8. MULTIVARIATNE METODE
8.1. Uvod
8.1.1. Zakaj jih uporabljati

- Multivariatne metode omogočajo sočasno analizo kakršnegakoli števila spremenljivk.
- Poseben problem predstavlja grafična predstavitev več kot treh spremenljivk hkrati - ena izmed nalog multivariatnih metod je zato zmanjšanje dimenzionalnosti podatkov, do predstavljive upodobljivosti.

8.1.1. Zakaj jih uporabljati

- Razlogi, da multivariatnih metod ne uporabimo?
  - Logistični – razpoložljivi so le podatki za eno ali dve spremenljivki.
  - Strokovni – le ena ali dve spremenljivki sta ustrezni glede na zastavljeni namen študije.
  - Izkustveni – ugotovili smo, da za določen izid testa (npr. razlikovanje dveh populacij) zadostuje že ena spremenljivka, da zavrnemo H_0.
  - Matematični – sočasna analiza več spremenljivk nudi dodatne podatke le, če so spremenljivke med seboj korelirane.

8.1.1. Zakaj jih uporabljati

- Analizo multivariatnih podatkov vedno začnemo enostavno, z uporabo točkovnih in škatla-črta (box and whisker) diagramov in bivariatnih grafov ter izračunom osnovnih opisnih statistik ter normalnosti porazdelitve za vsako spremenljivko.
- Multivariatne analize zahtevajo velike vzorce. Opazovanj naj bi bilo najmanj dvakrta toliko kot je spremenljivk.

8.1.2. Oznake

- Matrika podatkov
  - Podatke meritve p spremenljivk na n objektih zapišemo v n x p matriki, kjer vsaka vrstica ustreza enemu opazovanju.
  - Spremenljivke označimo X_1, ..., X_p in vrednosti X_i v i-tem opazovanju z x_{ij}, ter zapišemo matriko:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

8.1.2. Oznake

- N-to vrednost spremenljivke X_j zapišemo kozt stolpični (kolonski) vektor

$$x_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

- in i-to opazovanje kot vrstični vektor

$$x_j^T = (x_{i1} \quad x_{i2} \quad \dots \quad x_{ip})$$

8.1.3. Izbira metode

- Multivariatne posplošitve univariatne statistike
  - Postavljanje statističnih ugotovitev o populacijskih parametrih – najpogosteje testiranje enakosti populacijskih srednjih vrednosti.
- Multipla regresija
  - Iskanje in testiranje enačbe najboljšega prileganja, ki povezuje eno odvisno spremenljivko z več neodvisnimi.
- Diskriminanta funkcija
  - Znotraj seta podatkov poišče enačbo premice, ki najbolje loči dve ali več podskupin, ki ju določi uporabnik, ter na tej osnovi umesti nove podatke v ustrezno skupino.
- Glavne komponente in faktorska analiza
  - Znotraj podatkov poišče smeri največje variance in jih uporabi za razvrstitev podatkov v eno, dve ali tri dimenzije ter jih razlaga kot faktorje, ki vplivajo na podatke.
- Koeficienti podobnosti in clusterska analiza
  - Kvantificiranje podobnosti med pari objektov ali opazovanj ter nadaljna uporaba za izkustveno razvrstitev (empirično klasifikacijo).

8.3. MULTIPLA LINEARNA REGRESIJA

- Običajno odvisnost eno odvisne spremenljivke pojasnjujeta dve ali več regresorskih spremenljivk.
- Cilji so:
  - Združitev (povzetek) velikega števila podatkov.
  - Razumevanje proceasa, ki se odraža v opazovanem pojavu.
  - Ocenjevanje težko merljivih količin, ki so povezane z drugimi, lažje določljivimi.

8.3.1. Regresijski modeli

- Oblike regresijskih modelov so lahko:

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	3D
$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$	hiperprostor
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$	polinomna regresija
$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_1 x_2 + \varepsilon$	hiperprostor

- Kjer so členi ε naključne spremenljivke s srednjo vrednostjo nič in varianco σ^2, enako za vse vrednosti regresorskih spremenljivk.
- Pomemben pogoj je, da so nekorelirane.
- B_0, β_1, ..., β_k so delni regresijski koeficienti in β_i predstavlja spremembo Y zaradi spremembe X_i za eno enoto, kadar ostane vse ostalo konstantno.

8.3.1. Regresijski modeli

- Vsi navedeni modeli so v delnih regresijskih koeficientih β linearni, zato z njimi ne skušajmo obravnavati nelinearnih modelov.
- Prvi model opisuje točke v 3D prostoru, za katere iščemo ravnino najboljšega prileganja.
- Drugi in četrti predstavljata točke v hiperprostoru, ki jim skušamo prilagoditi hiperpovršino. Izbiramo jo tako, da čim boljzmanjšamo vsote kvadriranih odstopanj podatkovnih točk od nje.
- Tretji model je primer polinomne regresije. Čeprav vključuje kvadriranje je linearen, ker so delni regresijski koeficienti linearni.

8.3.2. Matrike podatkov ter vektorji podatkov in koeficientov

- Podatke neodvisne in odvisnih spremenljivk predstavimo v tabeli, imenovani podatkovna matrika:

$$\begin{array}{ccccc} Y & X_1 & X_2 & \dots & X_k \end{array}$$

$$\begin{pmatrix} y_1 & x_{11} & x_{12} & \dots & x_{1k} \\ y_2 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ y_n & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

8.3.2. Matrike podatkov ter vektorji podatkov in koeficientov

- Kadar potrebujemo vrednosti Y, jih predstavimo v obliki vektorja, kakor tudi delne regresijske koeficiente in člene napake:

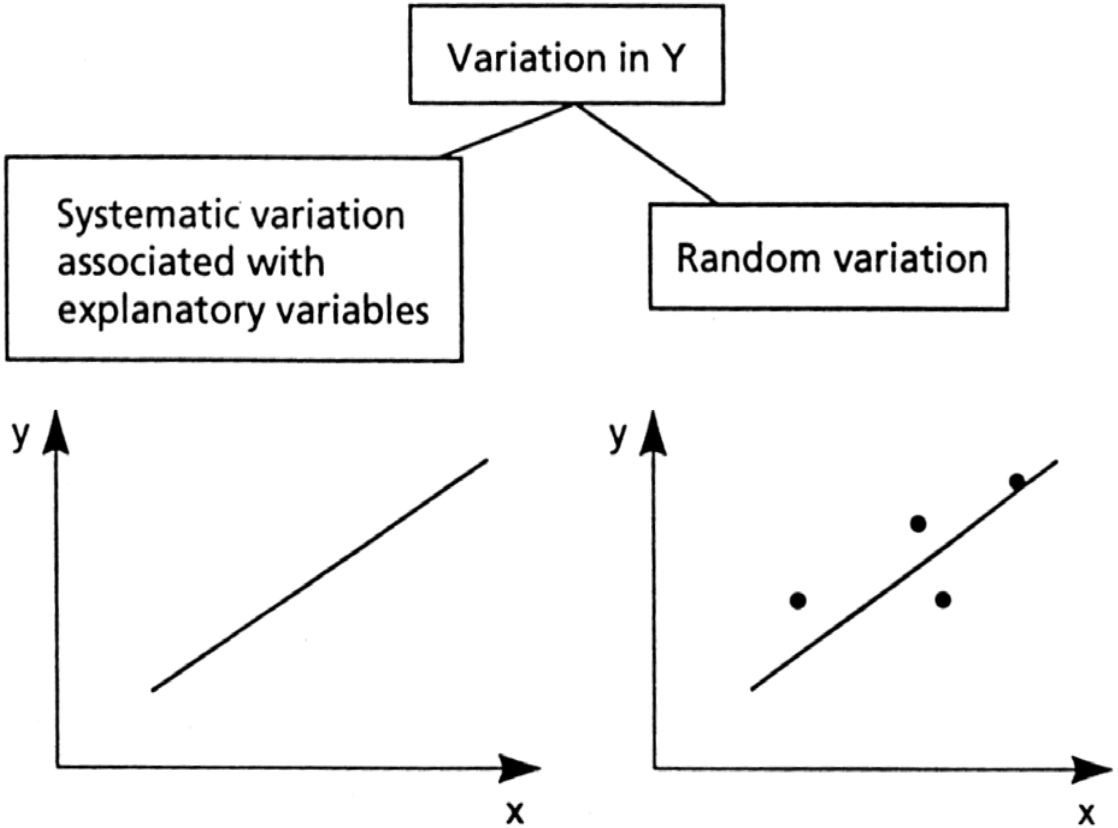
$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- Linearni model potem zapišemo:

$$y = X\beta + \varepsilon$$

8.3.3. Ocena parametrov

- Uporabimo princip najmanjših kvadratov – izberemo hiperravnino, ki zmanjša vsoto kvadratov preostankov.
- Varianco, merjeno s celotno popravljeno vsoto kvadratov vrednosti Y, razstavimo v del, imenovan regresijska vsota kvadratov, ki je sistematičen in drugi del, ki je naključen.



8.3.3. Ocena parametrov

- Ocene so nepristranske. Med seboj so v glavnem korelirane.
- Prilagojeni regresijski model zapišemo v matrični obliki:

$$\hat{y} = X\beta$$

- Preostanki so zapisani v vektorju:

$$z = y - \hat{y}$$

- Varianco ocenimo z:

$$s^2 = \frac{VK_{\text{preostankov}}}{n-k-1}$$

kjer je k število regresorskih spremenljivk v modelu.

8.3.3. Ocena parametrov

- Izračuni so podani v tabeli analize variance, kjer je k število odvisnih spremenljivk in N celotno število opazovanj.

VIR	ν	VK	s^2	F
Regresija $X_1, \ldots, X_k$	k	$VK_{regresije}$	$S^2_{regresije} = VK_{regresije}/k$	$s^2_{regresije}/s^2_{preostankov}$
Preostanki	n-k-1	$VK_{preostankov}$	$S^2_{preostankov} = VK_{preostankov}/n\text{-}k\text{-}1$	
Skupaj	n-1	PVK (Y)		

8.3.4. Testi modelov

- H_0: $\beta = 0$
- H_1: $\beta \neq 0$
- Kadar drži H_0 in je porazdelitev naključnih členov normalna, ima razmerje kvadratov srednjih vrednosti porazdelitev F s parametri k, n-k-1.
- Kadar je vsaj eden od regresorskih koeficientov različen od nič, je porazdelitev modificirana različica F, imenovana ne-centralna F, kjer so vrednosti razmerja kvadratov srednjih vrednosti visoke.

8.3.4. Testi modelov

- Pričakovanje regresijskega kvadrata srednje vrednosti je

σ^2+pozitiven člen, zaradi sistematične variance

- Testiramo s primerjavo razmerja povprečja vsote kvadratov z $F_{\alpha;k,n-k-1}$, kjer je α izbrana raven značilnosti.

8.3.5. Kriteriji izbire regresijskega modela

- Modeli so različni glede na število vključenih spremenljivk in načina izračuna regresije.
- V primeru dveh neodvisnih spremenljivk so možni trije modeli, v primeru treh že sedem in s štirimi 15.
- Najpomembnejši modeli so:
  - Prouči vse možne regresije
  - Postopna regresija
  - Napredujoča regresija
  - Obrnjeno opuščanje

8.3.5.1. Prouči vse možne regresije

- Nekateri kriteriji izbire modela so:

  1. Multipli regresijski koeficient R^2_p:

$$R_p^2 = \frac{\text{regresijska vsota kvadratov}}{\text{celotna popravljena vsota kvadratov}} = \frac{VK_{regresijska}}{PVK_{celotna}}$$

- Vrednost narašča z dodajanjem spremenljivk v model, dokler niso vključene vse spremenljivke, vendar se njihovi doprinosi zmanjšujejo.

8.3.5.1. Prouči vse možne regresije

2. Napaka povprečja vsote kvadratov naj bo majhna.
3. Prilagojeni R^2_p je multipli determinacijski koeficient, prilagojen stopnjam prostosti:

$$\overline{R}_p^2 = 1 - \frac{n-1}{n-p}\left(1 - R_p^2\right) = 1 - \frac{\overline{VK}_{\text{preostanki}}}{\overline{VK}_{celotni}} = 1 - \frac{s^2_{\text{preostanki}}}{s^2_{celotni}}$$

- Narašča z naraščanjem števila regresorjev, kadar je zmanjšanje napake vsote kvadratov dovol veliko, da uravnoteži izgubo stopnje prostosti.
- Metoda je enakovredna zmanjševanju povprečja vsote kvadratov preostankov.

8.3.5.1. Prouči vse možne regresije

4. c_p statistika omogoča izbiro modela, ki je najmanj asimetričen (pristranski).

$$c_p = \frac{\text{preostanki VK (p)}}{\text{preostanki } \overline{VK} \text{ (celoten model)}} - n - 2p$$

- Kadar ni pristranosti, je pričakovanje statistike p, zato nanesemo na graf vrednosti c_p proti p. Malo pristranosti izkazujejo vrednosti, ki so blizu premice, ki povezuje izhodišče s točko, ki jo dobimo, kadar vključimo vse vrednosti.

8.3.5.2. Postopna regresija

1. Preveri vse enostavne linearne regresije in izberi tisto, ki daje najvišjo vrednost statistike F.
2. Za vsako od preostalih spremenljivk izračunaj korelacijo z Y, upoštevajoč, da je prva spremenljivka že v modelu (to je delna korelacija):

$$F_j = \frac{\text{preostanki}\, VK(\beta_j | \beta_0, \beta_1)}{\text{preostanki}\, \overline{VK}\left(x_j, x_1\right)}$$

- kjer je števec zmanjšane vsote kvadratov preostankov, zaradi vključitve X_j, ko je X_1 že v modelu in imenovalec povprečje vsote kvadratov preostankov, z X_1 in X_2 v modelu.
- Spremenljivka X_j, ki da najvišjo vrednost te statistike vstopi v model, če F_j preseže mejo F_{in}, izbrano tako, da je verjetnost sprejetja nizka, ko X_j ne da značilne, dodatne informacije o Y.

