprii nenule este minim(*m, n*).

Evaluarea rezultatelor

Criterii de alegere a numărului de axe

Scopul primordial al analizei în componente principale este acela de a sintetiza varianța existentă la nivelul întregului set de date, de a scoate în evidență ceea ce este semnificativ. Componentele principale sintetizează informații independente intre ele dar nu la fel de importante. Astfel primul tip de informație, furnizat de componenta 1 este cel mai important pentru că este cel care generează cantitatea maximă de variabilitate, al doilea tip este mai puțin important, șamd. Problema care se pune este: câte tipuri de informație merită analizate, aprofundate? Geometric, problema constă în a determina numărul de axe alese pentru reprezentarea multidimensională astfel încât acoperirea informațională să fie satisfăcătoare.

1. ***Criteriul procentului de acoperire***

O preocupare importantă în evaluarea calității analizei este determinarea cantității de varianță explicată prin fiecare axă.

Deoarece criteriul de optim în alegerea axei *k* a fost maximizarea variantei pe axa respectivă, se poate scrie:

$\frac{1}{n}(a_k)^t X^t X a_k = (a_k)^t \alpha_k a_k = \alpha_k$.

Varianța explicată prin axa *k* este deci α_k. Știm că varianța totală este *m* (tabelul *X* este standardizat). Prin urmare, procentul de varianță explicat prin axa *k* este α_k/m , iar procentul de varianță explicat prin primele *k* axe este $v_k = \frac{\sum_{j=1}^{k} \alpha_j}{\sum_{i=1}^{m} \alpha_i}$. În cazul în care variabilele *X* sunt standardizate, $v_k = \frac{\sum_{j=1}^{k-1} \alpha_j}{m}$.

În mod similar, la analiza în spațiul variabilelor, la etapa *k*, corelația dintre variabila nouă C^k şi variabilele vechi este:

$$\sum_{j=1}^{m} R^2(C_k, X_j) = \frac{1}{n} \frac{(C_k)^t XX^t C_k}{(C_k)^t C_k} = \frac{(C_k)^t \alpha_k C_k}{(C_k)^t C_k} = \alpha_k.$$

Deci valoarea proprie α_k este tocmai suma coeficienților de determinare dintre componenta principală şi variabilele observate.

Dacă se notează cu *s* numărul axelor semnificative, conform criteriului procentului de acoperire, *s* este prima valoare pentru care $v_s>P$, unde *P* este procentul de acoperire ales.

2. ***Criteriul lui Kaiser***

Deoarece variabilele observate sunt standardizate și au varianță 1, apare firesc în această situație ca noile variabile, componentele principale, să fie considerate importante, semnificative, în măsura în care cumulează mai multă varianță decât o variabilă observată. Criteriul lui Kaiser recomandă reținerea acelor componente principale care au varianță mai mare decât 1.

3. ***Criteriul Cattell***

Acest criteriu se poate aplica în variantă grafică (vizual - Figura 2) sau analitică. În varianta grafică se determină primul unghi dintre două pante consecutive mai mare decât 180º. Se reţin doar valorile proprii de până în acel punct, inclusiv.

În varianta analitică se calculează diferențe de ordinul doi între valorile proprii:

$\varepsilon_k = \alpha_k - \alpha_{k+1}$, $k = 1, m\text{-}1$

$\delta_k = \varepsilon_k - \varepsilon_{k+1}$, $k=1, m\text{-}2$

şi se determină primul indice *s* astfel încât $\delta_s < 0$. Acest indice corespunde unui unghi între două pante determinate de dreptele (s-1,s) și (s, s+1) mai mare decât 180º.

Componentele principale semnificative după criteriul Cattell vor fi componentele C_j, $j = \overline{1, s+1}$.

