4. 1 STANDARDIZACIJA

Različne spremenljivke, ki jih želimo med seboj primerjati, moramo najprej prevesti na nek skupni imenovalec. Ko spremenljivke standardiziramo, lahko med seboj primerjamo standardizirane vrednosti različnih spremenljivk. V ta namen lahko uporabimo različne metode, vendar je najpogostejša med njimi naslednja: $z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$. Standardizirana vrednost je torej relativni odklon od aritmetične sredine. Poglavitna lastnost standardiziranih spremenljivk je, da imajo aritmetično sredino vedno enako nič, standardni odklon pa ena.

4. 2 MERE PODOBNOSTI IN MERE RAZLIČNOSTI

Mera podobnosti je preslikava, ki vsakemu paru enot (X,Y), priredi neko realno število $s:(X,Y)\to R$.
Mera različnosti je prav tako preslikava ki vsakemu paru enot (X,Y), priredi neko realno število $d:(X,Y)\to R$. Za uspešno razvrstitev mora mera različnosti zadoščati naslednjim pogojem:
nenegativnost; $d(X,Y)\geq 0$
$d(X,Y)=0$
simetričnost; $d(X,Y)=d(Y,X)$
Lahko pa zadošča še dvema dodatnima pogojema:
razločljivost; $d(X,Y)=0 \Rightarrow$ X=Y
trikotniška neenakost; $\forall Z: d(X,Y)\leq d(X,Y)+d(Z,Y)$
Če mera različnosti zadošča vsem petim pogojem, ji pravimo razdalja.
Z mero podobnosti ali mero različnosti pravzaprav urejamo množico neurejenih parov enot. Na podlagi urejenosti parov enot definiramo pojem enakovrednosti mer podobnosti oz. različnosti. Pravimo, da sta dve meri podobnosti ali različnosti enaki, če je urejenost parov enot, dobljena s prvo mero, enaka urejenosti parov enot z drugo mero podobnosti. Enako velja za mere različnosti.
Katero mero podobnosti oz. različnosti bomo izbrali, je odvisno od tipa podatkov, ki jih imamo. Za številske podatke se za mero različnosti najpogosteje uporablja evklidska razdalja. Za dve enoti (X,Y), kateri opisujejo številske spremenljivke, je evklidska razdalja definirana takole: $d(X,Y)=\sqrt{\sum_{i=1}^{m}(x_i-y_i)^2}$.
Vendar kot đže rečeno, obstajajo tudi druge razdalje kot na primer razdalja Manhattan, razdalja Minkowskega in razdalja Čebiševa.

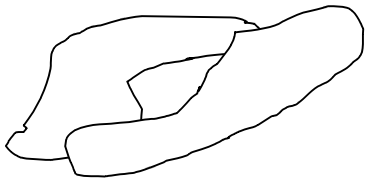
4. 3 RAZVRŠČANJE V SKUPINE

Z razvrščanjem v skupine uspemo večrazsežnosti prostor preslikamo v enorazsežnosti prostor, s čimer proučevano populacijo lahko predstavimo kot nekaj skupin med seboj podobnih objektov, kateri pa morajo biti med skupinami čimbolj različni. Objekt predstavlja neko živo bitje v vsej svoji kompleksnosti, enota pa je analitična definicija objekta oz. opis objekta. Množico iskanih skupin imenujemo razvrstitev. Temeljni koraki pri postopku razvrščanja v skupine so:
- izbira objektov;
- določitev množice spremenljivk, ki določajo enote;

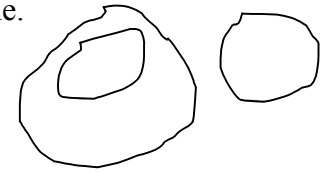
- računanje podobnosti med enotami;
- uporaba ustrezne metode razvrščanja v skupine;
- ocenitev dobljene rešitve.

4. 4 METODE HIERARHIČNEGA ZDRUŽEVANJA V SKUPINE

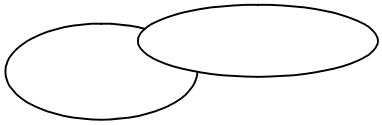
Največ metod združevanja v skupine temelji na zaporednem združevanju dveh ali več skupin v novo skupino. In med temi metodami so najpogostejše tiste, ki združijo vsakič po dve skupini. Metode združevanja v skupine se med seboj razlikujejo po tem, kako določajo mere različnosti d med novo skupino Cr in ostalimi skupinami. Poznamo Minimalno metodo, Maksimalno metodo, McQuittyjevo metodo, Povprečno metodo, Gowerjevo metodo, Wardovo metodo ipd. Nas bodo zanimale predvsem Minimalna, Maksimalna in pa Wardova metoda.
Minimalna metoda išče največjo povezanost oz. najmanjšo različnost med enotama ene in druge skupine. Zato je zlasti pomembna pri prepoznavanju dolgih »klobasastih« struktur in tudi neeliptičnih struktur.



Vendar pa je minimalna metoda neuporabna pri izrazito ločenih strukturah, saj se pojavi efekt veriženja. Takšen verižni učinek se na primer pojavlja pri tipično prekrivajočih se skupinah, ko minimalna metoda zazna dolgo, »klobasasto« strukturo, dejansko pa gre za dve prekrivajoči se skupini.
Maksimalna metoda znotraj skupin išče povezanost med objekti, torej ravno obratno kot minimalna metoda, zato dobro razkriva predvsem okroglaste skupine.



