ANALIZA GLAVNIH KOMPONENATA

- Analiza glavnih komponenata bavi se tumačenjem strukture matrice varijanci i kovarijanci skupa izvornih varijabli pomoću malog broja njihovih linearnih kombinacija.

- Osnovni ciljevi analize su:

- Redukcija podataka
- Interpretacija

Premda je p ulaznih varijabli odabrano kako bi se opisala varijablnost cijelog sustava, često je velik dio tog varijabiliteta opisan malim brojem k glavnih komponenata ($k<p$) .

Ako je to ispunjeno, k glavnih komponenata sadrži jednaku količinu informacija kao p ulaznih varijabli.

Stoga se početni skup podataka koji se sastoji od n mjerenja na p ulaznih varijabli može reducirati na skup od n mjerenja na k glavnih komponenata.

Analiza glavnih komponenata otkriva povezanost među varijablama i stoga dozvoljava interpretacije do kojih se inače bez ovako provedene analize ne bi došlo.

- Analiza glavnih komponenata često služi kao međukorak za provođenje drugih metoda kao primjerice:
- regresijske
- klaster
- ili faktorske analize.

Algebarski, glavne komponente su linearne kombinacije p slučajnih varijabli .

Geometrijski su te linearne kombinacije koordinatne osi novog koordinatnog sustava dobivenog rotacijom oko starog s glavnim komponentama kao koordinatnim osima.

- Kao što će se vidjeti, glavne komponente reprezentiraju smjer maksimalnog varijabiliteta i omogućuju jednostavniji opis kovarijančne strukture.
- Također će se vidjeti da glavne komponente ovise samo o matrici varijanci i kovarijanci (odnosno o korelacijskoj matrici) polaznih varijabli $X_1, X_2, \cdots, X_p$

Neka slučajni vektor $X' = [X_1, X_2, \cdots, X_p]$

ima matricu varijanci i kovarijanci Σ sa svojstvenim vrijednostima (eigenvalues, latent roots): $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

Promotrimo linearne kombinacije:

$$
\begin{aligned}
Y_1 &= a_1' X = a_{11} X_1 + a_{12} X_2 + \cdots + a_{1p} X_p \\
Y_2 &= a_2' X = a_{21} X_1 + a_{22} X_2 + \cdots + a_{2p} X_p \\
&\vdots \qquad\qquad\qquad\qquad\qquad \vdots \\
Y_p &= a_p' X = a_{p1} X_1 + a_{p2} X_2 + \cdots + a_{pp} X_p
\end{aligned}
$$

(1.1)

Odatle je:

$$Var(Y_i) = Var(a_i' X) = \quad a_i' \Sigma a_i \qquad\qquad i = 1,2,\cdots, p$$

$$Cov(Y_i, Y_k) = a_i' \Sigma a_k \qquad\qquad i, k = 1,2,\cdots, p$$

(1.2)

Glavne komponente su one linearne kombinacije $Y_1, Y_2, \cdots, Y_p$

čije su varijance što je moguće veće.

Prva glavna komponenta je linearna kombinacija s najvećom varijancom.

S obzirom da se varijanca $\boldsymbol{Var(Y_1) = a_1' \Sigma a_1}$ može povećati množenjem vektora konstantom, pažnja se ograničava na vektore koeficijenata duljine jedan.

Prva glavna komponenta = linearna kombinacija $\boldsymbol{Y_1 = a_1' X}$

koja maksimizira

$$\boldsymbol{Var(Y_1) = a_1' \Sigma a_1}$$ uz uvjet $$\boldsymbol{a_1' a_1 = 1}$$

Druga glavna komponenta = linearna kombinacija $\boldsymbol{Y_2 = a_2' X}$

koja maksimizira $\boldsymbol{Var(Y_2) = a_2' \Sigma a_2}$ uz uvjet $\boldsymbol{a_2' a_2 = 1}$ i

$$Cov(a_1' X, a_2' X) = 0$$

.

.

.

i-ta glavna komponenta = linearna kombinacija $Y_i = a_i' X$

koja maksimizira $Var(Y_i) = a_i' \Sigma a_i$ uz uvjet

$a_i' a_i = 1$ i $Cov(a_i' X, a_k' X) = 0 \quad za \quad k < i$

TEOREM 1

Neka je ∑ matrica varijanci i kovarijanci pridružena slučajnom vektoru:

$$X' = \left[X_1, X_2, \cdots, X_p\right]$$

Neka su parovi svojstvenih vrijednosti i svojstvenih vektora matrice ∑:

$$(\lambda_1, e_1), (\lambda_2, e_2), \cdots, (\lambda_p, e_p)$$

pri čemu vrijedi $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_i \geq \cdots \geq \lambda_p \geq 0$

Tada je i-ta glavna komponenta dana s:

$$Y_i = e_i' X = e_{i1} X_1 + e_{i2} X_2 + \cdots + e_{ip} X_p \qquad i = 1,2,\cdots, p \tag{1.3}$$

Uz takav izbor

$$Var(Y_i) = e_i' \Sigma e_i = \lambda_i \qquad i = 1,2,\cdots, p$$

$$Cov(Y_i, Y_k) = Cov(e_i' \Sigma e_k) = 0 \qquad za \quad i \neq k \tag{1.4}$$

Ako su neke svojstvene vrijednosti međusobno jednake izbor odgovarajućih koeficijenata vektora e_i, dakle i Y_i nije jednoznačan.