Exemplu. In tabelul 1 sunt prezentate 8 valori proprii reprezentând varianțele a 8 componente principale rezultate în urma aplicării metodei pe un set de date cu 8 variabile observate. Prima diferență δ mai mică decât 0 este -0.15 și corespunde componentei 3 (Tabelul 2). Deci numărul de componente reținute este 4 (Marcajul albastru din figura 2).

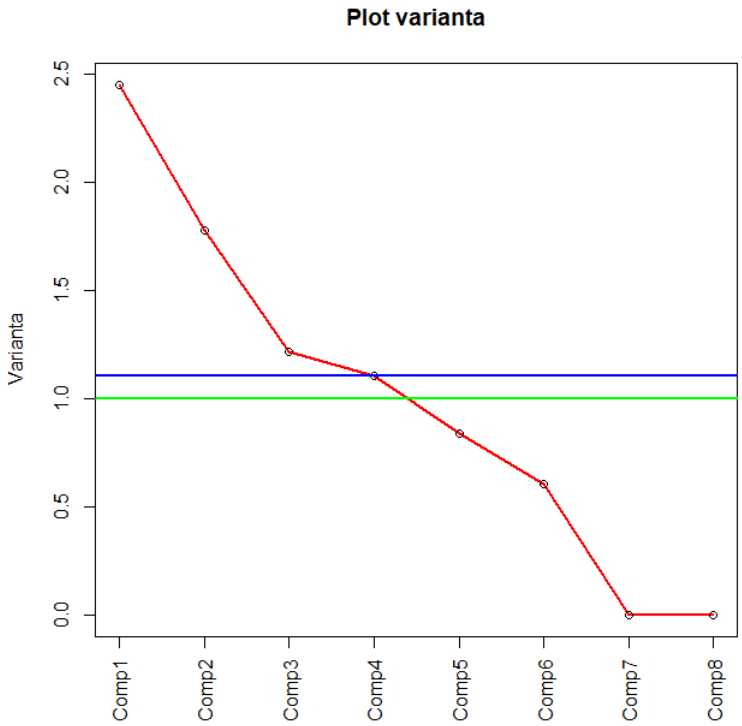

Figura 2. Alegerea numărului de axe. Criteriul Cattell

Număr valoare proprie	α	ε	δ
1	2.45	0.67	0.11
2	1.78	0.56	0.45
3	1.22	0.11	-0.15
4	1.11	0.26	0.03
5	0.84	0.24	-0.37
6	0.61	0.61	0.61
7	0.00	0.00	
8	0.00		

Tabelul 1. Alegerea numărului de axe. Criteriul Cattell

Scorurile

Scorurile sunt standardizări ale componentelor principale: $C_{ik}^s = \frac{C_{ik}}{\sqrt{\alpha_k}}$, i=1,n, k = 1,s, unde $\sqrt{\alpha_k}$ este abaterea standard a componentei C_k.

Calitatea reprezentării unui punct

Componentele principale constituie o nou spațiu de reprezentare a indivizilor numit *spațiul principal*. Baza acestui spațiu, vectorii unitari ai axelor, este constituită de vectorii proprii a_k, k = 1,m, iar coordonatele indivizilor în aceste noi axe sunt date de vectorii C_k, k =1,m. Așa cum am văzut un individ este reprezentat geometric de un punct într-un spațiu m-dimensional. Pătratul distanţelor de

la un punct de index *i* spre centrul de greutate al norului de date este $\sum_{k=1}^{m} c_{ik}^2$. Un individ este cu atât mai bine reprezentat pe o axă oarecare a_j cu cât c_{ij}^2 are o valoare mai mare în raport cu suma pătratelor proiecțiilor punctului pe celelalte axe, $\sum_{k=1}^{m} c_{ik}^2$.

Calitatea reprezentării individului *i* pe axa a_j este dată deci de raportul $Q_{ij}=\frac{c_{ij}^2}{\sum_{k=1}^{m} c_{ik}^2}$. Valoarea raportului este egală şi cu pătratul cosinusului unghiului dintre vectorul punctului *i* şi vectorul a_j. În figura 2 este prezentată o situație în care calitatea reprezentării instanței *i* pe axa a_1 este mai mare decât calitatea reprezentării pe axa a_2.

