DIJAGNOSTIČKA VALJANOST

Predviđanje pripadnosti grupi

Na šta se svodi problem predviđanja grupne pripadnosti?

Problem je vrlo sličan problemu koji smo rešavali u regresionoj analizi

- Kod linearne regresije tražimo kombinaciju prediktorskih varijabli koja će najviše korelirati sa kriterijumskom
- Kod predviđanja grupne pripadnosti tražimo kombinaciju prediktorskih varijabli koja će najbolje razlikovati dve grupe

Korelacija jedne kontinuirane i jedne binarne varijable

Point-biserijska korelacija (r_{pb})

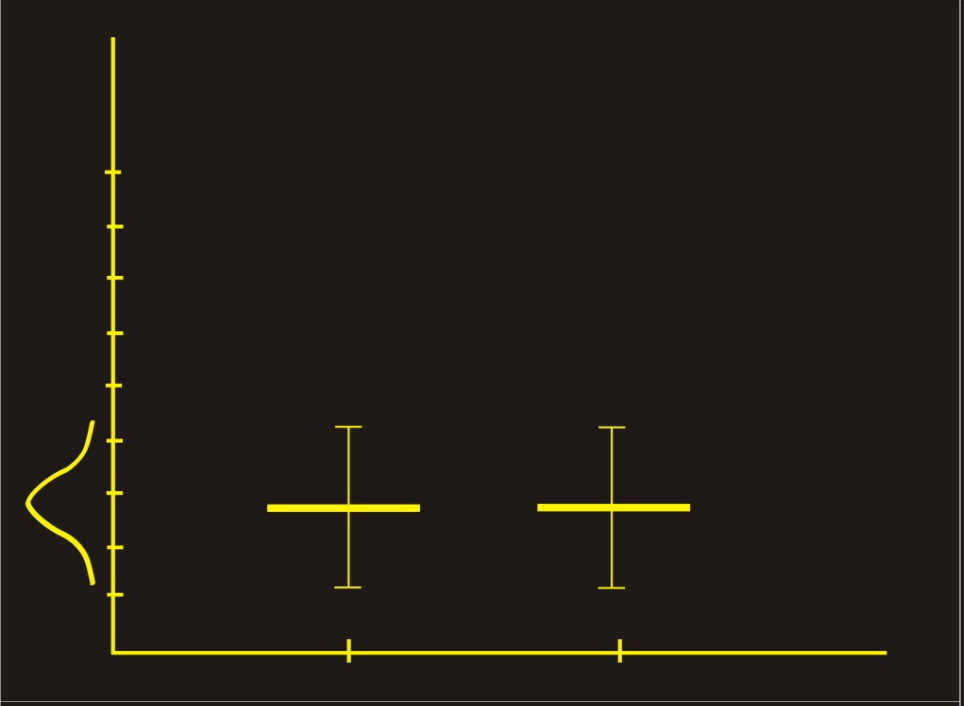
$$r_{pb} = \frac{(M_x - M_y) * \sqrt{pq}}{SD}$$

