Standardizovana (normalizovana) vrednost obeležja

Normalizovano odstupanje je mera varijacije koja pokazuje algebarsko odstupanje jedne vrednosti obeležja od aritmetičke sredine, izraženo u standardnim devijacijama.

Ova mera je pogodna za upoređivanje varijacija obeležja iz različitih numeričkih serija čija su obeležja izražena u različitim jedinicama mere.

U literaturi se vrlo često za normalizovano odstupanje upotrebljava naziv z-skor.

Formule za izračunavanje su sledeće:

[2-33] Za osnovni skup: $$U = \frac{X - \mu}{\sigma}.$$

[2-34] Za uzorak: $$U_u = \frac{X - \bar{x}}{\sigma_u}.$$

Analiza glavnih komponenti (Principal Components Analysis)

Definicija

Ovu tehniku je prvi put opisao Karl Pearson 1901. godine. Iako je vršio izračunavanja sa samo dve ili tri varijable Pearson je verovao da se analiza glavnih komponenti može upotrebiti i za rešavanje problema sa puno više promenljivih. Opis izračunavanja je dat mnogo kasnije od strane Htelling-a, 1933. godine. Međutim, i dalje su izračunavanja bila previše komplikovana i zamorna kada bi trebalo napraviti analizu sa većim brojem varijabli. Široka upotreba analize glavnih komponenti je usledila zapravo tek sa pojavom računara.

Analiza glavnih komponenti predstavlja jednu od najjednostavnih multivarijantih tehnika. Ona se primenjuje kada je velik broj varijabli u skupu redundantan, odnosno kada se više varijabli odnosi na istu dimenziju i kada ne pružaju nikakvu dodatnu informaciju koja već nije obuhvaćena nekom drugom varijablom. Geometrijski gledano, to znači da na prostoru od *k* dimenzija imamo *p* varijabli pri čemu je $k<p$. Očekuje se da će k najvećih glavnih komponenti biti dovoljno da objasni varijabilitet podataka u skupu.

Cilj analize je da se uzme p varijabli (X_1, X_2,...,X_p) i da se pronađe kombinacija istih da bi se izračunale nove varijable (Z_1,Z_2,...,Z_p) koje međusobno nisu u korelaciji i koje će opisivati varijacije podataka. Nepostojanje korelacije znači da nove varijable mere međusobno različite „dimenzije" podataka. i njihove varijanse su poređane u opadajući niz ($Var(Z_1) \geq Var(Z_2) \geq ... \geq Var(Z_p)$). Promenljive Z predstavljaju zapravo glavne komponente.

Kada se radi analiza glavnih komponenti, želja je da varijanse većine promenljivih Z budu toliko male da su zanemarljive. U tom slučaju, veći deo varijacija originalnih podataka se može adekvatno opisati sa svega nekoliko glavnih komponenti, čime se postiže određeni stepen uštede.

Analiza glavnih komponenti ne uspeva uvek u tome da veliki broj originalnih varijabli X smanji na mali broj izvedenih varijabli Z. Ako originalne varijable nisu u korelaciji, analiza neće postići nikakav rezultat. Najbolji rezultati se postižu kada su originalne varijable u visokoj korelaciji, bilo pozitivnoj ili negativnoj. Ako postoji takav slučaj, onda se može očekivati da će se skup od 20 originalnih varijabli redukovati na svega dve ili tri glavne komponente. Pored toga, korisna će biti i činjenica da je otkriven visok stepen redundantnosti kod originalnih varijabli.

Izvedena proenljiva Z predstavlja zapravo prosek standardizovanih vrednosti obeležja originalnih promenljivih i može se posmatrati kao indeks.

Procedura

Analiza počinje sa podacima o p varijabli za n jedinica posmatranja, kao što pokazuje tabela. Prva glavna komponenta je tada linearna kombinacija originalnih varijabli (X_1, X_2,...,X_p):

$$Z_1 = a_{11} X_1 + a_{12} X_2 + ... + a_{1p} X_p ,$$

koje variraju što je više moguće individualno, pod uslovom da je zadovoljen uslov:

$$a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1 .$$

Tabela: Izgled podataka za analizu glavnih komponenti

Redni broj	X_1	X_2	...	X_p
1	x_{11}	x_{12}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2p}
...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{np}

Varijansa izvedene promenljive Z_1, $Var(Z_1)$, je zbog postavljenog uslova maksimalna. Navedeni uslov je postavljen jer bez njega $Var(Z_1)$ bi mogla da se povećava jednostavnim povećavanjem bilo koje od vrednosti a_{1j}.