8.3.5.2. Postopna regresija

3. Spremenljivko, izbrano v prvem koraku, potem testiramo, ali jo obdržati v v modelu ali ne. Izračunamo:

$$F_j = \frac{\text{preostanki}\, VK(\beta_1 | \beta_0, \beta 2)}{\overline{VKE}(x_1, x_2)}$$

- Primerjamo jo z mejno vrednostjo F_{out}. X_1 izločimo, če je $F_1 < F_{out}$. F_{out} mora biti enaka (alui manjša) F_{in}.

8.3.5.2. Postopna regresija

4. Postopek nadaljujemo, dokler se nobena spremenljivka več ne uvrsti v model in nobene več ne izločimo. Na vsaki stopnji testiramo vsako prej izbrano spremenljivko, izločimo pa le eno, če sploh katero.

8.3.5.3. Napredujoča regresija

- Sledi enakemu postopku kot postopna regresija, le da predhodno izbranih spremenljivk ne testiramo, da bi ugotovili, kako nanje vplivajo kasneje dodane.

8.3.5.4. Obrnjeno opuščanje

- Vključimo vse regresijske kandidate in tistega z najnižjim delnim korelacijskim koeficientom z Y izločimo, vse ostale pa upoštevamo, če je delna statistika F nižja od F_{out}.
- Postopek nadaljujemo, dokler ni model tak, da nobena od spremenljivk nima delne statistike F nižje od F_{out}.

8.3.6. Preiskava primernosti

- Preverimo okoliščine, kakršna je na primer, da členi napake med seboj niso povezani:
  - Uporabimo metode, opisane za enostavno analizo linearne regresije.
  - Testiramo avtokorelacijo preostankov z Durbin – Watsonovim testom. Temelji na korelaciji zaporednih preostankov, z H_0: $\rho = 0$.

8.3.7. Multikolinearnost

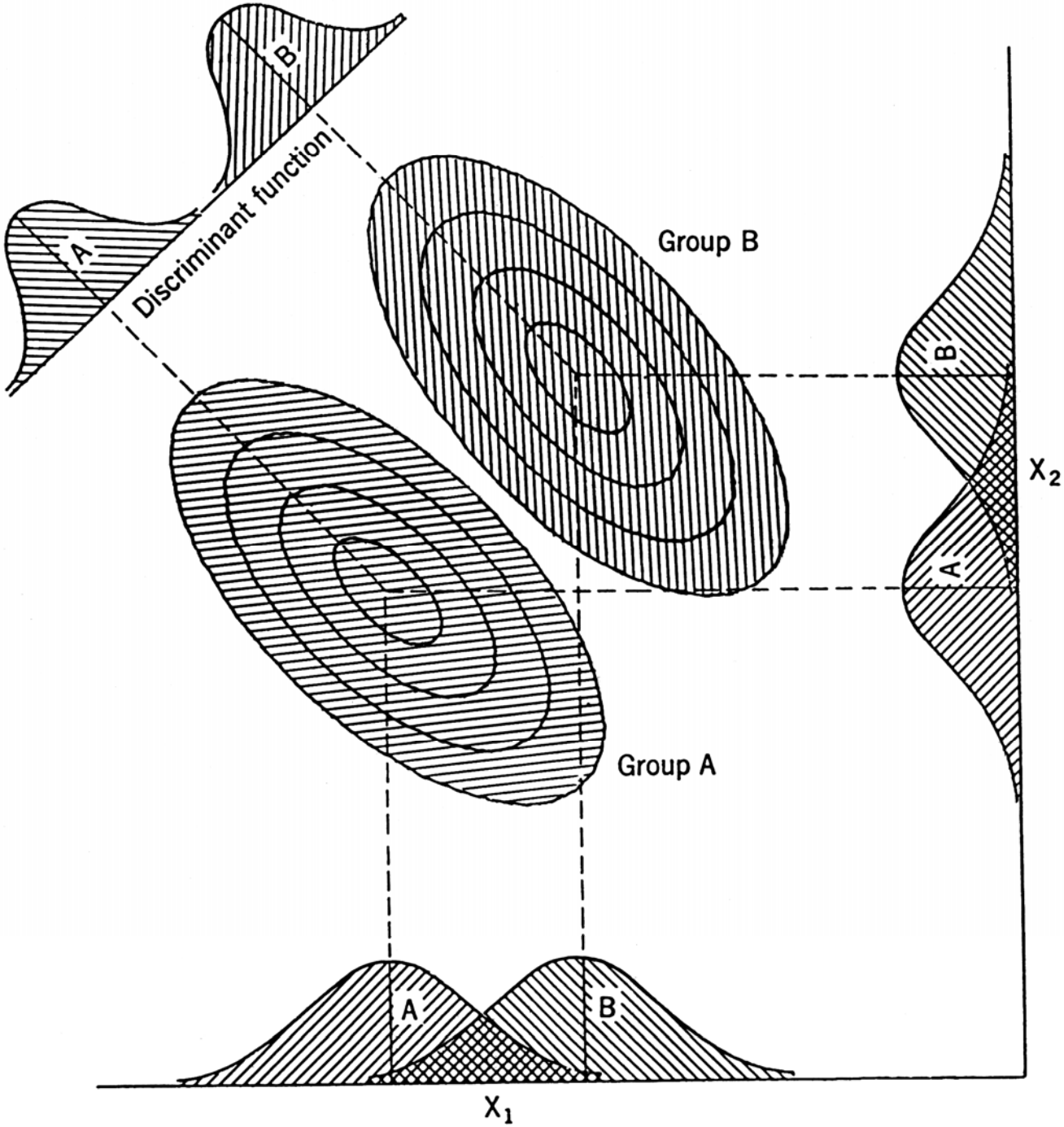
- Pojavi se, kadar so nekatere regresorske spremenljivke visoko korelirane in povzroči nekatere resne probleme pri regresijski analizi.
  - Variance in kovariance ocen delnih regresorskih koeficientov postanejo visoke, kar pomeni, da je njihova točnost nizka.
  - Delna regresorska koeficienta, dveh visoko koreliranih neodvisnih spremenljivk, sta enake velikosti, a različnih predznakov, ne glede na to, kakšne so vrednosti β_1 in β_2.
  - Računalnik ni sposoben pravilno izračunati nekaterih matrik. Rezultati so nesmiselni in nezanesljivi.

8.3.8. Vplivne vrednosti: Cookova razdalja

- Včasih so opazovanja v hiperprostoru od ostalih zelo oddaljena in močno vplivajo na izračunani model.
- Če se izkaže, da so napačna, jih odstranimo. Kadar nudijo koristne podatke, jih ohranimo.
- Odkrijemo jih s količino, imenovano Cookova razdalja, ki je kvadrat razdalje med $\hat{\beta}$, ki temelji na n opazovanjih in $\hat{\beta}$, če odstranimo i-to opazovanje.
- Vrednosti Cookove razdalje, večje od 1, nakazujejo vplivno opazovanje.

8.4. DISKRIMINANTNA ANALIZA
8.4.1. Načela in uporaba

- Diskriminantna analiza služi uvrščanju objektov v eno izmed dveh uvedenih skupin.
- Potrebujemo "poskusno skupino" podatkov znane uvrstitve, za vsako od kategorij, ki jo želimo ločevati.
- Poiščemo diskriminantno funkcijo – vektorje v smeri najboljšega ločevanja skupin.
- Podatki "poskusnih skupin", nanešeni vzdolž diskriminantne funkcije bodo kar najbolj mogoče ločeni.
- Ni nujno, da je diskriminantna funkcija vektor med srednjima vrednostima skupin – upoštevati je potrebno kovariančno strukturo skupin.
- Na osnovi diskriminantne funkcije uvrščamo nova opazovanja v eno od obeh skupin.


B
A
Discriminant function
Group B
Group A
B
A
X_2
A
B
X_1

8.4.1. Načela in uporaba

- Fosil želimo uvrstiti v eno izmed vrst A ali B le na podlagi dolžine.
- Obe vrsti sta normalno porazdeljnei z enako varianco σ^2 in različnima srednjima vrednostima μ_A in μ_B.
- Kadar je dolžina primerka x, bližje μ_A, ga bomo uvrstili v skupino A.
- Klasično bi to zapisali:

$x - \mu_B > \mu_A - x$ ali $x > \dfrac{\mu_A + \mu_B}{2}$

8.4.1. Načela in uporaba

- Zapišemo lahko tudi v obliki populacijskuh gostot:

$$f(x|\ \mu_A) > f(x|\ \mu_B)$$

oz. verjetnostjo (Likelihood)

$$L(\mu_A|x) > L(\mu_B|x)$$

8.4.1. Načela in uporaba

- Če pa npr. vemo, da je vrsta A sorazmerno redka, bomo ločnico pomaknili bližje k μ_A – dokaz na osnovi dolžine mora biti bolj prepričljiv, da ga uvrstimo v redko vrsto.
- Če označimo prioritetno verjetnost s Pr(A) oz. Pr(B), pravilo, da primerek uvrstimo v vrsto A zapišemo:

$L(\mu_A|x)\cdot Pr(A) > L(\mu_B|x)\cdot Pr(B)$ oz.

$$\frac{L(\mu_A|x)}{L(\mu_B|x)} > \frac{\Pr(B)}{\Pr(A)}$$

8.4.1. Načela in uporaba

- V praksi lažje delamo z logaritmi navedenih funkcij, zato neenakost zapišemo:

$\ln\{L(\mu_A|x)\} - \ln\{L(\mu_B|x)\} > \ln k$

kjer je

$$k = \ln\left\{\frac{\Pr(B)}{\Pr(A)}\right\}$$

8.4.2. Uporaba *p* spremenljivk za ločevanje dveh populacij

- Predpostavimo, da nek objekt pripada eni od dveh populacij, katerih vektorja srednje vrednosti sta μ_1 in μ_2 ter kovariančna matrika Σ.
- Iz linearne komponente spremenljivk izračunamo indeks, ki izhaja iz univariatne normalne porazdelitve (enko kot pri testiranju hipoteze o enkosti populacijskih povprečij).
- Uporabljena komponenta je tista, ki daje največjo razdaljo med srednjima vrednostima populacij.

8.4.2. Uporaba *p* spremenljivk za ločevanje dveh populacij

- Uteži, ki določajo diskriminantno funkcijo so podane z:

  $a^T = (\mu_1 - \mu_2)^T \cdot \Sigma^{-1}$

- Dejansko jih ocenimo iz:

  $$(\bar{x}_1 - \bar{x}_2)^T S^{-1}$$

  kjer so $\bar{x}_1$, $\bar{x}_2$ in S ocene obeh vektorjev srednje vrednosti in skupna kovariančna matrika obeh populacij.

8.4.2. Uporaba *p* spremenljivk za ločevanje dveh populacij

- Novemu opazovanju x izračunamo izid diskriminantne funkcije z množenjem matrike:

$$(\bar{x}_1 - \bar{x}_2)^T S^{-1} x$$

- Vrednost primerjamo z izidoma diskriminantne funkcije obeh vektorjev srednjih vrednosti vzorcev

$$(\bar{x}_1 - \bar{x}_2)^T S^{-1} \bar{x}_1 \qquad \text{in} \qquad (\bar{x}_1 - \bar{x}_2)^T S^{-1} \bar{x}_2$$

8.4.2. Uporaba *p* spremenljivk za ločevanje dveh populacij

- Primerek uvrstimo v skupino z vektorjem srednje vrednosti μ_1, če je izračunani izid diskriminantne funkcije bliže prvi kot drugi vrednosti, oz. v drugo skupino, če je obratno.

$$(\bar{x}_1 - \bar{x}_2)^T S^{-1}(x - \bar{x}_1) < (\bar{x}_1 - \bar{x}_2)^T S^{-1}(\bar{x}_2 - x)$$

- Kriterij uvrščanja primerka v skupino A, ob upoštevanju prioritetnih vrjetnosti Pr(A) in Pr(B) je:

$$\frac{(\bar{x}_1 - \bar{x}_2)^T S^{-1} x - (\bar{x}_1 - \bar{x}_2)^T S^{-1}(\bar{x}_1 + \bar{x}_2)}{2} > \ln\left\{\frac{\Pr(B)}{\Pr(A)}\right\}$$

8.4.3. Ocena linearne diskriminantne funkcije

- V praksi si o diskriminantni funkciji postavimo tri vprašanja:

1. Ali so izbrane srednje vrednosti izhodnih populacij dovolj različne, da omogočajo zanesljivo ločevanje?
2. Če so, ali lahko izpustimo nekatere spremenljivke in kljub temu ločimo populaciji?
3. Kako lahko primerjamo učinkovitost možnih diskriminantnih funkcij?

8.4.3.1. Test uporabnosti ločevanja

- Izračunamo testno statistiko

$$\frac{(n_1+n_2-p-1)}{p(n_1+n_2-2)}\cdot\frac{n_1n_2}{n_1+n_2}(\bar{x}_1-\bar{x}_2)^T S^{-1}(\bar{x}_1-\bar{x}_2)$$

- Izbrane spremenljivke sprejmemo kot uporabne za ločevanje, kadar vrednost testne statistike presega F_{α,p,n_1+n_2-p-1}.
- To ne pomeni, da so te spremenljivke dejansko tudi najustreznejše za ločevanje.