Wardova metoda ima vgrajeno strukturo elipsoida, zato dobro razkriva eliptične strukture. Elipsoid je zelo fleksibilna struktura, zato Wardova metoda daje zelo jasne in čiste slike razvrstitev v skupine.



V splošnem so metode hierarhičnega združevanja zelo priljubljene zaradi naslednjih značilnosti:
postopek je relativno preprost;
rezultat združevanja je mogoče nazorno prikazati z drevesom združevanja (dendogramom);
v splošnem postopek zahteva relativno malo računalniškega časa;
uporabniku ni potrebno vnaprej določiti števila skupin, kar je sploh najbolj cenjena lastnost metod nehierarhičnega združevanja v skupine.
Imajo pa tudi svojo slabo lastnost, so omejene s številom enot.

4. 5 DREVO ZDRUŽEVANJA V SKUPINE

Drevo združevanja v skupine ali dendogram je gafični prikaz potka združevanja po metodah hierarhičnega razvrščanja. Enote prestavljajo listi tega drevesa, točke združitve pa so sestavljene skupine: levi in desni naslednik vsake točke sta skupini, iz katerih je nova skupina nastala. Višina točke ali nivo združevanja je sorazmerna meri različnosti med skupinama.

4. 6 METODA VODITELJEV

Metoda voditeljev sodi med nehierarhične metode združevanja v skupine. Prednost te metode je v tem, da zmore razvrščati v skupine večje število enot; njena slabost pa je, da je potrebno vnaprej določiti, v koliko skupin razvrščamo enote.
Metoda voditeljev uporablja evklidsko razdaljo. Na začetku imamo množico vnaprej podanih predstavnikov posameznih skupin, to so voditelji. Metoda voditeljev nato priredi enote najbližjim voditeljem in poišče centroide, težišča tako dobljenih sredin. Centroidi so v bistvu aritmetične sredine. Ti centroidi nato postanejo novi voditelji in tem novim voditeljem so zopet prirejene najbližje enote. Postopek se konča, ko se nova množica voditeljev ne razlikuje od množice voditeljev, dobljene korak pred njo.
Začetno množico voditeljev lahko določamo na različne načine. Najpreprosteje je, če so določeni slučajno. Drugače pa lahko voditrelje tudi maksimalno razpršimo med proučevanimi enotami, in sicer tako, da za prvega voditelja izberemo enoto, ki je v središču vseh enot, za drugega najoddaljenejšo enoto od prvih dveh voditeljev itd. Najbolj učinovito pa je, če pred voditeljev, temeljito preučimo podatke in domneve o strukturi proučevanih pojavov.
Tudi pri metodi voditeljev skušamo dobiti čim boljšo razvrstitev, kar storimo tako, da postopek ponovimo večkrat z različnimi začetnimi množicami voditeljev. Ustreznost razvrstitev merimo z Wardovo kriterijsko funkcijo, kjer je d kvadrat evklidske razdalje. Wardova kriterijska funkcija ima namreč to lastnost, da monotono pada.

4. 7 HEVRISTIKA METOD ZDRUŽEVANJA V SKUPINE

Hevristika metod združevanja v skupine se kaže v učinku požrešnosti teh metod. Vemo da postopek združevanja poteka tako, da se na vsakem koraku združi dve najbližji skupini. Pri manjšem številu skupin, torej v naslednjih korakih, pa se lahko izkaže, da bi bilo bolj, če bi v prejšnjih korakih združevali drugače, vendar poti nazaj ni. Učinek »požrešnosti« metod združevanja se tako manj pozna na nižjih nivojih združevanja in bolj pri višjih. Torej so razvrstitve dobljene z rezanjem drevesa združevanja na višjih nivojih v splošnem manj zanesljive.
Kadar imamo izrazite strukture, do učinka »požrešnosti« ne pride, vendar pa v to ne moremo zagotovo prepričani, saj ne poznamo strukture podatkov. Zato je najbolje, če podatke analiziramo še z drugimi nehierarhičnimi metodami in tako preverimo stabilnost dobljene razvrstitve.

4. 8 KRITERIJSKA FUNKCIJA

S pomočjo kriterijske funkcije ocenjujemo primernost različnih razvrstitev. Torej: če imamo množico razvrstitev in za vsako razvrstitev izračunamo vrednost kriterijske funkcije, je najboljša tista razvrstitev, ki ima najmanjšo vrednost kriterijske funkcije.

Kriterijska funkcija je določena takole: $P(C) = \sum_{C \in C} \sum_{X \in C} d(X, T_C)$,kjer je Tc = $\left(\overline{X}_{C,} \overline{Y}_C\right)$ težišče skupine C ter različnost d evklidska razdalja.

Metoda glavnih komponent

Metoda glavnih komponent je ena najpogosteje uporabljenih multivariatnih metod. Osnoval jo je Karl Pearson (1901). Osnovna zamisel metode je opisati razpršenost *n* enot v *m* razsežnem prostoru (določen z *m* merjenimi spremenljivkami) z množico nekoreliranih spremenljivk – komponent, ki so linearne kombinacije originalnih merjenih spremenljivk. Nove spremenljivke so urejene od najpomembnejše do najmanj pomembne, kjer pomembnost pomeni, da prva glavna komponenta pojasnjuje kar največ razpršenosti osnovnih podatkov. Običajni cilj te analize je poiskati nekaj prvih komponent, ki pojasnjujejo večji del razpršenosti analiziranih podatkov. Analiza glavnih komponent omogoča povzeti podatke s čim manjšo izgubo informacij tako, da zmanjša razsežnost podatkov.