Druga glavna komponenta se izračunava na sledeći način:

$Z_2 = a_{21} X_1 + a_{22} X_2 + \ldots + a_{2p} X_p$.

Varijansa druge glavne komponente, Var(Z_2), ima maksimalnu vrednost jer je postavljen uslov:

$a_{21}^2 + a_{22}^2 + \ldots + a_{2p}^2 = 1$.

Dodatni uslov je da izvedene varijable Z_1 i Z_2 imaju korelaciju nula. Naredne glavne komponente se definišu na identičan način. Ako postoji *p* varijabli u sistemu, onda će postojati i *p* glavnih komponenti.

Da bi se dobili i koristili rezultati analize glavnih komponenti, nije potrebno znati kako se jednačine za glavne komponente izračunavaju. Međutim, potrebno je poznavati prirodu jednačina. Analiza glavnih komponenti podrazumeva pronalaženje ajgenvrednosti matrice kovarijansi uzorka. Matrica kovarijansi je simetrična i ima sledeći oblik:

$$C = \begin{bmatrix} c_{11} & c_{12} & \ldots & \ldots & c_{1p} \\ c_{21} & c_{22} & \ldots & \ldots & c_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ c_{p1} & c_{p2} & \ldots & \ldots & c_{pp} \end{bmatrix}$$

U matrici dijagonalni elementi c_{ii} su zapravo varijansa originalne promenljive X_i, dok su ostali elementi van dijagonale, c_{ij}, kovarijanse originalnih promenljivih X_i i X_j.

Varijanse glavnih komponenti su ajgenvrednosti matrice C. Postoji *p* ajgenvrednosti, od kojih su neke nula. Negativne ajgenvrednosti ne mogu biti negativne u matrici kovarijansi. Ajgenvrednost λ_i je zapravo varijansa glavne komponente Z_i, odnosno $\lambda_i = Var(Z_i)$.

Važna osobina ajgenvrednosti je da je njihov zbir jednak zbiru elemenata na dijagonali matrice C:

$\lambda_1 + \lambda_2 + \ldots + \lambda_p = c_{11} + c_{22} + \ldots + c_{pp}$.

Iz ovoga proizilazi da je suma varijansi originalnih varijabli jednaka sumi varijansi izvedenih varijabli.To znači da glavne komponente obuhvataju sve varijacije originalnih podataka.

Da neka od originalnih varijabli ne bi imala prejak uticaj na glavne komponente i stvarala pristrasnost rezultata, vrši se njihovo kodiranje tako da imaju aritmetičku sredinu nula i varijansu jednaku jedinici. Drugim rečima izračunavaju se standardizovane vrednosti iz originalnih podataka. Matrica C tada ima sledeći oblik:

$$C = \begin{bmatrix} 1 & c_{12} & \ldots & \ldots & c_{1p} \\ c_{21} & 1 & \ldots & \ldots & c_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{bmatrix}$$

$$\begin{vmatrix} \ldots & \ldots & \ldots & \ldots & \ldots \\ c_{p1} & c_{p2} & \ldots & \ldots & 1 \end{vmatrix}$$

U ovoj matrici je $c_{ij}=c_{ji}$. Drugim rečima, analiza glavnih komponenti je izvedena na korelacionoj matrici. U tom slučaju, suma dijagonalnih elemenata, time i suma ajgenvektora, jednaka je sa brojem originalnih varijabli (p).

Analiza glavnih komponenti se izvodi u sledećim koracima:

1. Vrši se standardizacija originalnih podataka tako da originalne varijable imaju aritmetičku sredinu jednaku nuli i varijansu jednaku jedinici. Ovaj korak se najčešće ne preskače iako ima slučajeva da se to čini kada se veruje da je važnost originalnih varijabli dobro iskazana kroz varijanse.
2. Izračunava se matrica kovarijansi C.
3. Izračunavaju se ajgenvrednosti $\lambda_1,\lambda_2,...,\lambda_p$ i odgovarajući ajgenvektori $a_1,a_2,...,a_p$. Glavna komponenta je tako iskazana preko koeficijenta a_i i varijanse λ_i.
4. Komponente koje se u modelu odnose na malu proporciju varijacija podataka se eliminišu. Na primer, ako prve dve komponente objašnjavaju 95% varijanse, onda se sve ostale eliminišu. Tada su prve dve komponente zapravo glavne komponente.