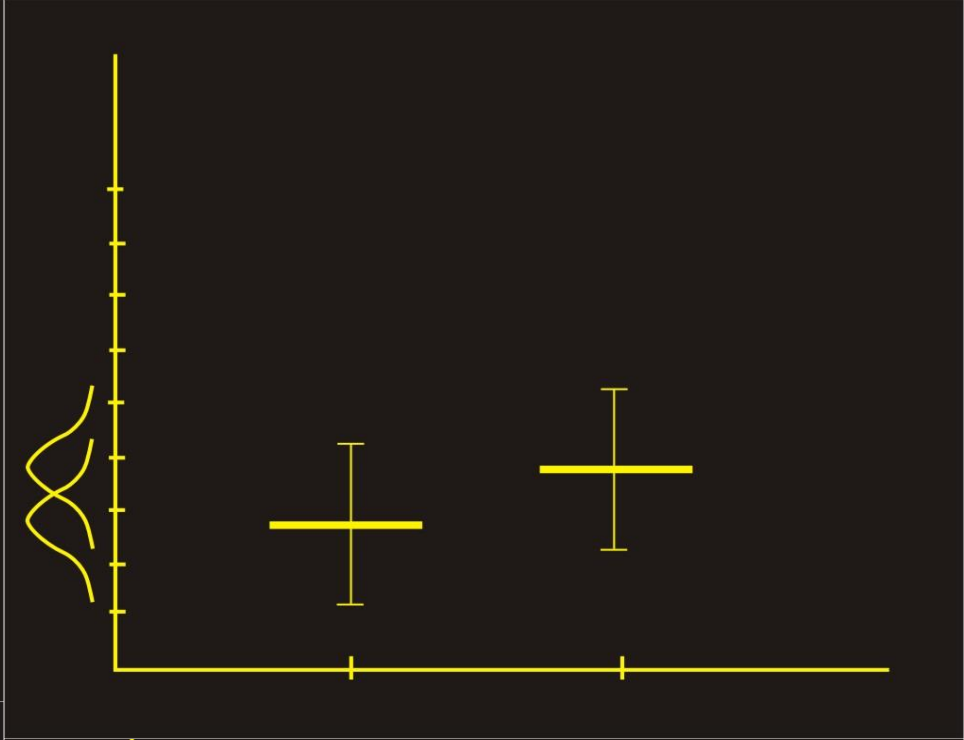
Mx - aritmetička sredina u grupi X

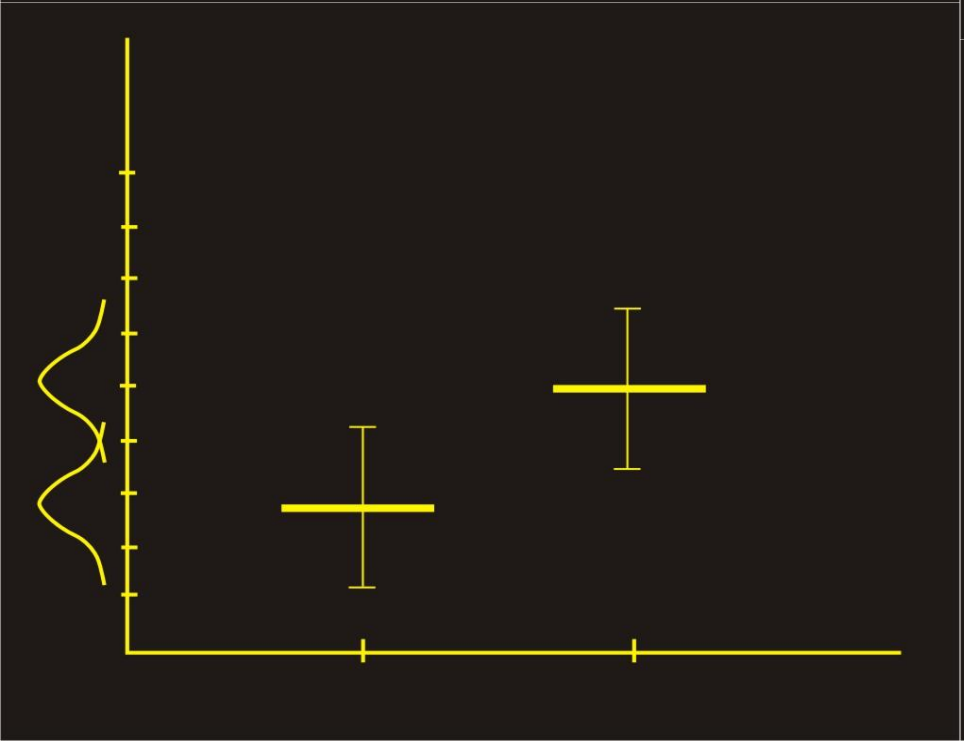
My - aritmetička sredina u grupi y

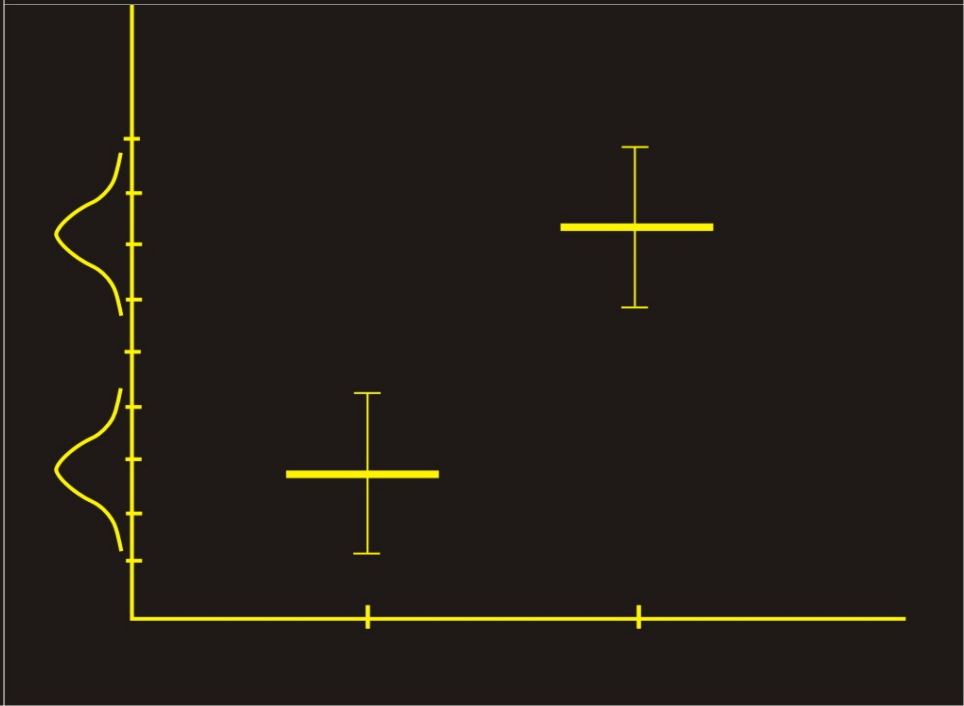
p i q - veličine grupa x i y, izražene kao proporcija

SD - standardna devijacija kontinualne varijable









Kada r_{pb} može imati vrednost 1?

Kada je razlika između grupa 2 standardne devijacije i kada je broj ispitanika u dve grupe jednak (oba uslova moraju biti zadovoljena)

Predviđene vrednosti

Ako je prediktorska varijabla kontinuirana, a kriterijumska binarna, onda će predviđene vrednosti imati vrednosti između 0 i 1

- One će predstavljati verovatnoću da ispitanik pripada jednoj ili drugoj grupi (žene/muškarci, pali/položili ispit i sl.)
- Ova verovatnoća je izračunata na osnovu njihovih skorova na kontinuiranoj varijabli

Koliko će predviđenih vrednosti biti kada predviđamo iz binarne u kontinuiranu varijablu?