Osnovna misel metode glavnih komponent je, da želimo poiskati take linearne kombinacije opazovanih spremenljivk, da kar se da močno korelirajo z opazovanimi spremenljivkami, oziroma pojasnijo kar se da veliko razpršenosti opazovanih spremenljivk. Zato pri metodi glavnih komponent določimo uteži pri linearni kombinaciji spremenljivk, tako da je varianca te linearne kombinacije največja. Torej želimo poiskati take uteži a_1, za katere bo varianca Y_1 največja:
var (Y_1) = var (Xa_1) = max

Ko izračunamo prvo komponento (Y_1) z največjo varianco, poiščemo drugo komponento tako, da je nekorelirana s prvo in ima zopet največjo varianco. Postopek ponavljamo do zadnje komponente.

Scree diagram

Obstaja več metod kako določimo najprimernejše število komponent. Ena najpogosteje uporabljenih je grafična predstavitev lastnih vrednosti. Na abscisno os nanesemo število komponent, na ordinatno pa lastne vrednosti posameznih komponent. Dobljene točke povežemo z ravnimi črtami. Na mestu kjer prihaja do loma grafa leži sugestija za število komponent. Tak graf imenujemo 'scree' diagram.

V literatiri obstajajo še nekatera druga pravila za določitev števila najpomembnejših komponent:

- izbrano število komponent naj pojasni vsaj 80% skupne variance;
- lastne vrednosti komponent naj bodo večje kot povprečne vrednosti lastnih vrednosti;
- odstotek pojasnjene variance zadnje vzete komponente naj bo vsaj 5.

Lastne vrednosti in lastni vektorji

Lastne vrednosti so variance komponent, lastni vektorji pa predstavljajo uteži. Osnova za pridobivanje uteži je korelacijska matrika. Lastni vektorji (λ) predstavljajo uteži in so pravokotni med seboj. Vsota diagonalnih členov matrike je enaka vsoti lastnih vrednosti matrike. Delež pojasnjene variance z j-to glavno komponento je $\lambda_{j/m}$.

4.OPIS UPORABLJENIH METOD

Pri metodi multiple regresije sem uporabila različne statistične metode:

4.1. REGRESIJA - UVOD

Regresijska analiza je statistična metoda, ki nam pomaga analizirati odnos med odvisno spremenljivko (LIKERT) ter eno ali več neodvisnimi spremenljivkami (SPOL, DRŽAVA, NAČRTIZO, IZOOČE, IZOMATI, VEROBRED, VEROBČUT). Raziskovalec najprej postavi teoretične predpostavke o odnosih med spremenljivkami, pravimo, da postavi regresijski model. Ta model nato raziskovalec testira na določenem vzorcu. S pomočjo regresijske analize oceni parametre regresijskega modela in statistični pomen tega modela. Poleg te, opisane vloge, pa ima regresijska analiza še napovedovalno vlogo. Ko smo namreč določen regresijski model sprejeli in ocenili njegove parametre, lahko iz vrednosti neodvisnih spremenljivk napovemo vrednost odvisne spremenljivke.

Pri samem postavljanju modela naleti raziskovalec na cel kup problemov. Smiselno mora izbrati odvisno spremenljivko in poiskati vse tiste neodvisne spremenljivke, ki nanjo pomembno vplivajo. Poleg tega mora dobro zadeti vrsto odvisnosti med temi spremenljivkami. Naloga same regresijske analize se torej omejuje predvsem na testiranje določenih predpostavk o modelu in tako šele posredno pomaga raziskovalcu pri razčiščevanju postavitve njegovega modela.

4.2. KONSTRUKCIJA LIKARTOVE LESTVICE

Likertovo lestvico konstruiramo iz večjega števila spremenljivk, ki nam vse merijo isti koncept. Merjeni koncept želimo zaobjeti v eno samo spremenljivko, ki jo konstruiramo na naslednji način: vrednosti vseh spremenljivk, ki jih vključimo v konstrukcijo nove spremenljivke seštejemo ter dobljeno vsoto delimo s številom spremenljivk. Pri tem pa je seveda potrebno paziti, da so vse spremenljivke "obrnjene v isto smer", oz. da vrednosti posamezne spremenljivke pri vsaki spremenljivki pomenijo oz. na enak način merijo isto. Nova (konstruirana) spremenljivka je intervalnega tipa.