Primer: Dve grupe pilića

Na jednoj farmi popisano je pet različitih telesnih dimenzija pilića (X_1 do X_5). Pilići su podeljeni u dve grupe: prva grupa (od rednog broj 1 do 21) je bila otporna na bolest, dok druga grupa (od rednog broja 22 do 49) nije bila. Originalni podaci se nalaze u tabeli. Podaci o koeficijentima korelacije koji su dati u tabeli ukazuju da su dovoljno visoki da bi se mogla izvesti analiza glavnih komponenti.

Tabela: Koeficijenti korelacije originalnih varijabli

Varijable	X_1	X_2	X_3	X_4	X_5
X_1	1,000	-	-	-	-
X_2	0,735	1,000	-	-	-
X_3	0,662	0,674	1,000	-	-
X_4	0,645	0,769	0,763	1,000	-
X_5	0,605	0,529	0,526	0,607	1,000

Tabela

Redni broj	Grupa	X_1	X_2	X_3	X_4	X_5

1	0	156	245	31.6	18.5	20.5
2	0	154	240	30.4	17.9	19.6
3	0	153	240	31.0	18.4	20.6
4	0	153	236	30.9	17.7	20.2
5	0	155	243	31.5	18.6	20.3
6	0	163	247	32.0	19.0	20.9
7	0	157	238	30.9	18.4	20.2
8	0	155	239	32.8	18.6	21.2
9	0	164	248	32.7	19.1	21.1
10	0	158	238	31.0	18.8	22.0
11	0	158	240	31.3	18.6	22.0
12	0	160	244	31.1	18.6	20.5
13	0	161	246	32.3	19.3	21.8
14	0	157	245	32.0	19.1	20.0
15	0	157	235	31.5	18.1	19.8
16	0	156	237	30.9	18.0	20.3
17	0	158	244	31.4	18.5	21.6
18	0	153	238	30.5	18.2	20.9
19	0	155	236	30.3	18.5	20.1
20	0	163	246	32.5	18.6	21.9
21	0	159	236	31.5	18.0	21.5
22	1	155	240	31.4	18.0	20.7
23	1	156	240	31.5	18.2	20.6
24	1	160	242	32.6	18.8	21.7
25	1	152	232	30.3	17.2	19.8
26	1	160	250	31.7	18.8	22.5
27	1	155	237	31.0	18.5	20.0
28	1	157	245	32.2	19.5	21.4
29	1	165	245	33.1	19.8	22.7
30	1	153	231	30.1	17.3	19.8
31	1	162	239	30.3	18.0	23.1
32	1	162	243	31.6	18.8	21.3
33	1	159	245	31.8	18.5	21.7
34	1	159	247	30.9	18.1	19.0
35	1	155	243	30.9	18.5	21.3
36	1	162	252	31.9	19.1	22.2
37	1	152	230	30.4	17.3	18.6
38	1	159	242	30.8	18.2	20.5
39	1	155	238	31.2	17.9	19.3
40	1	163	249	33.4	19.5	22.8
41	1	163	242	31.0	18.1	20.7
42	1	156	237	31.7	18.2	20.3
43	1	159	238	31.5	18.4	20.3
44	1	161	245	32.1	19.1	20.8
45	1	155	235	30.7	17.7	19.6
46	1	162	247	31.9	19.1	20.4
47	1	153	237	30.6	18.6	20.4
48	1	162	245	32.5	18.5	21.1
49	1	164	248	32.3	18.8	20.9

Prvi korak u analizi glavnih komponenti trebao bi da bude standardizacija svih vrednosti obeležja odnosno svih originalnih podataka. Na ovaj način se svim varijablama daje isti značaj u analizi. Kada se ne bi uradila standardizacija, varijable X_1 i X_2 bi imale većeg uticaja kod izračunavanja glavnih komponenti jer imaju velike numeričke vrednosti.

Matrica kovarijansi za standardizovane vrednosti je korelaciona matrica. Ajgenvrednosti te matrice su 3,616; 0,532; 0,386; 0,302 i 0,165. Zbir ovih vrednosti je tačno 5 koliko iznosi i zbir dijagonalnih elemenata u korelacionoj matrici.