Predviđene vrednosti

Šta se dešava kada umesto jedne kontinuirane varijable imamo skup kontinuiranih varijabli meren na najmanje intervalnom nivou?

GENERALIZACIJA

Rešavamo problem koji je identičan onom koji smo rešavali kod multiple regresije

Tražimo LINEARNU KOMBINACIJU varijabli koja najbolje razlikuje dve grupe ispitanika

- Time u stvari tražimo linearnu kombinaciju koja ima najveću point-biserijsku korelaciju sa binarnom varijablom

KANONIČKA DISKRIMINACIONA ANALIZA

Podsećanje

Multipla linearna regresija

Tražimo linearnu kombinaciju prediktora koja najbolje predviđa kriterijumsku varijablu (to jest najbolje korelira sa njom)

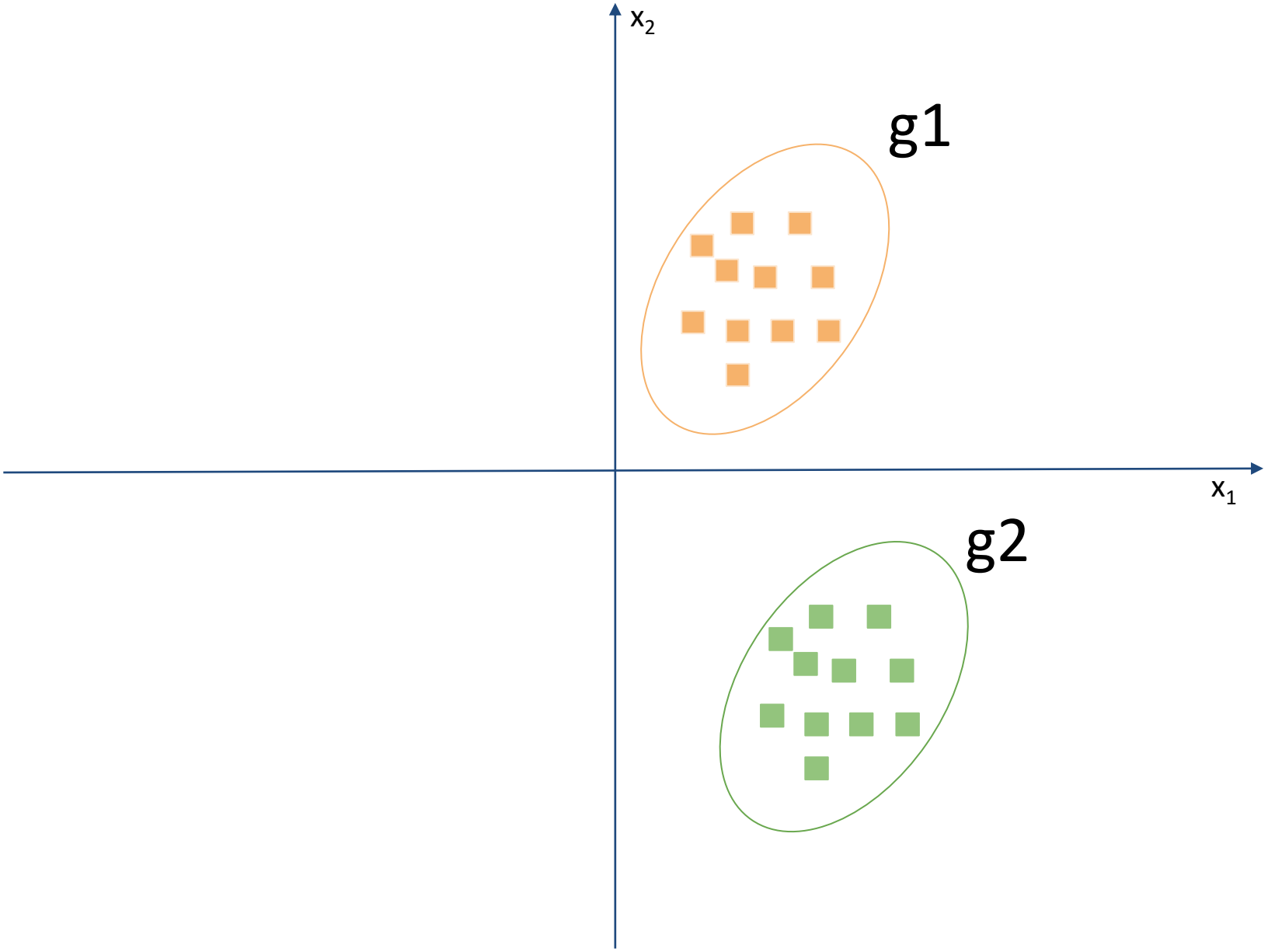
U KKA

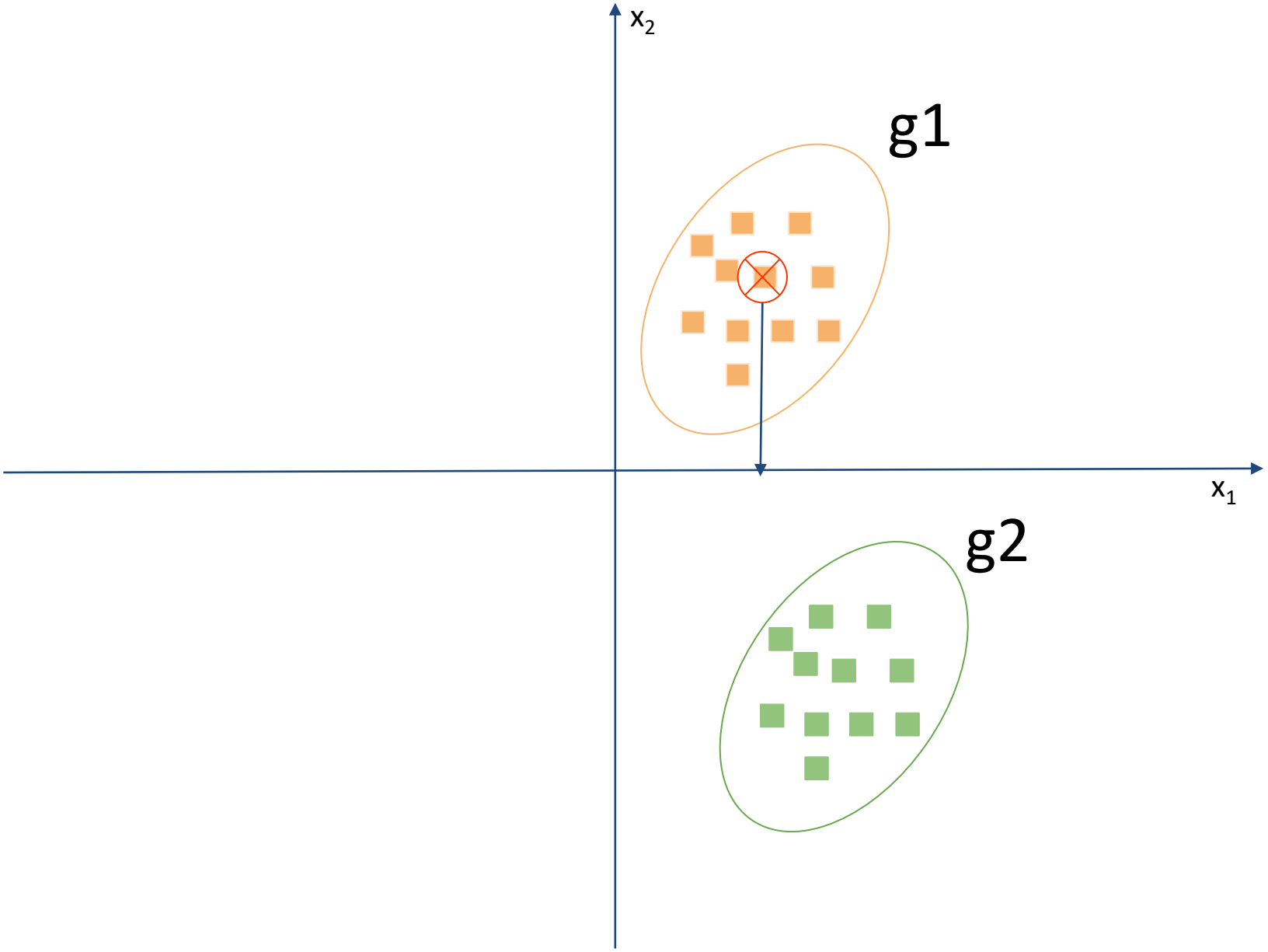
Tražimo linearne kombinacije varijabli u dva skupa koje najviše moguće međusobno koreliraju

