I *UVOD*

(1) **Multivarijaciona analiza** predstavlja skup statističkih metoda koje simultano analiziraju višedimenzionа merenja dobijena za svaku jedinicu posmatranja iz skupa objekata koji ispitujemo. Metode multivarijacone analize se dele na (2) i (3).

(2) **Metode zavisnosti**:

a) ***Multivarijaciona regresija*** – 2 slučaja, prvi kada se bavimo anlizom zavisnosti jedne promenljive od skupa drugih (metod višestruke regresije) i drugi kada skup zavisnih prom. sadrži više od jednog člana. Zadatak je oceniti ili predvideti sr.vrednosti zavisne (zavisnih) prom. na bazi poznatih vrednosti nezavisnih promenljivih.

b) ***Kanonička korelaciona analiza*** – uopštava predhodnu metodu, i njome se uspostavlja linearna zavisnost između skupova zavisnih i nezavisnih promenljivih.

c) ***Diskriminaciona analiza*** – bavi se problemom razdvajanja grupa i alokacijom opservacija u ranije definisane grupe. Omogućava identifikaciju promenljive koja je najviše doprinela razdvajanju grupa kao i predviđanje verov. da će objekat pripasti jednoj od grupa.

d) ***Multivarijaciona analiza varijanse*** – uticaj eksperimentalnih promenljivih na dve ili više zavisnih promenljivih.

e) ***Logit analiza*** – kod regresionog modela sa kvalitativnom zavisnom prom. (kada je u regresionom modelu zavisna promenljiva dihotomnog tipa) je zavisna promenljiva (tzv. logit funkcija) logaritam količnika verovatnoća da će dihotomna zavsna prom. uzeti jednu ili drugu vrednost.

(3) **Metode međusobne zavisnosti**

a) ***Analiza glavnih komponenti*** – metoda za redukciju većeg broja promenljivih na manji broj glavnih komponenti. Zadatak – konstruisanje linearne kombinacije glavnih komponenti tako da se obuhvati što veći iznos varijanse originalnog skupa prom.

b) ***Faktorska analiza*** – za razliku od analize g.k pretpostavlja postojanje odgovarajućeg statističkog modela kojim se originalna prom. iskazuje kao linearna kombinacija faktora plus greška modela, a zainteresovani smo za objašnjenje kovarijanse.

c) ***Analiza grupisanja*** – metoda za redukciju podataka koja je orjentisana ka redovima matrice podataka. Objekti se kombinuju u grupe relativno homogenih objekata.

d) ***Višedimenziono proporcionalno prikazivanje*** – metoda orjentisana ka objektima, koristii meru sličnosti odnosno razlike između njih u cilju njihovog prostornog prikazivanja.

e) ***Loglinearni modeli*** – omogućavaju ispitivanje međuzavisnosti kvalitativnih prom. koje formiraju višedimenzionalnu tabelu kontigencije. Kod tabela kontigencija logit f-ja se izražava preko ćelijskih frekvencija, za razliku od modela logističke regresije gde je iskazujemo preko skupa nezavisnih kval. ili kvan. promenljivih.

(4) **Vrste podataka i merne skale**

-***Skala odnosa*** - najčešće korišćena kod kvantitativnih promenljivih, osobine:
  a) količnik ma koje dve vrednosti ima smislenu interpretaciju
  b) rastojanje između dva objekta mereno na ma kom delu ove skale je jednako
  c) smisleno je rangirati od višeg ka nižim.

-***Intervalna skala*** - za razliku od skale odnosa, nema fiksni početak, poseduje osobine b i c.

-***Ordinarna skala*** – poseduje samo osobininu c, dakle samo rangiranje.

-***Nominalna skala*** – predstavlja najniži oblik merne skale, koristi kvalitativna obeležja. Nije omogućeno čak ni rangiranje.

(5) **Načini grafičkog prikaza multivarijacionih podataka** – zvezde, Andrewsove krive, Chernoffova lica.

(6) **Svojstva Andrewsove krive**:

a) Andr.kriva sredine multivarijacionih opservacija jednaka je sredini Andr.krivih pojedinih multivarijacionih opservacija;

b) odstojanje između dve Andr.krive proporcionalno je euklidskom odstojanju između odgovarajućih višedimenzionih tačaka.