Naredna tabela sadrži ajgenvektore standardizovane tako da suma njihovih kvadrata iznosi 1 za svaki ajgenvektor. Ovi ajgenvektori daju koeficijente za glavne komponente.

Tabela: Ajgenvrednosti i ajgenvektori korelacione matrice originalnih varijabli

Glavna komponenta	Ajgenvrednost	Ajgenvektori (koeficijenti glavnih komponenti)				
		X_1	X_2	X_3	X_4	X_5
Z_1	3,616	0,452	0,462	0,451	0,471	0,398
Z_2	0,532	−0,051	0,300	0,325	0,185	−0,877
Z_3	0,386	0,691	,0341	−0,455	−0,411	−0,179
Z_4	0,302	−0,420	0,548	−0,606	0,388	0,069
Z_5	0,165	0,374	−0,530	−0,343	0,652	−0,192

Ajgenvrednost ukazuje na udeo u ukupnoj varijansi koji je vezan za određenu glavnu komponentu. Na primer udeo prve glavne komponente (Z_1) u ukupnoj varijansi iznosi:

$$\frac{3{,}616}{5} \bullet 100 = 72{,}3\%$$

To znači da 72,3% varijacija podataka je posledica glavne komponente Z_1. Ostale komponente imaju udeo 10,6%; 7,7%; 6,0% i 3,3% respektivno. Uočava se da udeo komponenti postepeno opada. Prva komponenta je, naravno, daleko važnija i uticajnija od ostalih. Prema tome, jednačina prve glavne komponente iznosi:

$$Z_1 = 0{,}452\,X_1 + 0{,}462\,X_2 + 0{,}451\,X_3 + 0{,}471\,X_4 + 0{,}398\,X_5$$

gde su vrednosti od X_1 do X_5 standardizovane varijable. Uočava se da su koeficijenti svih varijabli približno jednaki, što upućuje na podjednaku važnost svih dimenzija kod pilića. Na osnovu dobijenih rezultata zaključuje se da 72,3% varijacija podataka dolazi usled razlike u veličini pilića.

Jednačina za drugu glavnu komponentu je:

$$Z_2 = -0{,}051\,X_1 + 0{,}300\,X_2 + 0{,}325\,X_3 + 0{,}185\,X_4 - 0{,}877\,X_5$$

Neki statistički programi mogu da daju obrnuti raspored predznaka:

$$Z_2 = 0{,}051\,X_1 - 0{,}300\,X_2 - 0{,}325\,X_3 - 0{,}185\,X_4 + 0{,}877\,X_5$$

U ovom slučaju se menja smer ali se i dalje meri isti aspekt kod podataka.

Kod druge komponente se uočava kontrast između promenljivih X_1 i X_5 sa jedne i promenljivih X_2, X_3 i X_4 sa druge strane. To znači da će Z_2 biti visoka ako su visoke vrednosti varijabli X_2, X_3 i X_4, a niska vrednost X_1 i X_5 i obrnuto. Niska vrednost koeficijenta varijable X_1 (−0,051) upućuje na mali značaj varijable na Z_2. Druge varijable se mogu interpetirati na sličan način.

Vrednosti glavnih komponenti mogu biti veoma korisne za dalju analizu. One su izračunate na osnovu standardizovane vrednosti obeležja. Na primer, za prvo pile iz uzorka originalne vrednosti obeležja su:

x_1= 156; x_2= 245 ; x_3= 31,6; x_4= 18,5; x_5= 20,5.

Aritmetičke sredine i standardne greške originalnih promenljivih su:

$\bar{x}_1 = 157{,}98$; $\sigma_1 = 3{,}654$; $\bar{x}_2 = 241{,}327$; $\sigma_2 = 5{,}068$; $\bar{x}_3 = 31{,}459$; $\sigma_3 = 0{,}795$; $\bar{x}_4 = 18{,}469$; $\sigma_4 = 0{,}564$; $\bar{x}_5 = 20{,}827$; $\sigma_5 = 0{,}991$.