4.3. MULTIPLA REGRESIJA

Regresijska funkcija $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m + E$ nam kaže, kakšen bi bil vpliv spremenljivk $X_1, X_2, \ldots X_m$ na spremenljivko Y, če razen teh vplivov ne bi bilo nobenih drugih vplivov. Spremenljivke $X_1, X_2, \ldots X_m$ so neodvisne spremenljivke, spremenljivka Y je odvisna spremenljivka, E pa je člen napake, ki mu včasih rečemo tudi motnja ali disturbanca. Za i-to enoto seveda velja: $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_m x_{mi} + e_i$. Regresijski model pa lahko zapišemo tudi matrično: Y = Xß + E, pri čimer je:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & \vdots & x_{1m} \\ 1 & x_{21} & \vdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \vdots & x_{nm} \end{bmatrix} \qquad \beta = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{bmatrix} \qquad E = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Ko zgradimo regresijski model, seveda nastopi vprašanje, koliko je regresijski model prilagojen podatkom. Če vemo, da za i-to enoto velja:
$y_i = y_i' + e_i$, pri čemer je y_i prava vrednost, y_i' regresijska ocena, e_i pa razlika, potem mora (po metodi najmanjših kvadratov) veljati:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - y_i')^2 = min$$

da bo regresijski model čimbolj točen.
Velja torej:

$$\sum_{i=1}^{N} (y_i - \bar{y})^2 = \sum_{i=1}^{N} (y_i' - \bar{y})^2 + \sum_{i=1}^{N} e_i^2$$

oz. celotna varianca (SST) = pojasnjena varianca (SSR) + nepojasnjena varianca (SSE).

Vektorsko lahko to zapišemo takole:

$$(Y - \bar{Y})'(Y - \bar{Y}) = (Y' - \bar{Y})'(Y' - \bar{Y}) + E'E$$

Determinacijski koeficient, ali kvadrat multiplega koeficienta korelacije, ki nam pove odstotek pojasnjene variance, analitično nato izračunamo takole:

$$R^2 = 1 - \frac{E'E}{(Y - \bar{Y})' - (Y - \bar{Y})}$$

Ker je pri determinacijskem koeficientu števec odvisen od števila neodvisnih spremenljivk, ga je potrebno popraviti:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1}$$

Celotni regresijski model testiramo z F-testom:

$$F = \frac{SSR / m}{SSE / (n-m-1)} = \frac{R^2 / m}{(1 - R^2) / (n-m-1)}$$

Manjša kot je statistična značilnost F-statistike, boljši je regresijski model.

4. OPIS UPORABLJENIH METOD

Pri faktorski metodi sem uporabila različne statistične metode:

FAKTORSKA ANALIZA

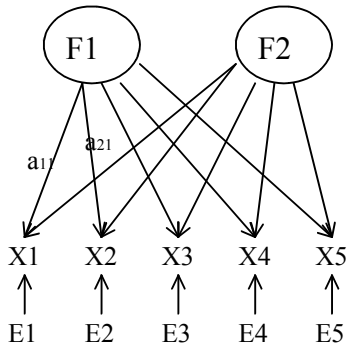
Faktorska analiza je ena izmed metod za redukcijo podatkov. Pri faktorski analizi gre za študijo povezav med spremenljivkami, in sicer tako, da poizkušamo najti novo množico spremenljivk, ki predstavljajo to, kar je skupnega opazovanim spremenljivkam. Množica novih spremenljivk mora biti seveda manjša od množice merjenih spremenljivk. Z drugimi besedami: faktorska analiza poizkuša poenostaviti kompleksnost povezav med množico opazovanih spremenljivk z razkritjem skupnih razsežnosti ali *faktorjev*, ki omogočajo vpogled v osnovno strukturo podatkov. Metoda je uporabna v vseh tistih primerih, ko zaradi različnih vzrokov neposredno merjenje neke spremenljivke ni možno. V tem primeru merimo samo indikatorje pojma oz. konstrukta, ki ga neposredno ne moremo meriti. S faktorsko analizo nato ugotovimo, ali so zveze med opazovanimi spremenljivkami (ali indikatorji) pojasnjlive z manjšim številom posredno opazovanih spremenljivk ali faktorjev.

SPLOŠNI FAKTORSKI MODEL

Osnova faktorskega modela je domneva, da med spremenljivkami X_i (i = 1, ..., m), F_r (r = 1, ..., k) in E_i (i = 1, ..., m) velja zveza:

; i = 1, ..., m in k < m

pri čemer so X_i merjene spremenljivke, F_r skupni faktorji, E_i pa specifični faktor, ki vpliva samo na X_i, a_{ir} pa je faktorska utež, ki kaže na vpliv faktorja F_r na X_i.



V matrični obliki splošni faktorski model zapišemo takole:
X = F A' + E

Na osnovi naslednjih predpostavk splošnega faktorskega modela:
specifični faktorji so pravokotni med seboj
(cov (E_i, E_r) = 0, če velja i ≠ j)
vsak specifični faktor E_i je pravokoten na vsak skupni faktor F_j
(cov (E_i, F_j) = 0, za vsak i in j)
skupni faktorji so pravokotni med seboj
(cov (F_i, F_r) = 0, če velja i ≠ j)
spremenljivke X_i, F_r in E_i naj bodo centrirane
(E(X_i) = E(F_r) = E(E_i) = 0)
lahko izpeljemo naslednjo faktorsko enačbo:
Σ = A A' + ψ
Enačbo lahko zapišemo tudi drugače:
$\sigma^2 = \sum_{j=1}^{k} a_{ij}^2 + \Psi_{ii}$, pri čemer je $\sum_{j=1}^{k} a_{ij}^2$ varianca skupnih faktorjev, Ψ_{ii} pa varianca specifičnih faktorjev (slednja mora biti seveda čim manjša).

