



Diskriminantna analiza

Multivariatna analiza

Uvod

- Z diskriminantno analizo (DA) poiščemo tako linearno kombinacijo merjenih spremenljivk, da bo maksimalno ločila vnaprej določene skupine in da bo napaka pri uvrščanju enot v skupine najmanjša.
- Pri diskriminantni analizi torej gre za iskanje tistih razsežnosti, ki kar najbolj pojasnjujejo razlike med skupinami (pojasnjevanje), in za kar se da dobro prirejanje enot vnaprej danim skupinam (napovedovanje).

Predpostavke linearne DA

1. $k \geq 2$.
2. Vsaj 2 enoti v vsaki skupini.
3. $p < n - 1$; p je število spremenljivk in n število vseh enot v vzorcu.
4. Nobena spremenljivka ne sme biti linearna kombinacija preostalih sprem. (multikolinearnost).
5. Pri statističnemu ocenjevanju se predpostavlja, da so v vsaki skupini enot (vzorcu) enote slučajno izbrane iz populacije, kjer se spremenljivke porazdeljujejo večrazsežno normalno.
6. Variančno-kovariančna matrika $p \times p$ je v vsaki populacijski skupini enaka.

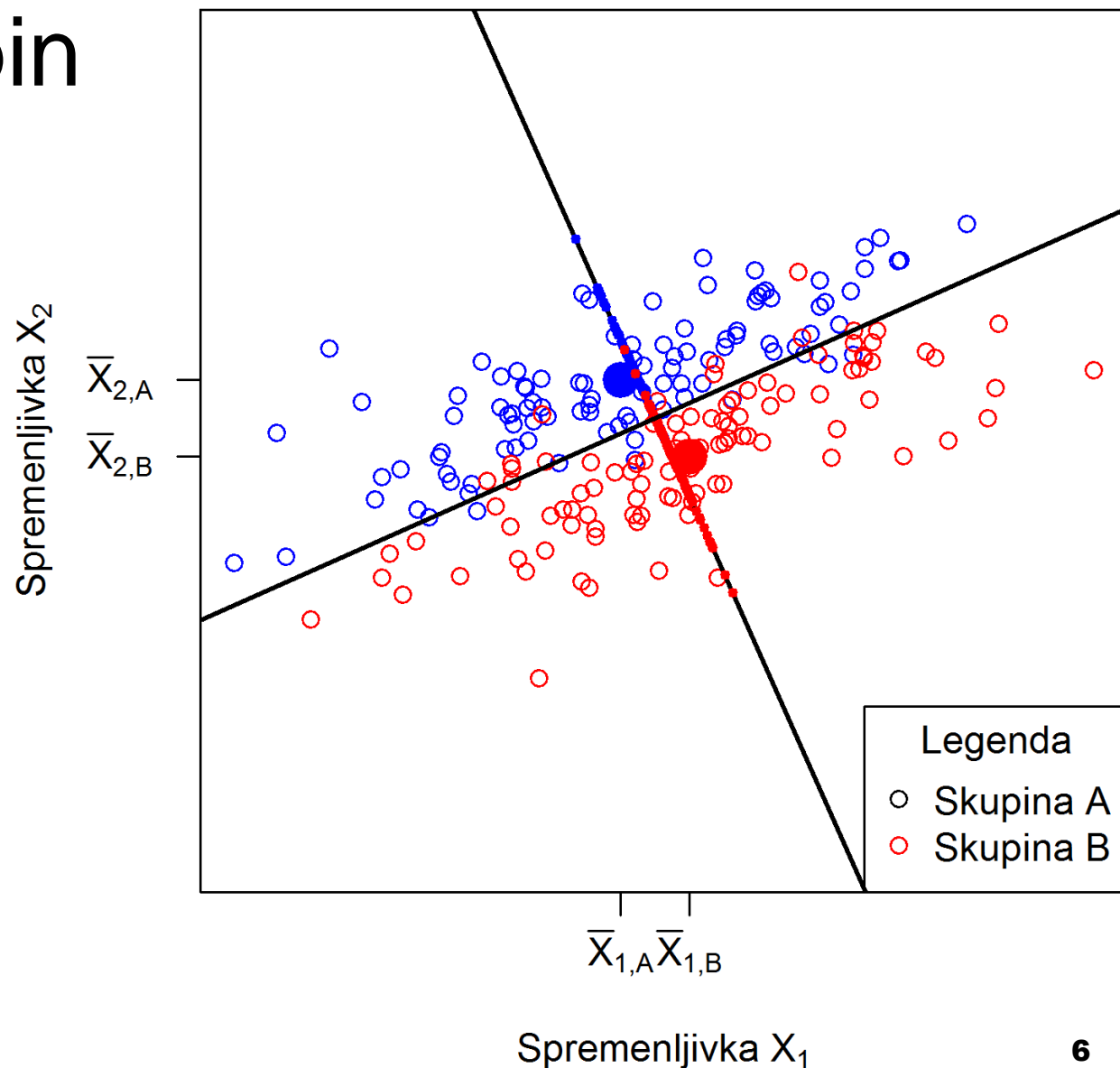
Predpostavka o enakosti kovariančnih matrik (po skupinah)

- Predpostavko testiramo z Box-ovim M testom.
- Testna statistika M meri, koliko so determinante kovariančnih matrik po skupinah različne od determinante skupne kovariančne matrike.
- Še posebej pri različnih velikostih skupin in večjem številu skupin ali spremenljivk, test ni robusten na odstopanja od normalnosti
- V večini primerov je preveč občutljiv (p vrednosti so premajhne \rightarrow prekmalu zavarča ničelno domnevo), razen če so večje variance v manjših skupinah

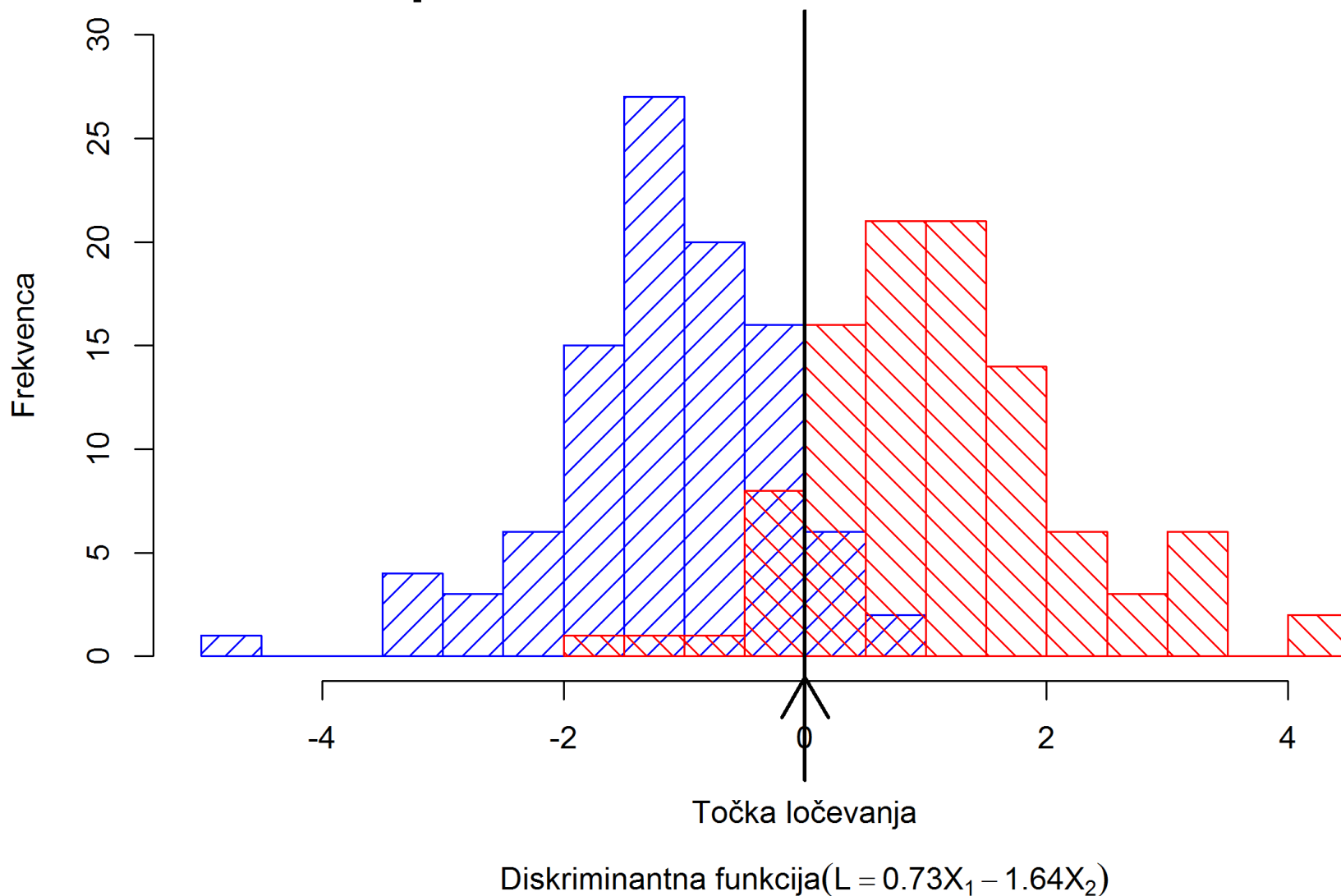
Kvadratna diskriminatna analiza

- Obstaja tudi kvadratna diskriminantna analiza, ki nima zadnje od prej omenjenih predpostavk
- Je pa pri tej metodi predpostavka o multivariatni normalni porazdelitvi pomembna tudi za samo klasifikacijo enot (in ne le pri testih)
- Primerna je torej takrat, ko so kovariančne matrike med skupinami različne, porazdelitev pa je v vseh skupinah multivariatno normalna.
- R jo podpira, SPSS pa ne v celoti. V SPSS-ju različne kovariančne matrike lahko upoštevamo pri uvrščanju, kar je pa enako le, če je $p \leq k - 1 \rightarrow$ uporabljajo se le „kvadratne klasifikacijske funkcije“.
- Podrobneje je ne bomo obravnavali

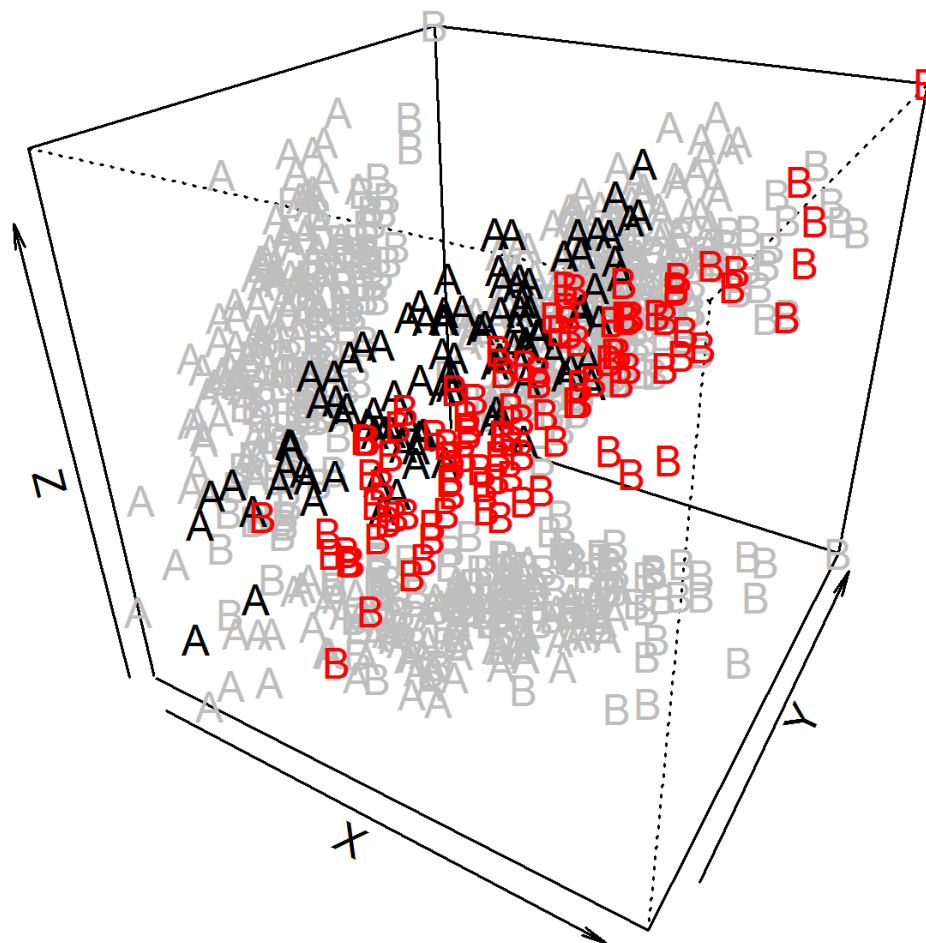
Diskriminantna analiza v primeru dveh skupin