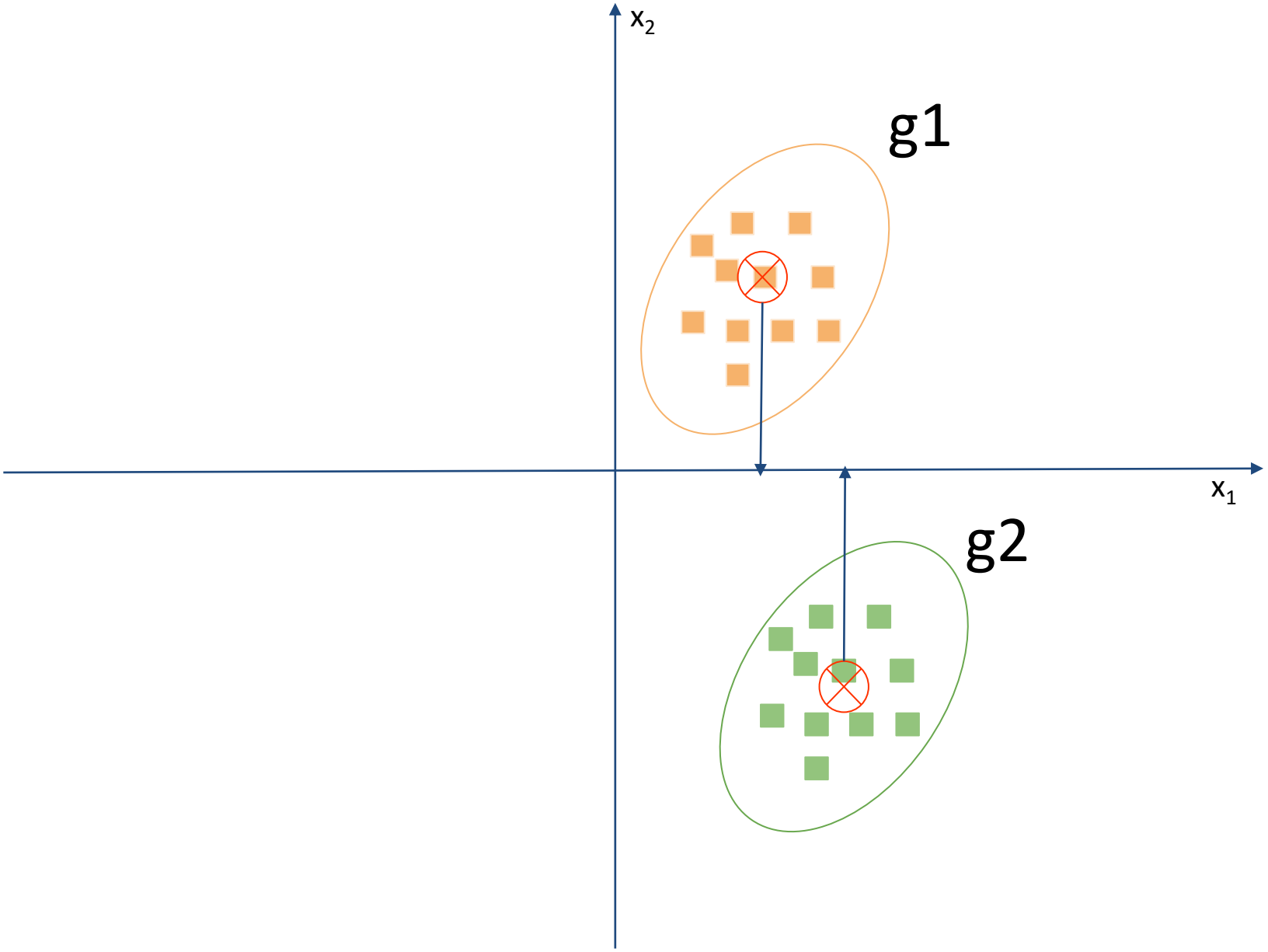
DEFINICIJA

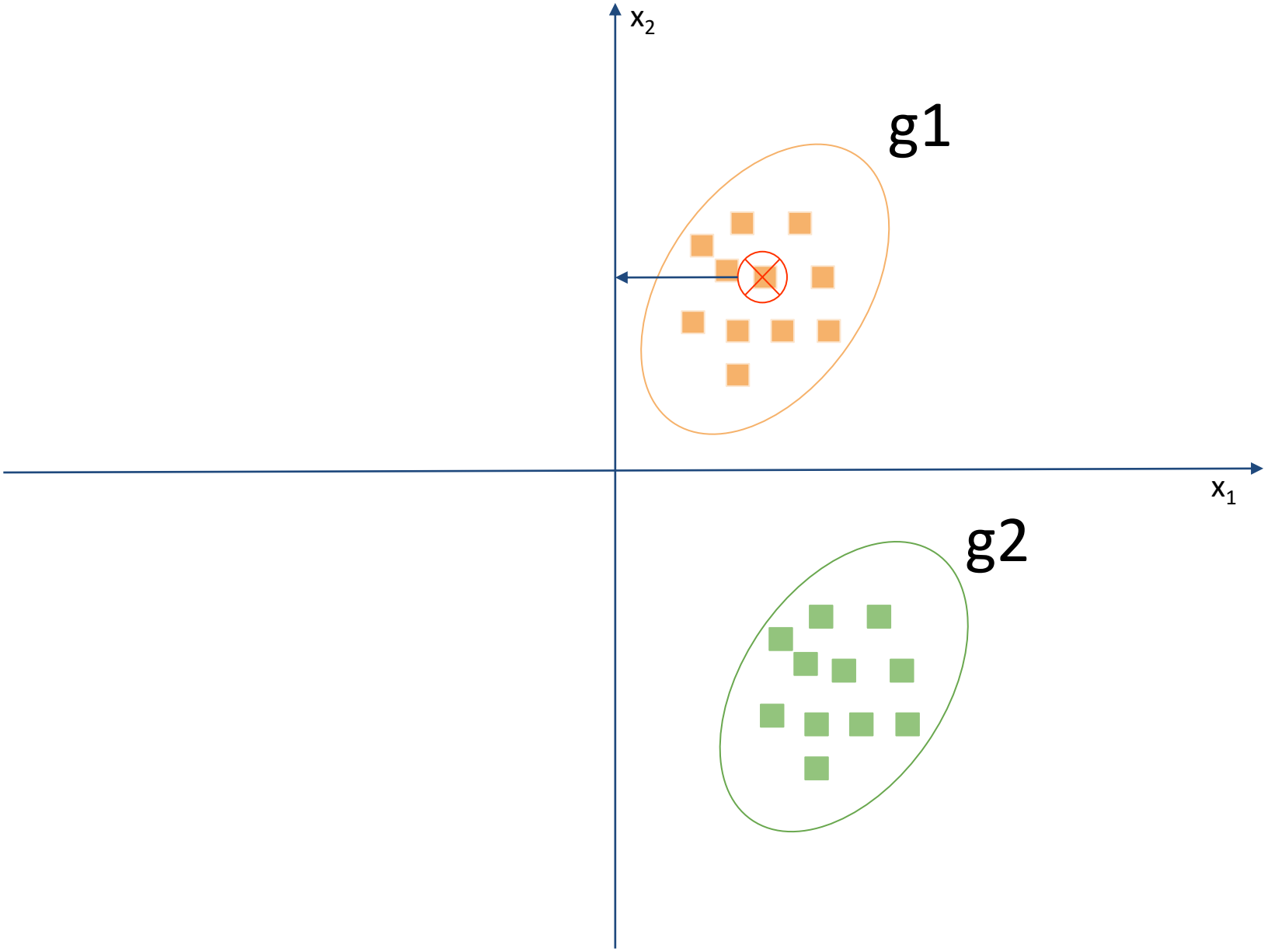
Kanonička diskriminaciona analiza je postupak pomoću kojeg tražimo linearnu kombinaciju prediktorskih varijabli koja najbolje razlikuje grupe