II *VIŠEDIMENZIONI RASPOREDI (RASPODELE)*

(1) **Uslovni raspored** – raspored jednog podskupa slučajnih promenljivih za date vrednosti drugog podskupa. **Uslovna f-ja gustine verovatnoće od X_2 za dato $X_1 = x_1^*$** je:

$$f_{X_1|X_2}(x_2|X_1 = x_1^*) = \frac{f_X(x_1^*, x_2)}{f_{X_1}(x_1^*)}$$

(2) **Veza između kovarijacione i korelacione matrice**:

$$\rho = \left(D^{1/2}\right)^{-1}\Sigma\left(D^{1/2}\right)^{-1} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & & \rho_{2p} \\ \vdots & & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} tj, \Longrightarrow \Sigma = D^{1/2}\rho D^{1/2}$$

(3) **Mahalanobisovo odstojanje** – ako na mesto matrice koeficijenata $[a_{ij}]$ stavimo inverznu uzoračku kovarijacionu matricu. ***Svojstvo***: eliminiše efekat korelacije promenljivih. Kvadrat Mahalanobisovog odstojanja dat je izrazom:

$$d^2 = [x_1 - \overline{x_1} \quad x_2 - \overline{x_2}] \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \overline{x_1} \\ x_2 - \overline{x_2} \end{bmatrix}$$

(4) **Generalizovana varijansa** – dve definicije – prva: ona je u uzorku jednaka determinanti uzoračke kovarijacione matrice $|S|$ ili druga: jednaka je tragu uzoračke kovarijacione matrice (zbiru elemenata na glavnoj dijagonali $tr(S)$).

(5) **Koeficijent korelacije**: $r_{jk} = \cos(\Theta_{jk})$ – na osnovu čega zaključujemo da je r jednak nuli ukoliko su vektori odstupanja međusobno ortogonalni, a jednak jedinici ako zaklapaju mali ugao, ako leže na istom pravcu ali su suprotno orjentisani onda je jednak -1.

(6) **Statističko odstojanje** (ono koje uzima u obzir varijansu i kovarijansu promenljivih) ma koje tačke uzorka od težišta roja tačaka predstavlja euklidsko odstojanje računato na osnovu standardizovanih koordinata tačaka u novom koordinatnom sistemu:

$$d = \sqrt{\left(\frac{x_1^* - \overline{x_1^*}}{\sqrt{S_{11}^*}}\right)^2 + \left(\frac{x_2^* - \overline{x_2^*}}{\sqrt{S_{22}^*}}\right)^2}$$

Može se pokazati da *kvadrat odstojanja iskazan preko originalnih koordinata* ima formu:

$$d^2 = a_{11}(x_1 - \overline{x_1})^2 + 2a_{12}(x_1 - \overline{x_1})(x_2 - \overline{x_2}) + a_{22}(x_2 - \overline{x_2})^2$$

(7) Kovarijaciona matrica se može iskazati i kao **očekivana vrednost slučajne matrice**, koja se dobija kao proizvod slučajnih vektora odstupanja od sredine: $E[(X - \mu)(X - \mu)^T] = \Sigma$

(8) **Koeficijent korelacije između dve slučajne promenljive X_j i X_k:**

$$\rho = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}}$$

(9) **Uzoračka kovarijaciona matrica** je simetrična matrica (pxp):

$$S = \frac{1}{n-1} X^{*T} X^{*} = \frac{1}{n-1}(X - \bar{X})(X - \bar{X})^T$$

(10) **Zapremina elipsoida** za $p = 2$ je u relaciji sa **generalizovanom varijansom**:

$$|S| = c^{-2p} K(p) * V^2_{elipsoida}$$

(11) **Nedostatak generalizovane varijanse kao sintetičkog pokazatelja varijabiliteta** višedimenzionog skupa podataka je u tome što zanemaruje kovarijacionu strukturu podataka te se za različite kovar.matrice može dobiti ista vrednost $|S|$. ***Preporuka*** je da se uz njenu vrednost prezentuje i sama kovar.matrica.