Standardizovane vrednosti se izračunavaju na sledeći način:

$$x_1 = \frac{156 - 157{,}98}{3{,}654} = -0{,}542; \quad x_2 = \frac{245 - 241{,}327}{5{,}068} = 0{,}725; \quad x_3 = \frac{31{,}6 - 31{,}459}{0{,}795} = 0{,}177;$$

$$x_4 = \frac{18{,}5 - 18{,}469}{0{,}564} = 0{,}055; \quad x_5 = \frac{20{,}5 - 20{,}827}{0{,}991} = -0{,}330 .$$

Vrednost prve glavne komponente za prvo pile se izračunava na sledeći način:

$$Z_1 = 0{,}452 \bullet (-0{,}542) + 0{,}462 \bullet 0{,}725 + 0{,}451 \bullet 0{,}177 + 0{,}471 \bullet 0{,}055 + 0{,}398 \bullet (-0{,}330) = 0{,}064$$

Vrednost druge glavne komponente za drugo pile se izračunava na sledeći način:

$$Z_2 = -0{,}051 \bullet (-0{,}542) + 0{,}300 \bullet 0{,}725 + 0{,}325 \bullet 0{,}177 + 0{,}185 \bullet 0{,}055 - 0{,}877 \bullet (-0{,}330) = 0{,}602$$

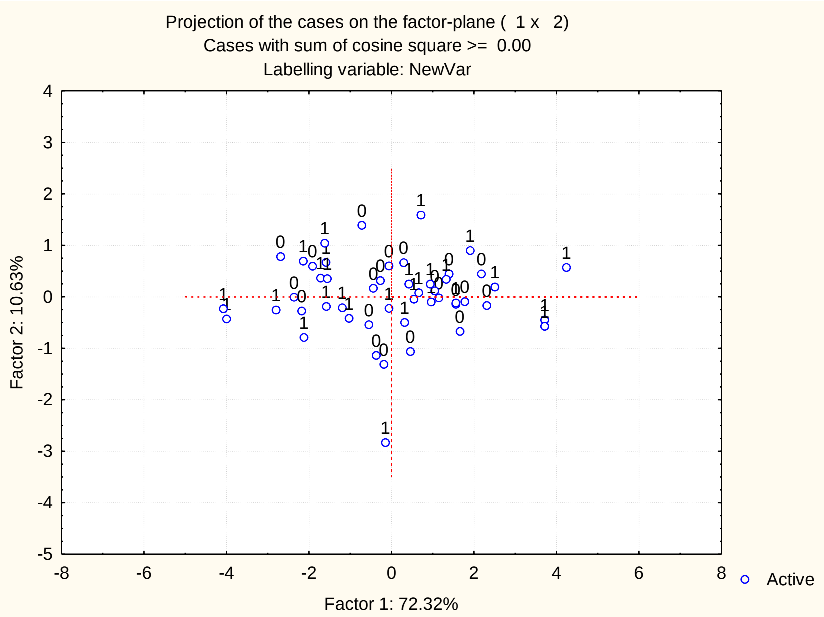
Ostale glavne komponente se izračunavaju na sličan način.

U nastavku analize, mogu se izračunati sve glavne komponente za sve jedinice posmatranja i zatim izvršiti upoređivanje između dve grupe pilića. Izračunate su aritmetičke sredine i standardne devijacije za obe grupe i prikazane u tabeli.

Tabela: Poređenje dve grupe jedinica posmatranja preko aritmetičkih sredina i standardnih devijacija glavnih komponenti

Glavna komponenta	Aritmetička sredina		Standardna devijacija	
	I grupa	II grupa	I grupa	II grupa
Z_1	−0,100	0,075	1,506	2,176
Z_2	0,004	−0,003	0,684	0,776
Z_3	−0,140	0,105	0,522	0,677
Z_4	0,073	−0,055	0,563	0,543
Z_5	0,023	−0,017	0,411	0,408

Kada se primeni t-test, nijedna od razlika aritmetičkih sredina dve grupe nije statistički značajna. Takođe, nijedna od razlika standardnih devijacija dve grupe nije statistički značajna kada se primeni F-test. Međutim, za testiranje razlika mogu da se primene i drugi testovi, kao, na primer, Levinov (Levene) test devijacije od medijane. Ovaj test otkriva razliku u varijacijama prve glavne komponente (Z_1) između dve grupe uz nivo signifikantnosti od 5%.



Slika Raspored jedinica posmatranja (dve grupe pilića) na osnovu vrednosti dve glavne komponente.