Dokaz:

$$\underbrace{\max}_{a \neq 0} \frac{a'\Sigma a}{a'a} = \lambda_1$$ (dostignuto ako je $a = e_1$)

No $e_1 e_1' = 1$, jer su svojstveni vektori normalizirani.

Odatle je:

$$\underbrace{\max}_{a \neq 0} \frac{a'\Sigma a}{a'a} = \lambda_1 = \frac{e_1'\Sigma e_1}{\underbrace{e_1'e_1}_{=1}} = e_1' \underbrace{\Sigma e_1}_{=\lambda_1 e_1} = \lambda_1 e_1' e_1 = \lambda_1 = Var(Y_1)$$

Slično:

$$\underbrace{\max}_{a \perp e_1, e_2, \cdots, e_k} \frac{a'\Sigma a}{a'a} = \lambda_{k+1} \qquad k = 1,2,\cdots, p$$ (dostignut ako je $a = e_{k+1}$, uz uvjet $e'_{k+1} e_i = 0$ za $i = 1,2,\cdots,k$)

$$\lambda_{k+1} = \frac{e'_{k+1}\Sigma e_{k+1}}{\underbrace{e'_{k+1} e_{k+1}}_{=1}} = e'_{k+1}\Sigma e_{k+1} = Var(Y_{k+1})$$

Svojstveni vektori su za različite svojstvene vrijednosti međusobno okomiti.

S obzirom da je

$$\Sigma e_k = \lambda_k e_k$$

Množenjem gornje jednadžbe s $e_i', i \neq k$ dobit će se:

$$Cov(Y_i, Y_k) = e_i' \underbrace{\Sigma e_k}_{\lambda_k e_k} = \lambda_k e_i' e_k = 0$$

Posljedica ovog teorema je da su glavne komponente nekorelirane i da su im varijance jednake svojstvenim vrijednostima matrice Σ

TEOREM 2

Neka $X' = \left[X_1, X_2, \cdots, X_p\right]$ ima matricu varijanci i kovarijanci Σ s parovima svojstvenih vrijednosti i svojstvenih vektora $(\lambda_1, e_1), (\lambda_2, e_2), \cdots, (\lambda_p, e_p)$, pri čemu vrijedi $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Neka su $Y_i = e_i' X = e_{i1} X_1 + e_{i2} X_2 + \cdots + e_{ip} X_p \qquad i = 1, 2, \cdots, p$ glavne komponente.

Tada vrijedi:

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

DOKAZ:

$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = tr(\Sigma)$ je trag matrice varijanci i kovarijanci.

Može se pisati

$\Sigma = P\Lambda P'$ pri čemu je Λ dijagonalna matrica svojstvenih vrijednosti, a matrica $P = \left[e_1, e_2, \cdots, e_p\right]$ je matrica svojstvenih vektora. Vrijedi $PP' = P'P = I$.

Trag od Σ:

$$tr(\Sigma) = tr(P\Lambda P') = tr(\Lambda PP') = tr(\Lambda) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Y_i),$$

što je trebalo dokazati.

kao posljedica ovog rezultata proporcija ukupne varijance protumačene k-tom glavnom komponentom je:

$$\text{(proporcija populacijske varijance objašnjena k-tom glavnom komponentom)} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad k = 1,2,\cdots,p \qquad (1.6)$$

Ako je velik dio varijance (80%-90%) protumačen jednom, dvije ili tri glavne komponente, tada te komponente mogu zamijeniti početnih p varijabli bez većeg gubitka informacija.

Promatra se veličina svake komponente vektora

$$e_i' = \begin{pmatrix} e_{i1} & \cdots & e_{ik} & \cdots & e_{ip} \end{pmatrix}$$

Komponenta e_{ik} mjeri važnost k-te varijable na i-toj glavnoj komponenti, neovisno o drugim varijablama.

Koeficijent e_{ik} proporcionalan je koeficijentu linearne korelacije između Y_i i X_k

TEOREM 3

Neka su

$$Y_1 = e_1' X$$
$$Y_2 = e_2' X$$
$$\vdots$$
$$Y_i = e_i' X$$
$$\vdots$$
$$Y_p = e_p' X$$

glavne komponente dobivene iz matrice varijanci i kovarijanci ∑

Tada su: $$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \qquad i,k = 1,2,\cdots,p$$

keficijenti linearne korelacije između Y_i i X_k

Ovdje su $(\lambda_1, e_1), (\lambda_2, e_2), \cdots, (\lambda_p, e_p)$ parovi svojstvenih

vrijednosti i svojstvenih vektora matrice $\sum$.