III *VIŠEDIMENZIONALNA NORMALNA RASPODELA*

(1) **Generalizovano odstojanje od x do μ** ($px1$ vektor očekivane vrednosti vektora X):

$$(x - \mu)^T \Sigma (x - \mu)$$

(2) **Ocena najveće verodostojnosti je sredina vektora**:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

a **ocena najveće verodostojnosti kovarijacione matrice Σ:**

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T = \frac{n-1}{n} S$$

(3) **Hipoteza sferičnosti** – nulta hipoteza, da su vandijagonalni elementi kovarijacione matrice su jednaki nuli, $H_0: \Sigma = \sigma^2 I$, se svodi na količnik aritmetičke i geometrijske sredine karakterističnih korena uzoračke kovarijacione matrice S:

$$n\, p \ln \frac{\frac{1}{p}\sum_{j=1}^{p} \lambda_j}{\left(\prod_{j=1}^{p} \lambda_j\right)^{\frac{1}{p}}}$$

Ako je tačna nulta hipozeza ova statistika ima hi-kvadrat asimptotsku raspodelu sa $\frac{(p-1)(p+2)}{2}$ stepeni slobode.

(4) **Maksimum funkcije gustine verovatnoće** (višedimenzionalne) **normalne raspodele** dostiže se za $x = \mu$.

(5) Statistika koja se koristi za testiranje hipoteze: $H_0: \mu = \mu_0$ poznata je pod nazivom **Wilksova lambda**. Na osnovu veze koja postoji između T^2 i Λ statistike:

$$\Lambda^{2/n} = \left(1 + \frac{T^2}{n-1}\right)^{-1}$$

možemo koristiti T^2 raspodelu prilikom primene Λ statistike u testu hipoteze. Kažemo da su T^2 test i test zasnovan na količniku verodostojnosti međusobom ekvivalentni.

VI *DISKRIMINACIONA ANALIZA*

(1) **Diskriminaciona analiza** – metod multivarijacione analize koji se bavi razdvajanjem različitih grupa i alokacijom opservacija u unapred definisane grupe.

(2) **Diskriminacioni skor** – broj koji se određuje linearnom kombinacijom nezavisnih promenljivih za svakog ispitanika ili objekta, koji se zatim transformiše u aposteriornu verovatnoću da ispitanik ili objekat potiče iz jedne od grupa (dobija se kao $p_k f_k(x)$)

(3) **Centroid** – sredine diskriminacionih skorova za dve grupe.

(4) **Fišerov diskriminacioni kriterijum**:

$$\frac{(\mu_{1Y} - \mu_{1Y})^2}{\sigma_Y^2} = \frac{(a^{\mathrm{T}}\mu_1 - a^{\mathrm{T}}\mu_2)^2}{a^{\mathrm{T}}\Sigma a} = \frac{(a^T\delta)^2}{a^{\mathrm{T}}\Sigma a}$$

zadatak je da se odredi vektor koeficijenata linearne kombinacije a tako da se maksimizira ovaj poslednji količnik.

(5) **Fišerova linearna diskriminaciona funkcija:** $Y = a^T X = (\mu_1 - \mu_2)^T \Sigma^{-1} X$

(6) **Uzoračka Fišerova linearna diskriminaciona f-ja** na osnovu realizovanih vrednosti slučajnog uzorka: $y = \hat{a}^T X = (\overline{x_1} - \overline{x_2})^T \overline{\mathrm{S}}^{-1} X$

(7) **Opšta sredina svih grupa** (populacija se sastoji od g grupa):

$$\mu = \frac{1}{g}\sum_{k=1}^{g} \mu_k$$

(8) R_1 je oblast koji sadrži one x za koje važi: $\frac{f_1(x)}{f_2(x)} \geq \left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]$ a R_2: $\frac{f_1(x)}{f_2(x)} < \left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]$. Oblasti R_1 i R_2 minimiziraju očekivane troškove pogrešne klasifikacije. Kada su ovi troškovi međusobno jednaki tada su: R_1: $\frac{f_1(x)}{f_2(x)} \geq \frac{p_2}{p_1}$ a R_2: $\frac{f_1(x)}{f_2(x)} < \frac{p_2}{p_1}$. Ako su i apriorne verovatnoće jednake među sobom, tj $\frac{p_2}{p_1} = 1$, to je najčešći slučaj u praktičnoj primeni. Za novu opservaciju x_0 računa se vrednost f-je gustine verovatnoće, pa ako je $f_1(x_0) \geq f_2(x_0)$ tada se alocira u π_1, a u suprotnom u π_2. Ovo pravilo se naziva **pravilo najveće verodostojnosti.**

(9) **Bayesov rizik** predstavlja aposteriorni očekivani gubitak. Na njemu zasnovan **Bajesov kriterijum alokacije** zahteva klasifikaciju opservacije u onu grupu koja ima najmanji Bayesov rizik. Osnovna varijanta Bayesovog pravila diskriminacije računa se samo sa apriornim verovatnoćama, a ne i troškovima.