Slika pokazuje raspored dve grupe pilića (elementi prve grupe su označeni sa 0) s obzirom na vrednost prve dve glavne komponente. Ovim grafičkim prikazom je obuhvaćeno 82,9% varijacija originalnih podataka. Slika jasno pokazuje kako su pilići sa ekstremnom veličinom određenih delova tela podložniji bolesti.

Primer: Zaposlenost u evropskim zemljama

Prikupljeni su podaci o procentu zaposlenih u devet industrijskih sektorau Evropi od 1989. do 1995. godine (tabela).

Tabela: Procenat radne snage zaposlen u devet grana industrije u 30 zemalja Evrope

Country	Group	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC

Belgium	EU	2.6	0.2	20.8	0.8	6.3	16.9	8.7	36.9	6.8
Denmark	EU	5.6	0.1	20.4	0.7	6.4	14.5	9.1	36.3	7
France	EU	5.1	0.3	20.2	0.9	7.1	16.7	10.2	33.1	6.4
Germany	EU	3.2	0.7	24.8	1	9.4	17.2	9.6	28.4	5.6
Ireland	EU	22.2	0.5	19.2	1	6.8	18.2	5.3	19.8	6.9
Greece	EU	13.8	0.6	19.8	1.2	7.1	17.8	8.4	25.5	5.8
Italy	EU	8.4	1.1	21.93	0	9.1	21.6	4.6	28	5.3
Luxembourg	EU	3.3	0.1	19.6	0.7	9.9	21.2	8.7	29.6	6.8
Netherlands	EU	4.2	0.1	19.2	0.7	0.6	18.5	11.5	38.3	6.8
Portugal	EU	11.5	0.5	23.6	0.7	8.2	19.8	6.3	24.6	4.8
Spain	EU	9.9	0.5	21.1	0.6	9.5	20.1	5.9	26.7	5.8
U.K.	EU	2.2	0.7	21.3	1.2	7	20.2	12.4	28.4	6.5
Austria	EFTA	7.4	0.3	26.9	1.2	8.5	19.1	6.7	23.3	6.4
Finland	EFTA	8.5	0.2	19.3	1.2	6.8	14.6	8.6	33.2	7.5
Iceland	EFTA	10.5	0	18.7	0.9	10	14.5	8	30.7	6.7
Norway	EFTA	5.8	1.1	14.6	1.1	6.5	17.6	7.6	37.5	8.1
Sweden	EFTA	3.2	0.3	19	0.8	6.4	14.2	9.4	39.5	7.2
Switzerland	EFTA	5.6	0	24.7	0	9.2	20.5	10.7	23.1	6.2
Albania	Eastern	55.5	19.4	0	0	3.4	3.3	15.3	0	3
Bulgaria	Eastern	19	0	35	0	6.7	9.4	1.5	20.9	7.5
Czech/Slovak Rep.	Eastern	12.8	37.3	0	0	8.4	10.2	1.6	22.9	6.9
Hungary	Eastern	15.3	28.9	0	0	6.4	13.3	0	27.3	8.8
Poland	Eastern	23.6	3.9	24.1	0.9	6.3	10.3	1.3	24.5	5.2
Romania	Eastern	22	2.6	37.9	2	5.8	6.9	0.6	15.3	6.8
USSR (form.)	Eastern	18.5	0	28.8	0	10.2	7.9	0.6	25.6	8.4
Yugoslavia (form.)	Eastern	5	2.2	38.7	2.2	8.1	13.8	3.1	19.1	7.8
Cyprus	Other	13.5	0.3	19	0.5	9.1	23.7	6.7	21.2	6
Gibraltar	Other	0	0	6.8	2	16.9	24.5	10.8	34	5
Malta	Other	2.6	0.6	27.9	1.5	4.6	10.2	3.9	41.6	7.2
Turkey	Other	44.8	0.9	15.3	0.2	5.2	12.4	2.4	14.5	4.4

Napomena: AGR - poljoprivreda, šumarstvo i ribarstvo; MIN - rudastvo; MAN - prerađivačka industrija; PS - Proizvodnja električne energije, gasa i vode; CON - građevinarstvo; SER - usluge; FIN - finansijski sektor; SPS – društvene i lične usluge; TC – transport i komunikacije.

Koeficijenti korelacije su dati u tabeli.