S tem smo varianco merjene spremenljivke X_i razbili na del, ki je pojasnjen s skupnimi faktorji in na specifično varianco. Delež variance, ki je pojasnjena s skupnimi faktorji, imenujemo tudi ***komunaliteta***, označujemo pa jo z h_i^2

$\sum_{j=1}^{k} a_{ij}^2 = h_i^2$

V prvem delu faktorske analize moramo najprej izračunati neznane parametre faktorskega modela: faktorske uteži A in specifične variance ψ. Pred tem pa je potrebno preveriti:
identifikabilnost faktorskega modela (ugotoviti moramo, ali faktorske uteži A in specifične faktorje ψ sploh lahko ocenimo)
Potreben (ne pa tudi zadosten) pogoj za identifikacijo faktorskega modela je:

, pri čemer je *m* število spremenljivk vključenih v faktorski model, *k* pa število faktorjev.

Če ta pogoj ni izpolnjen, je model *prefaktoriziran*, kar pomeni, da imamo faktorje, ki že opisujejo merske napake. Sum na prefaktorizacijo nastopi takrat, ko npr. korelacijski koeficienti padejo iz intervala [-1, 1] ali ko se pojavi negativna varianca.

enoličnost ocen parametrov (ali lahko te parametre ocenimo enolično - z eno samo oceno) Pri enoličnosti pa nastopi problem, da se parametra A sploh ne da enolično izračunati. Zato računamo v dveh korakih: najprej izračunamo ψ (zakoličimo skupni prostor - ocenimo komunalitete), nato pa na podlagi tega izračunamo A. Postopek ponovimo večkrat, dokler model ne skonvergira.

Pri tem delu faktorske analize lahko uporabimo več različnih metod. Kratek opis štirih najbolj pogostih, ki sem jih tudi uporabila v svoji nalogi, podajam v spodnji tabeli:

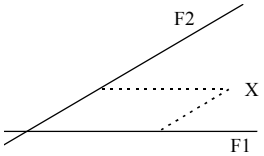
IME METODE	**OSNOVNI PRINCIP**	**OCENA KOMUNALITETE**
metoda glavnih osi	maksimizira varianco skupnih faktorjev	več načinov, kvadrat koeficienta multiple korelacije R^2
image	vsako spremenljivko regresira z ostalimi	iterativno
metoda najmanjšega verjetja	poišče najboljšo oceno za reprodukcijo variančno kovariančne-matrike Σ	iterativno
alfa	maksimizira generabilnost faktorjev	iterativno

ROTACIJE

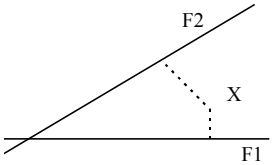
Drugi korak faktorske analize je rotacija. S pomočjo rotacije prečistimo strukturo. Bistvo rotiranja namreč je, da dobimo teoretično pomembne faktorje in čim enostavnejšo faktorsko strukturo. Če namreč dobljene rešitve ne moremo dobro interpretirati, lahko dobljeno rešitev v skupnem prostoru, ki je določen s skupnimi faktorji, transformiramo tako, da jo zarotiramo. Matematično to pomeni, da matriko A pomnožimo z transformacijsko matriko M:
$A^* = A\,M$.
Rešitev A* enako dobro reproducira originalne podatke kot prvotna rešitev A.

Za rotacijo se odločamo predvsem takrat, ko skupnih faktorjev ne moremo smiselno interpretirati - če so npr. projekcije iste spremenljivke precejšnje na več faktorjih, ali pa če imamo *splošen faktor* (projekcije vseh spremenljivk na prvi faktor so precejšnje). Ločimo dve vrsti rotacij:

pravokotne, kjer so rotirani faktorji neodvisni med seboj (znana metoda je npr. *varimax*, ki maksimizira varianco kvadratov uteži v vsakem faktorju in s tem poenostavlja strukturo po stolpcih);
poševne, kjer so rotirani faktorji odvisni med seboj, med njimi ni pravega kota in faktorji med seboj korelirajo (pri tem sem uporabila metodo *oblimin*). V primeru poševnih rotacij lahko spremenljivke (točke v poševnem koordinatnem sistemu) projiciramo na poševne faktorje na dva načina:
vzporedno, pri čemer dobimo *pattern uteži*, ki so parcialni koeficienti korelacije med spremenljivko in faktorjem in predstavljajo "suhi vpliv" spremenljivke na faktor;



pravokotno, s čemer dobimo *strukturne uteži*, ki so navadni koeficienti korelacije med spremenljivko in faktorjem.



V primeru pravokotnih faktorjev so *pattern* in *strukturne* uteži seveda enake.

Vsebinsko so poševne rotacije boljše, v praksi pa nastopijo problemi s kriterijsko funkcijo. Zato ponavadi najprej naredimo poševno rotacijo, nato pa pogledamo, kakšne so korelacije med faktorji. Če so korelacije med faktorji majhne (manj od 0,20), naredimo pravokotno rotacijo, sicer pa ne, saj bi bila v slednjem primeru struktura preveč vsiljena.