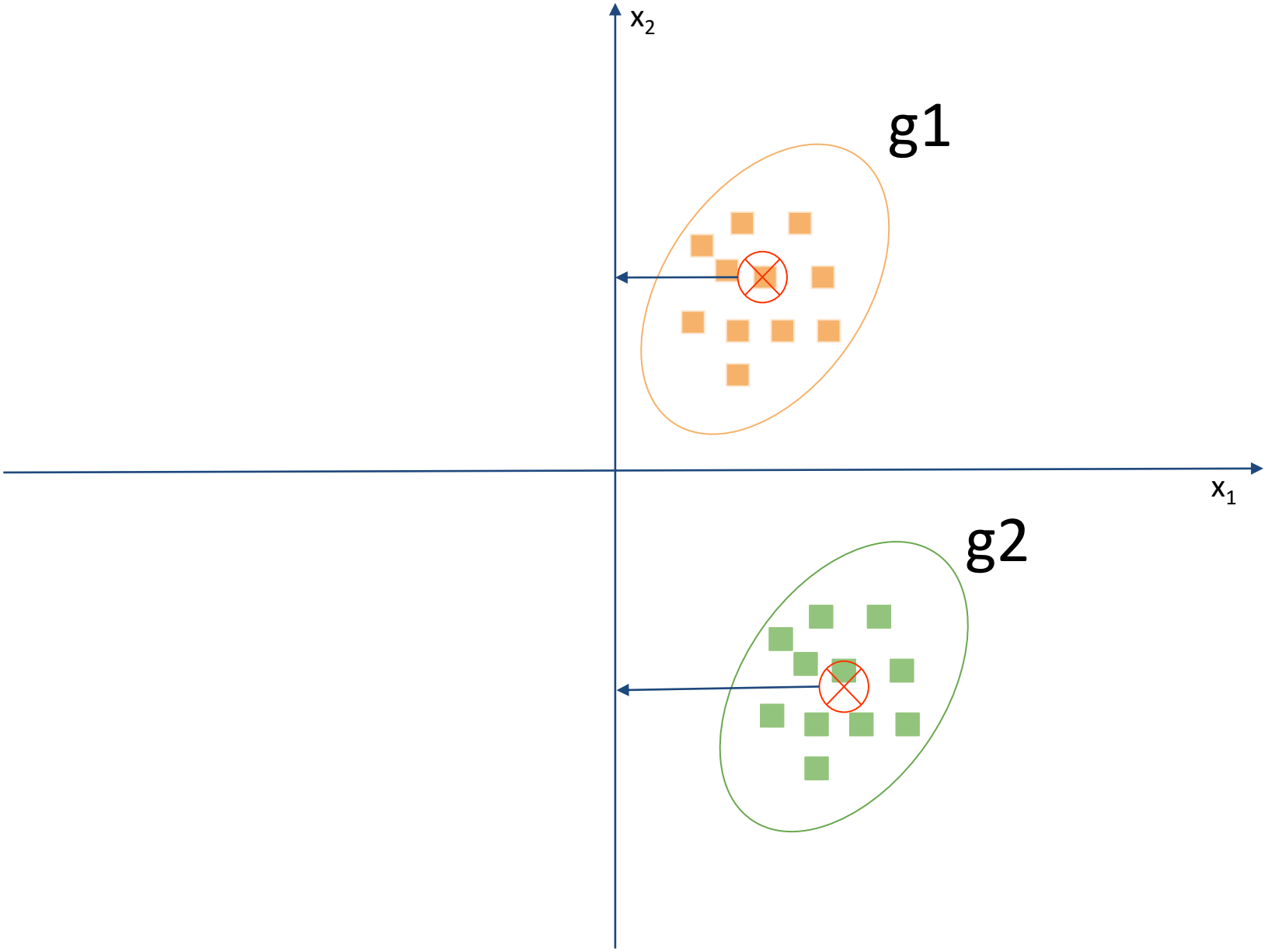
Ovu linearnu kombinaciju prediktora nazivamo **kanonička diskriminaciona funkcija (KDF)**