Tabela: Koeficijenti korelacije devet industrijskih sektora

Varijable	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
AGR	1,000	-	-	-	-	-	-	-	-

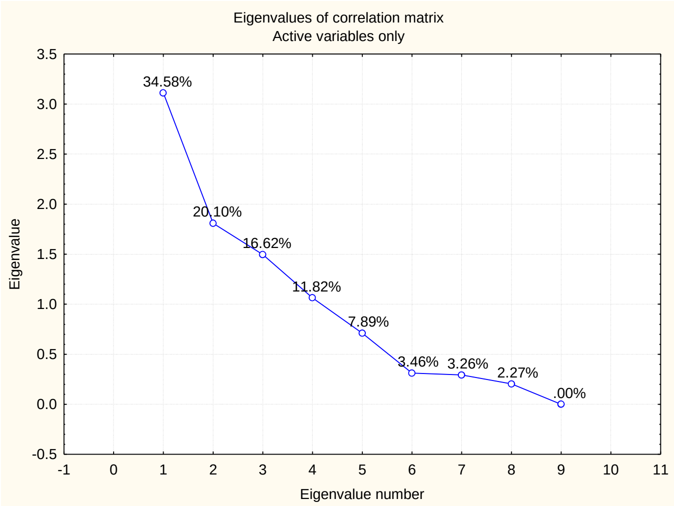
MIN	0,316	1,000	-	-	-	-	-	-	-
MAN	−0,254	−0,672	1,000	-	-	-	-	-	-
PS	−0,382	−0,387	0,388	1,000	-	-	-	-	-
CON	−0,349	−0,129	−0,034	0,165	1,000	-	-	-	-
SER	−0,605	−0,407	−0,033	0,155	0,473	1,000	-	-	-
FIN	−0,176	−0,248	−0,274	0,094	−0,018	0,379	1,000	-	-
SPS	−0,811	−0,316	0,050	0,238	0,072	0,388	0,166	1,000	-
TC	−0,487	0,045	0,243	0,105	−0,055	−0,085	−0,391	0,475	1,000

Ajgenvrednosti korelacione matrice, koje u zbiru iznose 9 zato što je to zbir elemenata dijagonale, date su u sledećoj tabeli:

Tabela: Pregled glavnih komponenti

Glavne komponente	Ajgenvrednosti	Udeo u ukupnoj varijansi
Z_1	3,112	34,6%
Z_2	1,809	20,1%
Z_3	1,496	16,6%
Z_4	1,063	11,8%
Z_5	0,710	7,9%
Z_6	0,311	3,5%
Z_7	0,293	3,3%
Z_8	0,204	2,3%
Z_9	0,000	0%

Poslednja ajgenvrednost je nula pa prema tome i pripadajuća glavna komponenta za sve jedinice posmatranja je takođe nula, sa varijansom jednakoj nuli. Slika pokazuje linijski dijagram sa procentualnim udelom glavnih komponenti.



Slika: Procentualni udeo glavnih komponenti u ukupnoj varijansi

Prva glavna komponenta obuhvata samo 35% varijacija originalnih podataka i čak četiri glavne komponente su potrebne da bi se obuhvatilo 83% varijacija. Stvar je procene koliko komponenti je važno. Obično se polazi od toga za šta će poslužiti rezultati u daljoj analizi. U ovom primeru, za prikazivanje osnovnih razlika između zemalja biće dovoljno u analizu uključiti prve dve komponente koje zajedno obuhvataju oko 55% varijacija originalnih podataka:

$$Z_1 = 0{,}51(AGR) + 0{,}37(MIN) - 0{,}25(MAN) - 0{,}31(PS) - 0{,}22(CON) - 0{,}38(SER) - 0{,}13(FIN) -$$

$$- 0{,}42(SPS) - 0{,}21(TC)$$

Vrednosti originalnih varijabli koje se koriste u jednačini su prethodno standardizovane tako da imaju aritmetičku sredinu nula i standradnu devijaciju jedan. Iz jednačine za Z_1 se uočava kontrast između promenljivih AGR i MIN sa jedne strane i ostalih promenljivih sa druge.