(10) **Minimax pristup** – koristi se u situacijama kada jedna od grupa ima malu apriornu verovatnoću. Prema njemu se minimizira maksimalna verovatnoća pogrešne klasifikacije.

(11) **Očekivani trošak pogrešne alokacije** ($P(k|i)$ je uslovna verovatnoća pogrešne klasifikacije elementa iz i-te populacije u k-tu populaciju)

$$C = \frac{1}{g}\sum_{i=1}^{g} p_1 \left[\sum_{\substack{k=1 \\ k\neq i}}^{g} P(k|i)c(k|i)\right]$$

(12) **Matrica konfuzije** – prikazuje broj ispravno i pogrešno klasifikovanih opservacija po grupama. Neka se problem alokacije odnosi na dve grupe π_1 i π_2 odakle smo uzeli uzorak od po n_1 i n_2 opservacije respektivno. Matrica konfuzija je data tabelom:

		Stvarna pripadnost	
		π_1	π_1
Predviđena pripadnost	π_1	n_{11}	n_{12}
	π_2	n_{21}	n_{22}
Veličina uzorka		$n_1 = n_{11} + n_{21}$	$n_2 = n_{12} + n_{22}$

Elementi na glavnoj dijagonali predstavljaju broj opservacija koje su ispravno alocirane, a van nje – one koje su neispravno alocirane.

(13) **Ocena stope greške kod matrice konfuzije**:

$$\hat{E} = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

(14) **Jackknife metod** – koristi se za ocenu pristrasnosti, varijanse i opštijih mera grešaka pri ocenjivanju. Za ma koju ocenu na osnovu slučajnog uzorka se može formirati jackknife ocena varijanse te ocene. Jackknife ocena stope greške je:

$$\hat{E}^{(J)} = \frac{{n_{12}}^{(J)} + {n_{21}}^{(J)}}{n_1 + n_2}$$

VII *GLAVNE KOMPONENTE*

(1) **Zadaci metoda glavnih komponenata** – određivanje linearne kombinacije originalnih promenljivih sa maksimalnom varijanso, i opštiji zadatak je određivanje nekoliko lin.kombinacija orig.prom. koje će, pored toga što imaju maksimalnu varijansu biti međusobno korelirane gubeći u što je manjoj meri informaciju sadržanu u skupu originalnih promenljivih.

(2) **Ciljevi metoda glavnih komponenti** – vrši redukciju originalnog skupa podataka i olakšava njegovu interpretaciju.

(3) **Glavne komponente** – nove promenljive (linearne kombinacije) u koje se originalne transformišu u postupku primene metoda glavnih komponenti. ***Prva glavna komponenta*** ($Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2$) konstruisana je tako da obuhvata janveći deo varijanse originalnog skupa podataka, a naredne onaj deo koji nije obuhvaćen predhodno izdvojenim glavnim komponentama.

(4) **Osobine glavnih komponenti**:

a) $E(Y_j) = 0,\ Var(Y_j) = \lambda_j,\ Cov(Y_i, Y_j) = 0, i \neq j$

b) $Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_p) \geq 0$

c) Generalizovane varijanse glavnih komponenata jednake su generalizovanim varijansama originalnog skupa promenljivih (po prvoj definiciji gen.varijanse dobija se da je $|\Lambda| = |A\Sigma A^T| = |A||\Sigma||A^T| = |\Sigma|$, a po drugoj: $tr(\Lambda) = tr(A\Sigma A^T) = tr(AA^T\Sigma) = tr(\Sigma)$

d) Na osnovu osobina karakterističnih vektora ($a_j^T a_j = 1$ i $a_i^T a_j = 0, i \neq j$) matrica A ima osobinu da je $A^T = A^{-1}$, pa se $Y = AX$ naziva ortogonalna transformacija ili rotacija, a matrica A ortogonalna matrica. Ona ima osobinu i da $|A| = \pm 1$. Transformacija se naziva ortogonalna jer se njome vrši rotacija

koordinatnih osa za izvestan ugao, pri čemu ose ostaju upravne među sobom, a ugao između ma koja dva vektora ostaje isti nakon transformacije.