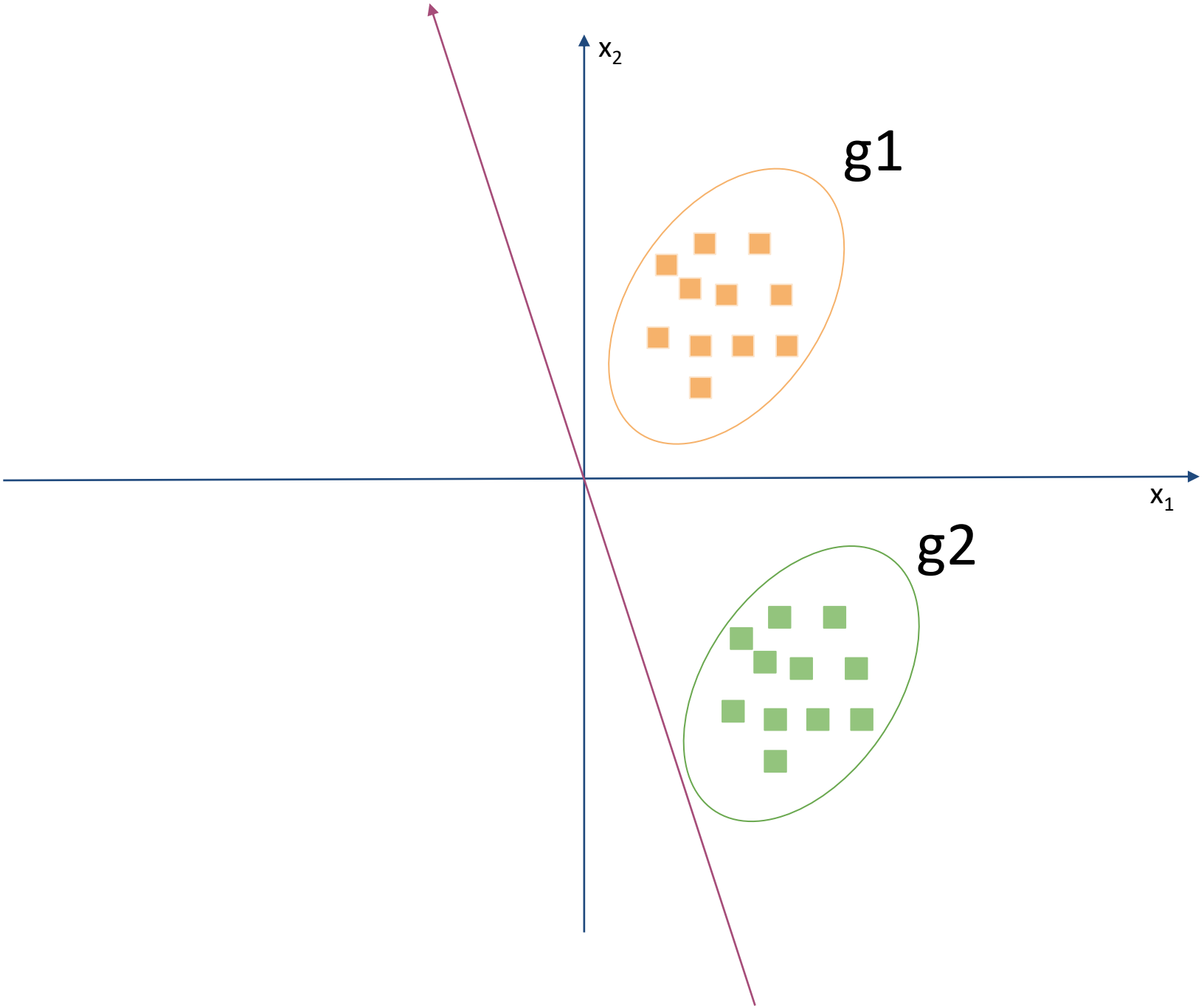

x2
g1
x1
g2

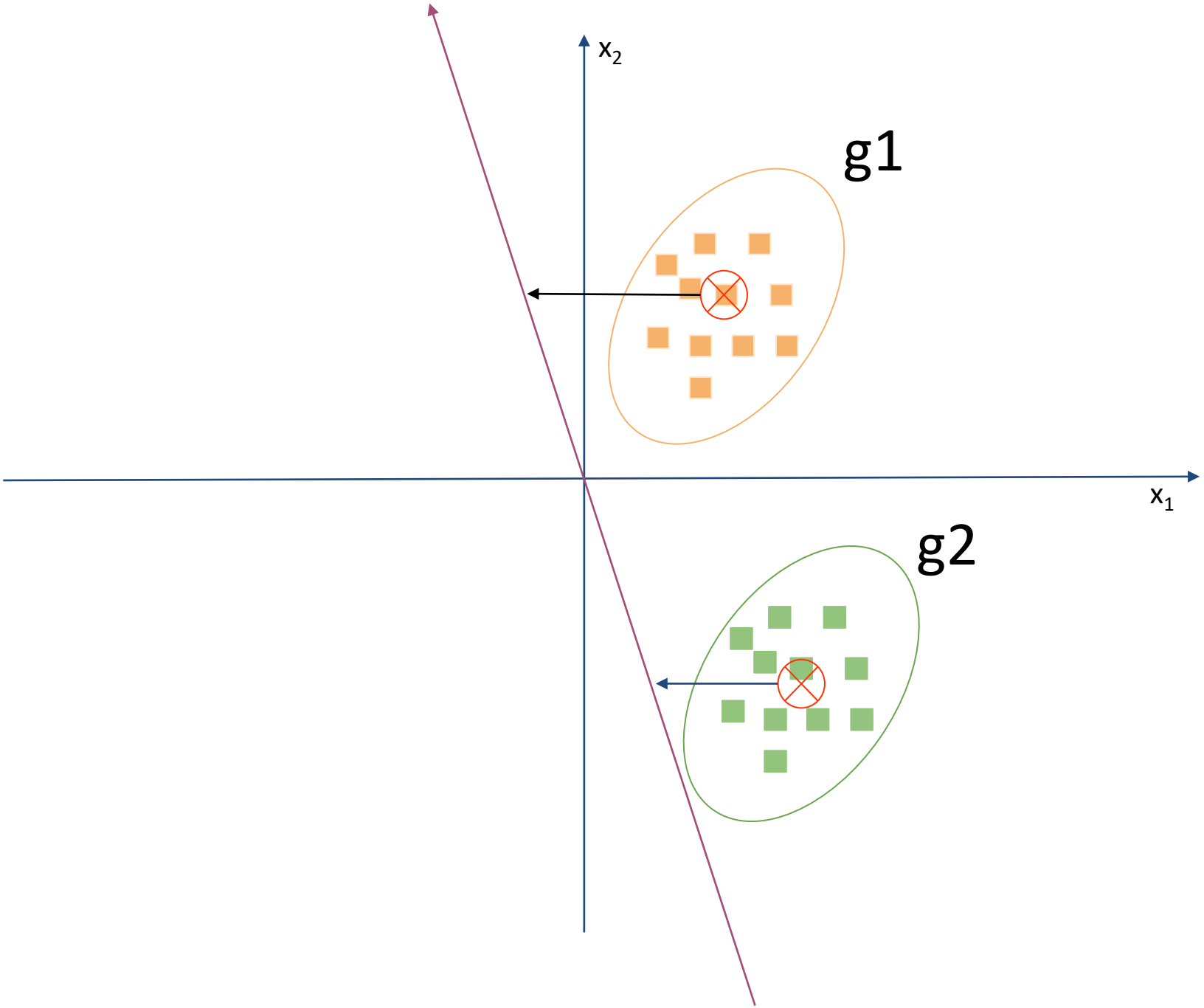

x2
g1