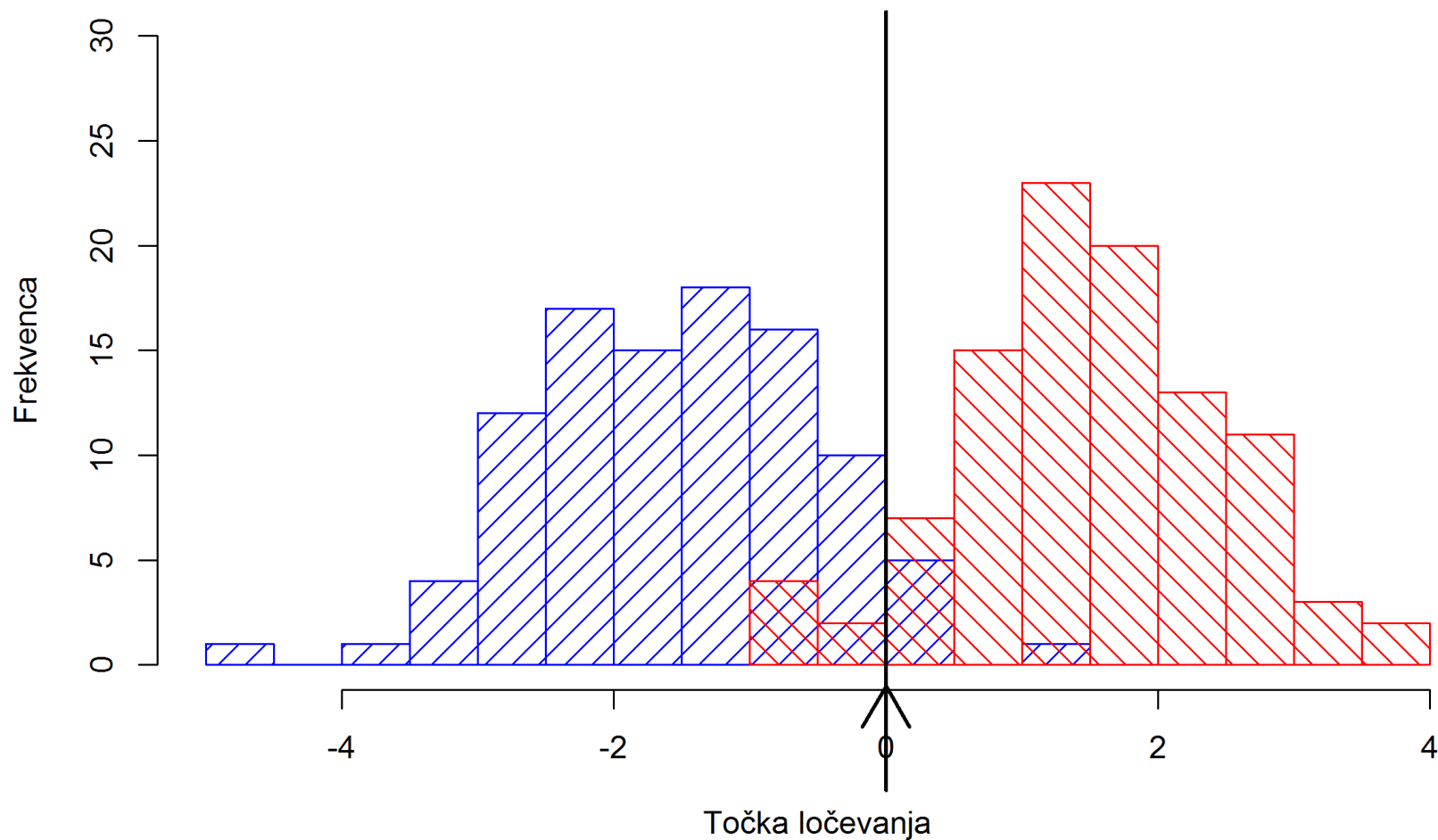
Diskriminantna analiza v primeru dveh skupin



Diskriminantna analiza v primeru dveh skupin – 3 spremenljivke



Diskriminantna analiza v primeru dveh skupin – 3 spremenljivke



Diskriminantna funkcija ($L = 0.53X + 1.04Y - 1.29Z$)

Diskriminantna analiza v primeru dveh skupin

- Recimo, da imamo dve skupini, G_1 in G_2 , kjer ima vsaka svoj vektor aritmetičnih sredin in kovariančno-variančno matriko.

Skupina	Vektor aritmetičnih sredin	Kovariančno-variančna matrika
G_1	μ_1	Σ_1
G_2	μ_2	Σ_2

- Predpostavka: $\Sigma_1 = \Sigma_2 = \Sigma$

- Fisher (1936) je definiral diskriminantno spremenljivko (funkcijo) Y kot linearno kombinacijo p merjenih spremenljivk X_i

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_pX_p = Xb$$

tako da je kvocient razlik aritmetičnih sredin diskriminantne spremenljivke v obeh skupinah (G_1 in G_2) glede na varianco diskriminantne spremenljivke znotraj skupine maksimalen.

- Aritmetični sredini diskriminantne spremenljivke v skupinah G_1 in G_2 sta: $\bar{Y}_1 = b'\mu_1$ in $\bar{Y}_2 = b'\mu_2$
- Varianca je: $\text{var}(Y_1) = \text{var}(Y_2) = b'\Sigma b$ (zaradi predpostavke $\Sigma_1 = \Sigma_2 = \Sigma$)

- Kvocient, ki naj bi bil maksimalen, pa je:

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\text{var}(Y_1)} = \frac{b' \mu_1 - b' \mu_2}{b' \Sigma b}$$

- To je pogoj, na osnovi katerega izračunamo uteži najboljše diskriminantne spremenljivke.
- Reševanje tega optimizacijskega problema privede do rešitve za b , ki je sorazmerna (rešitev je lahko pomnožena s poljubno konstanto)

$$\Sigma^{-1}(\mu_1 - \mu_2)$$

Vzorčne ocene

- Ponavadi imamo vzorčne podatke za vsako populacijo G_i , na osnovi katerih ocenimo parametre.
- Vzorčne ocene za μ_i so: $\bar{x}_i = (\bar{x}_{i1}, \bar{x}_{i2}, \dots, \bar{x}_{ip})$
- in za Σ (angl. pooled sample variance-covariance matrix):

$$S = \frac{(X_1'X_1 + X_2'X_2)}{n_1 + n_2 - 2} = \frac{S_1(n_1 - 1) + S_2(n_2 - 1)}{n_1 + n_2 - 2}$$

- kjer je n_1 število enot v vzorcu iz G_1 in n_2 število enot v vzorcu iz G_2 .
- Ocena uteži je tedaj: $\hat{b} = S^{-1}(\bar{x}_1 - \bar{x}_2)$

Centroid skupine

- Aritmetično sredino diskriminantne spremenljivke v določeni skupini imenujemo centroid skupine.
- Centroid skupine i je
$$\bar{Y}_i = b' \bar{x}_i$$

Pravila uvrščanja enot v skupine

- Denimo, da smo izračunali diskriminantno spremenljivko $Y = Xb$.
- Pravilo uvrščanja enote v optimalno skupino je ob enako velikih skupinah v populaciji tedaj, da i -to enoto (glede na p izmerjenih spremenljivk) uvrstimo v skupino
 - G1, če je njena vrednost na diskriminantni spremenljivki y_i
$$(y_i - \bar{Y}_1)^2 \leq (y_i - \bar{Y}_2)^2$$
 - ali v G2, če
$$(y_i - \bar{Y}_1)^2 > (y_i - \bar{Y}_2)^2$$

- Enakovreden pogoj uvrščanja je metoda srednje točke ali točke ločevanja.
- Če predvidevamo, da sta velikosti (deleža) skupin na populaciji enaka ($N_1 = N_2$ oz. $\pi_1 = \pi_2$), je ta

$$Y_c = \frac{\bar{Y}_1 + \bar{Y}_2}{2}$$

- Če predvidevamo, da je $\bar{Y}_1 > \bar{Y}_2$, lahko prejšnje pravilo zapišemo tako, da enoto uvrstimo v skupino:

□ G1, če drži

$$y_i \geq Y_c$$

oziroma

$$y_i - Y_c = y_i - 1/2 (\bar{Y}_1 + \bar{Y}_2) \geq 0$$

□ sicer pa v G2

- Če pa predvidevamo različne velikosti (oz. deleže/verjetnosti) skupin v populaciji, se izračun malce spremeni in sicer je potrebno upoštevati (predpostavljene) deleže obeh skupin na populaciji


$$\pi_1 = \frac{N_1}{N}, \pi_2 = \frac{N_2}{N}$$

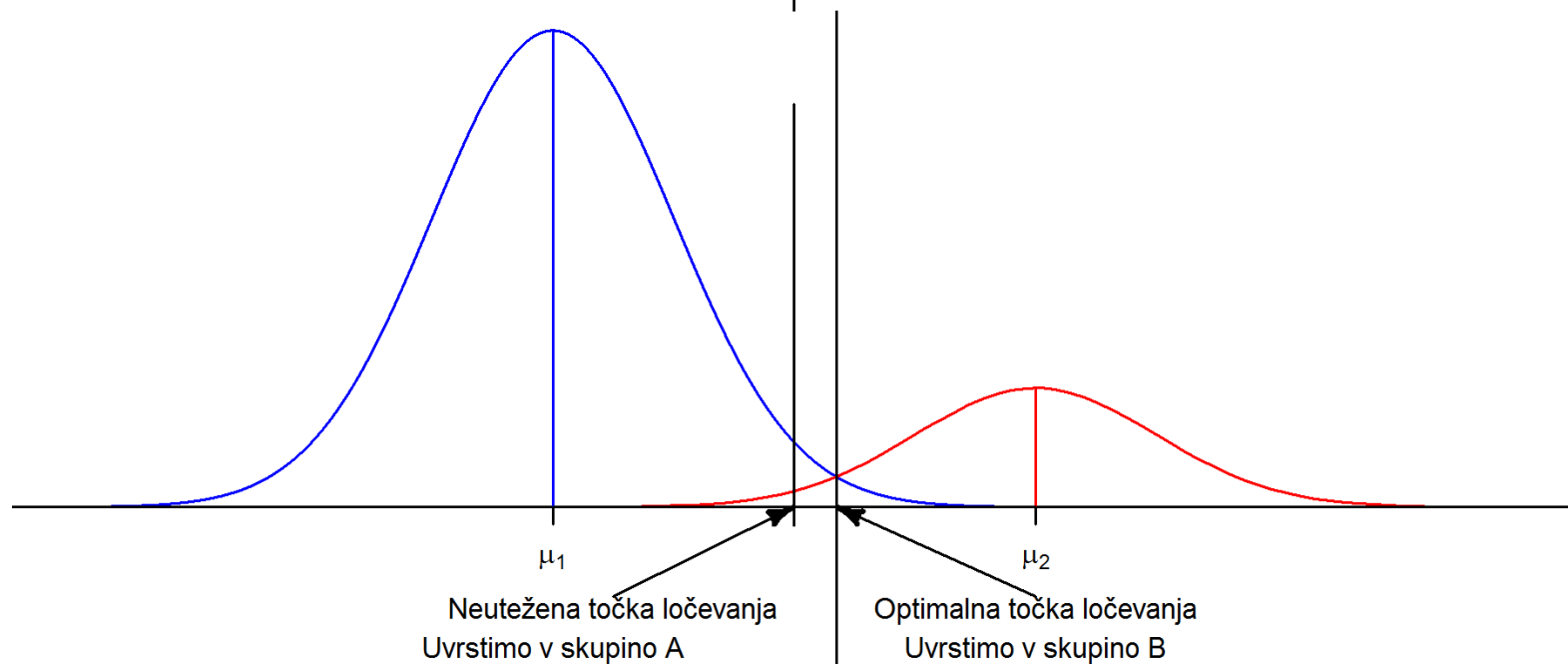
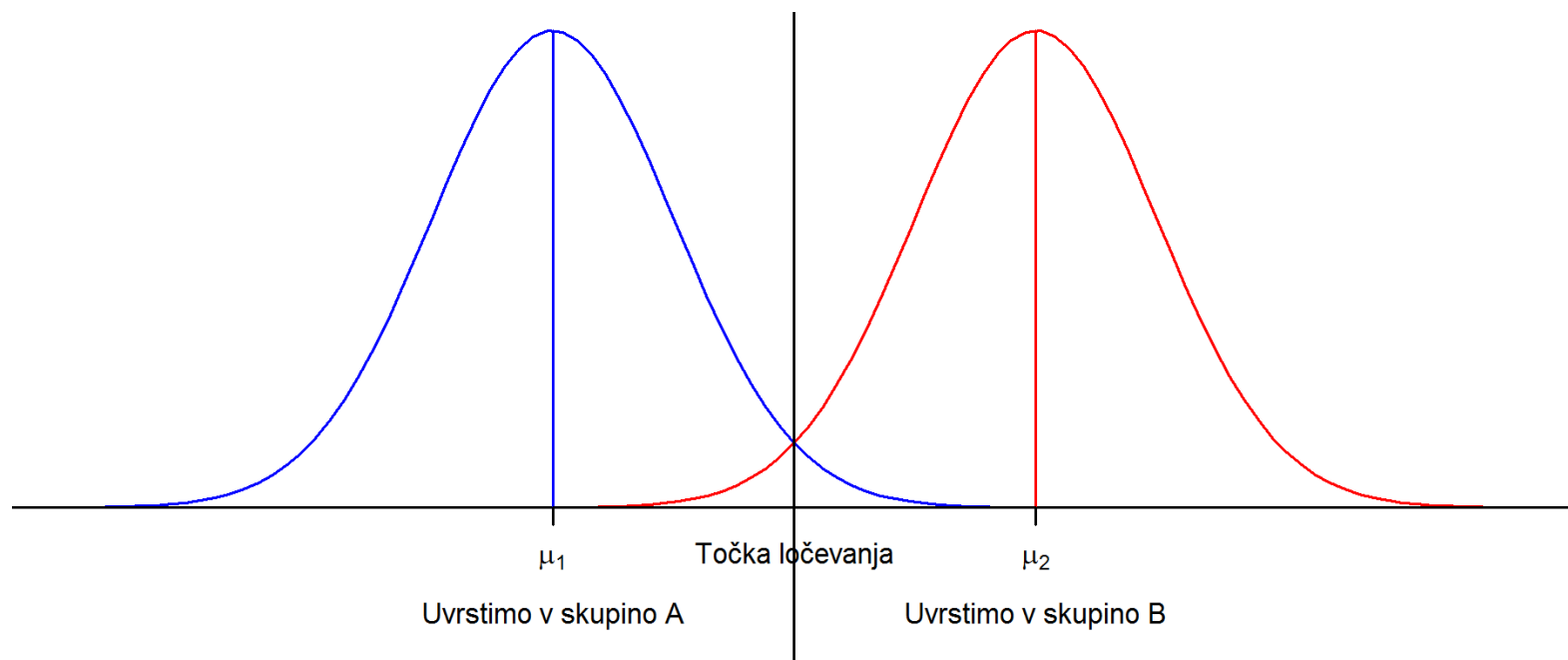
- Kot ocene pogosto vzamemo vzorčne deleže
- Pravilo, ki upošteva te verjetnosti (prior probabilities) je, da enoto uvrstimo v skupino

- G1, če drži

- $$1/2(y_i - \bar{Y}_1)^2 - \ln(\pi_1) \leq 1/2(y_i - \bar{Y}_2)^2 - \ln(\pi_2)$$

- sicer pa v G2

- 
- Enakovreden način uvrščanja so tudi Fisherjeve klasifikacijske linearne disriminatne funkcije, ki so še posebej priročne pri več skupinah.
 - Za vsako skupino se generira svoja klasifikacijska funkcija
 - Enoto razvrstimo v tisto skupino, pri kateri ima največjo vrednost klasifikacijske funkcije.