DOKAZ:

Stavi li se:

$a_k' = \left[0, \cdots 0, \underbrace{1}_{a_{kk}}, 0, \cdots, 0 \right]$ **tako da je** $X_k = a_k' X$ **, tada je:**

$$Cov(X_k, Y_i) = Cov(a_k' X, e_i' X) = a_k' \underbrace{\Sigma e_i}_{=\lambda_i e_i} = a_k' \lambda_i e_i = \lambda_i e_{ik}$$

Nadalje, kako je:

$Var(Y_i) = \lambda_i$ **i** $Var(X_k) = \sigma_{kk}$

to je koeficijent korelacije između Y_i **i** X_k **:**

$$\rho_{Y_i, X_k} = \frac{Cov(Y_i, X_k)}{\sqrt{Var(Y_i)} \cdot \sqrt{Var(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i,k = 1,2,\cdots,p$$

- **Premda korelacije između ulaznih varijabli i glavnih komponenata često pomažu pri interpretaciji komponenata, one mjere samo univarijatni doprinos jedne varijable na komponentu Y_i .**
- **Koeficijenti linearne korelacije ne pokazuju važnost individualne varijable na Y_i u prisustvu drugih varijabli.**

Zbog tog razloga statističari preporučuju da se samo koeficijenti e_{ik}, a ne koeficijenti korelacije koriste pri interpretaciji komponenata.

No, iako koeficijenti i korelacije, kao mjere značajnosti varijabli, vode do različitih rangiranja, u praksi varijable s relativno velikim koeficijentima (po apsolutnoj vrijednosti) imaju i relativno visoke korelacije.

Stoga dvije mjere značajnosti (važnosti) varijabli, od kojih je prva multivarijatna, a druga univarijatna daju često slične rezultate.

Imajući to u vidu, korisno je izračunati i koeficijente i korelacije kako bi se lakše interpretirale glavne komponente.

Glavne komponente izračunate polazeći od standardiziranih varijabli

Glavne komponente moguće je dobiti za p standardiziranih varijabli:

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}, \quad Z_2 = \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}}, \ \cdots, Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}} \tag{1.8}$$

što se u matričnoj notaciji može zapisati:

$$Z = (V^{-\frac{1}{2}})^{-1}(X - \mu) \tag{1.9}$$

Matrica $V^{\frac{1}{2}}$ je dijagonalna matrica standardnih devijacija originalnih varijabli:

$$V^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

Očito je:

$$E(Z) = 0 \text{ , a } Cov(Z) = (V^{-\frac{1}{2}})^{-1} \Sigma (V^{-\frac{1}{2}})^{-1} = \rho$$

Pri tom je:

$$\Sigma = E(X-\mu)(X-\mu)' = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix} \quad \rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}} \cdot \sqrt{\sigma_{kk}}}$$

- Glavne komponente mogu se izvesti pomoću svojstvenih vektora korelacijske matrice ρ od **X.**
- Svi se ranije izvedeni rezultati mogu primijeniti uz određena pojednostavljenja, s obzirom da su varijance standardiziranih varijabli jednake 1. Nastavit ćemo s oznakama $\boldsymbol{Y_i}$ za i-tu glavnu komponentu i $(\lambda_i, \boldsymbol{e}_i)$ za par i-te svojstvene vrijednosti i pridruženog i-tog svojstvenog vektora matrice Σ ili matrice ρ.

TEOREM 4

Neka je $Z' = (Z_1, Z_2, \cdots, Z_p)$ vektor standardiziranih varijabli s $Cov(Z) = \rho$.

Tada je i-ta glavna komponenta:

$$Y_i = e_i' Z = e_i' (V^{\frac{1}{2}})^{-1} (X - \mu) \qquad \mathbf{i = 1,2,\cdots,p}$$

Nadalje:

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p \qquad (1.10)$$

i

$$\rho_{Y_i Z_k} = e_{ik} \sqrt{\lambda_i} \qquad i, k = 1,2,\cdots,p$$

U ovom su slučaju $(\lambda_1, e_1), (\lambda_2, e_2), \cdots, (\lambda_p, e_p)$ parovi svojstvenih vrijednosti i svojstvenih vektora matrice ρ, s $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

DOKAZ

Dokaz teorema 4 slijedi iz teorema 1,2 i 3 ako se stavi $Z' = (Z_1, Z_2, \cdots, Z_p)$ umjesto

$X' = \left[X_1, X_2, \cdots, X_p\right]$, i matrica ρ umjesto matrice Σ.