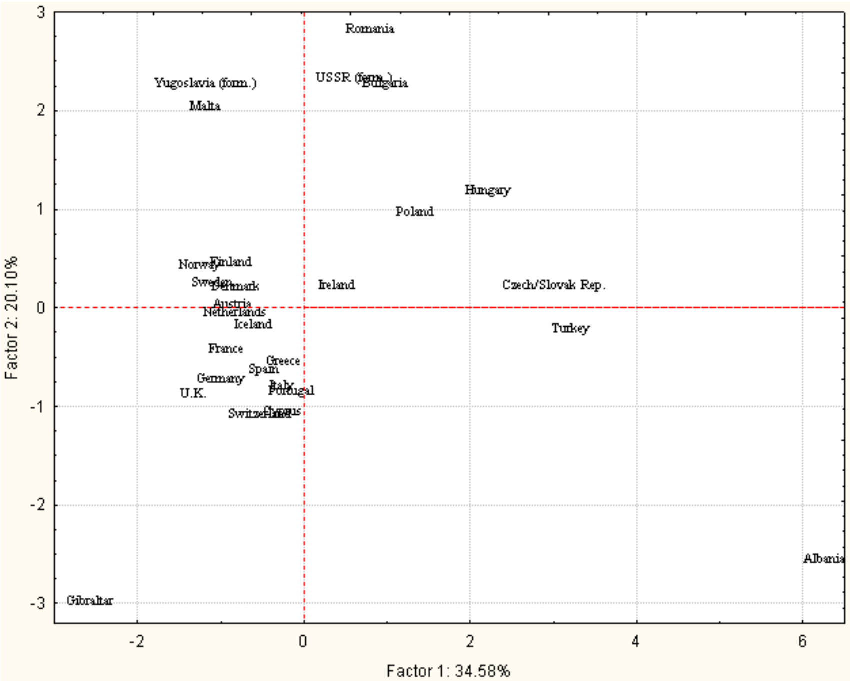
Druga komponenta iznosi:

$$Z_2 = -0{,}02(AGR) + 0{,}00(MIN) + 0{,}43(MAN) + 0{,}11(PS) - 0{,}24(CON) - 0{,}41(SER) - 0{,}55(FIN) +$$

$$+ 0{,}05(SPS) + 0{,}52(TC)$$

Ovde se uočava kontrast između varijabli MAN i TC sa jedne i CON, SER i FIN sa druge strane.

Slika pokazuje raspored zemalja na osnovu dve glavne komponente i na izuzetno jasan način pokazuje međusobni položaj zemalja.Većina zapadnoevropskih zemalja ima male negativne vrednosti za Z_1 i Z_2 dok Gibraltar i Albanija imaju specifičan pristup zapošljavanju.



Slika Raspored jedinica posmatranja na osnovu vrednosti dve glavne komponente.

Analiza glavnih komponenti u statističkom paketu STATISTICA

Razlika između statističkih paketa kada je u pitanju analiza glavnih komponenti je u tome što se kod nekih programa analiza glavnih komponenti nalazi u modulu za faktorsku analizu kao jedan tip faktorske analize, dok je kod drugih programa smeštena u poseban modul, kao što je slučaj i kod programa STATISTICA. Druga razlika je u tome što su promenjeni predznaci koeficijenata u jednačinama za glavne komponente što ne utiče na rezultate analize.

Koraci za izvođenje analize u programu su sledeći:

Pokretanje analize:

Statistics ▸ Multivariate Exploratory Technique ▸ Principal Components & Classification Analysis

Dobija se početni meni za analizu.

Definisanje varijabli:

Variables ▸ Variables for analysis

Variables ▸ Grouping variable

▸ OK

Za pokretanje izračunavanja potrebno je izabrati iz početnog menija opciju OK.

Dobijanje korelacione matrice:

Descriptives ▸ Correlation matrix

Ajgenvrednosti korelacione matrice:

Quick ▸ Eigenvalues

Skorovi za svaku jedinicu posmatranja

Advanced ▸ Cases ▸ Factor coordinates of cases

Grafički prikaz ajgenvrednosti:

Variables ▸ Scree plot

Ajgenvektori (koeficijenti u jednačinama glavnih komponenti):

Variables ▸ Eigenvectors

Koordinate za grafički prikaz jedinica posmatranja u dvodimenzionalnom sistemu sa dve glavne komponente:

Cases ▸ Factor coordinates of cases

Grafički prikaz jedinica posmatranja u dvodimenzionalnom sistemu sa dve glavne komponente:

Cases ▸ Plot case factor coordinates 2D (označiti prvo opciju Grouping labels da bi ispisao imena jedinica posmatranja)

Deskriptivna statistika:

Descriptives ▸ Summary descriptives