Klasifikacijska tabela

- Kvaliteto klasifikacije (uvrščanja) lahko ocenimo tudi z deležem (ne)pravilno uvrščenih enot. Uporabimo dobljeno diskriminantno funkcijo na podatkih, iz katerih je bila izračunana. Vsako enoto uvrstimo v eno od obeh skupin glede na pravilo uvrščanja. Rezultate uvrščanja lahko predstavimo z naslednjo tabelo:

Dejanski skupini	Število enot	Dobljeni skupini	
		G_1	G_2
G_1	n_1	n_{11}	n_{12}
G_2	n_2	n_{21}	n_{22}

- Delež pravilno uvrščenih enot je:


$$\frac{n_{11} + n_{22}}{n_1 + n_2}$$

- Ob predpostavki, da je v obeh skupinah enako število enot, je spodnja meja tega deleža 50%.
- Pri v naprej znanih (predpostavljanih) verjetnosti pa bi „na slepo“ pravilno razvrstili tak delež enot, kolikor je verjetnost večje skupine.
- Ocenjeni delež pravilno uvrščenih enot je optimistično pristranski, ker pri tem uporabljamo iste podatke za pravilo uvrščanja in oceno kakovosti uvrščanja.

Pričakovan delež pravilno razvrščenih enot pri slučajnem razvrščanju

- Delež enot, ki bi jih slučajno razvrstili v pravo skupini, če bi enote slučajno razvrščali v skupine (v skladu z deleži skupin), je dejansko (razen v primeru enako velikih skupine) manjši kot je deleža največje skupine
- Če so(sta) p_i verjetnosti ali deleži skupine, potem je pričakovan delež pravilno razvrščenih enot pri slučajnem razvrščanju enak

$$\sum_{i=0}^k p_i^2 \leq \max_i p_i$$

- 
- Za nepristranko oceno bi morali ocenjevati kakovost uvrščanja na **drugih** podatkih, kot smo jih uporabili za ocenjevanje diskriminantne funkcije. Dve možnosti sta (ena na naslednji strani):

- ☐ Podatke slučajno razdelimo na dva (približno enako velika) dela. Na enem ocenimo diskriminantno funkcijo (pravilo uvrščanja).

Slabosti metode je, da:

- Potrebujemo veliko podatkov
- Ocenjujemo kvaliteto pravila, ki smo ga ocenili le na polovici podatkov, v končni fazi pa bomo uporabljali pravilo, ocenjeno na vseh podatkih

- Ko preverjamo „pravilnost“ uvrščanja posamezne enote uporabimo za oceno diskriminantne funkcije vse enote razen te, ki jo uvrščamo (jackknife).
- Prednost metode je, da so pravila vedno generirana na skoraj vseh podatkih.
- Metoda je *skoraj* (torej ne popolnoma) nepristranska.
- Uporabljajo se lahko tudi druge oblike navzkrižnega preverjanja (jackknife je posebna oblika), več o tem drugo leto.

Primer

- Podatki so bili zbrani v okviru raziskave *Kakovost merjenja egocentričnih socialnih omrežij* (Ferligoj in drugi, 2000) leta 2000. Vzorec vsebuje 1033 prebivalcev Ljubljane. Analiza je bila narejena na 631 prebivalcih, ki so bili osebno intervjuvani.
- Na primeru bomo poskusili ugotoviti, katere razsežnosti ekstrovertiranosti najboljše pojasnjujejo razlike med spoloma.

ANOVE po vseh spremenljivkah

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
EKSTA Sem duša vsake družbe.	1,000	,103	1	613	,748
EKSTB Ne moti me, če sem v središču pozornosti.	,994	3,584	1	613	,059
EKSTE Na zabavah se pomenkujem z mnogo ljudmi vseh vrst	,998	1,262	1	613	,262
EKSTH Pogovore nacenjam jaz.	,999	,802	1	613	,371
EKSTIR Nerač/a pritegnem pozornost nase.	,997	1,937	1	613	,164
EKSTLR Sem redkobeseden(a).	,968	20,236	1	613	,000
EKSTNR V navzočnosti neznanih oseb sem molcec(a).	,999	,695	1	613	,405
EKSTOR Imam malo povedati.	,999	,574	1	613	,449
EKSTP Med ljudmi se pocutim sproščeno.	,999	,572	1	613	,450
EKSTRR Zadržujem se v ozadju.	1,000	,005	1	613	,943

Če med skupinami pri posamezni spremenljivki ni statistično značilnih razlik, ne moremo pričakovati, da bo dobro razlikovala med skupinami.

Enakost kovariančnih matrik

Box's Test of Equality of Covariance Matrices

Log Determinants		
E_SPOL spol ega	Rank	Log Determinant
1 moški	10	4,625
2 ženski	10	4,454
Pooled within-groups	10	4,625

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results		
Box's M		60,069
F	Approx.	1,073
	df1	55
	df2	1030950,832
	Sig.	,331

Tests null hypothesis of equal population covariance matrices.

- Box-ov M testira domnevo (eno izmed predpostavk), da sta kovariančni matriki na populaciji enaki
- Ker tu ne moremo zavrniti domneve, da sta matriki enaki, ne moremo trditi, da je predpostavka kršena

Kanonične diskriminantne funkcije

Summary of Canonical Discriminant Functions

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,058 ^a	100,0	100,0	,234

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,945	34,191	10	,000

- Kanonična korelacija med spremenljivko, ki predstavlja skupine (0/1) in „neodvisnimi“ spremenljivkami (diskriminantno funkcijo - DF) je 0,234.

- 1. diskriminantna funkcija je statistično značilna – ima statistično značilno različna povprečja po skupinah (podobno kot ANOVA)

Koeficienti diskriminatne funkcije (DF)

Canonical Discriminant Function Coefficients

	Function
	1
EKSTA Sem duša vsake družbe.	,046
EKSTB Ne moti me, če sem v središču pozornosti.	,217
EKSTE Na zabavah se pomenkujem z mnogo ljudmi vseh vrst	-,084
EKSTH Pogovore nacenjam jaz.	-,042
EKSTIR Nerač/a pritegnem pozornost nase.	,175
EKSTLR Sem redkobeseden(a).	-,730
EKSTNR V navzočnosti neznanih oseb sem molcec(a).	-,023
EKSTOR Imam malo povedati.	,114
EKSTP Med ljudmi se počutim sproščeno.	,312
EKSTRR Zadržujem se v ozadju.	,123
(Constant)	-,275

Unstandardized coefficients

- Surovi (nestandardizirani) koeficienti niso primerljivi med spremenljivkami z različnimi lestvicami oz. variabilnostjo, zato jih običajno ne interpretiramo.
- Za interpretacijo uporabljamo standardizirane koeficiente ali korelacije s DF (nekateri odsvetujejo)
- Nestandardizirani so uporabni za izračun vrednosti DF

Standardizirani koeficienti

Standardized Canonical Discriminant Function Coefficients

	Function
	1
EKSTA Sem duša vsake družbe.	,064
EKSTB Ne moti me, če sem v središču pozornosti.	,327
EKSTE Na zabavah se pomenkujem z mnogo ljudmi vseh vrst	-,115
EKSTH Pogovore nacenjam jaz.	-,054
EKSTIR Nerač/a pritegnem pozornost nase.	,257
EKSTLR Sem redkobeseden(a).	-1,036
EKSTNR V navzočnosti neznanih oseb sem molcec(a).	-,037
EKSTOR Imam malo povedati.	,157
EKSTP Med ljudmi se počutim sproščeno.	,309
EKSTRR Zadržujem se v ozadju.	,183

- Največji vpliv ima spremenljivka EKSTLR „Sem redkobeseden(a).“
- Ker je povprečje DF negativno pri ženskah, so ženske manj redkobesedne ($R \rightarrow$ spremenljivka je obrnjena).
- Spremenljivki EXSTB (ne moti pozornost) in EXSTP (sproščenost med ljudmi), ki kažeta na moške, tudi imata manjši vpliv.

Korelacije spremenljivk z DF

Structure Matrix

	Function
	1
EKSTLR Sem redkobeseden(a).	-,755
EKSTB Ne moti me, ce sem v središču pozornosti.	,318
EKSTIR Nerađ/a pritegnem pozornost nase.	,234
EKSTE Na zabavah se pomenkujem z mnogo ljudmi vseh vrst	-,189
EKSTH Pogovore nacenjam jaz.	-,150
EKSTNR V navzočnosti neznanih oseb sem molcec(a).	-,140
EKSTOR Imam malo povedati.	-,127
EKSTP Med ljudmi se pocutim sproščeno.	,127
EKSTA Sem duša vsake družbe.	-,054
EKSTRR Zadržujem se v ozadju.	,012

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

- Še vedno je najpomembnejša spremenljivka „Sem redkobeseden(a).“ (R), vrstni red ostalih spremenljivk pa je malce spremenjen, bolj podoben tistemu iz ANOVE

- Korelacije prikazujejo tako neposredne kot tudi posredne vplive in se zaradi koreliranosti spremenljivk razlikujejo od prejšnjih koeficientov.

Povprečja spremenljivk po skupinah

Functions at Group Centroids

E_SPOL spol ega	Function
	1
1 moški	,278
2 ženski	-,208

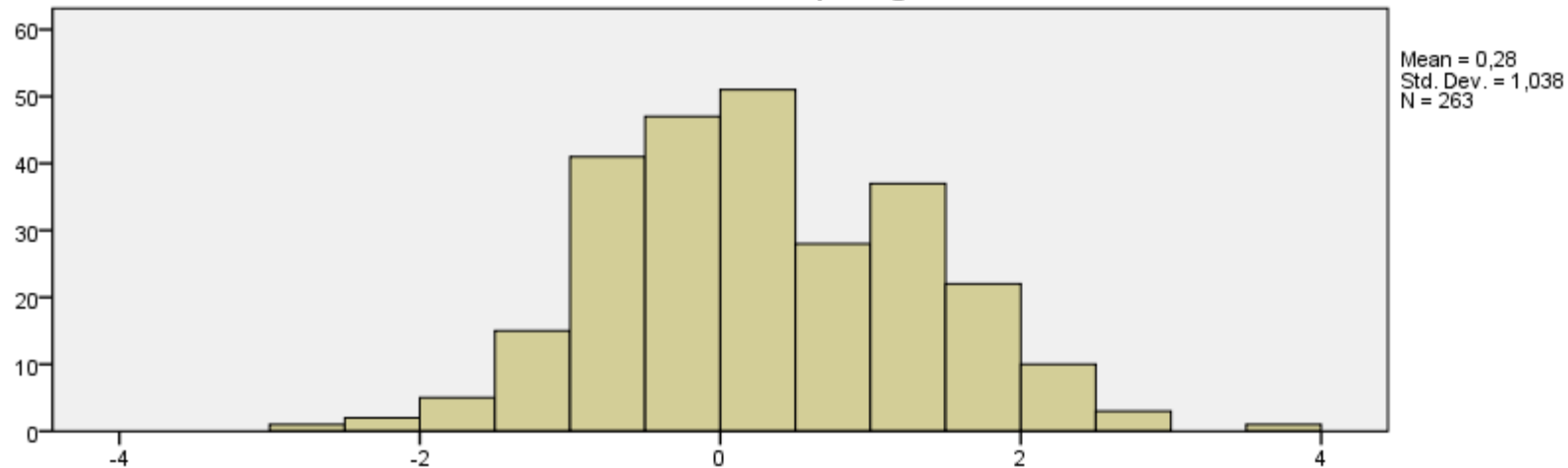
Unstandardized canonical
discriminant functions
evaluated at group means

- Moški imajo večje (pozitivno) povprečje, ženske pa manjše (negativno)
- Pomembno za interpretacijo prejšnjih tabel
- Pozitivni koeficienti kažejo, da večja vrednost spremenljivke kaže na moške, negativni pa da na ženske.

Porazdelitev vrednosti DF po skupinah

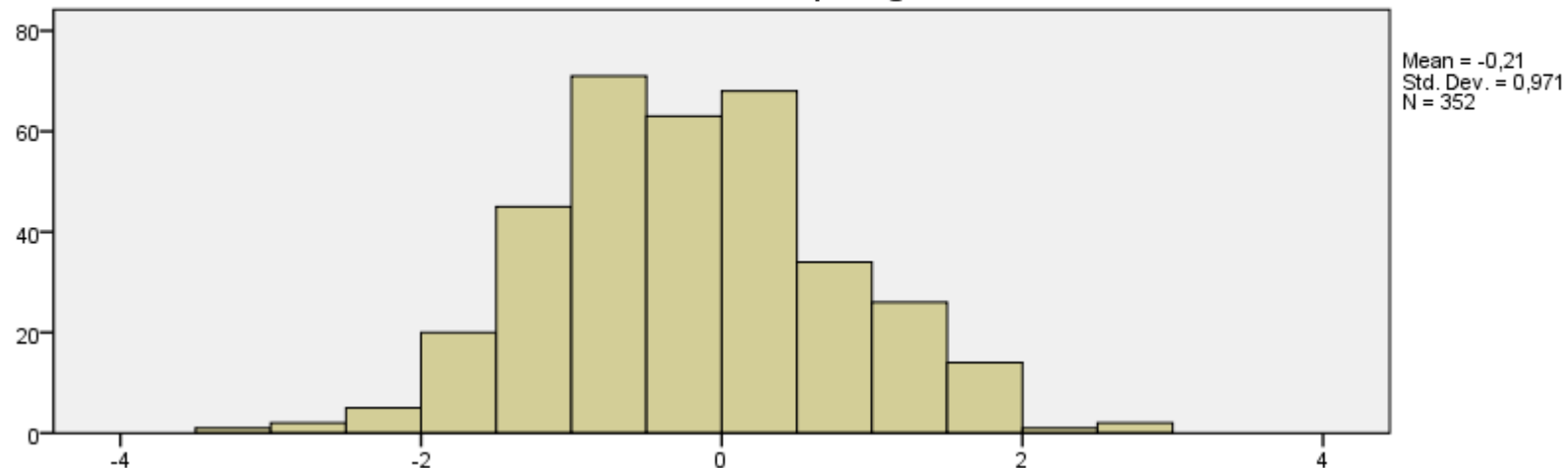
Canonical Discriminant Function 1

spol ega = moški



Canonical Discriminant Function 1

spol ega = ženski



Uvrščanje – enake začetne verjetnosti

Prior Probabilities for Groups

E_SPOL spol ega	Prior	Cases Used in Analysis	
		Unweighted	Weighted
1 moški	,500	263	263,000
2 ženski	,500	352	352,000
Total	1,000	615	615,000

- Tokrat smo predpostavili enake verjetnosti za skupine (spol) v populaciji

Fisherjeve klasifikacijske funkccije

Classification Function Coefficients

	E_SPOL spol ega	
	1 moški	2 ženski
EKSTA Sem duša vsake družbe.	,528	,506
EKSTB Ne moti me, če sem v središču pozornosti.	,687	,582
EKSTE Na zabavah se pomenkujem z mnogo ljudmi vseh vrst	,822	,863
EKSTH Pogovore nacenjam jaz.	,610	,631
EKSTIR Nerad/a pritegnem pozornost nase.	,880	,795
EKSTLR Sem redkobeseden(a).	,066	,421
EKSTNR V navzočnosti neznanih oseb sem molcec(a).	-,294	-,283
EKSTOR Imam malo povedati.	,782	,727
EKSTP Med ljudmi se počutim sproščeno.	3,462	3,310
EKSTRR Zadržujem se v ozadju.	-,130	-,189
(Constant)	-14,995	-14,845

Fisher's linear discriminant functions

- S pomočjo teh funkcije lahko izračunamo vrednosti funkcij za vsako skupino.
 - Enoto razvrstimo v skupino, ki ima večjo vrednost.
-
- Jih ne interpretiramo.