Iz (1.10) se vidi da je ukupna varijanca standardiziranih varijabli jednaka p (to je trag matrice ρ. Korištenjem relacije (1.6) s vektorom Z umjesto X, dobit će se da je proporcija ukupne varijance objašnjena k-tom glavnom komponentom od Z:

(proporcija populacijske varijance objašnjena k-tom glavnom komponentom)

$$= \frac{\lambda_k}{p} \qquad k = 1,2,\cdots,p \tag{1.11}$$

Pri čemu su $\lambda_k, k = 1,2,\cdots p$ svojstvene vrijednosti matrice ρ.

Primjer 1.1

Nakon snažne oluje 1. veljače 1898. u Herman Bumpus-ovom laboratoriju na Brown University na Rhode Islandu proučavan je veći broj (nastradalih) umirućih vrabaca.

Oko polovica ptica je uginula, a Bumpus je taj događaj tretirao kao priliku da ispita može li dati potporu Darwinovoj teoriji prirodne selekcije. Proveo je 8 morfoloških mjerenja na svakoj ptici, a također je izmjerio njihovu težinu. Rezultati 5 mjerenja za ptice ženskog spola predočeni su u datoteci Ptice.xls

Na osnovi prikupljenih podataka Bumpus je zaključio da ptice koje su stradale, nisu stradale slučajno, nego stoga jer su bile fizički diskvalificirane, a da su one ptice koje su preživjele, preživjele zato jer su imale određene fizičke karakteristike. Posebno je utvrdio da su preživjele ptice bile kraće, manje teške, te da su imale dulje kosti krila, dulje noge, dulju prsnu kost i veći moždani kapacitet od onih koje nisu preživjele. Također je zaključio da je proces selektivne eliminacije najjače povezan s ekstremnom varijablom jedinke, bez obzira na smjer varijacije. Zaključio je da je jednako opasno biti iznad određenog standarda organske izvrsnosti, kao i ispod tog standarda. Time je rečeno da se dogodila stabilizacija selekcije i da su jedinke s mjerenjima bliže prosjeku bolje preživjele od jedinki s mjerenjima daleko od prosjeka.

U doba dok je Bumpus pisao svoj rad razvoj multivarijatnih metoda je bio na početku. 1897. je Francis Galton predstavio koeficijent korelacije kao mjeru povezanosti među varijablama. Tek 56 godina kasnije Harold Hotelling je opisao metodu provođenja analize glavnih komponenata, koja se može primijeniti na Bumpusove podatke. Bumpus nije čak ni računao standardne devijacije, no njegove su podatke ponovo analizirali brojni autori i općenito potvrdili njegove zaključke.

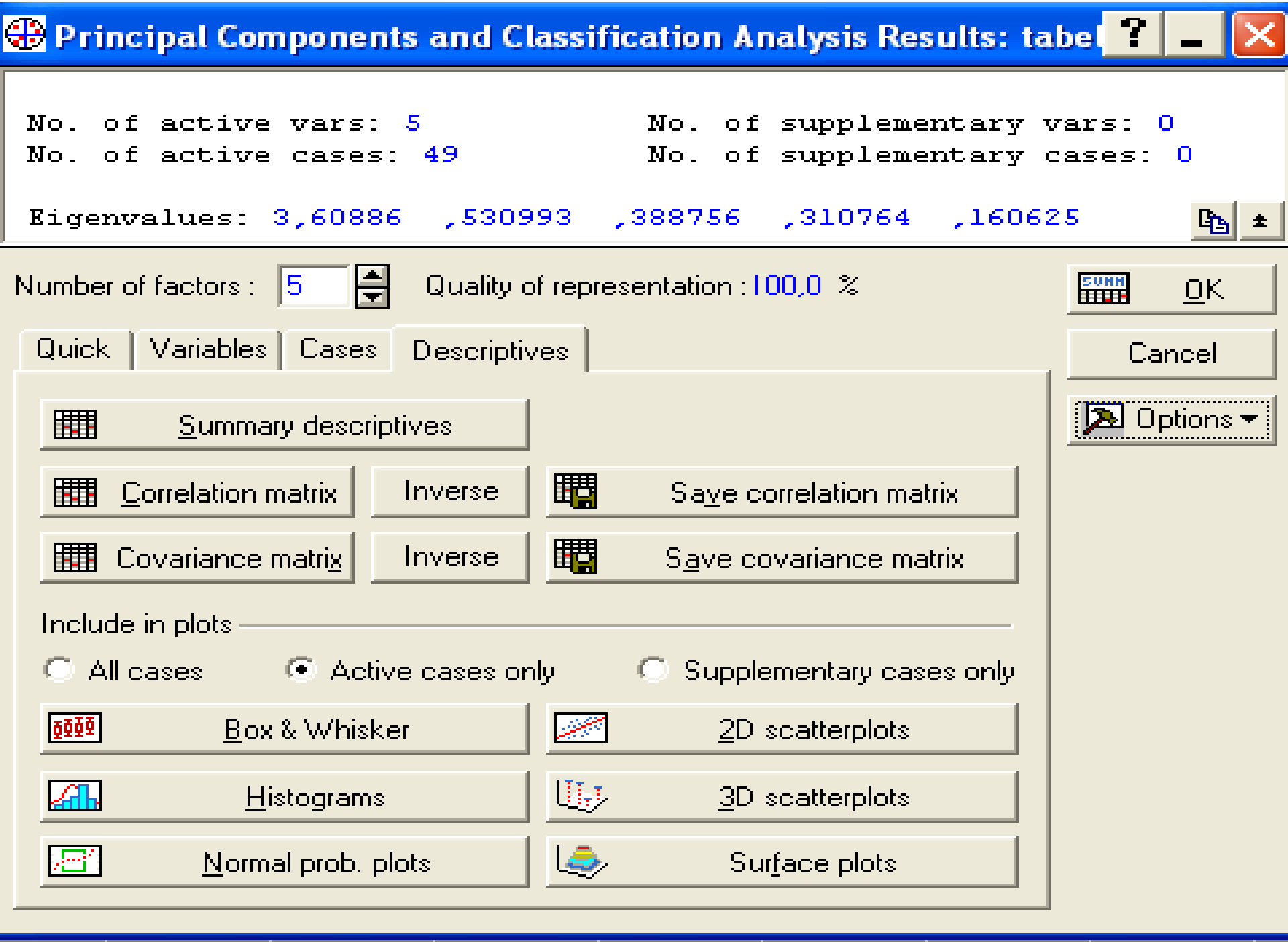
Odaberu li se opisani podaci kao primjer za ilustraciju multivarijatnih metoda, javljaju se slijedeća interesantna pitanja:

Na koji su način povezana različita mjerenja? Je li npr velika vrijednost jedne varijable povezana s velikom vrijednosti druge varijable?
Jesu li sredine varijabli preživjelih i uginulih jedinki sitatistički signifikantno različite?
Imaju li preživjele i uginule jedinke slični iznos varijacije za pojedine varijable?

Primjer 1.1

- **Svojstvene vrijednosti korelacijske matrice za 5 mjerenja na 49 ženskih vrabaca**

- X1=duljina tijela X2= opseg krila X3=duljina vrata i glave
  X4= duljina humerusa (nadlaktična kost) X5=duljina prsne kosti
- **Primjenom programskog paketa Statistica dobiveni su između ostalih slijedeći rezultati:**


Principal Components and Classification Analysis Results: tabe
No. of active vars: 5
No. of supplementary vars: 0
No. of active cases: 49
No. of supplementary cases: 0
Eigenvalues: 3,60886 ,530993 ,388756 ,310764 ,160625
Number of factors : 5
Quality of representation : 100,0 %
OK
Cancel
Options
Quick
Variables
Cases
Descriptives
Summary descriptives
Correlation matrix
Inverse
Save correlation matrix
Covariance matrix
Inverse
Save covariance matrix
Include in plots
All cases
Active cases only
Supplementary cases only
Box & Whisker
2D scatterplots
Histograms
3D scatterplots
Normal prob. plots
Surface plots

Svojstvene vrijednosti i pridruženi pokazatelji

Eigenvalues of covariance matrix, and related statistics (tabela1.1) Active variables only

	Eigenvalue	% Total	Cumulative	Cumulative
1	3,608862	72,17724	3,608862	72,1772
2	0,530993	10,61987	4,139855	82,7971
3	0,388756	7,77512	4,528612	90,5722
4	0,310764	6,21528	4,839375	96,7875
5	0,160625	3,21249	5,000000	100,0000

- **Svojstvene vrijednosti su varijance glavnih komponenata. Zbroj svojstvenih vrijednosti iznosi 5. U drugom je stupcu izračunat postotak ukupne varijance objašnjen svakom glavnom komponentom, a u 4. je dan kumulativni niz postotaka iz drugog stupca.**

Tako je npr prvom glavnom komponentom objašnjeno 72.18% ukupne varijance, drugom 10.62%, odnosno s prve dvije glavne komponente protumačeno je 82.80% ukupne varijance.

- **Drugi način gledanja na relativnu važnost pojedinih komponenata je uspoređivanje njihovih varijanci s varijancama ulaznih standardiziranih varijabli (koje su jednake 1).**

  Prva glavna komponenta ima varijancu 3.609 puta veću od varijance originalnih standardiziranih varijabli, druga ima varijancu samo 0.531 od varijance originalnih standardiziranih varijabli, a preostale glavne komponente objašnjavaju još manji dio varijacija. To potvrđuje važnost prve glavne komponente u odnosu na ostale.