DOLOČITEV FAKTORSKIH VREDNOSTI NA ENOTAH

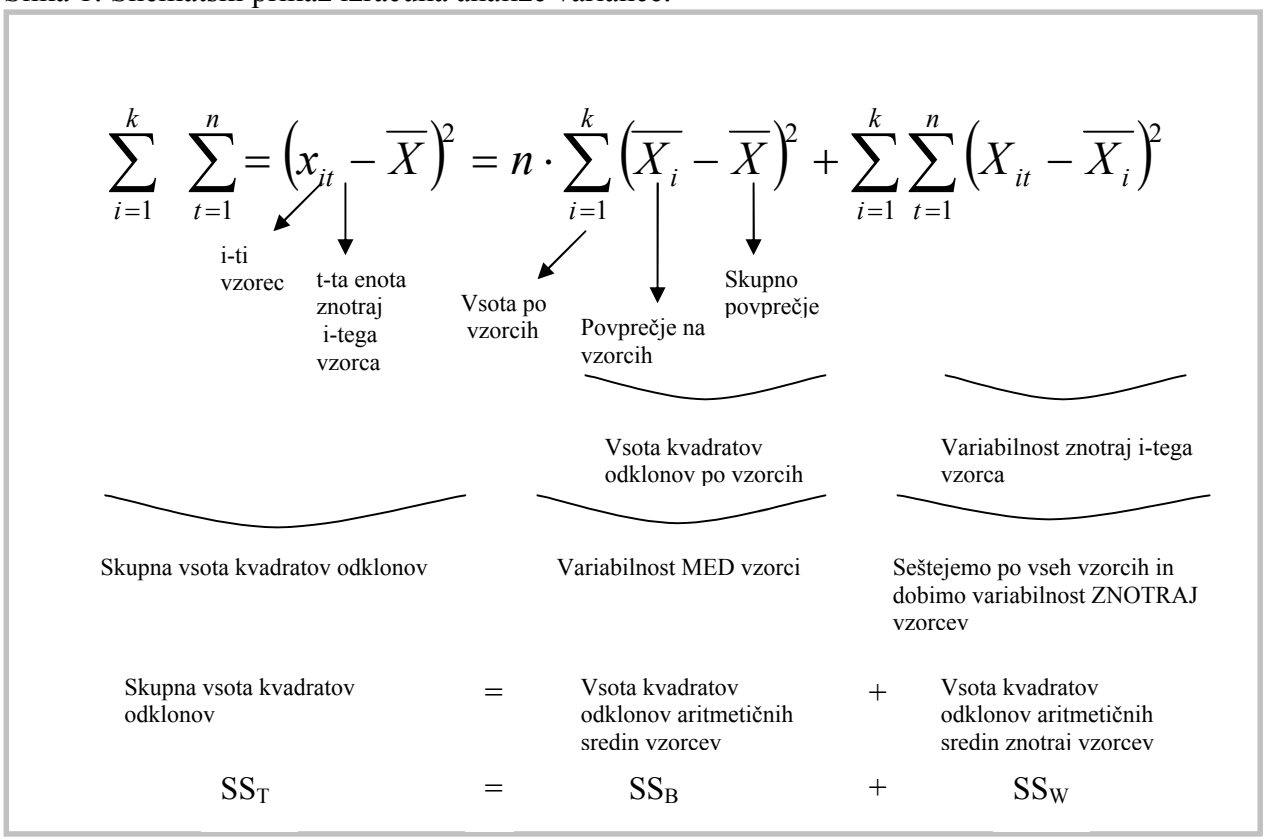
V tretji, zadnji fazi faktorske analize določimo še faktorske uteži na posameznih enotah. Eden izmed načinov za določitev teh uteži je regresijska ocena faktorske vrednosti. Pri tem dobimo ocenjeno faktorsko vrednost $\hat{F}$ in ne prave faktorske vrednosti F. Korelacije med $\hat{F}$ zato ne bodo take kot med F, lahko pa se spremeni tudi smer faktorja $\hat{F}$.

4. 0 OPIS UPORABLJENIH METOD

4. 1 ANALIZA VARIANCE

Pri analizi variance ugotavljamo, kako dobro lahko merjene spremenljivke ločimo med seboj. Ker variabilnost oziroma različnost merimo z varianco, se postopek imenuje analiza variance.

Slika 1: Shematski prikaz izračuna analize variance.

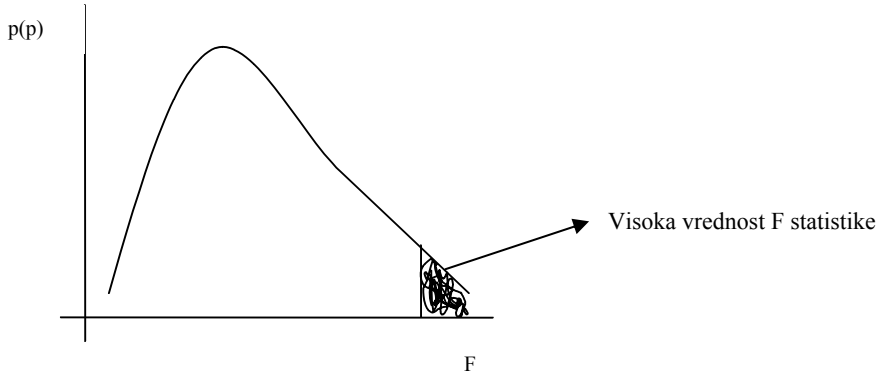


Varianco MED vzorci (skupinami) izračunamo tako: $\frac{SS_B}{k-1} = n\cdot s_{\overline{x}}^2 = \frac{n\cdot\sum_{i=1}^{k}\left(\overline{X_i}-\overline{X}\right)^2}{k-1}$.