Klasifikacijska tabela – vse enote

Classification Results^{b,c}

E_SPOL spol ega			Predicted Group Membership		Total
			1 moški	2 ženski	
Original	Count	1 moški	149	114	263
		2 ženski	143	209	352
	%	1 moški	56,7	43,3	100,0
		2 ženski	40,6	59,4	100,0
Cross-validated ^a	Count	1 moški	139	124	263
		2 ženski	145	207	352
	%	1 moški	52,9	47,1	100,0
		2 ženski	41,2	58,8	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 58,2% of original grouped cases correctly classified.

c. 56,3% of cross-validated grouped cases correctly classified.

Pravilno je bilo uvrščenih:

- 56,7% moških
- 59,4% žensk
- 58,2% vseh enot

Klasifikacijska tabela – jackknife

Classification Results^{b,c}

E_SPOL spol ega			Predicted Group Membership		Total
			1 moški	2 ženski	
Original	Count	1 moški	149	114	263
		2 ženski	143	209	352
	%	1 moški	56,7	43,3	100,0
		2 ženski	40,6	59,4	100,0
Cross-validated ^a	Count	1 moški	139	124	263
		2 ženski	145	207	352
	%	1 moški	52,9	47,1	100,0
		2 ženski	41,2	58,8	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 58,2% of original grouped cases correctly classified.

c. 56,3% of cross-validated grouped cases correctly classified.

Pravilno je bilo uvrščenih:

- 52,9% moških
- 58,8% žensk
- 56,3% vseh enot

- Rezultati so malo slabši kot prej, saj računanju DF za uvrščanje določene enote nismo upoštevali te enote.
- Približno tako dobro pričakujemo, da bi se diskriminante funkcije obnese na populaciji.

Uvrščanje – začetne verjetnosti izračunane iz velikosti skupin

Prior Probabilities for Groups

E_SPOL spol ega	Prior	Cases Used in Analysis	
		Unweighted	Weighted
1 moški	,428	263	263,000
2 ženski	,572	352	352,000
Total	1,000	615	615,000

- Tokrat smo predpostavili, da so verjetnosti za skupine (spol) enake kot v vzorcu
- Pri spolu ta predpostavka verjetno ni najbolj primerna
- Posledica tega bo, da bodo relativno bolje (glede na prej) uvrščene enote iz skupin, ki imajo večjo verjetnost
- V te skupine bo uvrščeno tudi več enot
- Spremeni se samo klasifikacija (npr. DF pa ne)

Klasifikacijska tabela – vse enote

Classification Results^{b,c}

E_SPOL spol ega			Predicted Group Membership		Total
			1 moški	2 ženski	
Original	Count	1 moški	90	173	263
		2 ženski	69	283	352
	%	1 moški	34,2	65,8	100,0
		2 ženski	19,6	80,4	100,0
Cross-validated ^a	Count	1 moški	87	176	263
		2 ženski	73	279	352
	%	1 moški	33,1	66,9	100,0
		2 ženski	20,7	79,3	100,0

Pravilno je bilo uvrščenih:

- 34,2% moških
- 80,4% žensk
- 60,7% vseh enot

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 60,7% of original grouped cases correctly classified.

c. 59,5% of cross-validated grouped cases correctly classified.

- Uvrščanje moških se je poslabšalo (56,7 → 34,2), žensk pa izboljšalo (59,4 → 80,4), saj so bile začetne verjetnosti za ženske večje.
- Skupaj je rezultat malce boljši (58,2 → 60,7)

Klasifikacijska tabela – jackknife

Classification Results^{b,c}

E_SPOL spol ega			Predicted Group Membership		Total
			1 moški	2 ženski	
Original	Count	1 moški	90	173	263
		2 ženski	69	283	352
	%	1 moški	34,2	65,8	100,0
		2 ženski	19,6	80,4	100,0
Cross-validated ^a	Count	1 moški	87	176	263
		2 ženski	73	279	352
	%	1 moški	33,1	66,9	100,0
		2 ženski	20,7	79,3	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 60,7% of original grouped cases correctly classified.

c. 59,5% of cross-validated grouped cases correctly classified.

Pravilno je bilo uvrščenih:

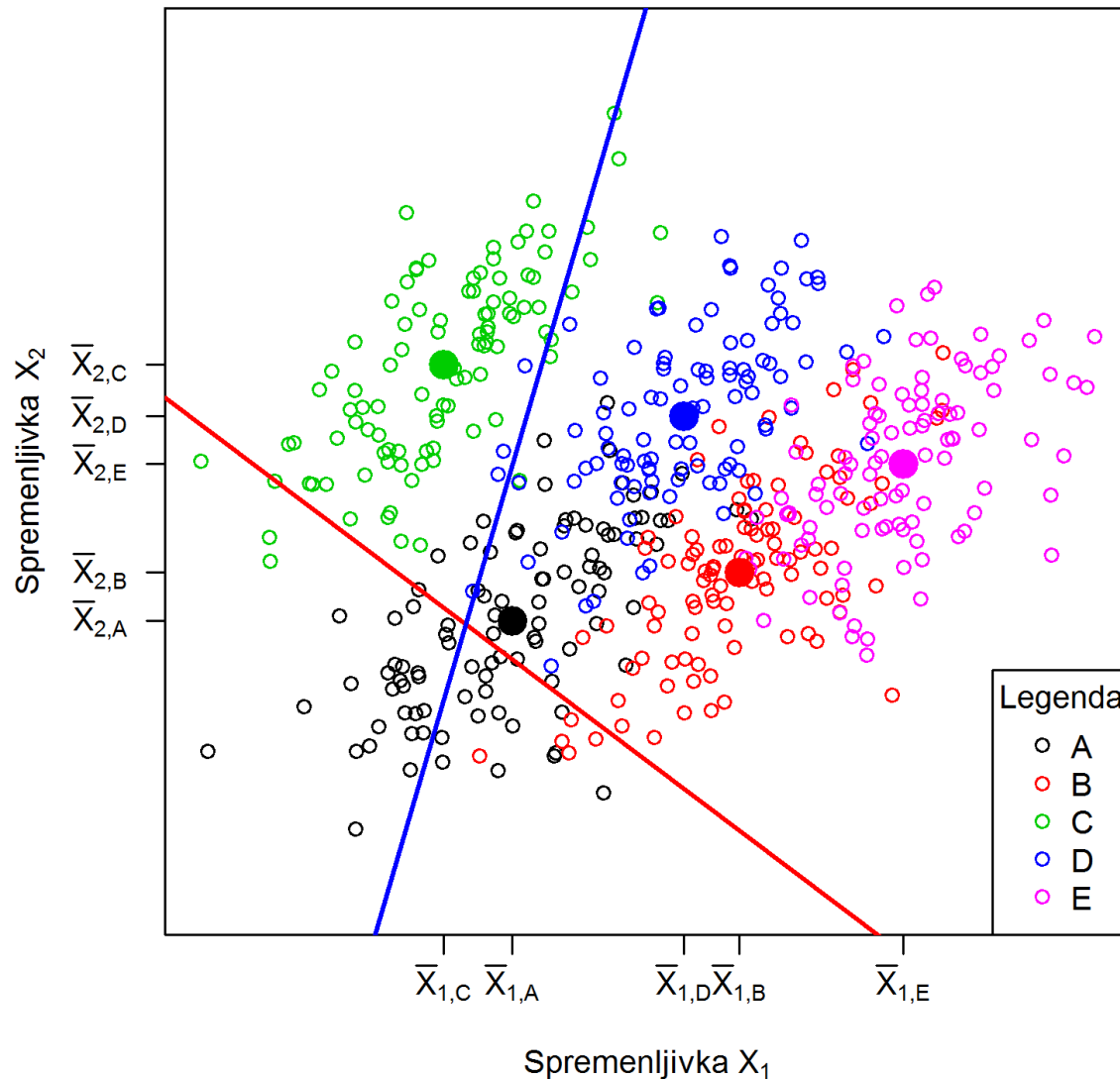
- 33,1% moških
- 79,3% žensk
- 59,5% vseh enot

- Rezultati so malo slabši kot pri originalnih enotah.
- V primerjavi z enakimi verjetnostmi so slabši pri moških, boljši pa pri ženskah in malo boljši, če upoštevamo vse enote skupaj.

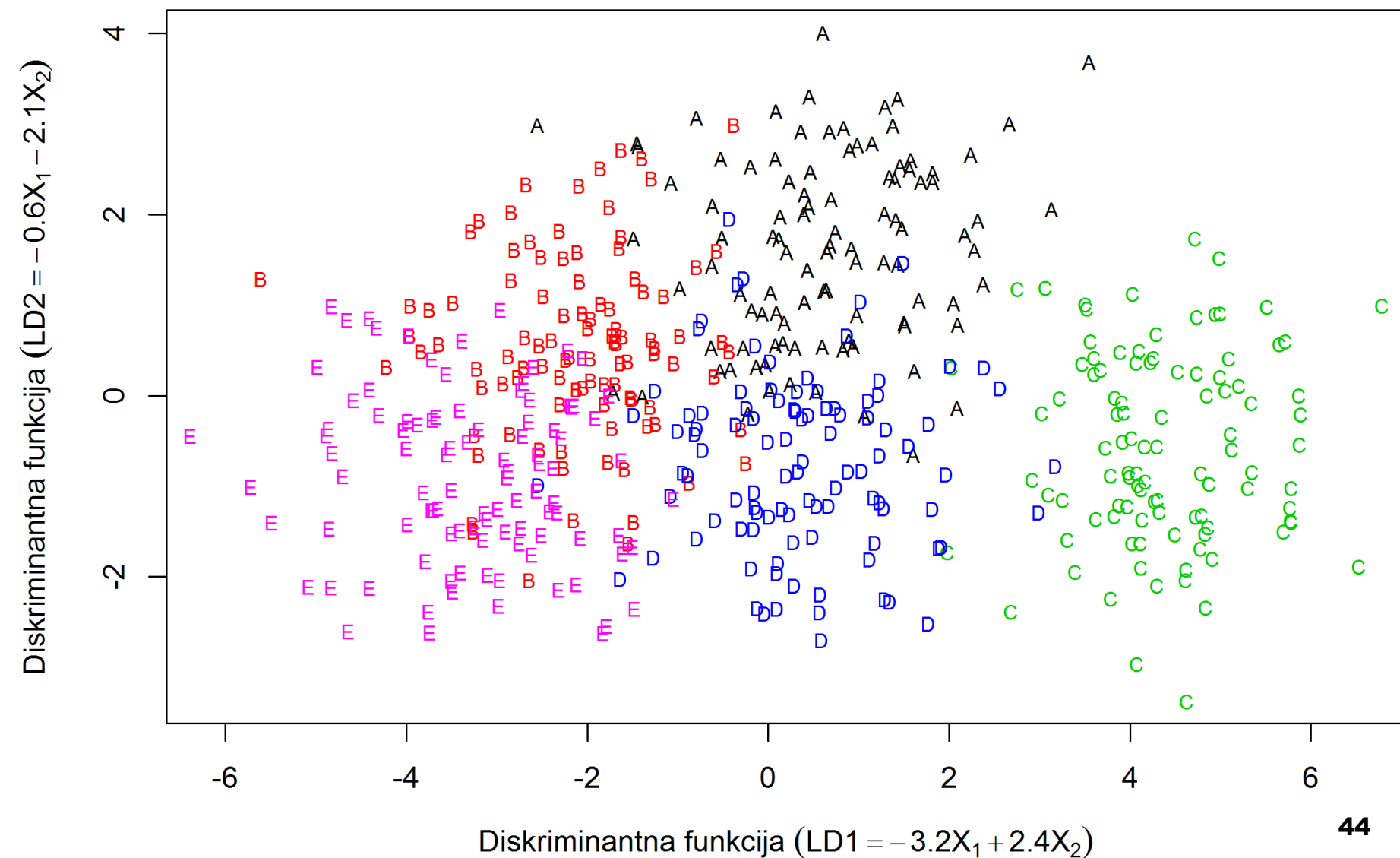
Diskriminantna analiza – več skupin

- V primeru več skupin razlike med skupinami popišemo z več diskriminantnimi spremenljivkami.
- Lahko jih je največ $\min(k - 1, p)$, kjer je k število skupin in p število spremenljivk.

Primer – več skupin



Enote v prostoru diskriminatnih funkcij



Postopek

- Predpostavimo, da imamo k skupin in v vsaki n_1, n_2, \dots, n_k enot.
- Matriko vsot kvadratov in produktov odklonov za i -to skupino pa označimo z W_i .
- Variabilnost znotraj skupin je potem:

$$W = W_1 + W_2 + \dots + W_k$$

- Variabilnost med skupinami pa izračunamo kot:

$$B = \frac{1}{k-1} \sum_{i=1}^k n \cdot p_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

, kjer je :

- \bar{x}_i vektor povprečij za i -to skupino,
 - p_i predhodna verjetnost za i -to skupino,
 - $\bar{x} = \sum_{i=1}^k p_i \bar{x}_i$ vektor skupnih povprečij.
- Zgornje formule veljajo za implementacijo R-u, v funkciji `lda{MASS}`.
- SPSS ne upošteva predhodnih verjetnosti pri ocenjevanju modela (le pri klasifikaciji enot). Formula za SPSS velja, če nastavimo $p_i = n_i/n$ ne glede na prave predhodne verjetnosti.