Svojstveni vektori korelacijske matrice

Eigenvectors of correlation matrix (tabela1.1_A) Active variables only

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
z1	-0,450380	-0,018718	-0,730426	0,347058	0,377936
z2	-0,461962	0,299784	-0,273386	-0,595008	-0,517635
z3	-0,452537	0,312784	0,390085	0,643980	-0,361248
z4	-0,470349	0,169774	0,465880	-0,329877	0,651219
z5	-0,397154	-0,884942	0,150092	-0,044577	-0,186097

Prve dvije glavne komponente su:

$$Y_1 = -0.450380 Z_1 - 0.461962 Z_2 - 0.452537 Z_3 - 0.470349 Z_4 - 0.397154 Z_5$$

$$Y_2 = -0.018718 Z_1 + 0.299784 Z_2 + 0.312784 Z_3 + 0.169774 Z_4 - 0.88494 Z_5 ,$$

pri čemu su varijable $Z_1, \cdots Z_5$ **standardizirane ulazne varijable.**

Koeficijenti uz varijable $Z_1, \cdots Z_5$ u prvoj glavnoj komponenti su približno jednaki, i ta komponenta je povezana s veličinom vrabaca. Iz dobivenih rezultata proizlazi da je 72.18% svih varijacija podataka povezano s razlikama u njihovoj veličini.

Druga glavna komponenta je kontrast između Z_2, Z_3, Z_4 s jedne i Z_5 s druge strane[1]. To znači da će Y_2 biti velik ako Z_2, Z_3, Z_4 poprimaju velike, a Z_5 malu vrijednost i obratno. Stoga Y_2 predočuje razlike u obliku vrabaca. Mali koeficijent uz Z_1 (ukupna duljina) znači da vrijednost te varijable ne utječe na Y_2.

---

[1] X2=duljina krila X3=duljina vrata i glave X4= duljina humerusa (nadlaktična kost) X5=duljina prsne kosti

- **Vrijednosti glavnih komponenata mogu se koristiti za daljnje analize. One se računaju iz standardiziranih varijabli.**

U prvom retku Tabele1.1 su mjerenja za prvog vrapca itd.:

Tabela 1.1 Mjerenja tijela ženki

PTICA	X1	X2	X3	X4	X5
1	156	245	31,6	18,5	20,5
2	154	240	30,4	17,9	19,6
3	153	240	31,0	18,4	20,6

U slijedećoj tabeli tabeli navedene su sredine (Means) i standardne devijacije (Std. Dev.) ulaznih varijabli:

Summary statistics (tabela1.1)

	Mean	Std. Dev.
X1	157,8980	3,709475
X2	241,3265	5,067822
X3	31,4592	0,794753
X4	18,4694	0,564286
X5	20,8265	0,991374

Standardizirane vrijednosti mjerenja za prvog vrapca su:

$$Z_{11} = \frac{(X_{11} - \mu_1)}{\sigma_{11}} = \frac{(156 - 157.8980)}{3.709475} = -0.51165$$

$$Z_{12} = \frac{(X_{12} - \mu_2)}{\sigma_{22}} = \frac{(245 - 241.3265)}{5.067822} = 0.724862$$

$$Z_{13} = \frac{(X_{13} - \mu_3)}{\sigma_{33}} = \frac{(31.6 - 31.4592)}{0.794753} = 0.177182$$

$$Z_{14} = \frac{(X_{14} - \mu_4)}{\sigma_{44}} = \frac{(18.5 - 18.4694)}{0.564286} = 0.05425$$

$$Z_{15} = \frac{(X_{15} - \mu_5)}{\sigma_{55}} = \frac{(20.5 - 20.8265)}{0.001374} = -0.32937$$

Vrijednost prve glavne komponente za prvog vrapca je:

$$Y_{11} = -0.450380 \cdot (-0.51165) - 0.461962 \cdot 0.724862 - 0.452537 \cdot 0.177182 -$$
$$- 0.470349 \cdot 0.05452 \approx -0.079$$

$$Y_2 = -0.018718 Z_1 + 0.299784 Z_2 + 0.312784 Z_3 + 0.169774 Z_4 - 0.88494 Z_5$$

pa je vrijednost druge glavne komponente za prvog vrapca:

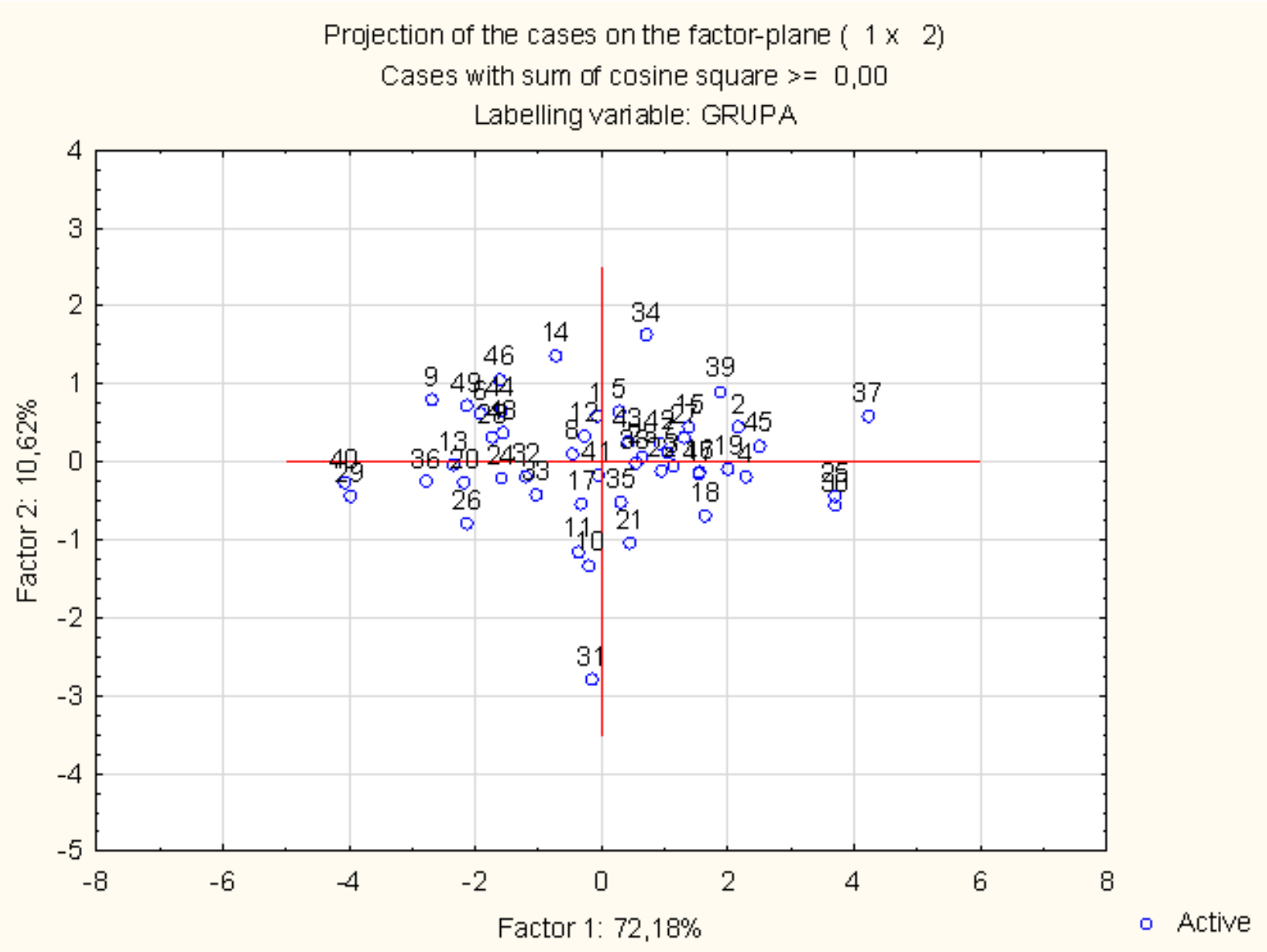
$$Y_{12} = -0.018718 \cdot (-0.51165) + 0.299784 \cdot 0.724862 + 0.312784 \cdot 0.177182 +$$
$$+ 0.169774 \cdot 0.0552 - 0.88494 \cdot (-0.32937) \approx 0.583$$

Vrijednosti (za prva četiri vrapca) su predočene u slijedećoj tabeli:

Factor coordinates of cases, based on correlations (tabela1.1_A)

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
1,000000	-0,07931	0,58298	0,22051	-0,497983	-0,535969
2,000000	2,16325	0,44789	-0,33656	-0,679181	-0,207070
3,000000	1,12565	-0,05314	0,71905	-0,623828	-0,192366
4,000000	2,29093	-0,18266	0,24726	0,191980	-0,471100