Varianca ZNOTRAJ vzorcev (skupin) pa je enaka: $\frac{SS_W}{k\cdot(n-1)} = s_W^2 = \frac{\sum_{i=1}^{k}\sum_{t=1}^{n}\left(x_{it}-\overline{X_i}\right)^2}{k\cdot(n-1)}$

Za testno statistiko vzamemo F statistiko, ki se porazdeljuje po F porazdelitvi z dvema prostostnima stopnjama; F(m_1 = k-1; m_2 = k (n-1)). S pomočjo F statistike skušamo zavreči ničelno domnevo, in sicer da se aritmetične sredine po vzorcih (skupinah) ne razlikujejo. Če je vrednost F statistike visoka sledi, da so razlike med povprečji in variabilnosti med vzorci visoke, torej ničelno domnevo lahko zavrnemo.

Slika 2: F porazdelitev.



4. 2 DISKRIMINANTNA ANALIZA

Diskriminantna analiza je posplošitev analize variance. Razlike med analizo variance in diskriminantno analizo je v tem, da imamo pri analizi variance le eno skupino oziroma le en vzorec, pri diskriminantni analizi pa imamo vsaj dva vzorca (skupini). Drugače pa imamo povsod več merjenih spremenljivk.
Z diskriminantno analizo tako ugotavljamo, kako dobro lahko naše merjene spremenljivke ločijo skupine med seboj. Torej iščemo razlike med skupinami. Cilj je poiskati takšno linearno kombinacijo merjenih spremenljivk, da bo nova spremenljivka ločila vnaprej določene skupine tako, da bo napaka pri uvrščanju enot v skupine najmanjša.

4. 3 PREDPOSTAVKE DISKRIMINANTNE ANALIZE

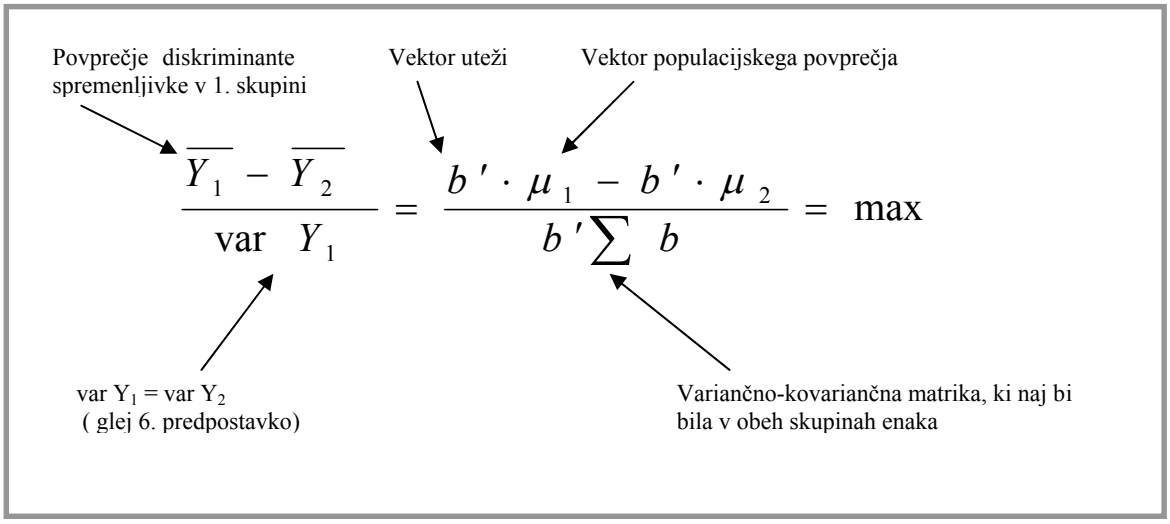
Imeti moramo vsaj dve skupini ($k \geq 2$).
V vsaki skupini morata biti vsaj dve enoti.
Število spremenljivk mora biti približno trikrat manjše kot število enot (p< n-2; n=število enot v vzrocu).
Spremenljivke moraja biti vsaj intervalnega tipa.
Nobena spremenljivka ne sme biti linearna kombinacija preostalih spremenljivk, sicer nastopi problem multikolinearnosti. V primeru multikolinearnosti je determinanta korelacijske matrike enaka nič in tako ne moremo naredit inverzne matrike.
Variančno-kovariančna matrika je za vsako skupino enot (približno) enaka.
Pri statističnem ocenjevanju se predpostavlja, da so v vsaki skupini enot spremenljivke dobljene iz populacije z večrazsežno normalno porazdelitvijo spremenljivk.

4. 4 DISKRIMINANTNA ANALIZA V PRIMERU DVEH SKUPIN

Diskriminantna spremenljivka oziroma funkcija (Y) je linearna kombinacija merjenih spremenljivk: $Y = b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p = X\,b$; p – število spremenljivk in b – uteži.

Definirana je tako, da je kvocient razlik povprečij diskriminantne spremenljivke v obeh skupinah, glede na varianco diskriminantne spremenljivke v skupini maksimalen:

Slika 3: Diskriminantna analiza.



Iz danih vzorčnih podatkov ponavadi lahko ocenimo μ_i in $\sum_i$:

$\overline{X}_i = \left(\overline{X}_{i1}, \overline{X}_{i2}, ..., \overline{X}_{ip}\right)$

$S = \frac{1}{n_1 + n_2 - 2}\left(X_1' \cdot X_1 + X_2' \cdot X_2\right)$ - obtežena variančna-kovariančna matrika.

Ob tem velja omeniti, da če so spremenljivke standardizirane, je variančno-kovariančna matrika kar korelacijska matrika.