8.4.3.2. Vključevanje in opuščanje spremenljivk

- Izmed spremenljivk X_1 do X_p želimo preveriti, ali naj p-to spremnljivko obdržimo ali zavrnemo.
- Izračunamo Mahalanobisovi razdalji D^2_p in $D^2_{p-,1}$ za podatke z in brez proučevane spremenljivke ter testiramo značilnost spremembe razdalje D^2_p - $D^2_{p-,1}$:

$$\frac{(n_1+n_2-p-1)(D_p^2-D_{p-1}^2)}{\left\{\frac{(n_1+n_2-2)(n_1+n_2)}{n_1 n_2}\right\}+D_{p-1}^2}$$

8.4.3.2. Vključevanje in opuščanje spremenljivk

- Mahalanobisovi razdalji oblike sta:

$$(\bar{x}_1 - \bar{x}_2)^T S^{-1} (\bar{x}_1 - \bar{x}_2)$$

- Če izračunana vrednost presega kritično $F_{\alpha,1,n_1+n_2-p-1}$, p-to spremenljivko obdržimo, ker bi se razdalja v nasprotnem primeru značilno zmanjšala.

8.4.3.3. Verjetnost napačne uvrstitve

- Vrednost diskriminatorja je odvisna od deleža pravilno uvrščenih objektov.

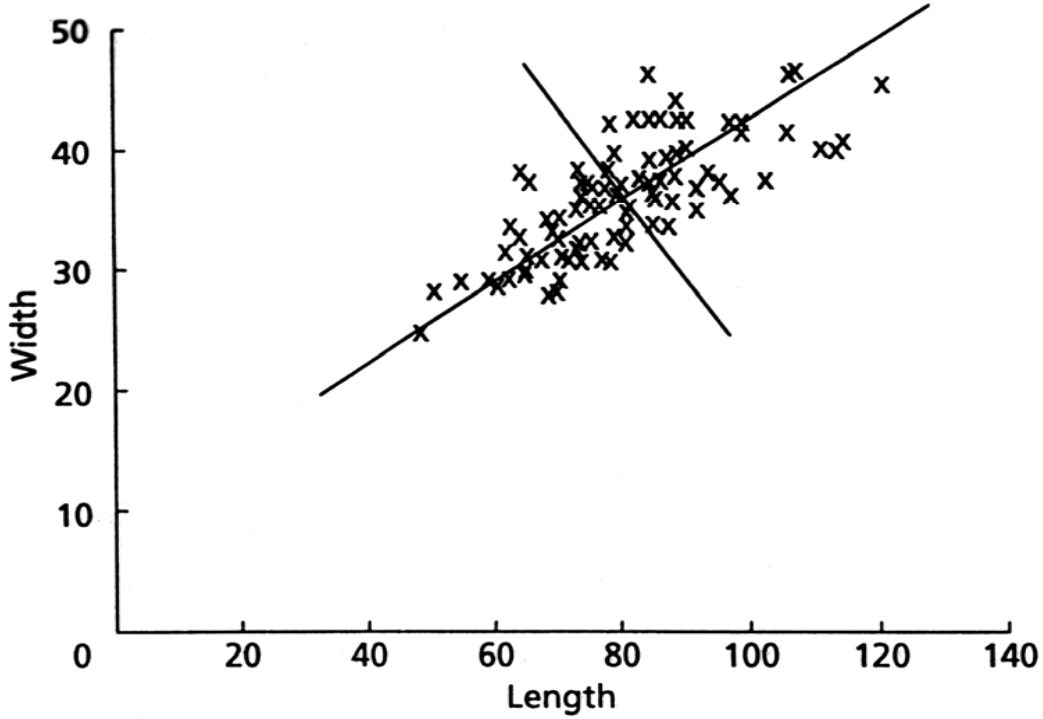
1. Potrebujmo veliko zbirko podatkov. Del razpoložljivih podatkov uporabimo za izračun uteži, preostale objekte pa uvrstimo na podlagi dobljene diskriminantne funkcije. Ugotovimo delež napačnih uvrstitev v vsako skupino.
2. Vsak objekt po vrsti uvrstimo in izračunavamo uteži iz preostalih (n-1) opazovanj. Postopek zahteva precej časa, ker moramo n-krat oceniti parametre porazdelitve za vsak izračun.

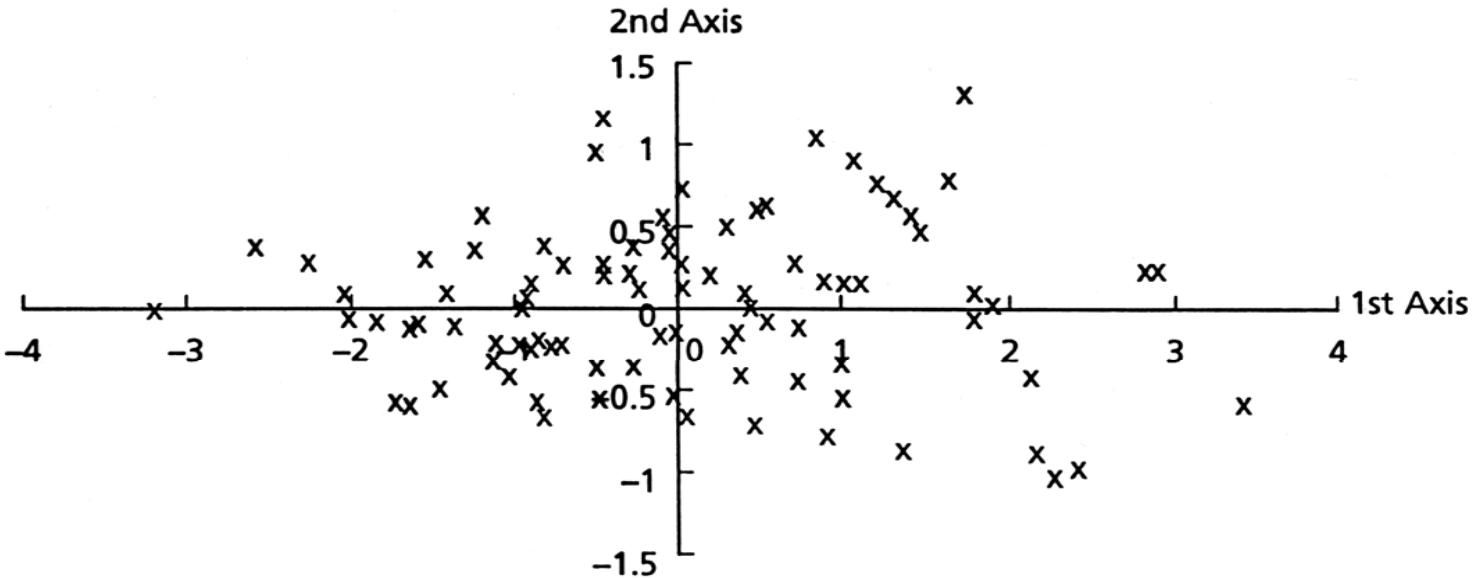
8.4.3.4. Zahtevane predpostavke

- 1. Podatki izhajajo iz multivariatnih normalnih porazdelitev.
  - Metoda je uspešna tudi kadar podatki niso normalno porazdeljeni; nezanesljivi so le testi upoštevanja ali izključevanja spremenljivke.
  - Analiza je bolj občutljiva na odstopanja od normalne porazdelitve kot v primeru uporabe linearne diskriminantne funkcije.
- 2. Kovariančni matriki sta enaki.
  - Kadar to ne drži, vključuje diskriminantna funkcija kvadratne in linearne člene in jo je zato pogosto težko interpretirati.
  - Kadar uporabimo dovolj velike vzorce, neeenaki kovariančni matriki ne vplivata premočno na testno funkcijo.

8.5. METODE LASTNIH (EIGEN) VEKTORJEV
8.5.1. Analiza glavnih komponent (PC)
8.5.1.1. Uvod

- Iščemo vektor, ki poteka vzdolž “dolge osi” oblaka podatkov ter drugi vektor, ki je nanj pravokoten.
- Vektorja postaneta nov referenčni okvir podatkov.
- Količina variance v smeri prve osi je najvišja možna, v smeri druge pa najmanjša.


Width
Length
0
10
20
30
40
50
20
40
60
80
100
120
140


2nd Axis
1st Axis
1.5
1
0.5
0
−0.5
−1
−1.5
−4
−3
−2
−1
0
1
2
3
4

8.5.1.1. Uvod

- Linearne glavne kompoenete (Principal Components – PC) imajo številne uporabne lastnosti:
  - Večino celotne variance p spremenljivk razložimo z razmeroma majhnim številom k novih spremenljivk. Dimenzionalnost podatkov se zmanjša iz p na k.
  - So nekorelirane, kar olajša preverjanje in analize podatkov.

8.5.1.1. Uvod

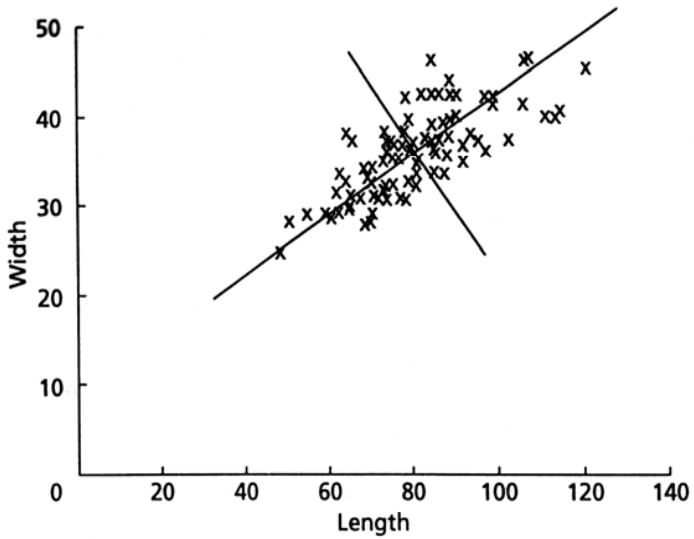
- Pomembni omejitvi sta:
  - Če so prvotni podatki nekorelirani, uporaba tehnike nima smisla.
  - Kadar je število spremenljivk majhno, odvisnost razberemo vizualno iz korelacijske matrike.

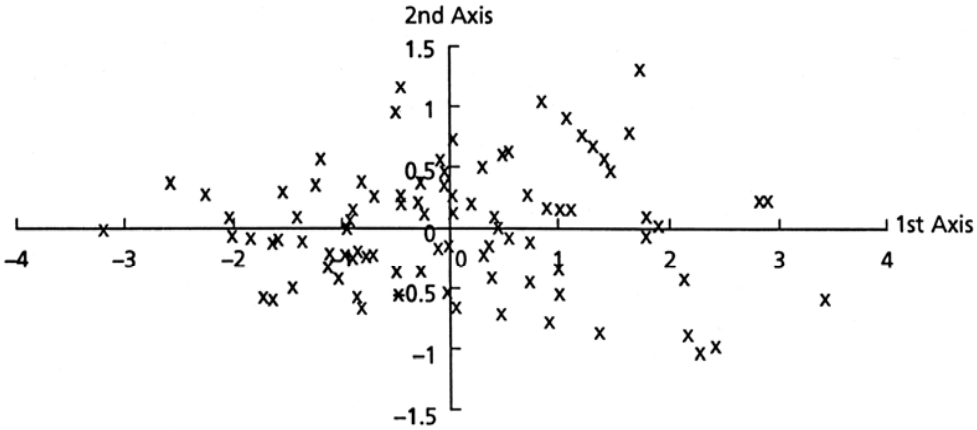
$$\begin{array}{c} \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0{,}85 & & & \\ 0{,}14 & 0{,}21 & & \\ 0{,}23 & 0{,}19 & 0{,}90 & \\ 0{,}78 & 0{,}95 & 0{,}25 & 0{,}32 \end{pmatrix} \end{array}$$

$$\begin{array}{c} \\ 2 \\ 5 \\ 4 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 5 & 4 \end{array} \\ \begin{pmatrix} 0{,}85 & & & \\ 0{,}78 & 0{,}95 & & \\ 0{,}23 & 0{,}19 & 0{,}32 & \\ 0{,}14 & 0{,}21 & 0{,}25 & 0{,}90 \end{pmatrix} \end{array}$$

8.5.1.2. Geometrijska razlaga

- Skušamo definirati manj-dimenzijonalni prostor, s katerim pojasnimo skoraj vso varianco.
- Proces je enak vrtenju osi v p-dimenzionalnem prostoru v takih smereh, da je projekcija točk nanje taka, do so čim bolj razpotegnjene.
- Rotacije, ki ohranijo med osema pravi kot, so ortogonalne.
- Izhodišče osi premaknemo v tisti točko p-dimenzionalnega prostora, ki ustreza srednjim vrednostim vzorca.





8.5.1.3. Definicija komponent

- Prva glavna komponenta (PC), označena z Y_1, ima največjo možno varianco:

  $Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$ ali $Y_1 = a_1^T X$

- Vsota kvadratov uteži mora biti 1:

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

- Varianco zapišemo:

  $Var(Y_1) = a_1^T S_x a_1$

  kjer je S_x vzorčna kovarianca prvotnih spremenljivk.

8.5.1.3. Definicija komponent

- Druga PC je linearna komponenta:

  $Y_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$

- katere varianca je naslednja najvišja:

  $Var(Y_2) = a_2^T S_x a_2$

- Komponenti nista korelirani.