- Diskriminantni kriterij, ki ga je potrebno maksimizirati, je podoben kot v primeru dveh skupin (Fischerjev kriterij):

$$\frac{\text{variabilnost med skupinami}}{\text{variabilnost znotraj skupin}} = \max$$

- Varianca diskriminantne spremenljivke $Y = Xb$ je $b' \Sigma b$
- Variabilnost med skupinami je potem $b' B b$
- in variabilnost znotraj skupin $b' W b$
- Diskriminantni kriterij, ki ga je potrebno maksimizirati, je potem: $\frac{b' B b}{b' W b} = \lambda = \max$

- Najboljšo rešitev dobimo tako, da izračunamo lastne vrednosti in lastne vektorje matrike $W^{-1}B$.
- Lastne vrednosti so λ_i .
- Možno je dobiti $r = \min(k - 1, p)$ rešitev.
- Lastni vektorji (označimo jih z b) so koeficienti diskriminantnih funkcij (lahko jih je največ r).
- Normaliziramo jih tako, da velja:

$$b'Sb = I \text{ (identična matrika)}$$

Kjer je $S = W / (n - k)$ – Varianca znotraj skupin

- Največja lastna vrednost in pripadajoči lastni vektor, katerega elementi predstavljajo diskriminantne uteži, določata prvo diskriminatno spremenljivko.
- Relativna vrednost lastne vrednosti λ_i predstavlja indeks pomembnosti posamezne diskriminantne spremenljivke:

$$\frac{\lambda_i}{\sum_{j=1}^p \lambda_j}$$

Pravila uvrščanja enot v skupine

- Denimo, da smo izračunali diskriminantne spremenljivke $Y = Xb$ (Y je matrika, ki ima toliko stolpcev, kolikor diskriminantnih spremenljivk smo izračunali) oziroma
- $Y_i = Xb_i$, kjer je b_i vektor koeficientov za i -to diskriminanto spremenljivko
- Enoto uvrstimo (ob predlagani normalizaciji) v skupino, katere centroid (povprečje) v prostoru diskriminantnih spremenljivk ji je najbližje.

- Pravilo uvrščanja enote v optimalno skupino je ob enako velikih skupinah v populaciji je torej, da j -to enoto (glede na p izmerjenih spremenljivk) uvrstimo v k , za katero velja (v primeru r diskriminatnih spremenljivk):

$$D_k^2(y_j) = \sum_{i=1}^r (y_{j,i} - \bar{y}_{k,i})^2 \leq \sum_{i=1}^r (y_{j,i} - \bar{y}_{l,i})^2 = D_l^2(y_j)$$

, za vsak $l \neq k$, kjer je j oznaka enote, k in l pa sta oznaki skupin.


- V primeru predpostavke različno velikih skupin na populaciji (različnih „predhodnih“ (prior) verjetnosti) moramo te razdalje popraviti za te verjetnosti in sicer tako, da izberemo skupino k za katero velja:

$$\frac{1}{2} D_k^2(y_j) - \ln(p_k) \leq \frac{1}{2} D_l^2(y_j) - \ln(p_l)$$

, za vsak $l \neq k$, kjer je j oznaka enote, k in l pa sta oznaki skupin.

- Na podlagi teh razdalj lahko izračunamo tudi izračunamo verjetnosti, da posamezna enota pade v odločeno skupino in sicer:

$$D_k^*(y_i) =$$
$$= \left(\frac{1}{2} D_k^2(y_i) - \ln(p_k) \right) - \min_h \left(\frac{1}{2} D_h^2(y_j) - \ln(p_h) \right)$$
$$p_k(y_i) = \frac{e^{(-D_k^*(y_i))}}{\sum_h e^{(-D_h^*(y_i))}}$$

- 
- Enakovreden način uvrščanja so tudi Fisherjeve klasifikacijske linearne disriminatne funkcije, ki so še posebej priročne pri več skupinah.
 - Za vsako skupino se generira svoja klasifikacijska funkcija.
 - Enoto razvrstimo v tisto skupino, pri kateri ima največjo vrednost klasifikacijske funkcije.

Predhodne verjetnosti skupin (prior probabilities)

- Pri vseh implementacijah LDA na predhodne verjetnosti vplivajo na klasifikacijo enot (napovedi).
- Odvisno od implementacije pa lahko vplivajo tudi na izračun uteži (pri več kot 2 skupinah), saj se lahko upoštevajo tudi pri izračunu skupnega povprečja in posledično matrik T in B . Npr., SPSS jih ne upošteva, R (funkcija `lda` iz `MASS`) pa jih.
- Če so to prave verjetnosti skupin v populaciji → jih moramo upoštevati
- Če jih želimo uporabiti le kot „utež“ pri uvrščanju enot (npr. zaradi različne pomembnosti različnih vrst „napak“) → bolje, da jih ne upoštevamo

Zveza med diskriminantno analizo in kanonično korelacijsko analizo

- Iz nominalne spremenljivke, ki določa k skupin v diskriminantni analizi, naredimo $(k - 1)$ 'dummy' (umentih) spremenljivk.
- Po drugi strani imamo p merjenih spremenljivk. Za ti dve skupini spremenljivk lahko izračunamo $\min(k - 1, p)$ kanoničnih rešitev.
- Te dobimo z izračunom lastnih vrednosti in lastnih vektorjev matrike $\Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX}$.
- Lastne vrednosti označimo z λ_j^{kka} .

- Kot vemo imamo v primeru diskriminantne analize imamo k skupin in p merjenih spremenljivk.
- Za oceno diskriminatnih spremenljivk računamo lastne vrednosti in lastne vektorje matrike $W^{-1}B$.
- Lastne vrednosti označimo z λ_j^{da} .

- Velja zveza:

$$\lambda_j^{da} = \frac{\lambda_j^{kka}}{1 - \lambda_j^{kka}} (/100)$$

- Enake so tudi dobljene uteži (pred normalizacijo v diskriminatni analizi)

Ocenjevanje kvalitete modela

- Kvaliteto modela zopet ocenjujemo enako kot v primeru dveh skupin.
- Za (skoraj) nepristransko oceno moramo uporabiti za oceno uporabiti druge podatke kot smo jih uporabili za ocenjevanje modela
→ Jacckknife
- Upoštevamo, da bi „na slepo“ pravilno razvrstili tak delež enot, kolikor je verjetnost največje skupine (če so skupine enako verjetne $p = \frac{1}{k}$).
- Pri več skupinah lahko torej pričakujemo manjši delež pravilno uvrščenih enot.

Primer

- Podatki so bili zbrani v okviru raziskave *Kakovost merjenja egocentričnih socialnih omrežij* (Ferligoj in drugi, 2000) leta 2000. Vzorec vsebuje 1033 prebivalcev Ljubljane. Analiza je bila narejena na 631 prebivalcih, ki so bili osebno intervjuvani.
- Na primeru bomo poskusili ugotoviti, katere razsežnosti emocionalne stabilnosti najbolj pojasnjujejo razlike med zakonskimi stanovi (samski, izvenzakonska zveza, poročen, ločen, vdovec).

ANOVE po vseh spremenljivkah

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
EMOCC Redkokdaj sem potr(a).	,994	,995	4	618	,410
EMOCDR Zlahka me kaj vrže iz tira.	,993	1,106	4	618	,353
EMOCF Sem vecidel sproščen(a).	,994	,884	4	618	,473
EMOCGR Zlahka me kaj razdraži.	,998	,276	4	618	,894
EMOCJR Zlahka me kaj vznemiri.	,994	,901	4	618	,463
EMOCKR Sem zaskrbljene narave.	,963	5,855	4	618	,000
EMOCMR Velikokrat sem muhasto razpoložen(a).	,974	4,202	4	618	,002
EMOCQR Moje razpoloženje se pogosto menja.	,985	2,285	4	618	,059
EMOCSR Pogosto sem potr(a).	,963	5,912	4	618	,000
EMOCTR Zlahka se me poloti napetost.	,990	1,512	4	618	,197

Če med skupinami pri posamezni spremenljivki ni statistično značilnih razlik, ne moremo pričakovati, da bo dobro razlikovala med skupinami.

Enakost kovariančnih matrik

Box's Test of Equality of Covariance Matrices

Log Determinants

D10 zakonski stan	Rank	Log Determinant
1 samski-a	10	3,187
2 izvenzakonska zveza	10	4,008
3 poročen	10	4,685
4 ločen-a	10	2,081
5 vdovec, vdova	10	3,888
Pooled within-groups	10	4,421

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		321,406
F	Approx.	1,380
	df1	220
	df2	119589,541
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

- Box-ov M testira domnevo (eno izmed predpostavk), da sta kovariančni matriki na populaciji enaki
- Ničelno domnevo o enakosti kovariančnih matrik lahko zavrnamo pri tveganju $< 0,05\%$ → predpostavka je kršena.
- Razmisliti bi bilo potrebno o uporabi kvadratne DA (mi ne bomo)

Kanonične diskriminantne funkcije

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,130 ^a	65,5	65,5	,339
2	,045 ^a	22,6	88,1	,207
3	,019 ^a	9,5	97,7	,136
4	,005 ^a	2,3	100,0	,068

a. First 4 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 4	,827	116,481	40	,000
2 through 4	,935	41,329	27	,038
3 through 4	,977	14,361	16	,572
4	,995	2,827	7	,901

- Prva kanonična korelacija med „skupinami“ in „neodvisnimi“ spremenljivkami je 0,339, druga pa 0,207.
- Samo prvi dve diskriminantni funkciji statistično značilno razlikujeta med skupinami

Koeficienti diskriminatne funkcije (DF)

Canonical Discriminant Function Coefficients

	Function			
	1	2	3	4
EMOCC Redkokdaj sem potr(a).	,127	,131	,096	,293
EMOCCR Zlahka me kaj vrže iz tira.	-,060	-,335	,225	,457
EMOCF Sem vecidel sproščen(a).	,222	-,017	,471	-,120
EMOCGR Zlahka me kaj razdraži.	,182	,027	,232	-,384
EMOCJR Zlahka me kaj vznemiri.	-,116	,039	-,547	,421
EMOCKR Sem zaskrbljene narave.	-,321	-,335	-,255	-,164
EMOCMR Velikokrat sem muhasto razporežen(a).	,348	-,127	-,254	-,197
EMOCQR Moje razporeženje se pogosto menja.	,333	,181	-,132	,206
EMOCSR Pogosto sem potr(a).	-,472	,715	,024	-,166
EMOCTR Zlahka se me poloti napetost.	-,173	-,043	,470	-,125
(Constant)	-,354	-1,410	-1,742	-,540

Unstandardized coefficients

- Surovi (nestandardizirani) koeficienti niso primerljivi med spremenljivkami z različnimi lestvicami oz. variabilnostjo, zato jih običajno ne interpretiramo.

- Za interpretacijo uporabljamo standardizirane koeficiente ali korelacije s DF (nekateri odsvetujejo)
- Nestandardizirani so uporabni za izračun vred. DF₆₄

Koeficienti diskriminatne funkcije (DF)

Canonical Discriminant Function Coefficients

	Function			
	1	2	3	4
EMOCC Redkokdaj sem potr(a).	,127	,131	,096	,293
EMOCDR Zlahka me kaj vrže iz tira.	-,060	-,335	,225	,457
EMOCF Sem vecidel sproščen(a).	,222	-,017	,471	-,120
EMOCGR Zlahka me kaj razdraži.	,182	,027	,232	-,384
EMOCJR Zlahka me kaj vznemiri.	-,116	,039	-,547	,421
EMOCKR Sem zaskrbljene narave.	-,321	-,335	-,255	-,164
EMOCMR Velikokrat sem muhasto razporežen(a).	,348	-,127	-,254	-,197
EMOCQR Moje razpoloženje se pogosto menja.	,333	,181	-,132	,206
EMOCSR Pogosto sem potr(a).	-,472	,715	,024	-,166
EMOCTR Zlahka se me poloti napetost.	-,173	-,043	,470	-,125
(Constant)	-,354	-1,410	-1,742	-,540

Unstandardized coefficients

- Največji negativni koeficient pri prvi DF ima spremenljivka „Pogosto sem potr. (R)“ (EMOCSR), kar velike pa še EMOCMR, EMOCQR in EMOCKR(-).

- Pri drugi DF so najpomembnejše spremenljivke EMOCSR (++) , EMOCKR(-) in EMOCDR(-)
- **Opozorilo:** Jaz tu uporabljam oznake zaradi pomanjkanja prostora.

Standardizirani koeficienti

Standardized Canonical Discriminant Function Coefficients

	Function			
	1	2	3	4
EMOCC Redkokdaj sem potr(a).	,183	,190	,138	,424
EMOCDR Zlahka me kaj vrže iz tira.	-,096	-,535	,359	,728
EMOCF Sem vecidel sproščen(a).	,246	-,019	,522	-,132
EMOCGR Zlahka me kaj razdraži.	,276	,041	,353	-,584
EMOCJR Zlahka me kaj vznemiri.	-,174	,059	-,824	,634
EMOCKR Sem zaskrbljene narave.	-,504	-,525	-,400	-,257
EMOCMR Velikokrat sem muhasto razpoložen(a).	,505	-,184	-,368	-,286
EMOCQR Moje razpoloženje se pogosto menja.	,503	,273	-,199	,312
EMOCSR Pogosto sem potr(a).	-,594	,899	,030	-,208
EMOCTR Zlahka se me poloti napetost.	-,249	-,062	,677	-,181

- 1. DF definirajo predvsem spremenljivke o (ne) zaskrbljenosti in potrtosti (-) in (ne) menjavanju razpoloženja (+).

- Pri 2. je najpomembnejša spremenljivka (ne) potrtost (+), ki jih sledita (ne) zaskrbljenost in „vrže iz tira“ (-)

Korelacije spremenljivk z DF

Structure Matrix

	Function			
	1	2	3	4
EMOCKR Sem zaskrbljene narave.	-,514*	-,269	-,114	-,052
EMOCMR Velikokrat sem muhasto razpoložen(a).	,446*	,053	-,244	-,100
EMOCSR Pogosto sem potr(a).	-,375	,665*	,047	,120
EMOCF Sem vecidel sproščen(a).	,066	,071	,510*	,016
EMOCTR Zlahka se me poloti napetost.	-,231	-,012	,384*	,117
EMOCDR Zlahka me kaj vrže iz tira.	-,098	-,248	,249	,655*
EMOCJR Zlahka me kaj vznemiri.	-,168	,003	-,216	,526*
EMOCC Redkokdaj sem potr(a).	,073	,308	,182	,434*
EMOCQR Moje razpoloženje se pogosto menja.	,280	,305	-,029	,310*
EMOCGR Zlahka me kaj razdraži.	,096	-,077	,112	,145*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

- S 1. DF najbolj korelirajo spremenljivke o (ne) zaskrbljenosti in potrtosti (-) in o (ne) muhavosti (+)
- Pri 2. DF je podobno kot prej (+ redka potrtost in menjavanje razpoloženja)

- Opomnik: korelacije prikazujejo neposredne in posredne učinke.