(5) **Kovarijansa između X i Y je**: $Cov(X,Y) = Cov(X,AX) = \Sigma A^T = (A^T \Sigma A)A^T = A^T \Lambda$

(6) **Kovarijansa između k-te originalne promenljive i j-te glavne komponente**:

$$\rho_{X_k Y_j} = \frac{Cov(X_k Y_j)}{\sqrt{Var(X_k)}\sqrt{Var(Y_j)}} = \frac{\lambda_j \alpha_{jk}}{\sqrt{\sigma_{kk}}\sqrt{\lambda_j}} = \alpha_{jk}\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}$$

-U matričnom zapisu: $\rho_{XY} = \Lambda^{1/2} A D^{1/2}$

(7) **LR testovi** (Likelihood ratio) – na kojima je zasnovana statistika testa za hipoteze: $H_0: \lambda_{k+1} = \lambda_{k+2} = \lambda_p$ protiv H_1 koja kaže da su barem dva međusobno različita, a koja je konstruisana uz predpostavku o normiranosti. Ova statistika glasi ($\hat{\lambda}_j$ su karakteristični koreni uzoračke kovarijacione matrice):

$$LR = \left(\frac{\prod_{j=k+1}^{p} \hat{\lambda}_j}{\left(\frac{1}{p-k}\sum_{j=1}^{p} \hat{\lambda}_j\right)^{p-k}} \right)^{\frac{n}{2}}$$

Na osnovu ovoga zaključujemo da je test zasnovan na poređenju geometrijske i aritmetićke sredine poslednja $(p-k)$ karakteristična korena.

(8) **Izbor broja glavnih komponenata**

a) prvi pristup polazi od apriornog fiksiranja kumulativne proporcije ukupne varijanse objašnjene izdvojenim skupom glavnih komponenata. Izabira se, u zavisnosti od problema, npr. 80% ukupne varijanse, pa se broj zadržanih komponenti povećava sve dok se ne dostigne ova unapred zadata granična vrednost.

b) drugi sugeriše zadržavanje svih glavnih komponenata čija je varijansa (λ_j) veća od prosečne vrednosti $\bar{\lambda} = \sum_{j=1}^{p} \frac{\lambda_j}{p}$. Ako umesto kovarijacione koristimo korelacionu, tada je prosečna vrednost varijanse jednaka jedinici, što znači da kriterijum glasi: zadržati one komponente kod kojih je varijansa (λ_j) veća od 1 (Kajzerov kriterijum).

c) treći pristup koristi geometrijsku sredinu u kriterijumu izbora. Ako vrednost generalizovane varijanse ($\prod_{j=1}^{p} \lambda_j$) dignemo na stepen $\frac{1}{p}$ dobićemo geometrijsku sredinu kar.korena. Zadržavaju se one komponente čiji je kar.koren veći od geometrijske sredine svih kar.korena.

d) četvrti pristup se zasniva na grafičkom prikazu vrednosti kar.korena prema njihovom rednom broju. Ovaj dijagram naziva se scree test.

(9) Ozbiljan **problem koji se pri korišćenju glavnih komponenata u diskriminacionoj analizi javlja** ogleda se u tome što prve glavne komponente ne moraju biti dobri diskriminatori između grupa. U opštem slučaju nema garancija da će razdvajanje između grupa biti u pravcu glavne komponente sa visikom varijansom. Tako se izostavljanjem glavnih komponenata sa niskom varijansom može odbaciti važna informacija koja se tiče varijacija između grupa.

(10) **Multikolinearnost** – pojava visoke linearne zavisnosti između regresora u višestrukom regresionom modelu.

(11) **Nestandardne opservacije** (outlier) – one koje su na neki način nekonzistetne sa preostalim delom podataka.