8.5.1.3. Definicija komponent

- Izračunamo lahko toliko komponent, kot je v setu podatkov koreliranih spremenljivk.
- Dejansko so vse vzorčne kovariance in korelacije različne od nič, tako da iz p spremenljivk izračunamo p komponent.
- Celotna varianca komponent je tako enaka kot za prvotne spremenljivke – celotno varianco pojasnimo le z vsemi komponentami.
- Vendar je običajno večina variance nakopičena v prvih nekaj komponentah, in sicer do take mere, da z njimi pojasnimo večino variance.

8.5.1.4. Izračun uteži (obremenitev) a_i

- Uteži ali obremenitve izračunamo ali iz kovariančne ali iz korelacijske matrike.
- Če imajo spremenljivke različne merilne skale je primernejša korelacijska matrika.
- Rezultate izračunane iz kovariančne matrike pa je lažje interpretirati.
- Uteži, izračunane na en ali drugi način, se razlikujejo in jih ni možno transformirati iz enega tipa matrike v drugega.

8.5.1.4. Izračun uteži (obremenitev) a_i

- Zaporedje p transformacij združimo v eni enačbi:

$$Y = AX$$

- Kjer je A matrika, katere vrstice so vektorji uteži

$$\begin{array}{ccc} a_1^T = (a_{11} & \dots & a_{1p}) \\ \vdots & \vdots & \vdots \\ a_p^T = (a_{p1} & \dots & a_{pp}) \end{array}$$

8.5.1.4. Izračun uteži (obremenitev) a_i

- Nove spremenljivke Y_i so nekorelirane in njihove variance podaja $a_1^T S a_i$.
- Kovariančna matrika PCje potem

$$AS_xA^T = \begin{pmatrix} Var(Y_1) & 0 & \dots & \dots & 0 \\ 0 & Var(Y_2) & 0 & \dots & 0 \\ \vdots & 0 & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \dots & 0 & Var(Y_p) \end{pmatrix}$$

ali $AS_xA^T = S_y$

8.5.1.4. Izračun uteži (obremenitev) a_i

- V matematičnem jeziku se vrstice A (in enako stolpci A^T) imenujejo lastni (Eigen) vektorji matrike s_x in elementi v vodilni diagonali S_y lastne (Eigen) vrednosti.
- V statističnem izrazoslovju so lastni vektorji uteži a_{ij}, s katerimi obremenimo orvotne spremenljivke; lastne vrednosti pa so variance novih, nekoreliranih spremenljivk.

8.5.1.5. Zadetki (scores) glavnih komponent

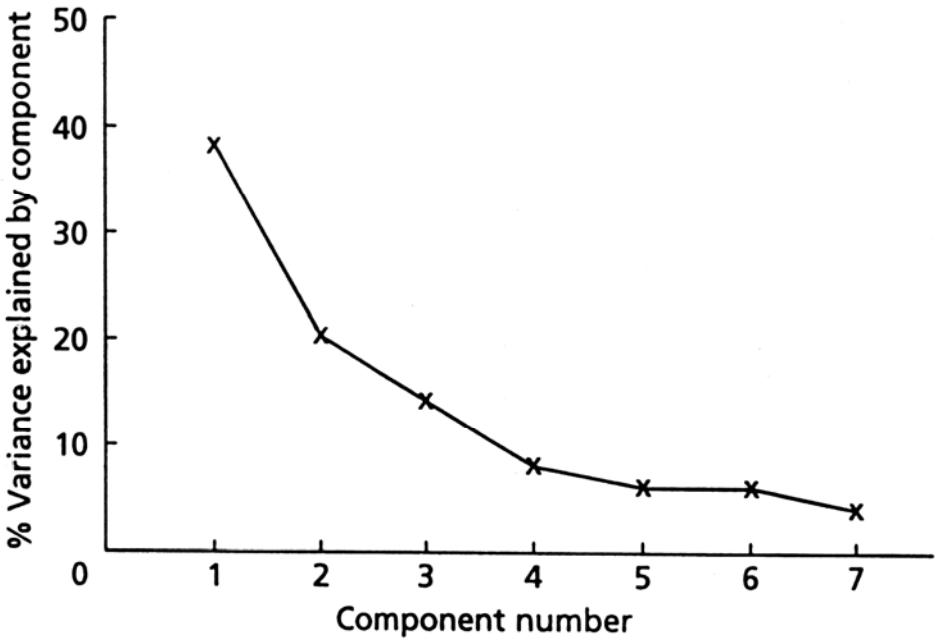
- Za vsako opazovanje izračunamo linearno kombinacijo vrednosti iz prvotnih spremenljivk tako, da izračunamo n vrednosti novih spremenljivk z uporabo uteži a_{ij}.

$$Y_{kr} = a_{k1}x_{k1} + a_{k2}x_{k2} + \ldots + a_{kp1}x_{kp}$$

- Vrednosti so k-ti zadetki PC, ki ustrezajo novim koordinatam.
- Običajno izdelamo razsevni diagram PC zadetkov, zlasti PC1 proti PC2.

8.5.1.6. Zmanjšanje dimenzionalnosti prostora

- Kriteriji, da neko komponento opustimo, so:
  - Dvig odstotnega deleža pojasnjene celotne variance, ko vključimo naslednjo najpomembnejšo komponento.
  - Delež celotne pojasnjene variance.
  - Komponente ne upoštevamo, če je njena varianca nižja od povprečja vseh komponent.
  - Približna enakost zadnjih lastnih vrednosti (varianc).



8.5.1.7. Korelacija med spremenljivkami in komponentami

- Kvadrat vzorčnega korelacijskega koeficienta med spremenljivko X_i in glavno komponento Y_i je mera deleža variance X_i, pojasnjena z j-to glavno komponento (PC):

$$r_{ij}^2 = \frac{a_{ij}^2 \cdot Var(Y_j)}{s_{ii}}$$

kjer je s_{ii} vzorčna varianca spremenljivke X_i.

8.5.1.8. Vključevanje in opuščanje spremenljivk

- Komponenta z nizko lastno vrednostjo – varianco, običajno ni pomembna. Zato opustimo spremenljivko, ki je močno obtežena na tej PC.
  - Opusti spremenljivko z najvišjo absolutno vrednostjo koeficientov v komponenti in ponovi postopek za vsako od (p – k) najmanjših lastnih vrednosti.
  - PC, katerih prispevki so pod 0,7, običajno opustimo. Ohraniti pa moramo vsaj štiri PC.

8.5.1.9. Nekatere običajne geološke razlage PC

- Geokemija in petrologija
  - PC1 je pogosto močna in popolnoma razložljiva z zaprtostjo podatkov – vpliv najbolj obogatene prvine na razredčenje vseh ostalih.
  - PC so posledica povezanosti skupine prvin, ki predstavljajo geološki (ali antropogen) izvor.

STATISTICA - [Data: ISTRA_tla.sta (20v by 269c)]

File Edit View Insert Format Statistics Graphs Tools Data Window Help

Add to Workbook Add to Report

Vars Cases

	1 NewVar	2 Al	3 Ca	4 Fe	5 Mg	6 Mn	7 Na	8 Ti	9 Ba	10 Co	11 Cr	12 Cu	13 Ga	14 La	15 Ni	16 Pb	17 Sr	18 V	19 Zn	20 Zr
153a	b	13	0,50	4,00	0,40	0,07	0,80	0,30	350,00	20,00	135,00	35,00	18,00	25,00	75,00	25,00	90,00	90,00	80,00	220,00
153a	b	13	0,50	4,00	0,40	0,07	0,80	0,30	350,00	20,00	135,00	35,00	18,00	25,00	75,00	25,00	90,00	90,00	80,00	220,00
153b	b	13	0,40	4,00	0,50	0,06	0,70	0,20	295,00	24,00	125,00	30,00	15,00	20,00	70,00	22,00	75,00	80,00	70,00	210,00
153b	b	13	0,40	4,00	0,50	0,06	0,70	0,20	295,00	24,00	125,00	30,00	15,00	20,00	70,00	22,00	75,00	80,00	70,00	210,00
154	a	9,50	7,00	6,20	1,00	0,30	0,50	0,50	370,00	14,00	110,00	58,00	19,00	40,00	80,00	52,00	80,00	120,00	110,00	200,00
155	a	13	13	2,50	0,90	0,06	1,20	0,30	1300,00	9,00	130,00	120,00	24,00	30,00	63,00	20,00	480,00	105,00	50,00	230,00
156	i	3,50	0,70	1,50	0,60	0,06	0,40	0,09	200,00	8,00	70,00	78,00	17,00	30,00	50,00	50,00	50,00	25,00	210,00	125,00
157	a	5,00	7,20	2,20	0,60	0,09	0,90	0,20	750,00	16,00	130,00	100,00	24,00	30,00	115,00	52,00	220,00	140,00	120,00	120,00
158	i	13	13	7,00	1,50	0,08	0,80	0,50	250,00	15,00	120,00	40,00	20,00	33,00	300,00	40,00	1000,00	150,00	130,00	180,00
158	i	13	13	7,00	1,50	0,08	0,80	0,50	250,00	15,00	120,00	40,00	20,00	33,00	300,00	40,00	1000,00	150,00	130,00	180,00
158r	i	13	13	6,00	1,30	0,09	0,70	0,40	300,00	12,00	110,00	60,00	19,00	50,00	250,00	36,00	800,00	90,00	50,00	180,00
158r	i	13	13	6,00	1,30	0,09	0,70	0,40	300,00	12,00	110,00	60,00	19,00	50,00	250,00	36,00	800,00	90,00	50,00	180,00
159	a	9,50	3,50	5,00	1,00	0,09	0,30	0,40	400,00	14,00	130,00	25,00	20,00	30,00	40,00	80,00	130,00	120,00	45,00	120,00
160	a	13	1,00	5,00	1,00	0,08	0,70	0,50	400,00	13,00	160,00	40,00	20,00	50,00	125,00	30,00	150,00	120,00	40,00	200,00
161	a	3,00	0,70	1,00	1,30	0,05	0,05	0,04	100,00	20,00	52,00	30,00	15,00	40,00	27,00	60,00	30,00	30,00	125,00	50,00
162	b	5,10	1,50	4,40	0,60	0,07	0,60	0,16	180,00	25,00	115,00	19,00	20,00	150,00	65,00	22,00	110,00	60,00	40,00	100,00
162	b	5,10	1,50	4,40	0,60	0,07	0,60	0,16	180,00	25,00	115,00	19,00	20,00	150,00	65,00	22,00	110,00	60,00	40,00	100,00
162r	b	5,60	1,60	5,00	0,70	0,08	0,70	0,20	210,00	28,00	120,00	19,00	25,00	120,00	65,00	25,00	95,00	60,00	50,00	220,00
162r	b	5,60	1,60	5,00	0,70	0,08	0,70	0,20	210,00	28,00	120,00	19,00	25,00	120,00	65,00	25,00	95,00	60,00	50,00	220,00
163	b	7,00	9,00	1,90	0,70	0,07	0,80	0,11	400,00	15,00	160,00	36,00	14,00	30,00	59,00	35,00	210,00	50,00	40,00	100,00
163	b	7,00	9,00	1,90	0,70	0,07	0,80	0,11	400,00	15,00	160,00	36,00	14,00	30,00	59,00	35,00	210,00	50,00	40,00	100,00
165a	a	9,00	1,50	5,30	0,80	0,04	0,90	0,60	320,00	8,00	100,00	72,00	32,00	50,00	63,00	120,00	140,00	150,00	130,00	250,00
165b	a	7,00	1,50	3,90	1,00	0,06	1,00	0,50	700,00	8,00	82,00	90,00	33,00	40,00	53,00	105,00	170,00	200,00	106,00	210,00
166	i	6,00	13	0,90	0,70	0,10	0,70	0,20	330,00	9,00	75,00	22,00	16,00	70,00	16,00	30,00	650,00	60,00	40,00	100,00
166	i	6,00	13	0,90	0,70	0,10	0,70	0,20	330,00	9,00	75,00	22,00	16,00	70,00	16,00	30,00	650,00	60,00	40,00	100,00
167a	i	10,00	13	2,60	1,00	0,10	0,75	0,50	400,00	9,00	125,00	100,00	19,00	50,00	50,00	46,00	350,00	70,00	100,00	150,00
167a	i	10,00	13	2,60	1,00	0,10	0,75	0,50	400,00	9,00	125,00	100,00	19,00	50,00	50,00	46,00	350,00	70,00	100,00	150,00
167b	i	8,00	13	2,00	0,80	0,08	0,60	0,50	400,00	9,00	145,00	87,00	16,00	60,00	40,00	89,00	320,00	60,00	110,00	140,00
167b	i	8,00	13	2,00	0,80	0,08	0,60	0,50	400,00	9,00	145,00	87,00	16,00	60,00	40,00	89,00	320,00	60,00	110,00	140,00
168	i	9,30	9,00	3,00	0,70	0,80	0,70	0,20	800,00	13,00	110,00	90,00	18,00	80,00	43,00	51,00	120,00	100,00	64,00	140,00
168	i	9,30	9,00	3,00	0,70	0,80	0,70	0,20	800,00	13,00	110,00	90,00	18,00	80,00	43,00	51,00	120,00	100,00	64,00	140,00
169	i	10,00	1,00	5,00	1,20	0,07	1,00	0,60	600,00	33,00	120,00	50,00	23,00	20,00	90,00	20,00	110,00	100,00	60,00	140,00
169	i	10,00	1,00	5,00	1,20	0,07	1,00	0,60	600,00	33,00	120,00	50,00	23,00	20,00	90,00	20,00	110,00	100,00	60,00	140,00
170	i	5,00	0,10	3,10	0,30	0,08	0,10	0,20	400,00	31,00	100,00	25,00	15,00	90,00	60,00	31,00	40,00	60,00	105,00	100,00
171a	a	13	5,00	2,60	1,20	0,20	0,30	0,50	200,00	9,00	70,00	60,00	20,00	100,00	60,00	35,00	250,00	80,00	45,00	100,00
171b	a	9,50	6,00	3,10	1,00	0,10	0,20	0,40	250,00	14,00	90,00	100,00	23,00	90,00	64,00	80,00	220,00	60,00	40,00	100,00
172	a	5,00	13	1,00	1,10	0,20	2,00	0,08	700,00	9,00	90,00	160,00	25,00	50,00	45,00	25,00	1000,00	40,00	120,00	70,00
173	a	6,00	1,70	3,00	1,00	0,07	0,90	0,20	400,00	10,00	106,00	30,00	28,00	44,00	25,00	90,00	150,00	125,00	78,00	250,00
173r	a	7,00	1,30	3,50	0,90	0,07	1,10	0,30	350,00	12,00	100,00	40,00	31,00	25,00	30,00	100,00	150,00	120,00	88,00	250,00
174a	a	7,50	0,70	2,50	0,40	0,01	0,70	0,50	300,00	16,00	90,00	40,00	20,00	30,00	63,00	75,00	60,00	70,00	64,00	300,00
174b	a	8,00	0,50	2,40	0,50	0,03	0,60	0,40	260,00	25,00	90,00	30,00	17,00	40,00	55,00	60,00	80,00	60,00	50,00	200,00

Principal Components...