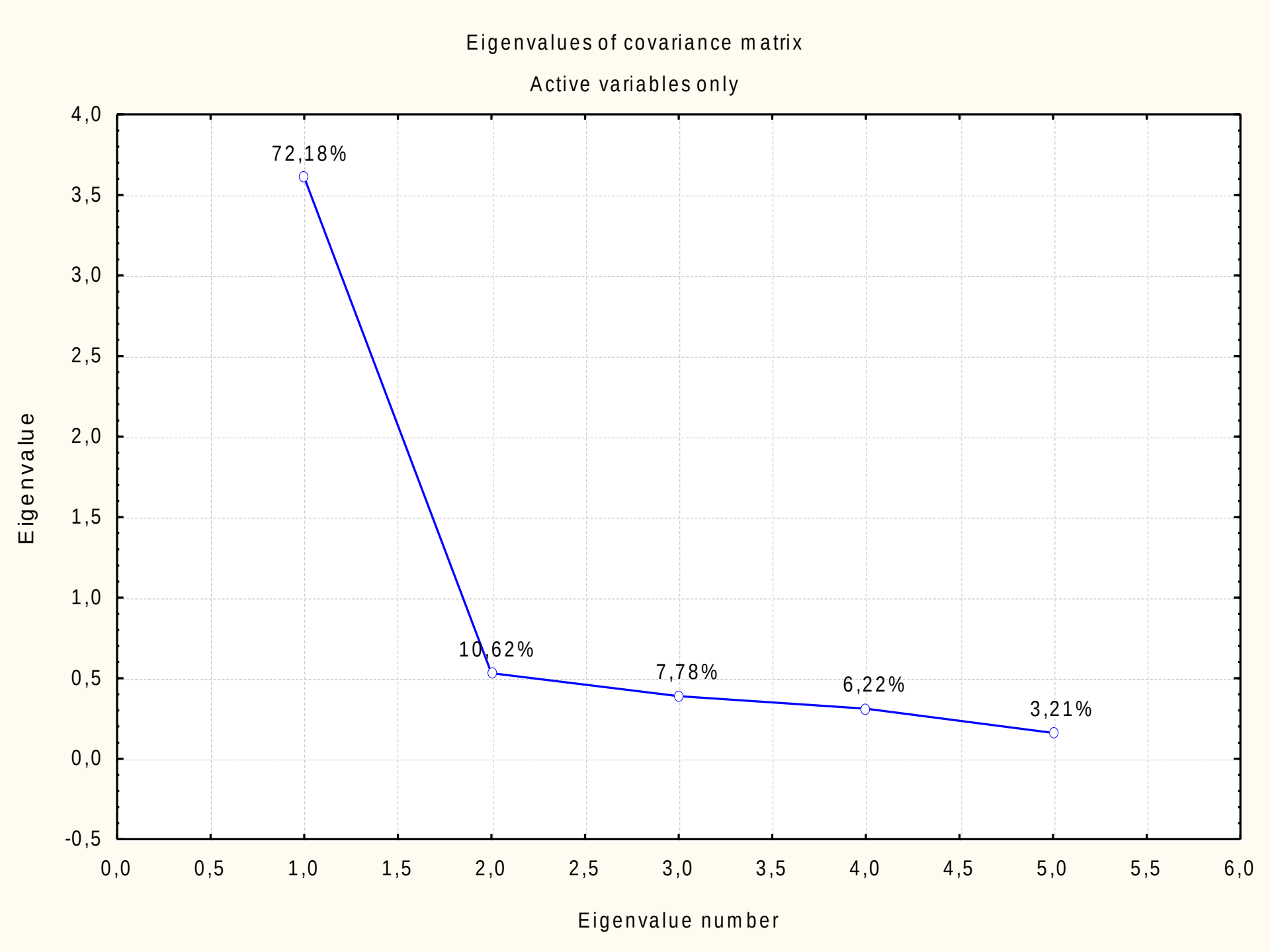
- **Promatrane ptice pokupljene su nakon snažne oluje. Prvih 21 vrabaca se oporavilo i preživjelo, a ostalih 28 je uginulo. Pitanje, pokazuju li preživjeli i uginuli vrapci bilo kakve razlike. Sa stajališta analize glavnih komponenata može se promatrati dijagram rasipanja za 49 vrijednosti prve i druge glavne komponente podijeljene u dvije grupe: preživjeli (označeni plavim krugom) i uginuli (označeni crvenim kvadratom):**


Projection of the cases on the factor-plane (1 x 2)
Cases with sum of cosine square >= 0,00
Labelling variable: GRUPA
Factor 2: 10,62%
Factor 1: 72,18%
Active

- **Na dijagramu rasipanja se jasno vidi da ptice s ekstremnim vrijednostima na prvoj (a jednako tako i na drugoj komponenti) nisu preživjele.**

Broj glavnih komponenata

- Uvijek se postavlja pitanje: Koliko glavnih komponenata treba zadržati?
- Ne postoji konačni odgovor na to pitanje. Pomoć pri donošenju odluke može pružiti scree-dijagram. To je dijagram koji dužinama povezuje točke u ravnini, čija je apscisa jednaka rednom broju svojstvene vrijednosti, a ordinata njenoj veličini. S obzirom da su svojstvene vrijednosti poredane u padajući niz, dobivena izlomljena linija je opadajuća. Smatra se da je broj glavnih komponenata koje ostaju određen točkom na pregibu iza koje su svojstvene vrijednosti male i koje se značajno ne razlikuju.


Eigenvalues of covariance matrix
Active variables only
72,18%
10,62%
7,78%
6,22%
3,21%
Eigenvalue
Eigenvalue number
4,0
3,5
3,0
2,5
2,0
1,5
1,0
0,5
0,0
-0,5
0,0
0,5
1,0
1,5
2,0
2,5
3,0
3,5
4,0
4,5
5,0
5,5
6,0

- **U promatranom primjeru pregib je za i=2. Svojstvene vrijednosti iza su male, te se mogu zadržati prve dvije glavne komponente.**