VIII *FAKTORSKA ANALIZA*

(1) **Faktorska analiza** – metod multivarijacione analize koji se koristi za opis međusobme zavisnosti velikog broja promenljivih korišćenjem manjeg broja osnovnih, ali neopažljivih slučajnih promenljivih poznatih kao faktori. Utemeljivač faktorske analize – Čarls Spirman.

(2) **Razlike između faktorske i analize glavnih komponenti**

-dok je analiza glavnih komponenata koncentrisana na dijagonalne elemente kovarijacione matrice (varijanse) u faktorskoj nas interesuju vandijagonalni (kovarijanse)

-za razliku od analize glavnih komponenata, faktorska pretpostavlja postojanje promenljive i manjeg broja zajedničkih faktora;

-an.gl.komp. izučava ukupan varijabilitet skupa podataka, a faktorska polazi od razlaganja promenljive na dva dela: zajednički i specifični;

-faktorska izučava deo varijacija koji je zajednički za sve promenljive, a an.gl.komp. ukupan varijabilitet.

(3) **Faktorska analiza** polazi od razlaganja promenljive na dva dela – zajednički i specifični.

(4) Spirman predlaže redukciju problema sa $p = 4$ na $p = 1$, tako što ćemo rezultate svih testova (X_i) prikazati u obliku modela: $X_i = \beta_i F + \varepsilon_i$, gde je F ***zajednički faktor***, β_i ***faktorska opterećenja***, a ε_i ***slučajne greške*** (***specifični faktori***).

(5) **Model faktorske analize**:

$$\begin{matrix} (X_1 - \mu_1) = \beta_{11}F_1 + \beta_{12}F_2 + \cdots + \beta_{1m}F_m + \varepsilon_1 \\ (X_1 - \mu_1) = \beta_{12}F_1 + \beta_{22}F_2 + \cdots + \beta_{2m}F_m + \varepsilon_2 \\ \vdots \\ (X_p - \mu_p) = \beta_{p1}F_1 + \beta_{p2}F_2 + \cdots + \beta_{pm}F_m + \varepsilon_p \end{matrix} \quad \Leftrightarrow X - \mu = BF + \varepsilon$$

(6) **Generalizovana varijansa od X** je:

$$tr(\Sigma) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p}\sum_{j=1}^{m} {\beta_{ij}}^2 + \sum_{i=1}^{p} \psi_i$$

a **ukupan komunalitet od X** je:

$$h = \sum_{i=1}^{p} h_i^2 = \sum_{i=1}^{p}\sum_{j=1}^{m} {\beta_{ij}}^2 \implies tr(\Sigma) = h + tr(\psi)$$

formiranjem sledećeg količnika dobija se **proporcija uk.komunaliteta koji se može pripisati j-tom zajedničkom faktoru**: $\frac{\sum_{i=1}^{p} {\beta_{ij}}^2}{h}$.

(7) **Metodi ocenjivanja u faktorskoj analizi** – metod glavnih komponenti (faktora) i metod najveće verodostojnosti.

(8) **Matrica ocenjenih faktorskih opterećenja**:

$$\hat{B} = \left[\sqrt{\hat{\lambda}_1\hat{\alpha}_1} \vdots \sqrt{\hat{\lambda}_2\hat{\alpha}_2} \vdots \cdots \vdots \sqrt{\hat{\lambda}_p\hat{\alpha}_p}\right]$$

(8) **Matrica reziduala** se dobija kao rezultat uzoračke kovarijacione matrice sa prvih m faktora:

$$S - (\hat{B}\hat{B}^T + \hat{\Psi})$$

(9) **Korelaciona matrica** se razlaže po modelu $\rho = BB^T + \Psi$, gde su dijagonale $\rho_{ii} = 1 = h_i^2 + \psi_i$. Iz ovoga se formira nova matrica koja predstavlja korelacionu matricu zajedničkih faktora – **redukovana korelaciona matrica**:

$$\rho - \Psi = BB^T = \begin{bmatrix} h_1^2 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & h_2^2 & & \rho_{2p} \\ \vdots & & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & h_p^2 \end{bmatrix}$$

(10) **Uslov jednoznačnosti** – $B^T\Psi^{-1}B$ je dijagonalna matrica.