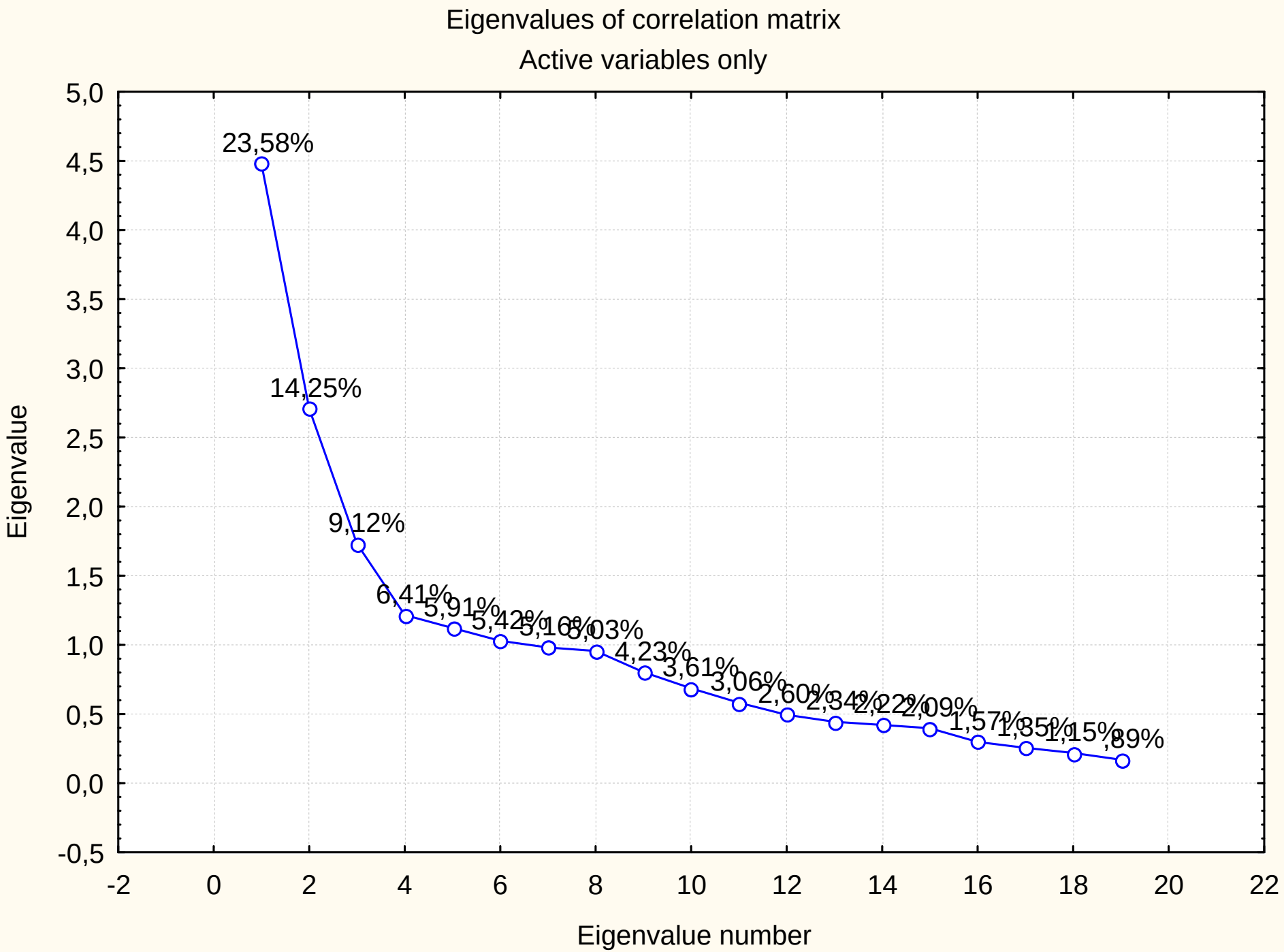
For Help, press F1

Sel:OFF Weight:OFF CAPS NUM REC

Microsoft PowerPoint - [statistika13 multivariatne.ppt]

start Yahoo! Messenger Microsoft PowerPoint ... STATISTICA - [Data: ... 9:08

Value number	Eigenvalues of correlation matrix, and related statist Active variables only			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	4,480274	23,58039	4,48027	23,5804
2	2,708076	14,25303	7,18835	37,8334
3	1,731875	9,11513	8,92022	46,9485
4	1,216996	6,40524	10,13722	53,3538
5	1,122404	5,90739	11,25962	59,2612
6	1,030146	5,42182	12,28977	64,6830
7	0,980965	5,16297	13,27074	69,8460
8	0,956515	5,03429	14,22725	74,8803
9	0,803844	4,23076	15,03110	79,1110
10	0,686418	3,61272	15,71751	82,7238
11	0,582014	3,06323	16,29953	85,7870
12	0,494052	2,60028	16,79358	88,3873
13	0,444420	2,33905	17,23800	90,7263
14	0,421840	2,22021	17,65984	92,9465
15	0,397048	2,08972	18,05689	95,0362
16	0,298882	1,57306	18,35577	96,6093
17	0,255877	1,34672	18,61165	97,9560
18	0,218472	1,14985	18,83012	99,1059
19	0,169882	0,89412	19,00000	100,0000


Eigenvalues of correlation matrix
Active variables only
Eigenvalue
Eigenvalue number
5,0
4,5
4,0
3,5
3,0
2,5
2,0
1,5
1,0
0,5
0,0
-0,5
-2
0
2
4
6
8
10
12
14
16
18
20
22
23,58%
14,25%
9,12%
6,41%
5,91%
5,42%
5,16%
5,03%
4,23%
3,61%
3,06%
2,60%
2,34%
2,22%
2,09%
1,57%
1,35%
1,15%
,89%

STATISTICA - [Workbook1* - Factor coordinates of the variables, based on correlations (ISTRA_tla.sta)]

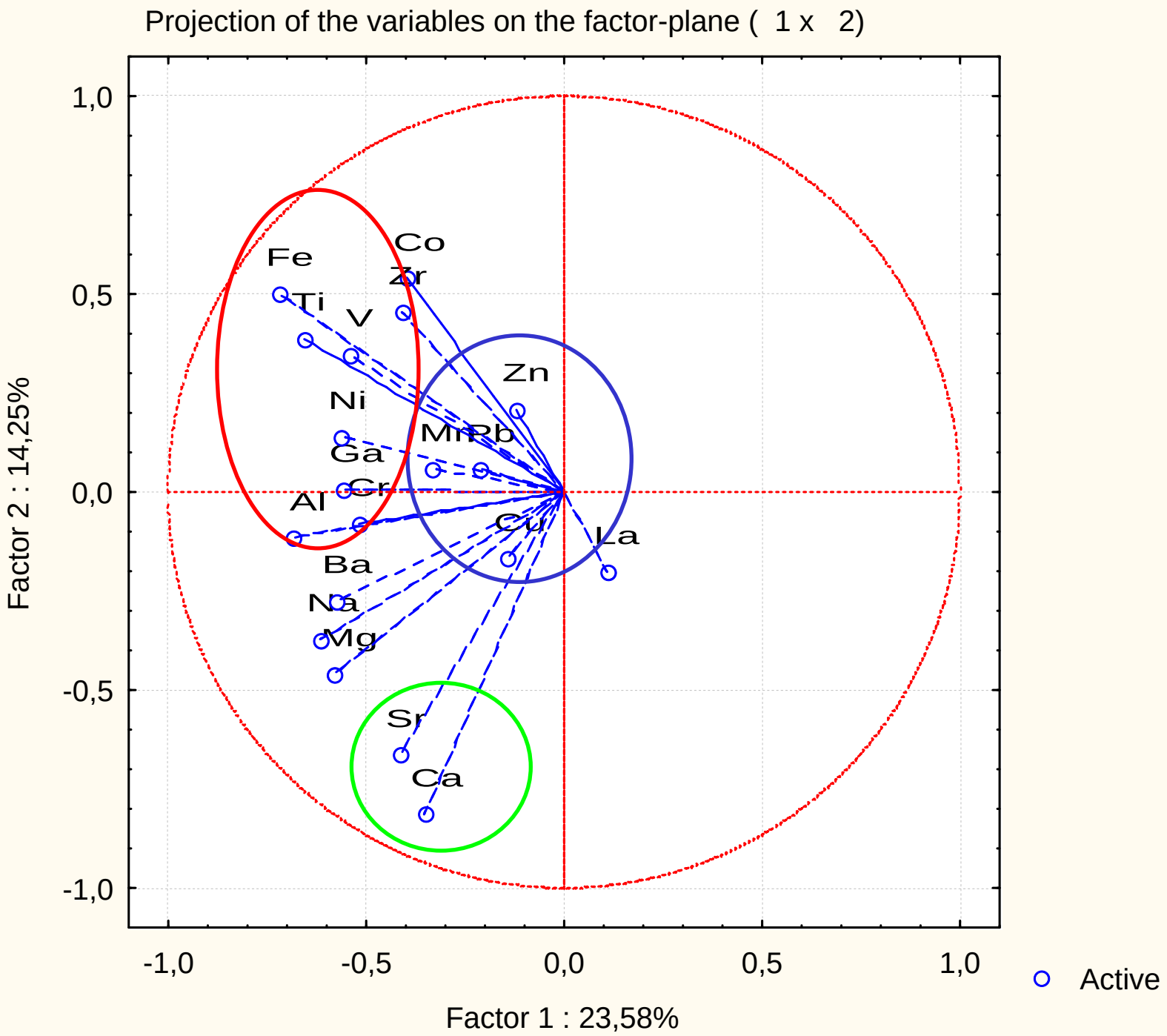
File Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help

Add to Workbook Add to Report

Arial 10

Factor coordinates of the variables, based on correlations (ISTRA_tla.sta)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Fact.10
Al	-0,683069	-0,116579	-0,049399	0,074706	0,218641	-0,266509	-0,082316	-0,114847	-0,346609	-0,168120
Ca	-0,353817	-0,814236	-0,038455	0,063471	-0,067112	-0,104199	-0,027605	0,068873	0,073226	-0,066202
Fe	-0,719705	0,501129	-0,099754	-0,049039	-0,085816	-0,207311	-0,112803	-0,054575	-0,095808	0,010265
Mg	-0,580556	-0,459788	-0,048774	-0,123549	-0,115892	0,008251	-0,223286	-0,173030	-0,085971	0,306247
Mn	-0,332025	0,058153	-0,042443	-0,395963	-0,363053	0,397067	-0,274870	0,373456	-0,288498	-0,243561
Na	-0,616491	-0,371379	0,037333	0,007724	0,148293	0,280558	0,264973	-0,121202	-0,237990	0,095096
Ti	-0,655301	0,388279	-0,006032	0,140818	0,231817	-0,024531	-0,120515	0,154834	0,221240	0,022186
Ba	-0,577660	-0,274565	-0,149848	0,036326	0,264450	0,242909	0,082683	0,366458	0,179292	-0,140677
Co	-0,397600	0,539048	-0,279380	0,004461	-0,213705	0,061495	0,275309	0,188772	-0,216938	0,113893
Cr	-0,515696	-0,080462	-0,273347	0,216563	-0,286559	0,357642	0,246809	-0,084333	0,247820	0,346464
Cu	-0,143969	-0,168747	0,776000	0,046581	-0,109282	0,000576	0,177648	0,319061	-0,019222	-0,044904
Ga	-0,559165	0,006009	0,384706	-0,363444	0,333711	0,103019	-0,013794	-0,387591	-0,116769	0,061751
La	0,108901	-0,201701	-0,181868	-0,736354	-0,282877	-0,067976	0,253594	-0,231752	0,249423	-0,156479
Ni	-0,562204	0,139728	0,028522	0,200908	-0,547138	-0,432041	-0,014351	-0,097314	-0,016996	-0,044459
Pb	-0,215088	0,056218	0,780190	-0,134717	-0,144628	-0,143297	0,127727	0,144181	0,137211	0,193209
Sr	-0,416727	-0,664050	-0,142759	0,167511	-0,059965	-0,203665	0,016987	0,037518	0,178793	-0,261102
V	-0,539923	0,348137	-0,035102	-0,257725	0,153334	-0,034335	-0,397443	0,078449	0,422926	0,051251
Zn	-0,122175	0,209779	0,348859	0,367121	-0,252277	0,458697	-0,206578	-0,447162	0,131910	-0,330652
Zr	-0,410306	0,454341	-0,033068	-0,023466	0,172312	-0,045345	0,541764	-0,083568	0,093640	-0,297435


Projection of the variables on the factor-plane (1 x 2)
Fe
Co
Zr
Ti
V
Zn
Ni
Mn
Pb
Ga
Cr
Al
Cu
La
Ba
Na
Mg
Sr
Ca
1,0
0,5
0,0
-0,5
-1,0
-1,0
-0,5
0,0
0,5
1,0
Factor 2 : 14,25%
Factor 1 : 23,58%
Active

8.5.1.9. Nekatere običajne geološke razlage PC

- Paleontologija
  - Kadar so podatki meritve različnih anatomskih delov fosilov, je PC1 skoraj vedno mogoče razložiti kot velikost (npr. juvenilni in odrasli osebki).
  - Ko s PC1 odstranimo dejavnik velikosti, ostale PC podajajo obliko, morda povezano z načinom življenja.