Povprečja spremenljivk po skupinah

Functions at Group Centroids

D10 zakonski stan	Function			
	1	2	3	4
1 samski-a	-,396	-,117	,013	-,062
2 izvenzakonska zveza	-,362	-,017	,141	,167
3 poročen	,151	,219	-,042	-,005
4 ločen-a	,317	-,424	-,357	,066
5 vdovec, vdova	,734	-,267	,261	-,028


Unstandardized canonical discriminant functions evaluated at group means

- Prva DF loči med samskimi in „izvenzakonskimi zvezami“ na eni strani in ločenimi ter še posebej vdovci na drugi (poročeni so vmes).
- 2. DF loči med poročenimi na eni strani in vdovci ter še posebej ločenimi na drugi (ostali so vmes).

Skupna interpretacija

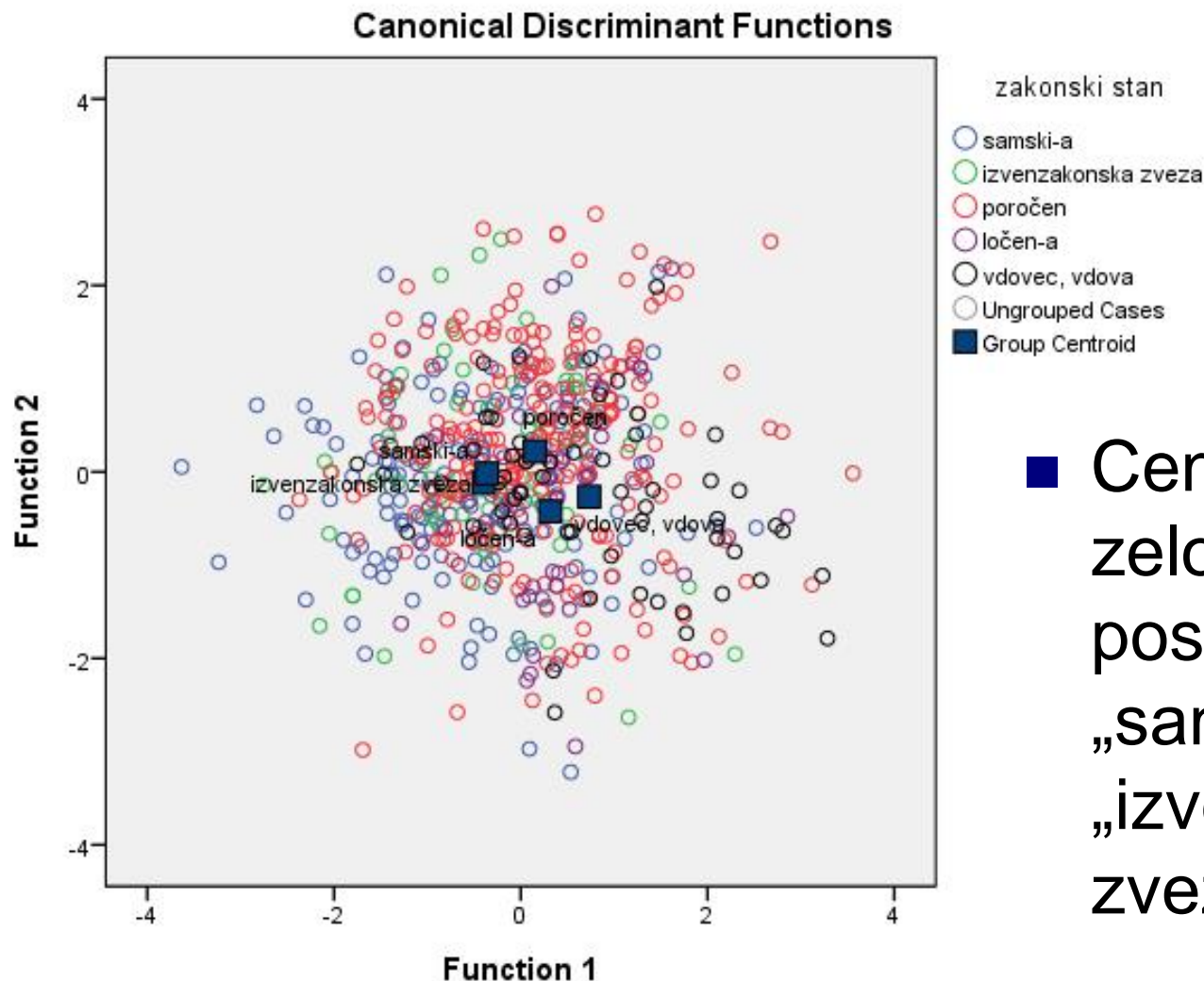
- Pri interpretacij DF se osredotočimo na prvi dve, saj samo ti dve stat. znač. ločita med skupinami.
- Prva (DF1) loči skoraj nekako po starosti oz. po možnem „poteku“ samski→izvenzakonska zveza→poročen→(ločen ali vdovec)
- Zanimivo je, da je korelacija med DF1 in starostjo statistična značilna ($p < 0,05\%$) in znaša 0,32.

- Ob nespremenjenih ostalih spremenljivkah najbolj proti „zgodnjim“ stanjem kaže spremenljivka „Sem pogosto potr. (R)“ in „Sem zaskrbljene narave. (R)“, proti „poznejšim“ pa „Velikokrat sem muhasto razpoložen(a). (R)“ in „Moje razpoloženje se pogosto menja. (R)“
- Če pa gledamo spremenljivke brez kontroliranja za ostale spremenljivke je interpretacija podobna, le vrstni red spremenljivk je malce spremenjen.

- 
- Druga DF (DF2) najbolj loči med poročenimi in ločenimi, s tem da so najbližje ločenim vdovci (nekako „ločeni kot posledica smrti“), najbližje poročenim pa tisti v izvenzakonski zvezi. Samski so na sredini.
 - Torej nekako loči tiste, ki živijo z nekom od tistih, ki ne, kjer so najbolj ločeni tisti, ki so se jasno odločili (ali njihov partner) za eno izmed možnosti (poročili/ločili).

- Ob nespremenjenih ostalih spremenljivkah najbolj proti „poročenim“ stanjem kaže spremenljivka „Pogosto sem potr. (R)“ (zelo močno) in v veliko manjši meri „Moje razpoloženje se pogosto menja. (R)“, proti „ločenim“ pa „Zlahka me kaj vrže iz tira. (R)“ in „Sem zaskrbljene narave. (R)“
- Če pa gledamo spremenljivke brez kontroliranja za ostale spremenljivke je interpretacija podobna, le kakšna podobna spremenljivka pride zraven.
- Torej, „poročeni“ so v primerjavi z „ločenimi“ so manj potrti, a bolj „razdražljivi“ in zaskrbljeni.

Enote v prostoru DF



- Centoridi so zelo skupaj, še posebej pa „samski“ in „izvenzakonska zveza“

Uvrščanje – enake začetne verjetnosti

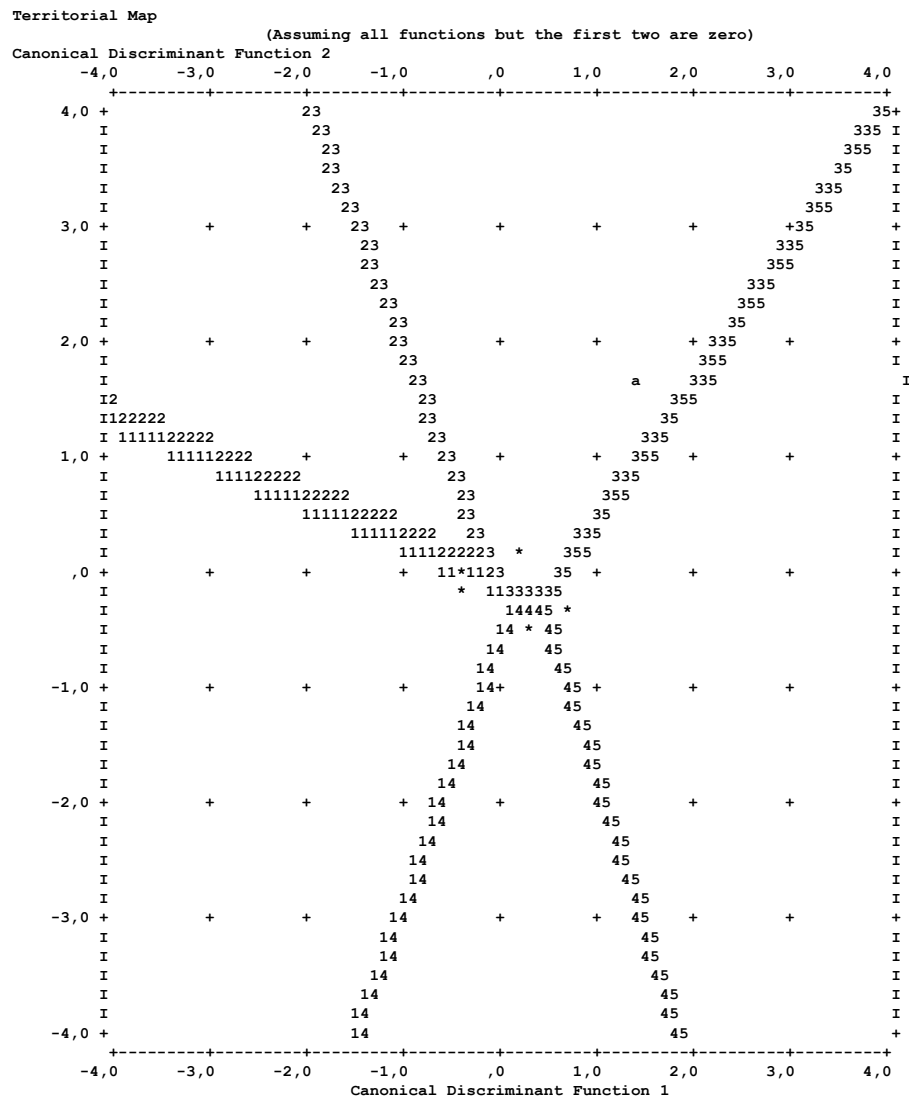
Prior Probabilities for Groups

D10 zakonski stan	Prior	Cases Used in Analysis	
		Unweighted	Weighted
1 samski-a	,200	185	185,000
2 izvenzakonska zveza	,200	68	68,000
3 poročen	,200	265	265,000
4 ločen-a	,200	46	46,000
5 vdovec, vdova	,200	59	59,000
Total	1,000	623	623,000

- Tokrat smo predpostavili enake verjetnosti za skupine (zakonski stan) v populaciji oz. smo skupine obravnavali enakovredno.
- Ker so med skupinami velike razlike, bi bilo smiselno predpostaviti tudi različne skupine, saj vzorec verjetno predstavlja tudi porazdelitev v populaciji

„Zemljevid“ prostora diskriminativnih funkcij

Kam bi razvrstili
posamezno enoto



Symbols used in territorial map

Symbol	Group	Label
1	1	samski-a
2	2	izvenzakonska zveza
3	3	poročen
4	4	ločen-a
5	5	vdovec, vdova
*		Indicates a group centroid

Klasifikacijska tabela – vse enote

Classification Results^{b,c}

Type=Original

D10 zakonski stan		Predicted Group Membership					Total
		1 samski-a ^a	2 izvenzakonska zveza ^a	3 poročen ^a	4 ločen-a ^a	5 vdovec, vdova ^a	
Count	1 samski-a	67	37	27	33	21	185
	2 izvenzakonska zveza	13	19	17	11	8	68
	3 poročen	43	57	75	43	47	265
	4 ločen-a	8	6	7	15	10	46
	5 vdovec, vdova	3	13	11	11	21	59
	Ungrouped cases	0	1	0	0	0	1
%	1 samski-a	36,2	20,0	14,6	17,8	11,4	100,0
	2 izvenzakonska zveza	19,1	27,9	25,0	16,2	11,8	100,0
	3 poročen	16,2	21,5	28,3	16,2	17,7	100,0
	4 ločen-a	17,4	13,0	15,2	32,6	21,7	100,0
	5 vdovec, vdova	5,1	22,0	18,6	18,6	35,6	100,0
	Ungrouped cases	,0	100,0	,0	,0	,0	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 31,6% of original grouped cases correctly classified.

c. 27,9% of cross-validated grouped cases correctly classified.

Klasifikacijska tabela – jackknife

Classification Results^{b,c}

Type=Cross-validated

D10 zakonski stan		Predicted Group Membership					Total
		1 samski-a ^a	2 izvenzakonska zveza ^a	3 poročen ^a	4 ločen-a ^a	5 vdovec, vdova ^a	
Count	1 samski-a	60	42	27	35	21	185
	2 izvenzakonska zveza	19	11	17	12	9	68
	3 poročen	43	57	71	45	49	265
	4 ločen-a	9	7	8	11	11	46
	5 vdovec, vdova	3	13	11	11	21	59
%	1 samski-a	32,4	22,7	14,6	18,9	11,4	100,0
	2 izvenzakonska zveza	27,9	16,2	25,0	17,6	13,2	100,0
	3 poročen	16,2	21,5	26,8	17,0	18,5	100,0
	4 ločen-a	19,6	15,2	17,4	23,9	23,9	100,0
	5 vdovec, vdova	5,1	22,0	18,6	18,6	35,6	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 31,6% of original grouped cases correctly classified.

c. 27,9% of cross-validated grouped cases correctly classified.