(11) **Kriterijum jediničnog korena** – u modelu zadržavamo onoliko zajedničkih faktora koliko ima karakterističnih korena uzoračke korelacione matrice koji su veći od jedinice. Znači da se zadržavaju oni faktori koji doprinose objašnjenju barem koliko i jedna promenljiva. Dobre osobine ovog kriterijuma su u slučaju kada je broj promenljivih između 20 i 50, ako je više od 50 ovaj kriterijum izdvaja suviše veliki broj zajedničkih faktora, ako je manje od 20 kriterijum je konzervativan jer izdvaja suviše mali broj zaj.faktora.

(12) **Rotacija faktora**. Postoji više različitih matrica fakt.opterećenja koje generišu istu kovarijacionu matricu. Primenjeni postupak ortogonalne transformacije matrice u geometrijskom smislu predstavlja rotaciju osa za izvestan ugao, pri čemu njihov mešusobni odnos ostaje nepromenjen, ostaju ortogonalne. Ovakav postupak se naziva (ortogonalna) rotacija faktora, a ako se napusti zahtev da rotirani vektori moraju biti ortogonalni – neortogonalne rotacije faktora (oblique rotacije).

(13) **Jednostavna struktura** – najčešće korišćeni kriterijum koji opredeljuje izbor ugla za koji ćemo rotirati faktore. Koristi se u cilju lakše interpretacije rešenja modela faktorske analize.

(14) **Kaiserov varimax metod** – najčešće korišćeni analitički metod ortogonalne rotacije faktora. Postupak je iterativan: izdvojeni faktori se posmatraju po parovima i vrši se njihova rotacija dok se ne postigne max vrednost varimax kriterijuma za prvi par faktora. Zatim se prvi rotirani sa trećim nerotiranim faktorom, rotira do postizanja max varimax kriterijuma. Postupak se ponavlja dalje sve dok se svih $\frac{m(m-1)}{2}$ parova faktora ne rotiraju (ciklus). Ciklus se ponavlja sve dok se ne postigne da su svi uglovi dobijeni za parove faktora manji od unapred izabrane vrednosti, koja predstavlja kriterijum konvergencije.

(15) **Quatrimax kriterijum** – prema kome se kao indikator jednostavnosti strukture uzima suma varijansi kvadrata svih elemenata matrice $\hat{\Gamma}$. Ova metoda obično rezultira u opštem faktoru jer se varijansa računa na osnovu svih elemenata matrice fakt.opterećenja (za razliku od varimax-a gde se varijansa opterećenja računa za svaku kolonu posebno, pa se dobija veći broj zajedničkih faktora).

(16) **Orthomax** – metod predstavlja ponderisani presek sirovog varimax i quatrimax kriterijuma.

IX *ANALIZA GRUPISANJA*

(1) **Analiza grupisanja** – metod multivarijacione analize koji se koristi za grupisanje objekata u grupe, zako da su objekti unutar grupe sličniji međusobom, a između grupa znatno različiti.

(2) **Postupci grupisanja** – hijerarhijski i nehijerarhijski metodi.

(3) U analizi grupisanja na osnovu nxp matrice podataka formuramo nxn matricu bliskosti čiji elementi mere stepen sličnosti ili razlike između svih parova profila iz matrice podataka.

(4) **Hijerarhijsko drvo** – struktura kod koje se u svakoj iteraciji objekti pridružuju predhodno formiranim grupama ili sa drugim objektom formiraju novu grupu.

(5) **Načini formiranja hijerarhijske strukture** – udruživanjem, kada se pri formiranju grupa krećemo od grana drveta ka njegovom korenu, i deobom, kada se ide u suprotnom smeru, od korena ka granama.

(6) **Ciljevi analize grupisanja** – istraživanje podataka, redukcija podataka, generisanje hipoteza i predviđanje

(7) Za meru bliskosti p_{rs} kažemo da predstavlja **meru sličnosti između objekata r i s** ako ispunjava:

a) uslov normiranosti: $0 \leq p_{rs} \leq 1$ za sve objekte r i s

b) $p_{rs} = 1$, samo ako su r i s identični

c) uslov simetričnosti: $p_{rs} = p_{sr}$

(8) **Pearsonov koeficijent korelacije** – da bi mogao da se korsti kao mera sličnosti postupa se na dva načina, ili se korsiti njegova apsolutna vrednost, ili se koeficijentu doda jedinica, pa se dobijena vrednost podeli sa 2 da bi se ispunio uslov normiranosti.