8.5.2. FAKTORSKA ANALIZA

- Osnova metode je, da je celotno spremenljivost podatkov mogoče prikazati z manjšim številom (q) faktorjev, ki nosijo informacijo o preiskovani snovi.
- Poiščemo le q vektorjev, medtem, ko je pri PCA število PC vedno enako številu vseh spremenljivk.
- Varianco v q smereh zvišamo do najvišje možne tako, da osi vrtimo iz osnovne rešitve PC lastnih vektorjev.

8.5.2. FAKTORSKA ANALIZA

- Metoda VARIMAX ohrani pravokotnost osi. Vrtenje osi je takšno, da poišče najvišje možne variance ob upoštevanju zahteve ortogonalnosti.
- Metoda OBLIQUE dovoljuje, da se osi zavrtijo iz pravokotnega položaja, zaradi česar postanejo do neke mere korelirane.

8.5.2. FAKTORSKA ANALIZA

- Kritike faktorske analize:
  - Ni jasnega kriterija za število faktorjev.
  - Težko je sistematično preiskati vse smeri multidimenzionalnega prostora.
  - Težko se prepričamo, da je dobljeni rezultat res najboljša možna rešitev.
  - Različni računalniški programi nudijo številne različne možnosti izračuna, za katere ne vemo vedno ali so res najustreznejši za naše podatke. Različni programi tako lahko dajo za isti set podatkov različne rešitve.

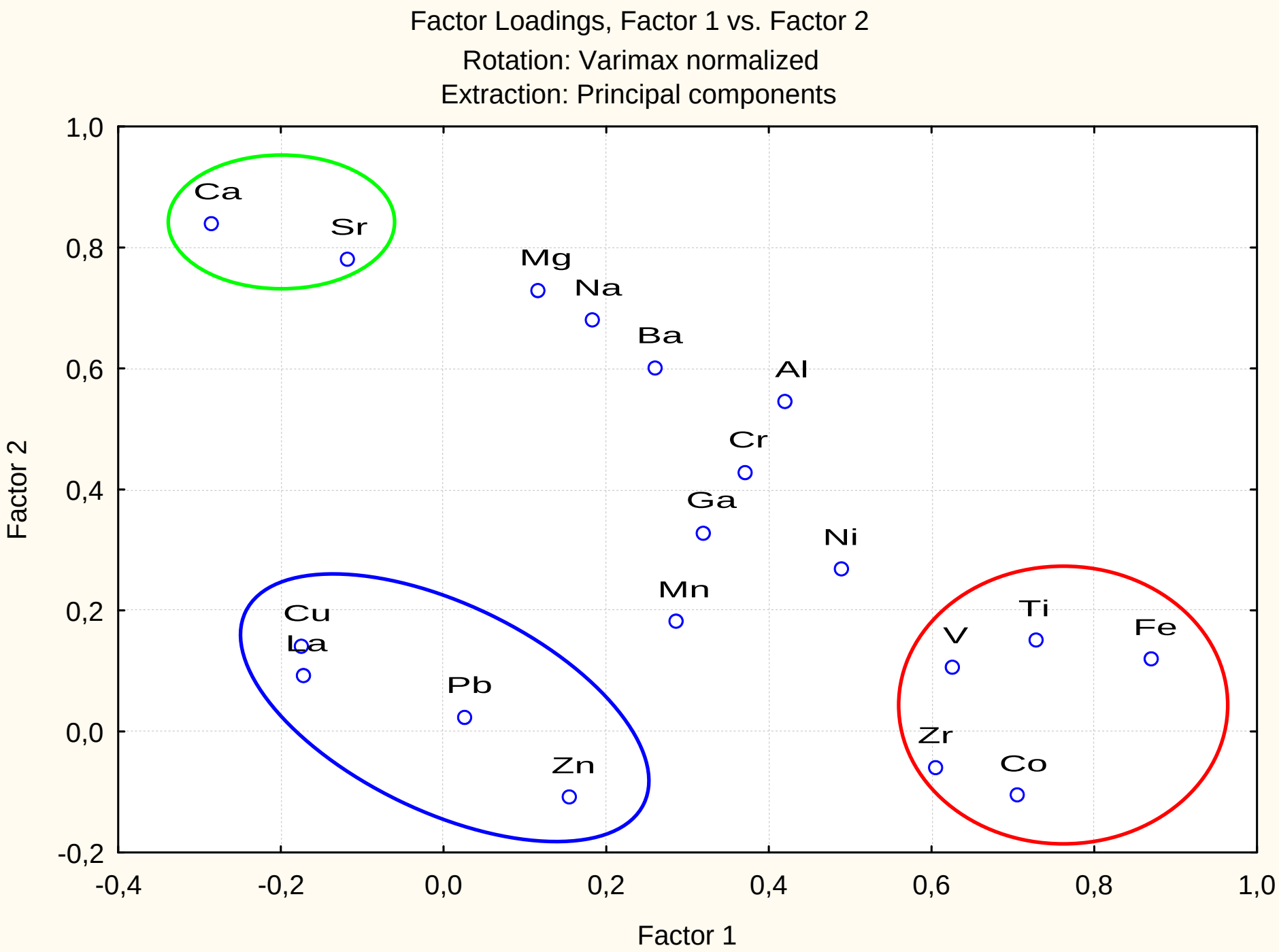
STATISTICA - Workbook1* - [Factor Loadings

File Edit View Insert Format Statistics Graphs

Arial 10

Factor Loadings (Varimax normalized)
Extraction: Principal components
(Marked loadings are > ,700000)

Variable	Factor 1	Factor 2	Factor 3
Al	0,417230	0,546997	0,096575
Ca	-0,287044	0,840808	-0,017112
Fe	0,868544	0,120688	0,100609
Mg	0,113993	0,731710	0,049344
Mn	0,283801	0,182991	0,037351
Na	0,180435	0,681959	0,147487
Ti	0,727062	0,151677	0,169088
Ba	0,257840	0,603113	-0,036115
Co	0,703713	-0,104599	-0,143383
Cr	0,370505	0,430742	-0,155956
Cu	-0,176865	0,143883	0,774206
Ga	0,318777	0,330882	0,499599
La	-0,173808	0,094090	-0,215769
Ni	0,486785	0,269481	0,163801
Pb	0,023703	0,024974	0,810515
Sr	-0,119306	0,782321	-0,093505
V	0,624428	0,107129	0,112077
Zn	0,152968	-0,107704	0,381624
Zr	0,603252	-0,057979	0,092712
Expl.Var	3,556612	3,488659	1,874952
Prp.Totl	0,187190	0,183614	0,098682


Factor Loadings, Factor 1 vs. Factor 2
Rotation: Varimax normalized
Extraction: Principal components
Factor 2
Factor 1
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,2
0,0
0,2
0,4
0,6
0,8
1,0
Ca
Sr
Mg
Na
Ba
Al
Cr
Ga
Ni
Mn
Cu
La
Pb
Zn
Ti
V
Fe
Zr
Co

8.6. MERE PODOBNOSTI IN ANALIZA ZDRUŽEVANJA (CLUSTER ANALYSIS)

8.6.1. Načela

- Osnova analiz je mera podobnosti oz. različnosti med objekti (opazovanji).
- Podobnost ugotavljamo z upoštevanjem več spremenljivk.
- Podatke izrazimo v obliki koeficientov podobnosti, ki podajajo numerično podobnost med ustreznimi dvojicami objektov.

8.6.2. Izračuni podobnosti

- Tri glavne kategorije mere podobnosti so:

1 . Distančni koeficienti ali koeficienti razdalje.

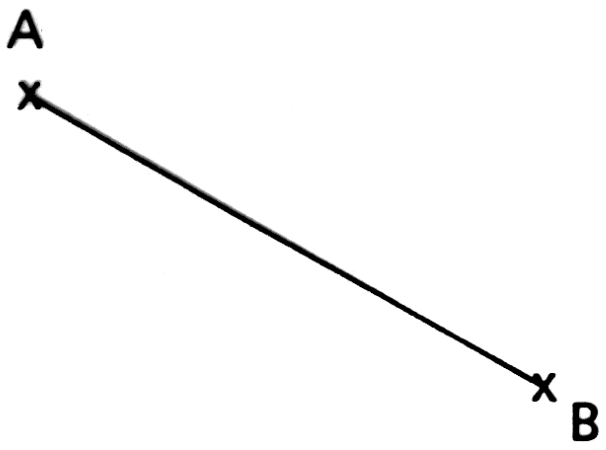
2. Korelacijski koeficienti.

3. Asociacijski (združevalni) koeficienti.

8.6.2.1. Koeficienti razdalje

- Euklidska razdalja

  je neposredna ravna črta:



$$d_{AB} = \sqrt{\sum_{i=1}^{m} (x_{iA} - x_{iB})^2}$$

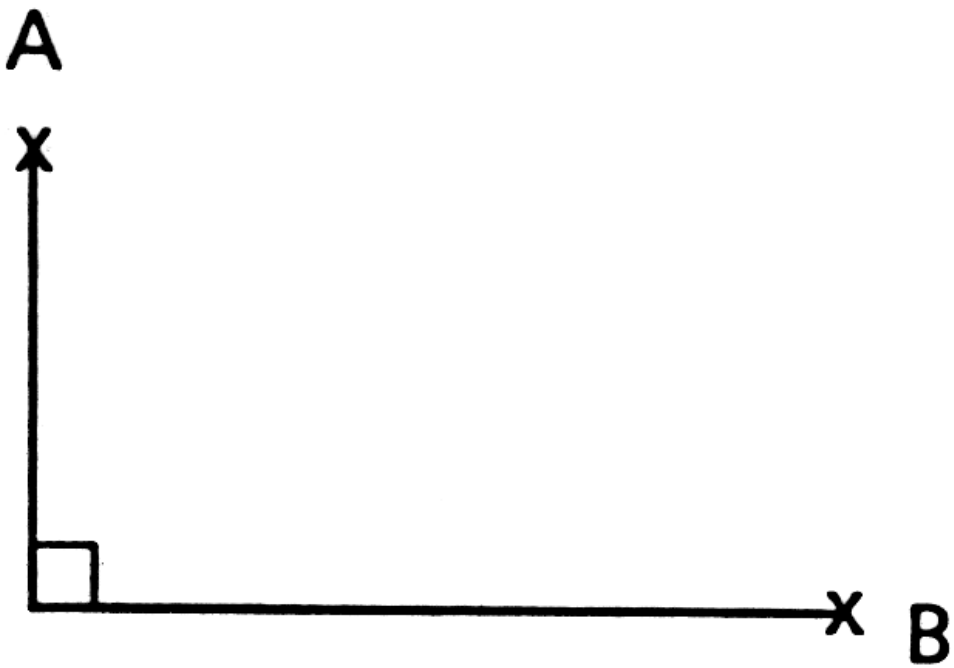
oz. ko ne želimo, da mera obvezno naraste zaradi več spremenljivk:

$$d_{AB} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_{iA} - x_{iB})^2}$$

8.6.2.1. Koeficienti razdalje

- Manhattanska razdalja
  je enostavna vsota absolutnih razlik.

$$d_{AB} = \frac{1}{m}\sum_{i=1}^{m}\left|x_{iA} - x_{iB}\right|$$



8.6.2.2. Korelacijski koeficient podobnosti

- Podobnost obravnavamo v smislu razmerij med vrednostmi različnih spremenljivk.
- Uporabimo korelacijski koeficient, le da namesto povezav dveh spremenljivk v več objektih, obravnavamo dva objekta z več spremenljivkami.
- Srednje vrednosti in standardna odklona objektov A in B izračunamo iz vseh spremenljivk, korelacijski koeficient med objektoma pa:

$$r_{AB} = \frac{\sum_{i=1}^{m} (x_{iA} - \bar{x}_A)(x_{iB} - \bar{x}_B)}{s_A s_B (n-1)}$$

8.6.2.3. Združevalni koeficient

- Uporabni so za binarne podatke (je ali ni prisoten) v stratigrafiji.
- Podatki za primerjavo vzorcev A in B so omejeni, vendar pa lahko sklepamo, da je njuna podobnost popolna, če so prisotnosti (1) in odsotnosti (0) preiskovanih lastnosti enake, oz. da gre za različna vzorca, če ni nobenega ujemanja.
- Podatke podamo v kontingenčni tabeli (tabeli možnosti):
  - a - število lastnosti, prisotnih v obeh vzorcih
  - b - število lastnosti, prisotnih v vzorcu A, odsotnih v B
  - c - število lastnosti, prisotnih v vzorcu B, odsotnih v A
  - d - število lastnosti, odsotnih v obeh vzorcih

		Vzorec B	
		prisoten	odsoten
Vzorec A	prisoten	a	b
	odsoten	c	d

8.6.2.3. Združevalni koeficient

- Asociacijske koeficiente izračunamo iz vrednosti a, b, c in d.
- Jaccardov asociacijski koeficient ne upošteva, da tudi skupna odsotnost lastnosti (d), kaže na podobnost. Je delež vseh ujemajočih se sprmenljivk:

$$J_{AB} = \frac{a}{a+b+c}$$

8.6.3. Stratigrafska korelacija
8.6.3.1. Podatki in matrika podobnosti

- Vsako fosilno vrsto obtežimo z indeksom – relativna biostratigrafska vrednost (RBV).
- Klasične lastnosti dobrega conskega fosila so:
  - Omejen stratigrafski razpon.
  - Široka geografska razširjenost.
  - Facialna neodvisnost.