Uvrščanje – začetne verjetnosti izračunane iz velikosti skupin

Prior Probabilities for Groups

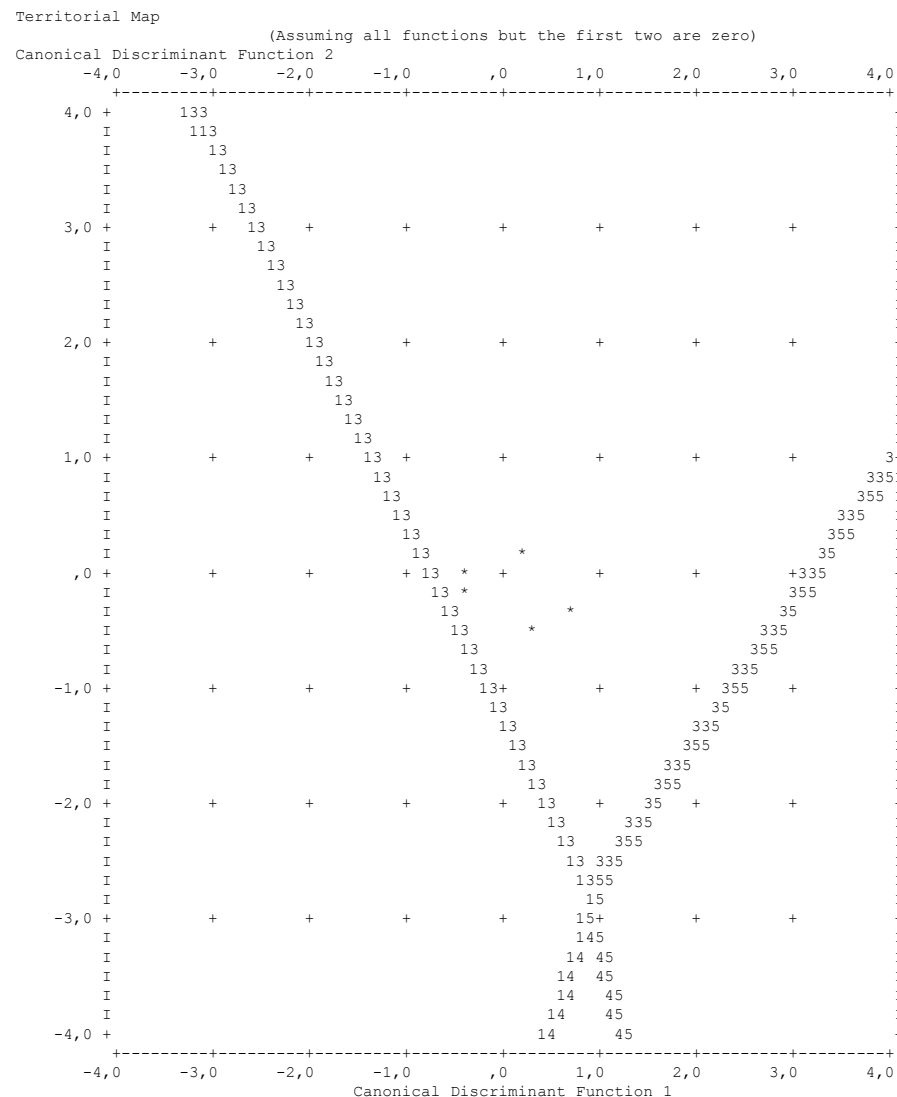
D10 zakonski stan	Prior	Cases Used in Analysis	
		Unweighted	Weighted
1 samski-a	,297	185	185,000
2 izvenzakonska zveza	,109	68	68,000
3 poročen	,425	265	265,000
4 ločen-a	,074	46	46,000
5 vdovec, vdova	,095	59	59,000
Total	1,000	623	623,000

- Tokrat smo predpostavili, da so verjetnosti za skupine enake kot v vzorcu
- Pri zakonskem stanu (in tem vzorcu) je to smiselno.
- Posledica tega bo, da bodo relativno bolje (glede na prej) uvrščene skupine, ki imajo večjo verjetnost
- V te skupine bo uvrščeno tudi več enot
- Spremeni se samo klasifikacija (npr. DF pa ne)

„Zemljevid“ prostora diskriminativnih funkcij

Kam bi razvrstili
posamezno enoto

- Ker se spremenijo
pravila za uvrščanje, se
spremeni tudi
„zemljevid“



Symbols used in territorial map

Symbol	Group	Label
1	1	samski-a
2	2	izvenzakonska zveza
3	3	poročen
4	4	ločen-a
5	5	vdovec, vdova
*		Indicates a group centroid

Klasifikacijska tabela – vse enote

Classification Results^{b,c}

Type=Original

D10 zakonski stan		Predicted Group Membership					Total
		1 samski-a ^a	2 izvenzakonska zveza ^a	3 poročen ^a	4 ločen-a ^a	5 vdovec, vdova ^a	
Count	1 samski-a	83	0	100	0	2	185
	2 izvenzakonska zveza	27	0	38	0	3	68
	3 poročen	48	0	207	1	9	265
	4 ločen-a	12	0	32	1	1	46
	5 vdovec, vdova	8	0	43	1	7	59
	Ungrouped cases	1	0	0	0	0	1
%	1 samski-a	44,9	,0	54,1	,0	1,1	100,0
	2 izvenzakonska zveza	39,7	,0	55,9	,0	4,4	100,0
	3 poročen	18,1	,0	78,1	,4	3,4	100,0
	4 ločen-a	26,1	,0	69,6	2,2	2,2	100,0
	5 vdovec, vdova	13,6	,0	72,9	1,7	11,9	100,0
	Ungrouped cases	100,0	,0	,0	,0	,0	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 47,8% of original grouped cases correctly classified.

c. 44,9% of cross-validated grouped cases correctly classified.

Klasifikacijska tabela – jackknife

Classification Results^{b,c}

Type=Cross-validated

D10 zakonski stan		Predicted Group Membership					Total
		1 samski-a ^a	2 izvenzakonska zveza ^a	3 poročen ^a	4 ločen-a ^a	5 vdovec, vdova ^a	
Count	1 samski-a	78	0	104	0	3	185
	2 izvenzakonska zveza	27	0	38	0	3	68
	3 poročen	56	0	197	2	10	265
	4 ločen-a	13	0	32	0	1	46
	5 vdovec, vdova	8	0	45	1	5	59
%	1 samski-a	42,2	,0	56,2	,0	1,6	100,0
	2 izvenzakonska zveza	39,7	,0	55,9	,0	4,4	100,0
	3 poročen	21,1	,0	74,3	,8	3,8	100,0
	4 ločen-a	28,3	,0	69,6	,0	2,2	100,0
	5 vdovec, vdova	13,6	,0	76,3	1,7	8,5	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 47,8% of original grouped cases correctly classified.

c. 44,9% of cross-validated grouped cases correctly classified.

Klasifikacija – primerjava

- Z upoštevanjem velikosti skupin se sicer % pravilno uvrščenih enot zelo poveča (31,6→47,8 originalno oz. 27,9→44,9 jackknife).
- Vendar pa je pri vseh skupinah enoti največji delež enot pri upoštevanju velikosti skupin uvrščen med „poročene“, ker so daleč največja skupina (42,5% vseh enot).
- V skupino „Izvenzakonska zveza“ ni celo uvrščen nihče, v skupino „ločen“ pa le 3 enote.

Primer 2: Napovedni problem

- Ali lahko napovemo zakonski stan na podlagi spola, starosti, emocionalne stabilnosti (likert), ekstrovertiranost (likert) in izobrazbe (kot numerična spremenljivka)
- Predpostavili bomo take deleže kot v vzorcu
- Uporabljene funkcije so v datoteki [LDAandCChelperFunctions.R](#) (povezava deluje le, če ste prijavljeni v spletno učilnico).

Enakost kovariančnih matrik

- Preverimo predpostavko o enakosti kovariančnih matrik

```
BoxMTest(X=ego[cmp,neodSprem], cl=ego$D10[cmp], alpha=0.05,  
test="F")
```

MBox	F	df1	df2	p
155.5433	1.7872	84	117937.39	0.0000

Covariance matrices are significantly different.

- Očitno kovariančne matrike po skupinah niso enake.

Linearna diskriminantna analiza

```
ldaZakStan<-ldaPlus(x=ego[cmp,neodSprem],  
grouping = ego$D10[cmp],CV = TRUE,  
pred=TRUE)
```

- Funkcija `ldaPlus` je nadgradnja funkcije `lda {MASS}` z dodatnimi izpisi in dodatno možnostjo.
- Dodatna možnost: `usePriorBetweenGroups=TRUE` (privzeta vrednost). `TRUE` (privzeta vrednost) da enak rezultat kot `lda {MASS}`, če pa jo nastavimo na `FALSE`, predhodne verjetnosti (prior) uporablja le za uvrščanje/klasificiranje enot in ne za iskanje rešitve (kot nekateri drugi statistični paketi, npr. SPSS)

Kanonične diskriminantne funkcije

`ldaZakStan$eigModel`

`ldaZakStan$sigTest`

Eigenvalues	%	Cum %	Cor	Sq. Cor
1.083	91.706	91.7	0.721	0.520
0.052	4.382	96.1	0.222	0.049
0.039	3.320	99.4	0.194	0.038
0.007	0.592	100.0	0.083	0.007

	WilksL	F	df1	df2	p
1 to 4	0.436	23.30	24	2084	0.000
2 to 4	0.909	3.89	15	1651	0.000
3 to 4	0.956	3.44	8	1198	0.001
4 to 4	0.993	1.40	3	600	0.243

- Prva kanonična korelacija je izrazito največja, kar 0,721 in pojasni 91,7 razlik med povprečji
- Prve tri DF statistično značilno razlikujejo med skupinami

Originalni keoficienti

`ldaZakStan$scaling`

	LD1	LD2	LD3	LD4
moski	-0.389	1.840	0.911	-0.072
STAROST	0.077	0.006	0.011	-0.006
ekstLikert	0.031	0.096	0.272	-0.843
emocLikert	-0.036	-0.075	0.074	0.713
VEL_OM	0.006	0.071	-0.093	-0.176
izoNum	0.010	0.302	-0.645	0.090

- Zaradi različnih merskih lestvic spremenljivk jih težko interpretiramo, zato raje interpretiramo standardizirane.

Standardizirani koeficienti - ver1

`ldaZakStan$standCoefTotal`

	LD1	LD2	LD3	LD4
moski	-0.192	0.911	0.451	-0.036
STAROST	1.427	0.108	0.196	-0.117
ekstLikert	0.025	0.076	0.214	-0.663
emocLikert	-0.032	-0.066	0.065	0.629
VEL_OM	0.018	0.215	-0.280	-0.534
izoNum	0.014	0.404	-0.863	0.121

- Standardizirani s skupnimi sd-ji, kot da bi izvedli funkcijo na standardiziranih spremenljivkah
- 1. DF definira predvsem starost
- Pri 2. je najpomembnejša spremenljivka spol, deloma je pomembna tudi izobrazba
- Pri tretji je najpomembnejša (-) izobrazba, sledi (+) spol

Standardizirani koeficienti – ver2 (SPSS)

ldaZakStan\$standCoefWithin

	LD1	LD2	LD3	LD4
moski	-0.188	0.888	0.440	-0.035
STAROST	1.002	0.076	0.137	-0.082
ekstLikert	0.024	0.075	0.212	-0.656
emocLikert	-0.032	-0.067	0.066	0.630
VEL_OM	0.018	0.215	-0.280	-0.533
izoNum	0.014	0.398	-0.850	0.119

- Standardizirani s „pooled“ sd-ji znotraj skupin (kot SPSS)
- Vsebinsko se interpretacija ne spremeni (ni pa vedno tako)
- 1. DF definira predvsem starost
- Pri 2. je najpomembnejša spremenljivka spol, deloma je pomembna tudi izobrazba
- Pri tretji je najpomembnejša (-) izobrazba, sledi (+) spol

Korelacije spremenljivk z DF

ldaZakStan\$corr

	LD1	LD2	LD3	LD4
moski	-0.104	0.882	0.428	0.139
STAROST	0.980	0.160	0.095	0.064
ekstLikert	-0.155	0.035	0.089	-0.537
emocLikert	-0.026	0.143	-0.022	0.508
VEL_OM	0.029	0.145	-0.357	-0.593
izoNum	0.055	0.454	-0.802	0.183

- S 1. DF korelira predvsem starost
- Pri 2. DF je podobno kot prej (++ spol, + izobrazba)
- Pri 3. DF spet podobno kot prej (-- izobrazba, + spol)
- Opomnik: korelacije prikazujejo neposredne in posredne učinke.

Povprečja spremenljivk po skupinah


ldaZakStan\$centroids

	LD1	LD2	LD3	LD4
samski-a	-1.213	-0.117	0.111	-0.050
izvenzakonska zveza	-0.879	0.003	-0.132	0.223
poročen	0.533	0.230	-0.001	-0.018
ločen-a	0.574	-0.336	-0.586	-0.078
vdovec, vdova	2.124	-0.425	0.286	0.054

- 1. DF loči med samskimi in „izvenzakonskimi zvezami“ na eni strani in poročenimi, ločenimi ter še posebej vdovci na drugi.
- 2. DF loči med poročenimi na eni strani in vdovci ter ločenimi na drugi (ostali so vmes).
- 3. DF loči predvsem med vdovci in ločenimi.