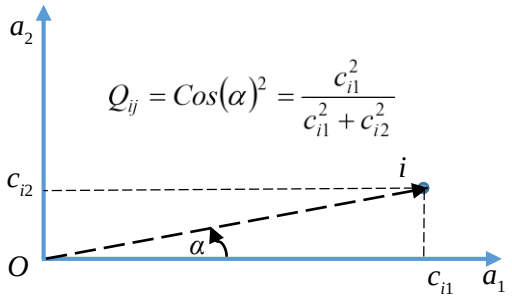

Figura 2. Valoarea cosinus pentru o instanță *i* în raport cu axa a_1 într-un spațiu bidimensional

Contribuția indivizilor la varianța axelor

Varianța explicată prin axa a_j este $\frac{1}{n}\sum_{i=1}^{n} c_{ij}^2 = \alpha_j$. Partea din această varianță datorată individului *i* este $\frac{c_{ij}^2}{n}$. Contribuția individului *i* la varianța axei a_j este deci : $\beta_{ij} = \frac{c_{ij}^2}{n \cdot \alpha_j}$.

Comunalitățile

Se numește comunalitate (communality) cantitatea de varianță explicată ***în comun*** de către un grup de componentele principale. Am văzut că o componentă principală, C_k, preia o cantitate de varianță egală cu α_k, iar suma coeficienților de determinare dintre această componentă și variabilele observate este tot α_k. Comunalitatea unei variabile X_j în raport cu primele *s* componente principale este deci suma coeficienților de determinare dintre variabilă și aceste componente principale: $\sum_{k=1}^{s} R(X_j, C_k)^2$. Pentru *s* = *m* această valoare este 1. Cele *m* componente principale preiau integral informația din *X*.

Calculul coeficienților de corelație

Coeficienții de determinare dintre o variabilă observată X_j şi componenta principală C_r se calculează astfel:

$R^2(C_r, X_j) = \frac{\mathrm{Cov}(C_r, X_j)^2}{Var(C_r)Var(X_j)} = \frac{Cov(C_r, X_j)^2}{\alpha_r}$, deoarece $\mathrm{Var}(C_r)=\alpha_r$, iar $\mathrm{Var}(X_j)=1$.

Matriceal, obținem vectorul coeficienților de corelație dintre variabilele observate şi C_r:

$$R_r = \frac{\frac{1}{n} X^t C_r}{\sqrt{\alpha_r}} = \frac{\frac{1}{n} X^t X a_r}{\sqrt{\alpha_r}} = \frac{\alpha_r a_r}{\sqrt{\alpha_r}} = a_r\sqrt{\alpha_r}\ .$$

Aceste corelații sunt numite ***corelații factoriale*** (*factor loadings*).

Variabile și indivizi suplimentari

În unele situații poate prezenta interes reprezentarea grafică a unuia sau mai multor indivizi sau a uneia sau mai multor variabile care nu figurau inițial în tabelul de observații inițial. Coordonatele acestor indivizi/variabile suplimentare se calculează în același mod

ca pentru indivizii/variabilele inițiale. Dacă se notează cu $Y = \begin{bmatrix} y_{11} & y_{12} & \dots & y_{1m} \\ y_{21} & y_{22} & \dots & y_{2m} \\ \dots & & & \\ y_{q1} & y_{q2} & \dots & y_{qm} \end{bmatrix}$ matricea de observații cu setul de q instanțe suplimentare, coordonatele acestor instanțe în spațiul principal sunt calculate astfel:

$C_Y = Y \cdot A$,

unde A este matricea vectorilor proprii, coeficienții legăturii liniare dintre componentele principale și variabilele observate.