8.6.3.1. Podatki in matrika podobnosti

- Kvantificiramo jih:

$$\text{vertikalni razpon } V = \frac{\text{razpon, ki ga zaseda takson}}{\text{celotna obravnavana stratigrafska sekvenca}}$$

$$\text{geografska razširjenost } G = \frac{\text{število sekcij, v katerih se pojavlja takson}}{\text{celotno število obravnavanih sekcij}}$$

$$\text{facialna neodvisnost } F = \frac{\text{število faciesov, v katerih se pojavlja takson}}{\text{celotno število faciesov}}$$

8.6.3.1. Podatki in matrika podobnosti

- Dober conski fosil ima nizek V ter visoka G in F.
- Indeksi RBV so sestavljeni tako, da kvantificirajo to kombinacijo lastnosti:
- McCammonov RBV:

$$RBV_1 = F(1-V) + (1-F)G$$

- Browerjev RBV:

$$RBV_2 = (F(1-V) + (1-V)G)/2$$

- Razpon RBV je od 0 (neuporaben) do 1 (popolen conski fosil).

8.6.3.1. Podatki in matrika podobnosti

- Netehtan RBV, kakršen je:

  RBV = (G + F + (1-V))/3

  bi bil visok pri prostorsko in časovno povsod navzočih fosilih, ki so kot conski fosili neuporabni.
- Jaccardov koeficient lahko ponovno definiramo kot:

  S = a/(a + b + c)
- Kjer je:
  - a – vsota RBV vrst, prisotnih v obeh vzorcih
  - b - vsota RBV vrst, prisotnih v vzorcu A, odsotnih v B
  - c - vsota RBV vrst, prisotnih v vzorcu B, odsotnih v A
  - d - vsota RBV vrst, odsotnih v obeh vzorcih

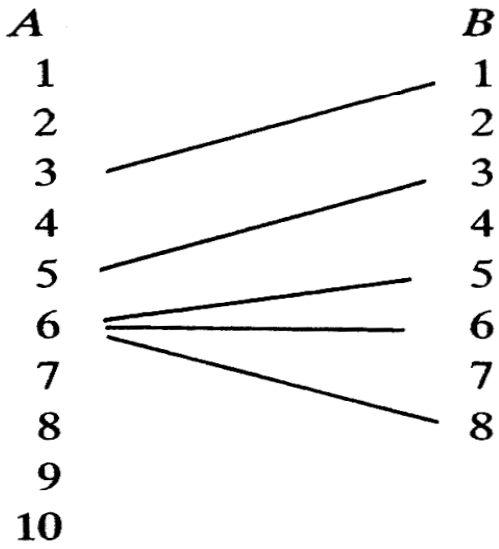
8.6.3.2. Stratigrafska korelacija iz matrike podobnosti

- Korelacijske povezave se ne smejo križati.
- Dve glavni vrsti algoritmov, ki upoštevata navedeno omejitev sta:
  - Lateralno (bočno, stransko) sledenje-
  - Zarezovanje (slotting)

8.6.3.2.1. Lateralno sledenje

- Poiščemo največjo podobnost v matriki (najnižji koeficient) in povežemo ustrezni par vzorcev.
- Poiščemo naslednjo največjo od preostalih podobnosti. Če ta ne seka obstoječe korelacijske povezave, para povežemo.
- Postopek nadaljujemo, dokler obstajajo možne povezave.
- Prednost metode je njena enostavnost.
- Pomankljivost so uteži, pripisane največjim podobnostim – če je ena izmed njih napačna, so preostale korelacije neuporabne.

Borehole A	Borehole B: 1	2	3	4	5	6	7	8
1	0.75	1.89	1.05	1.31	1.36	1.52	1.44	1.43
2	0.66	1.47	0.63	0.89	0.93	1.09	1.01	1.01
3	$\boxed{0.60^{10}}$	1.63	0.63	1.25	1.30	1.46	1.02	1.37
4	0.88	1.08	0.78	$\boxed{0.50^{6}}$	0.86	0.93	0.94	0.93
5	1.48	1.88	$\boxed{0.45^{4}}$	0.90	0.66	0.70	$\boxed{0.43^{3}}$	0.79
6	1.51	1.68	1.05	0.79	$\boxed{0.38^{1}}$	$\boxed{0.42^{2}}$	0.86	$\boxed{0.51^{7}}$
7	2.06	2.29	1.19	0.93	$\boxed{0.52^{8}}$	$\boxed{0.56^{9}}$	1.01	0.65
8	2.26	2.48	1.02	1.13	0.72	$\boxed{0.56^{9}}$	0.64	0.65
9	1.50	1.58	0.99	$\boxed{0.49^{5}}$	$\boxed{0.56^{9}}$	0.66	0.98	0.63
10	1.72	2.35	1.14	1.06	0.91	0.72	0.99	0.64



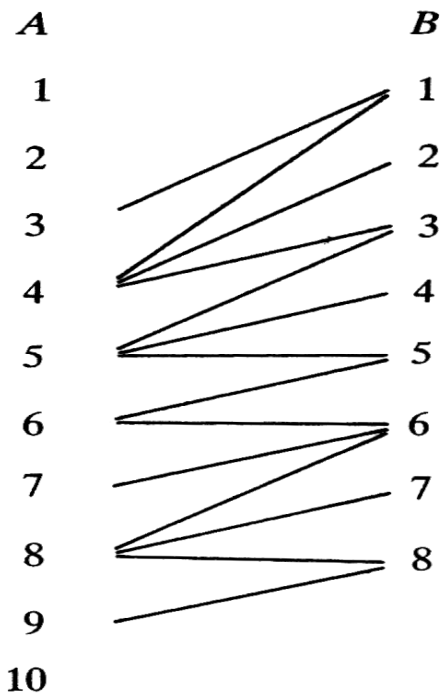
8.6.3.2.2. Zarezovanje

- Omejitev križanja je upoštevana z določilom, da položaj para v smeri desno-zgoraj proti levo-spodaj pomeni križanje, vse ostale možnosti (levo-zgoraj – desno-spodaj, gor-dol, levo-desno) pa, da križanja ni.
- Pari popolnoma koreliranih sekcij si bodo čez celotno matriko sledili s povezavami zgoraj-levo do spodaj-desno.
- Cilj zarezovanja je poiskati najboljšo takšno pot. Kriterij je, da naj bo lelotna vsota podobnosti največja.

8.6.3.2.2. Zarezovanje

- Začnemo zgoraj-desno (ali spodaj-levo) in vsak korak po matriki ima dve možnosti. Odločitev je vedno le za korak naprej in ne upošteva vseh možnosti.
- Zarezovanje je bolj grobo od lateralnega sledenja, ker je manj občutljivo na anomalne vrednosti.
- Problemi so:
  - Pristop k podatkom je asimetričen in odločitev, da vstopimo v matriko zgoraj-desno je umetna.
  - Rezultat lahko pokvari en sam anomalen vzorec v eni od sekcij, kar lahko ustvari zid majhnih podobnosti skozi celotno matriko, vzdolž katerga teče pot in ga ne seka.

	Borehole *B*							
Borehole *A*	1	2	3	4	5	6	7	8
1	0.75	1.89	1.05	1.31	1.36	1.52	1.44	1.43
2	0.66	1.47	0.63	0.89	0.93	1.09	1.01	1.01
3	**0.60**	1.63	0.63	1.25	1.30	1.46	1.02	1.37
4	**0.88**	**1.08**	**0.78**	0.50	0.86	0.93	0.94	0.93
5	1.48	1.88	**0.45**	**0.90**	**0.66**	0.70	0.43	0.79
6	1.51	1.68	1.05	0.79	**0.38**	**0.42**	0.86	0.51
7	2.06	2.29	1.19	0.93	0.52	**0.56**	1.01	0.65
8	2.26	2.48	1.02	1.13	0.72	**0.56**	**0.64**	**0.65**
9	1.50	1.58	0.99	0.49	0.56	0.66	0.98	**0.63**
10	1.72	2.35	1.14	1.06	0.91	0.72	0.99	0.64



8.6.4. CLUSTERSKA (ZDRUŽEVALNA) ANALIZA

8.6.4.1. Načela numerične klasifikacije

- V geologiji uporabljamo predvsem dve vrsti numerične klasifikacije:

1. Ločevanje glede na absolutne vrednosti.
2. Združevanje glede na relativne lastnosti vzorca.

8.6.4.1.1. Ločevanje glede na absolutne vrednosti

- Prostor je s črtami ali ravninami razdeljen v oddelke (trikotni in kvadratni diagrami).
- Podatkov za takšno klasifikacijo ni potrebno statistično analizirati.
- Prednosti:
  - Grobost – klasifikacijo lahko uporabimo vsesplošno in nedvoumno.
  - Ločnice postavimo tako, da imajo teoretični pomen.
- Slabosti:
  - Položaji ločnic so pogosto umetni in nenaravni.
  - Dva podobna primerka se lahko uvrstita na različni strani ločnice in bosta zato različno poimenovana.
  - Obvladovanje razdelitve, definirane z več kot dvema ali tremi spremenljivkami, je zelo okorno ali nemogoče.

8.6.4.1.2. Združevanje glede na relativne lastnosti vzorca

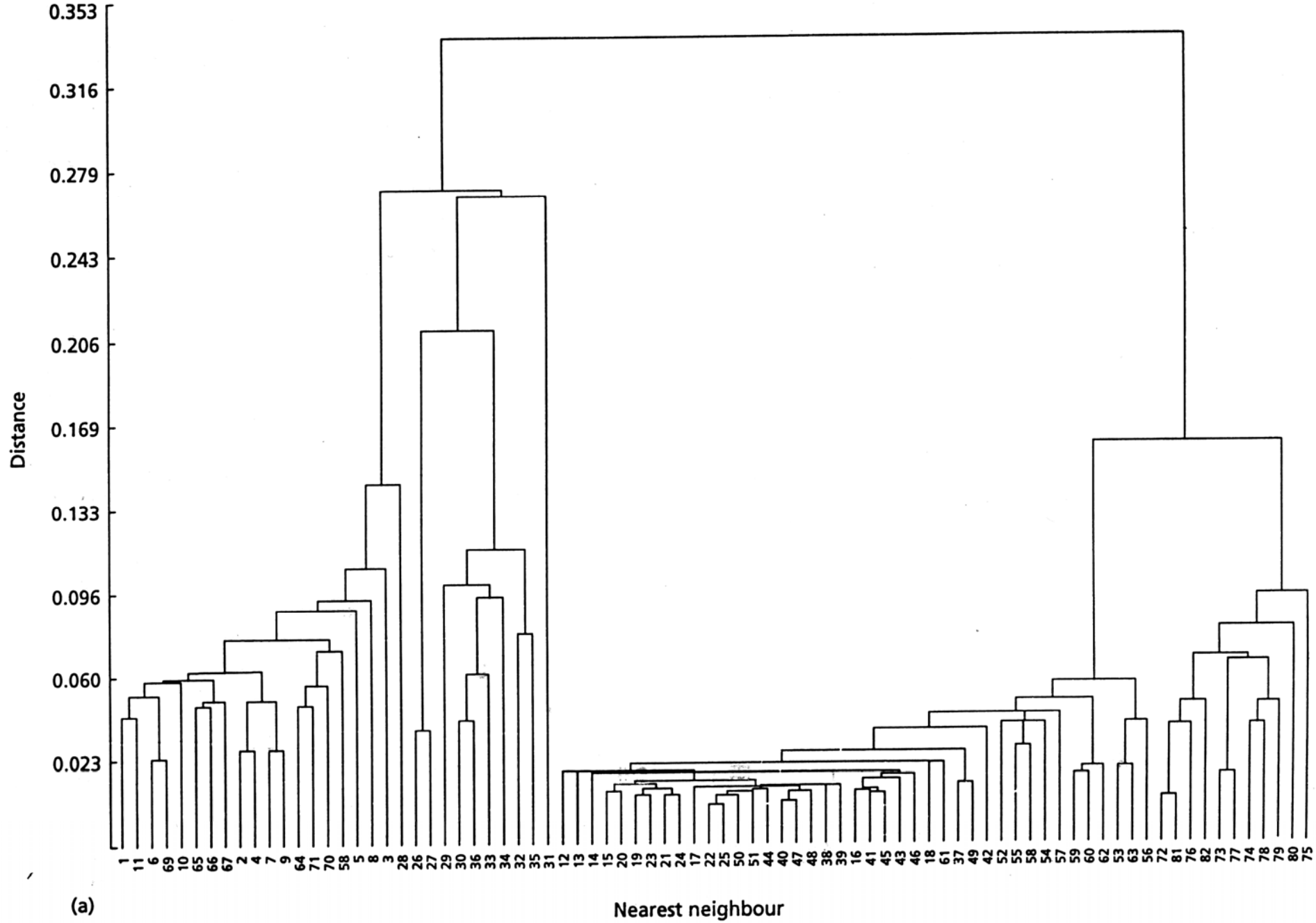
- Posamezne kategorije znotraj klasifikacije temeljijo na nakopičenih skupinah točk.
- Točki, ki sta v prostoru lastnosti (spremenljivk) v isti skupini, sta si bolj podobni kot točki iz različnih skupin.
- Prednosti:
  - Jasne, unimodalne skupine opazovanj se ne uvrstijo v različne kategorije.
  - Meje med skupinami so v tistih delih multivariatnega prostora, kjer je malo točk – so naravne in zato lahko posledica nekega procesa.
  - Upoštevamo lahko vse spremenljivke.
- Slabosti:
  - Nestabilnost – če dodamo nova opazovanja se lahko pojavijo nove skupine ali spremenijo meje prejšnih.
  - Uvrstitev novega primerka v skupino ni tako enostavno.