Skupna interpretacija

- Pri interpretacij DF se osredotočimo predvsem na prvo, ki je najpomembnejša, podamo pa še za 2. in 3., ki tudi pri 5-% tveganju stat. znač. ločita med skupinami.
- Prva (DF1) je skoraj definirana s starostjo in logično uredi skupine po „poteku“ življenja: samski → izvenzakonska zveza → poročen → (ločen ali vdovec)

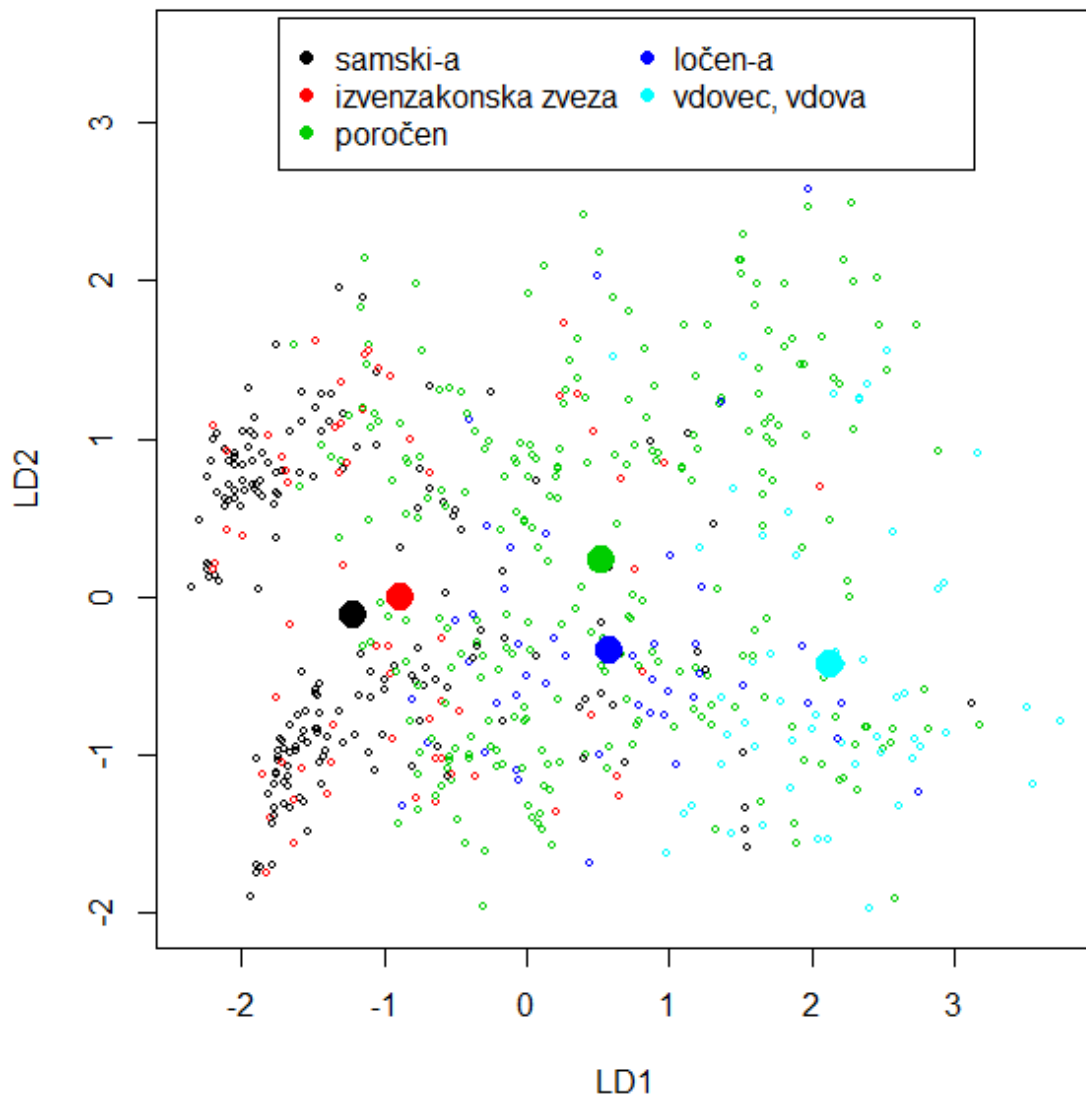
- 
- Moški spol in v manjši meri višja izobrazba kažeta proti poročenim in stran od ločenih in vdovcev (moških vdovcev je bistveno manj kot žensk, saj moški umirajo mlajši).
 - Višja izobrazba in v manjši meri ženski spol pa bolj kažeta proti ločenim.

Enote v prostoru DF - koda

```
par(mar=c(4,4,1,1)+0.1)
plot(ldaZakStan$pred$x[,1:2], col=ego$D10[cmp], cex=0.5,
ylim=c(-2,3.5))
legend(x="top",ncol=2,bg="transparent", col=1:5, pch=19,
legend = levels(ego$D10[cmp]), inset = 0.01)
tmp<-aggregate(ldaZakStan$pred$x,by =
list(ego$D10[cmp]),FUN=mean)
points(tmp[, 2:3], col=1:5, cex=2, pch=19)

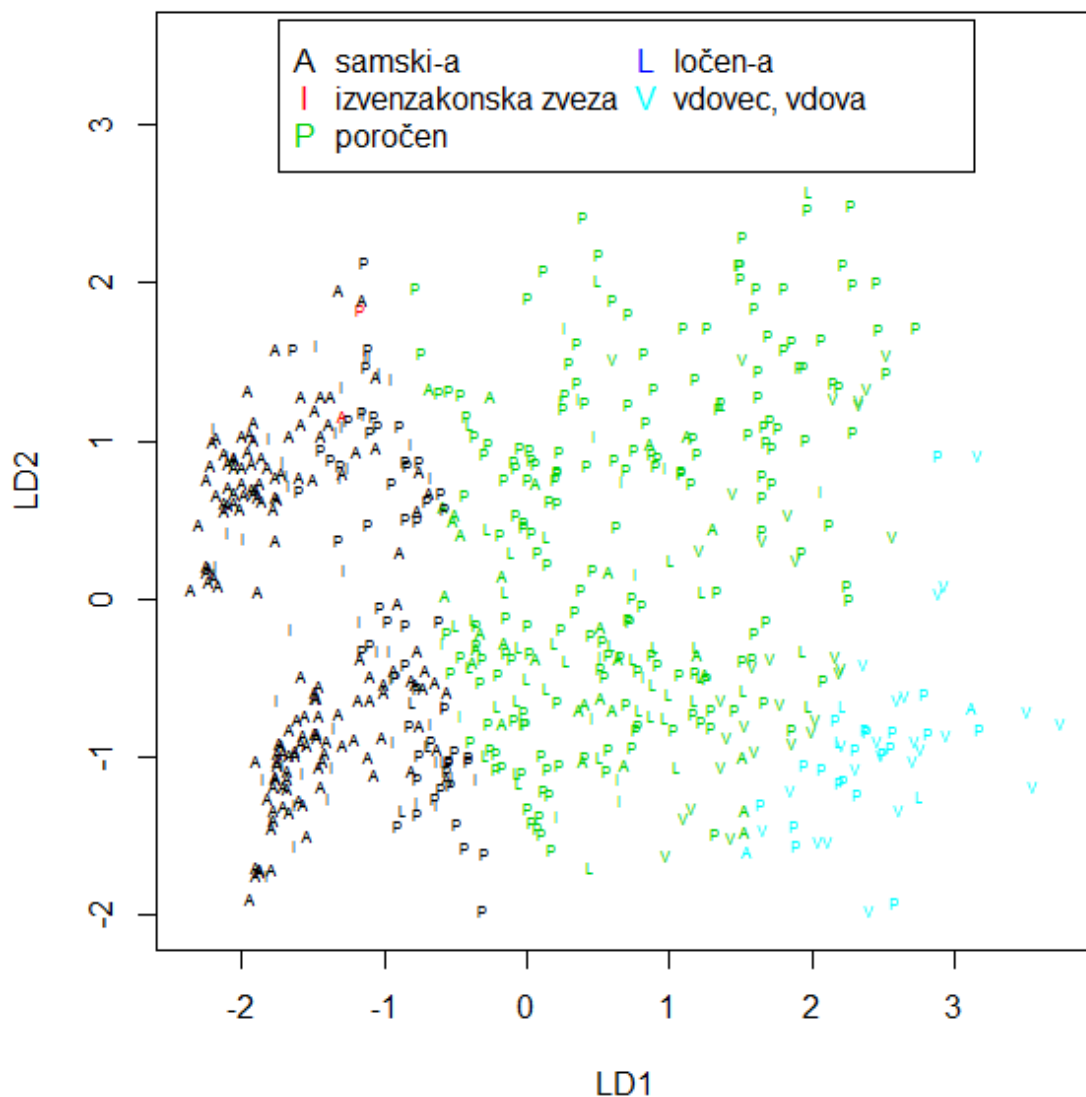
par(mar=c(4,4,1,1)+0.1)
plot(ldaZakStan$pred$x[,1:2], col=ldaZakStan$pred$class,
pch=c("A","I","P","L","V")[ego$D10[cmp]], cex=0.5, ylim=c(-
2,3.5))
legend(x="top",ncol=2,bg="transparent", col=1:5,
pch=c("A","I","P","L","V"), legend = levels(ego$D10[cmp]),
inset = 0.01)
```

Enote v prostoru DF



- Barve predstavljajo dejanske skupine
- Že prva DF precej dobro loči samske in izvenzakonske (mladi) od poročenih in ločenih (srednja leta) ter lete od vdovcev (stari)

Enote v prostoru DF

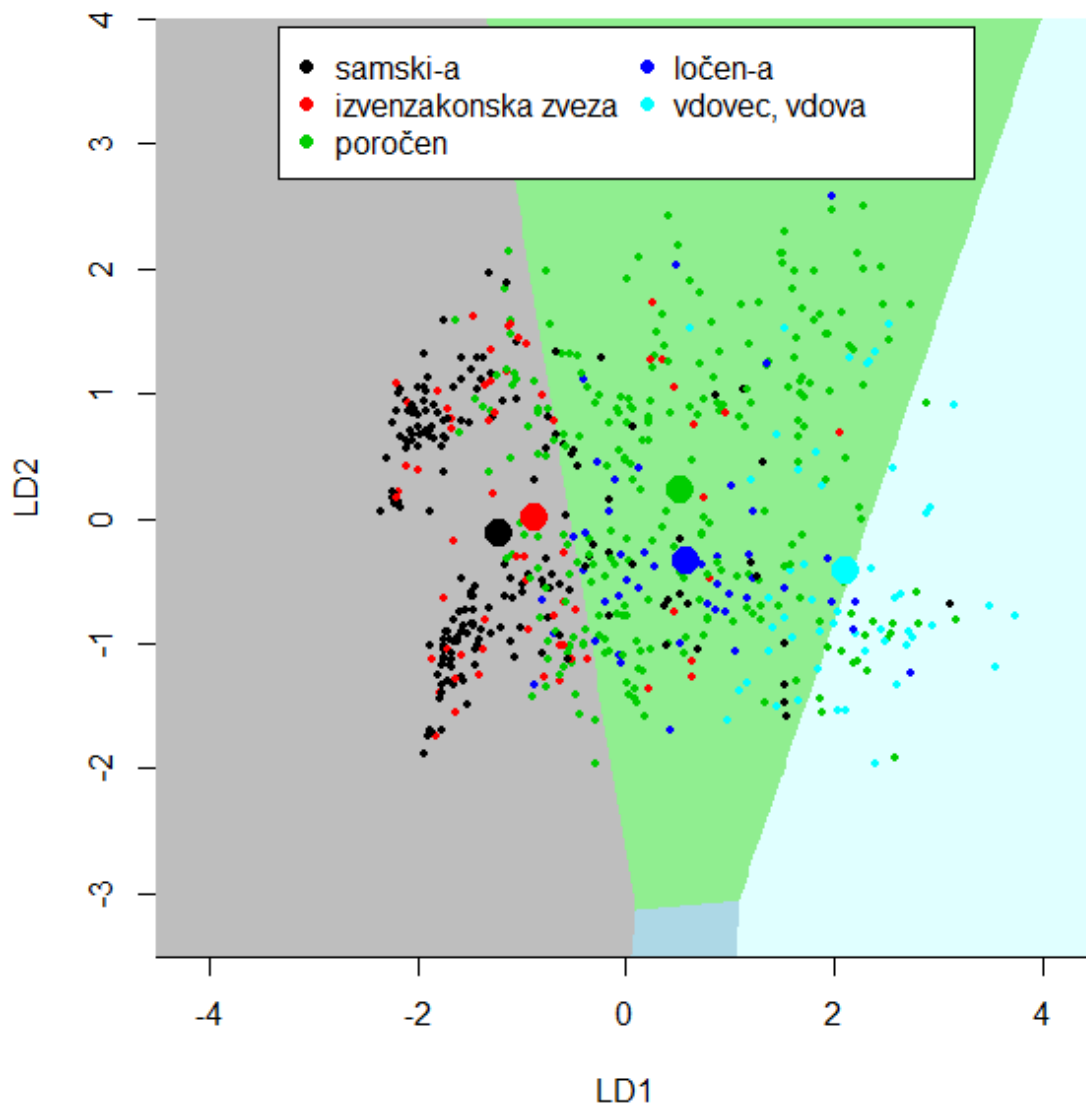


- Barve predstavljajo napovedane skupine
- Črke predstavljajo dejanske skupine
- Legenda je „pravilna“ le za pravilno klasificirane enote

Enote v prostoru DF - koda

```
par(mar=c(4,4,1,1)+0.1)
mapLda(ldaZakStan,xlim=c(-4.5,4.5), ylim=c(-
3.5,4),npoints=501,
col=c("gray","pink","lightgreen","lightblue","lightcyan"))
points(ldaZakStan$pred$x[,1:2], col=ego$D10[cmp],cex=0.5,
pch=19)
legend(x="top",ncol=2, col=1:5, pch=19, legend =
levels(ego$D10), inset = 0.01)
tmp<-aggregate(ldaZakStan$pred$x,by =
list(ego$D10[cmp]),FUN=mean)
points(tmp[, 2:3], col=1:5, cex=2, pch=19)
```

Enote v prostoru DF



- Barva ozadja predstavlja „napoved“ na podlagi samo prvih dveh DF.

Klasifikacijska tabela – vse enote

```
ldaZakStan$class$orgTab; ldaZakStan$class$perTab;  
ldaZakStan$class$corPer
```

Odstotek pravilno razvrščenih = 58.5

Frekvence

	samski-a	izvenzakonska zveza	poročen	ločen-a	vdovec, vdova	Sum
samski-a	148	1	32	0	2	183
izvenzakonska zveza	49	0	16	0	0	65
poročen	54	1	184	0	20	259
ločen-a	3	0	40	0	3	46
vdovec, vdova	0	0	31	0	23	54
Sum	254	2	303	0	48	607

Odstotki

	samski-a	izvenzakonska zveza	poročen	ločen-a	vdovec, vdova	Sum
samski-a	80.87	0.546	17.5	0	1.09	100
izvenzakonska zveza	75.39	0.000	24.6	0	0.00	100
poročen	20.85	0.386	71.0	0	7.72	100
ločen-a	6.52	0.000	87.0	0	6.52	100
vdovec, vdova	0.00	0.000	57.4	0	42.59	100

Klasifikacijska tabela – jackknife

```
ldaZakStan$classCV$orgTab; ldaZakStan$classCV$perTab;  
ldaZakStan$classCV$corPer
```

Odstotek pravilno razvrščenih = 58.3

Frekvence

	samski-a	izvenzakonska zveza	poročen	ločen-a	vdovec, vdova	Sum
samski-a	148	1	31	0	3	183
izvenzakonska zveza	49	0	16	0	0	65
poročen	55	1	183	0	20	259
ločen-a	3	0	40	0	3	46
vdovec, vdova	0	0	31	0	23	54
Sum	255	2	301	0	49	607

Odstotki

	samski-a	izvenzakonska zveza	poročen	ločen-a	vdovec, vdova	Sum
samski-a	80.87	0.546	16.9	0	1.64	100
izvenzakonska zveza	75.39	0.000	24.6	0	0.00	100
poročen	21.24	0.386	70.7	0	7.72	100
ločen-a	6.52	0.000	87.0	0	6.52	100
vdovec, vdova	0.00	0.000	57.4	0	42.59	100