Dacă se notează cu $Z = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1p} \\ z_{21} & z_{22} & \dots & z_{2p} \\ \dots & & & \\ z_{n1} & z_{n2} & \dots & z_{np} \end{bmatrix}$ matricea de observații cu setul de p variabile suplimentare, corelațiile dintre acestea și variabilele observate vor fi calculate în matricea:

$R_{ZC} = \begin{bmatrix} R(Z_1, C_1) & R(Z_1, C_2) & \dots & R(Z_1, C_m) \\ R(Z_2, C_1) & R(Z_2, C_2) & \dots & R(Z_2, C_m) \\ \dots & & & \\ R(Z_p, C_1) & R(Z_p, C_2) & \dots & R(Z_p, C_m) \end{bmatrix}$,

unde $R(Z_j, C_k),\ j = \overline{1, p},\ k = \overline{1, m}$ reprezintă coeficientul de corelație dintre variabila suplimentară Z_j și componente C_k.

Analiza în componente principale ponderată

Presupunem că ponderea indivizilor este diferită de $\frac{1}{n}$. Fie p_i, ponderea asociată individului i, $0 < p_i < 1$, $\sum_{i=1}^{n} p_i = 1$.

Fie P matricea ponderilor: $P = \begin{bmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & p_n \end{bmatrix}$.

Criteriul de optim în spațiul indivizilor devine următorul: $Maxim \sum_{i=1}^{n} p_i c_{ik}^2$.

$\sum_{i=1}^{n} p_i c_{ik}^2 = C_k^t P C_k = a_k^t X^t P X a_k$. Deci criteriul de optim va fi:

$$\begin{cases} \underset{a_k}{Maxim}\ a_k^t X^t P X a_k \\ a_k^t a_k = 1 \\ a_k^t a_j = 0, \quad j = \overline{1, k-1} \end{cases}$$

În spațiul variabilelor suma coeficienților de determinare dintre componenta principală și variabilele observate va fi:

$$\sum_{j=1}^{m} R^2(X_j, C_k) = \sum_{j=1}^{m} \frac{Cov(X_j, C_k)^2}{Var(X_j) Var(C_k)} = \sum_{j=1}^{m} \frac{C_k^t P X_j X_j^t P C_k}{C_k^t P C_k} = \frac{C_k^t P X X^t P C_k}{C_k^t P C_k} =$$

$$\frac{C_k^t \sqrt{P} \sqrt{P} X X^t \sqrt{P} \sqrt{P} C_k}{C_k^t \sqrt{P} \sqrt{P} C_k} = \frac{C_k^t \sqrt{P} X X^t P \sqrt{P} C_k}{C_k^t \sqrt{P} \sqrt{P} C_k} = \frac{(C_k^{''})^t X X^t P C_k^{''}}{(C_k^{''})^t C_k^{''}}$$

unde $\sqrt{P}$ este matricea diagonală a rădăcinilor pătrate din ponderi iar $C_k^{''} = C_k \sqrt{P} = \sqrt{P} C_k$.

Deci $C_k^{''}$ este vector propriu al matricei XX^tP (curs 3):

$XX^t P C_k^{''} = \beta_k C_k^{''} \Rightarrow XX^t P C_k \sqrt{P} = \beta_k C_k \sqrt{P}$

$\Rightarrow XX^t P C_k = \beta_k C_k$, deoarece $\sqrt{P}$ este o matrice pătratică inversabilă

C_k este vector propriu al matricei XX^tP corespunzător valorii proprii β_k

În concluzie, pentru analiza în componente principale ponderată vectorii a_k vor fi calculați ca vectori proprii succesivi ai matricei $X^t P \cdot X$ pentru o rezolvare a modelului în spațiul instanțelor, iar componentele principale ca vectori proprii succesivi ai matricei $X \cdot X^t P$ pentru o rezolvare în spațiul variabilelor.