(9) **Euklidsko odstojanje** predstavlja specijalan slučaj odstojanja Minkovskog (za $\lambda = 2$) koje je dato izrazom:

$$M = \left[\sum_{j=p}^{p} \left|x_{rj} - x_{sj}\right|^{\lambda}\right]^{1/\lambda}$$

A za $\lambda = 1$ dobijamo **Menhetn odstojanje** (city block).

(10) **Mahalanobisovo odstojanje** – vodi računa i o kovarijacionoj strukturi podataka – multivarijaciona mera odstojanja. Svojstvo – eliminiše efekat koreliranosti promenljivih.

(11) **Grafik profila** – koristi se za analizu razlike dva profila. Grafički prikaz je moguće dati ako su promenljive istovrsne ili je predhodno izvršena njegova standardizacija.

(12) Pri **poređenju dva profila** možemo porediti njihove nivoe, stepen varijabiliteta i njihov oblik.

(13) **Meru sličnosti** možemo kostruisati na osnovu mere odstojanja. Neka je izračunato Euklidsko odstojanje (d_{rs}) između dva objekta. Korišćenjem mere odstojanja definišemo meru sličnosti objekata r i s: $p_{rs} = \frac{1}{1+d_{rs}}$.

(14) **Koeficijenti sličnosti**:

a) ***jednostavni koeficijent podudarnosti*** (Sokala i Mičenera) – on daje isti ponder paru 0-0 kao i paru 1-1; mera sličnosti je:

$$\frac{a+d}{a+b+c+d}$$

b) ***Jaccardov koeficijent*** isključuje broj promenljivih kod kojih oba objekta uzimaju vrednost nula, koristi se kada je potrebno utvrditi sličnost objekta sa obzirom na prisustvo svojstva, a zanemariti odsustvo svojstva. Mera sličnosti je:

$$\frac{a}{a+b+c}$$

(15) Mere sličnosti i razlike između grupa – **metodi hijerarhijskog udruživanja**: metodi povezivanja (ima ih 3), metoda centroida, metoda minimalne sume kvadrata (Wardova metoda).

(16) **Metode povezivanja**: ***jednostruko povezivanje*** (metod najbližeg suseda, odstojanje između grupa se meri na osnovu najmanjeg odstojanja parova objekata iz posmatrane dve grupe), ***potpuno povezivanje*** (metod najudaljenijeg suseda – odstojanje između dve grupe određuje se na osnovu najvećeg odstojanja između parova objekata između dve grupe) i ***prosečno povezivanje*** (odstojanje se određuje na osnovu prosečnog odstojanja svih parova objekata iz dve posmatrane grupe).

(17) **Metod centroida** – dve grupe se udružuju u novu grupu ako su njihovi centroidi najmanje udaljeni međusobom u odnosu na međusobnu udaljenost svih mogućih parova grupa koje postoje na posmatranom nivou udruživanja.

(18) **Wardov metod** – dve grupe se spajaju u jednu ako je njihovim udruživanjem došlo do najmanjeg povećanja sume kvadrata unutar grupa u odnosu na povećanje sume kvadrata do koga je došlo udruživanjem ma koje druge dve grupe na posmatranom nivou udruživanja.

(19) Metode centroida i Wardov metod zahtevaju da se koriste kvadrat euklidskog odstojanja.

(20) **Izvedenu matricu odstojanja** formiramo iz: dendrograma ili rezultata analize grupisanja.

(21) **Kofenetički koeficijent –** koeficijent korelacije između originalnih i izvedenih mera odstojanja. Njegova vrednost bliska jedinici sugeriše da su formirane grupe objekata kvalitetno rešenje problema grupisanja.

(22) **Koeficijent konkordanse** se izračunava prema izrazu:

$$\gamma = \frac{S(+) - S(-)}{S(+) + S(-)}$$

gde $S(+)$ pokazuje koliko puta parovi objekata koji nisu u istoj grupi imaju veće odstojanje od drugih parova koji su u istoj grupi, a $S(-)$ obrnuto.

(23) **K-means metod** (metod k-sredina) – nehijerarhijska metoda grupisanja prema kojoj objekat pridružujemo grupi koja ima najbliži centroid.