8.6.4.2. Algoritmi združevanja v skupine

- Med različnimi metodami ločevanja skupin so:
  - Metode delitve, ki poiščejo s točkami redko zaseden dele in tam potegnejo meje skupin.
  - Metode gostote, ki določijo skupine tako, da v multivariatnem prostoru poiščejo nakopičenja točk.
  - Metode povezovanja, ki povezujejo eno za drugo najbližje točke.

8.6.4.2.1. Povezovanje najbližjih sosedov ali enojno povezovanje

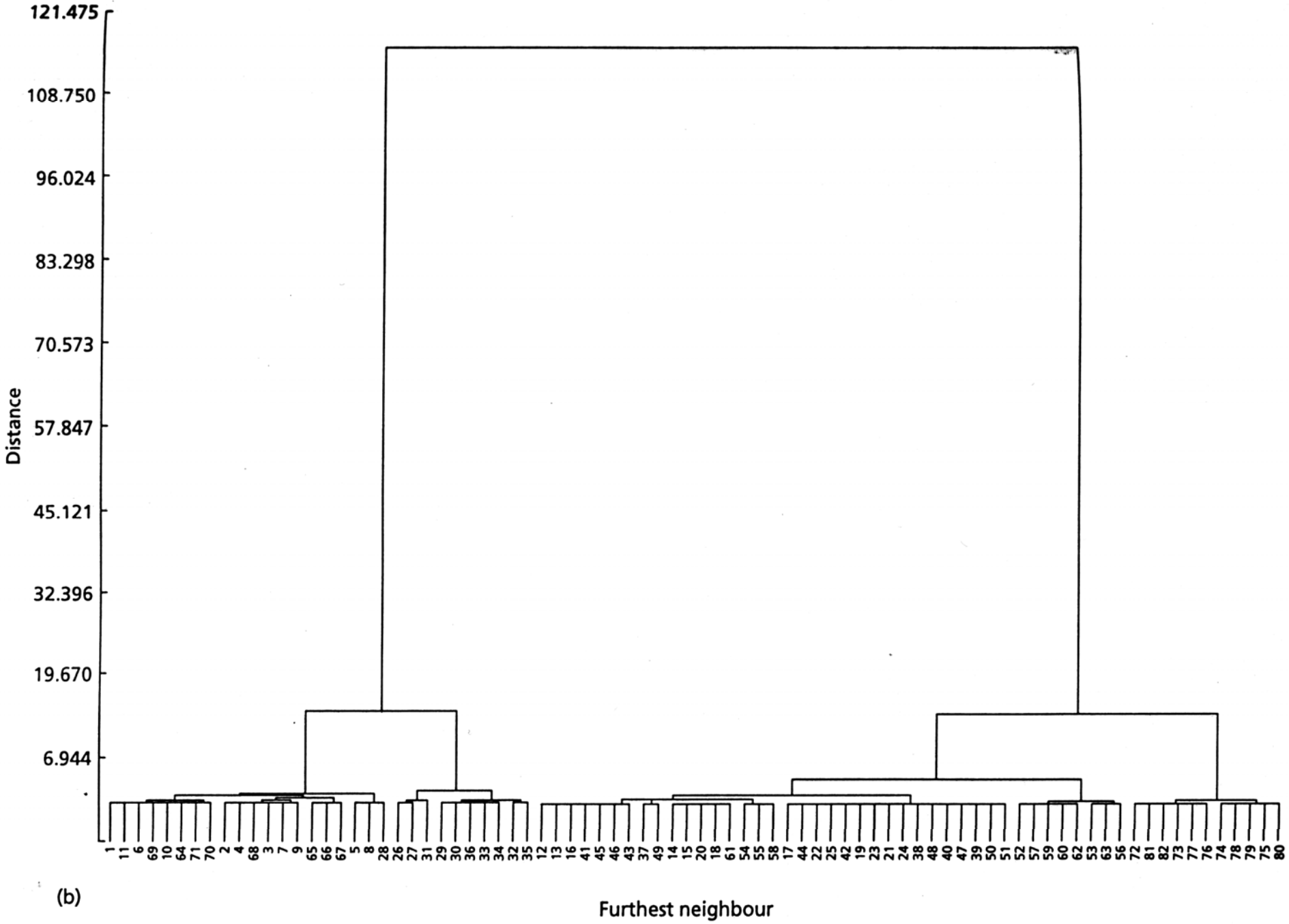
- Podobnost med točko in novo skupino je enaka podobnosti med to točko in najbolj podobno točko v kupini.
- Rezultat povezovanja prikažemo z dendrogramom.
- Ta mera oddaljenosti je najmanj primerna in je za večini točk, ki jih vsebuje skupina, manjša od resnične razdalje.
- Pogosto ji zato pripisujemo popačenje multivariatnega prostora s krčenjem.
- Na ta način zlahka povežemo končni točki v oblaku točk, razpotegnjenem v eni smeri, čeprav sta si točki na nasprotnih koncih take skupine bistveno različni.
- Učinek imenujemo veriženje in teži k rezulatu, ki je neurejen hierarhični sistem, nenaravnega videza.


Distance
0.353
0.316
0.279
0.243
0.206
0.169
0.133
0.096
0.060
0.023
1 11 6 69 10 65 66 67 2 4 7 9 64 71 70 58 5 8 3 28 26 27 29 30 36 33 34 32 35 31 12 13 14 15 20 19 23 21 24 17 22 25 50 51 44 40 47 48 38 39 16 41 45 43 46 18 61 37 49 42 52 55 58 54 57 59 60 62 53 63 56 72 81 76 82 73 77 74 78 79 80 75
Nearest neighbour

(a)

8.6.4.2.2. Povezovanje najoddaljenejših sosedov ali celotno povezovanje

- Postopek je obraten metodi najbližjega soseda – za podobnost med točko in skupino upošteva najmanjšo podobnost med vsemi možnimi pari podobnosti.
- Navidezne razdalje med skupinami so največje možne, zato metoda razredči prostor in rezultat so hipersferične skupine.
- Celo malo elipsoidne unimodalne skupine točk bo posteopek verjetno razbil in rezultat bo nakazoval več modusov (skupin).


121.475
108.750
96.024
83.298
70.573
57.847
45.121
32.396
19.670
6.944
Distance
1 11 6 69 10 64 71 70 2 4 68 3 7 9 65 66 67 5 8 28 26 27 31 29 30 36 33 34 32 35 12 13 16 41 45 46 43 37 49 14 15 20 18 61 54 55 58 17 44 22 25 42 19 23 21 24 38 48 40 47 39 50 51 52 57 59 60 62 53 63 56 72 81 82 73 77 76 74 78 79 75 80
Furthest neighbour

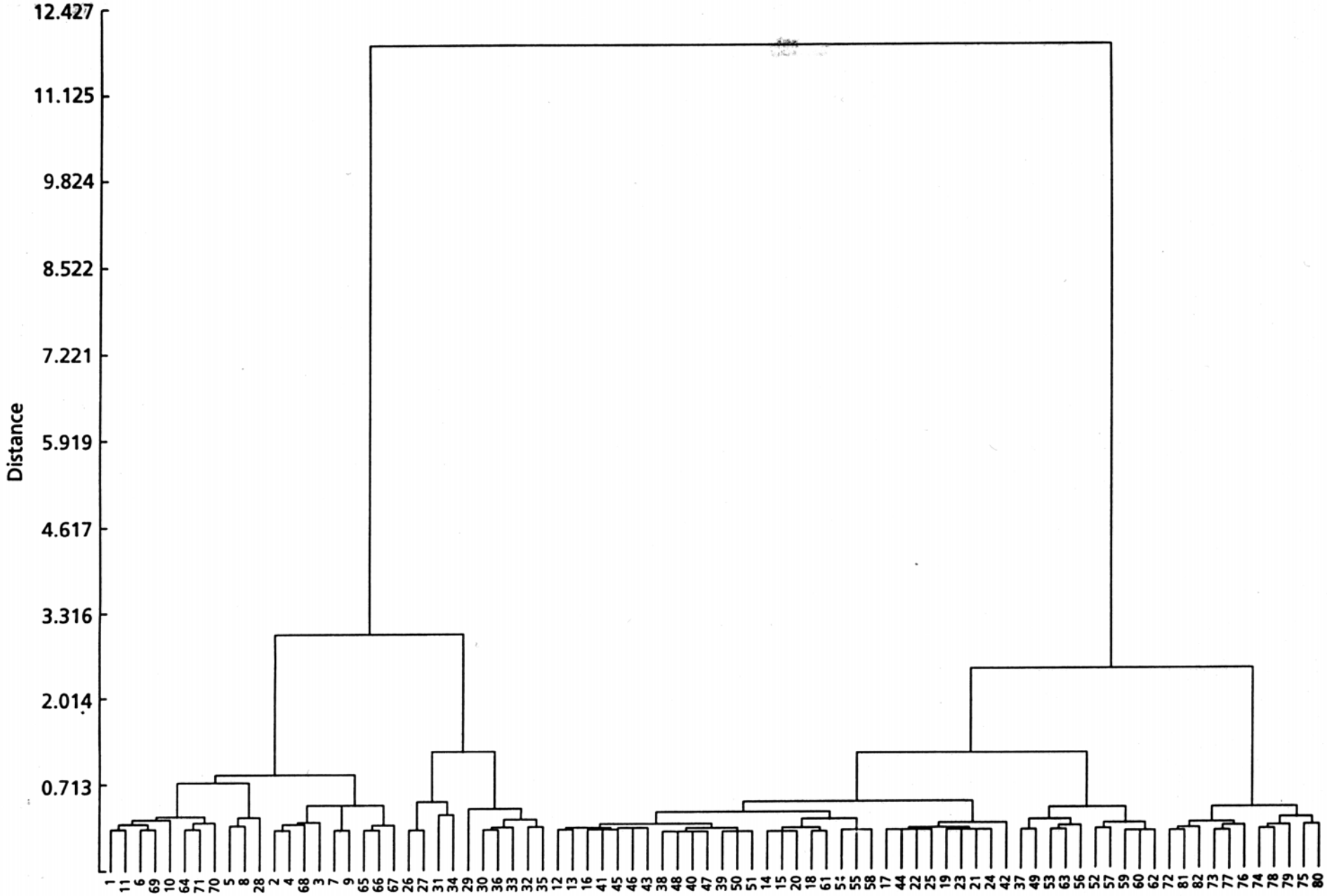
(b)

8.6.4.2.3. Povezovanje povprečij

- Kriterij preračunavanja podobnosti je v tem primeru med prej opisanima.
- Rezultat združuje oboje, srednjo vrednost obeh razdalj in oddaljenost centroida.
- Kadar so skupine večje, je možnih več permutacij, zato obstaja več, nekoliko različnih algoritmov.

8.6.4.2.4. Wardova metoda

- Kriterij izbire povezave je najmanjše povečanje vsote kvadriranih odstopanj od srednjih vrednosti skupin.
- V geologiji daje najboljše rezultate (običajno v kombinaciji z evklidsko razdaljo, kot mero podobnosti).



(c)

Ward's method

8.6.4.3. Dendrogram ali hierarhično drevo

- Dendrogram je enodimenzionalen v smislu, da je na eno os nanešen koeficient podobnosti, druga dimenzija pa le ureja niz objektov in topološki prikaz povezav z navpičnicami, ki združujejo ustrezna opazovanja/skupine glede na podobnost.
- Fenonska črta na levi strani dendrograma lahko loči od ene do n skupin.

8.6.4.4. Način Q in R

- Način Q pomeni, da rezultat združuje objekte (opazovanja, vrstice), na osnovi vrednosti spremenljivk.
- V načinu R združujemo spremenljivke, glede na njihove vrednosti v objektih,

8.6.4.5. Kritike clusterske analize

- Na vsakem od korakov moramo sprejemati odločitve, kar ruši objektivnost postopka.
- Objektivnost ponovno uvedemo, če so izbire nakjučne ali temeljijo na splošnih priporočilih, za katera se odločimo pred analizo in potem tudi sprejmemo rezultate!

8.6.4.5. Kritike clusterske analize

- Rezultatov analize ni enostavno testirati:
  - Ne moremo uporabiti običajnih testovza značilnosti razlik med vzorci, ker celo pri popolnoma naključnih podatkih povezovalni algoritmi zagotavljajo popolnoma ločene skupine, ki se ne prekrivajo in bodo zato vedno značilno različne.
  - Ni jasno, ali naj neko skupino ocenimo kot značilno ali pa je pomembna potencialna nenaključnost vključena v celotni strukturi dendrograma.