Oceno uteži torej dobimo tako: $\hat{b} = S^{-1}\left(\overline{X}_1 - \overline{X}_2\right)$

Izračunana diskriminantna spremenljivka na koncu je: Y = X b.

4. 5 PRAVILA UVRŠČANJA ENOT V SKUPINE

Enote lahko v optimalne skupine uvrščamo po naslednjem pravilu: $Y_k - \overline{Y}_1 \le Y_k - \overline{Y}_2$. To pomeni, da k-to enoto uvrstimo v prvo skupino, če je razlika med k-to enoto in povprečjem prve skupine manjša kot razlika med k-to enoto in povprečjem druge skupine.

Sicer pa lahko za uvrščanje enot v optimalne skupine uporabimo tudi metodo srednje točke oziroma točko ločevanja skupin.

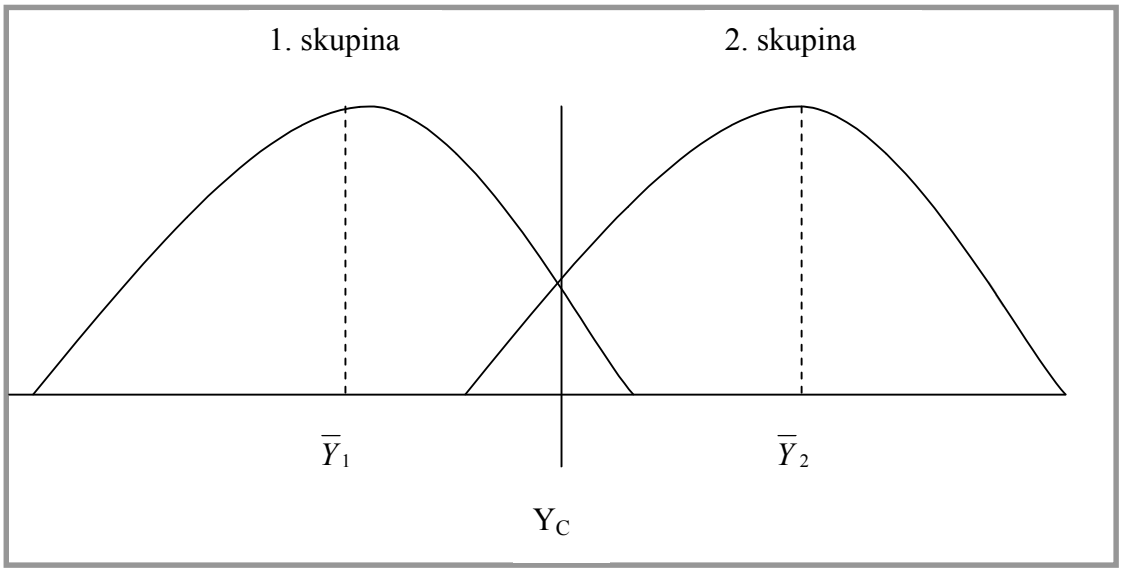
Če imamo dve enako veliki skupini oziroma vzorca točko ločevanja izračunamo po naslednji formuli: $Y_C = \frac{\overline{Y}_1 + \overline{Y}_2}{2} = \frac{\hat{b}' \cdot \overline{X}_1 + \hat{b}' \cdot \overline{X}_2}{2} = \frac{\left(\overline{X}_1 - \overline{X}_2\right)' S^{-1}\left(\overline{X}_1 + \overline{X}_2\right)}{2}$.

Za dve različno veliki skupini pa točko ločevanja izračunamo po formuli:

$Y_C^* = \frac{n_2 \cdot \overline{Y}_1 + n_1 \cdot \overline{Y}_2}{n_1 + n_2}$.

Ob uvrščanju enot v skupini, po metodi srednje točke upoštevamo pravilo: če ima i-ta enota vrednost diskriminantne spremenljivke manjšo od Y_C^* jo uvrstimo v prvo skupino, drugače pa v drugo.

Slika 4: Centroidi.



4. 6 KLASIFIKACIJSKA TABELA

Klasifikacijska tabela nam pove, kako dobro zmorejo merjene spremenljivke ločevati skupine. Ko izračunamo diskriminantno spremenljivko, vsako enoto ponovno uvrstimo v eno od obeh skupin. S klasifikacijski tabelo nato prikažemo enote glede na njihovo dejansko (a priori) in izračunano (aposteriori) uvrstitev.

Slika 5: Klasifikacijska tabela.

		Izračunana uvrstitev		
Dejanska uvrstitev (znani podatki)		1. skupina	2. skupina	
	1. skupina	a	b	n_1
	2. skupina	c	d	n_2
		m_1	m_2	

V klasifikacijski tabeli sta c in b napaki. Odstotek pravilno uvrščenih enot torej izračunamo tako: $\frac{a+d}{n_1+n_2}\cdot 100$.

Spodnja meja kvalitete razvrščanja je odstotek enot, ki bi bile pravilno razvrščene ob naključnem razvrščanju. V primeru dveh skupin je tako spodnja meja 50 % enot, v primeru treh pa 33,3 %.

4. 7 DISKRIMINANTNA ANALIZA NA VEČ SKUPINAH

Če želimo poiskati različnost med več skupinami (torej tremi ali več), razlike med skupinami popišemo z več diskriminantnimi spremenljivkami. Za dane podatke lahko izračunamo največ min(p, k-1) diskriminantnih spremenljivk; p – število spremenljivk in k – število skupin.