x1
g2


x2
g1
x1
g2


x2
g1
x1
g2


x2
g1
x1
g2

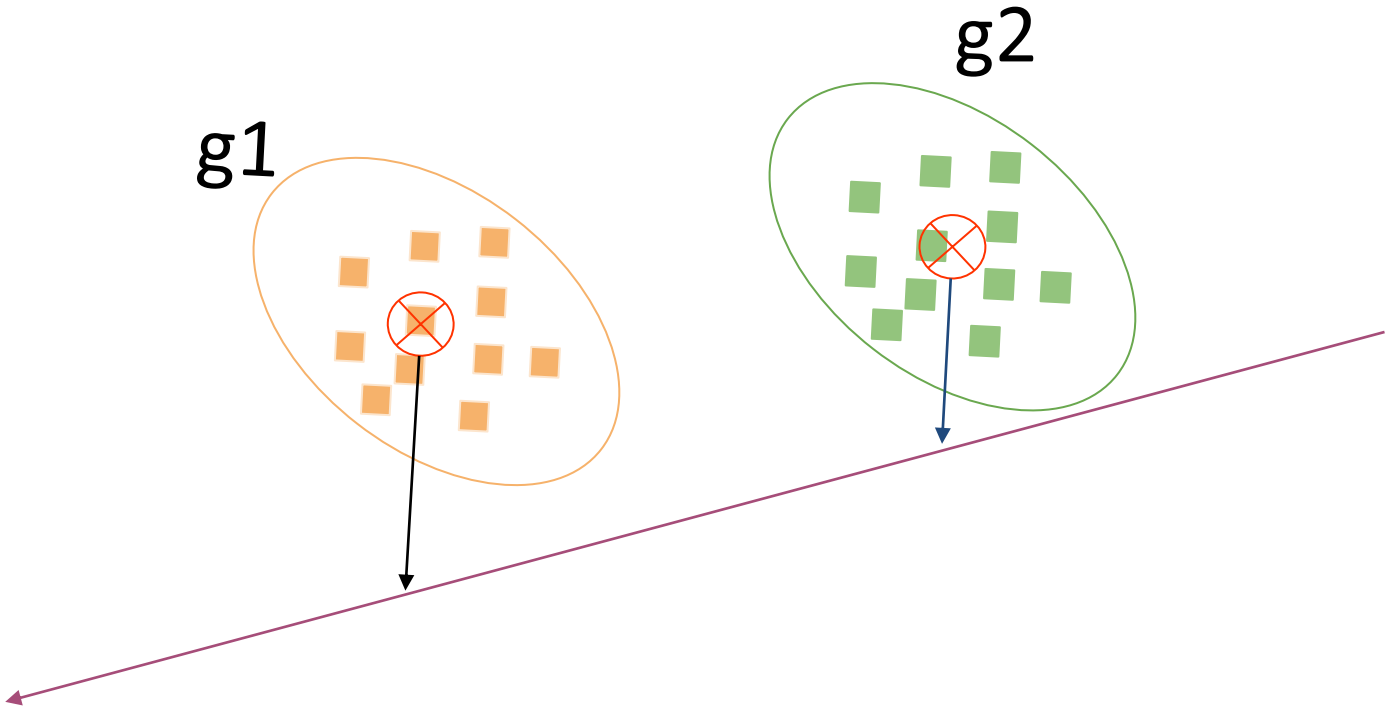

x2
g1
x1
g2

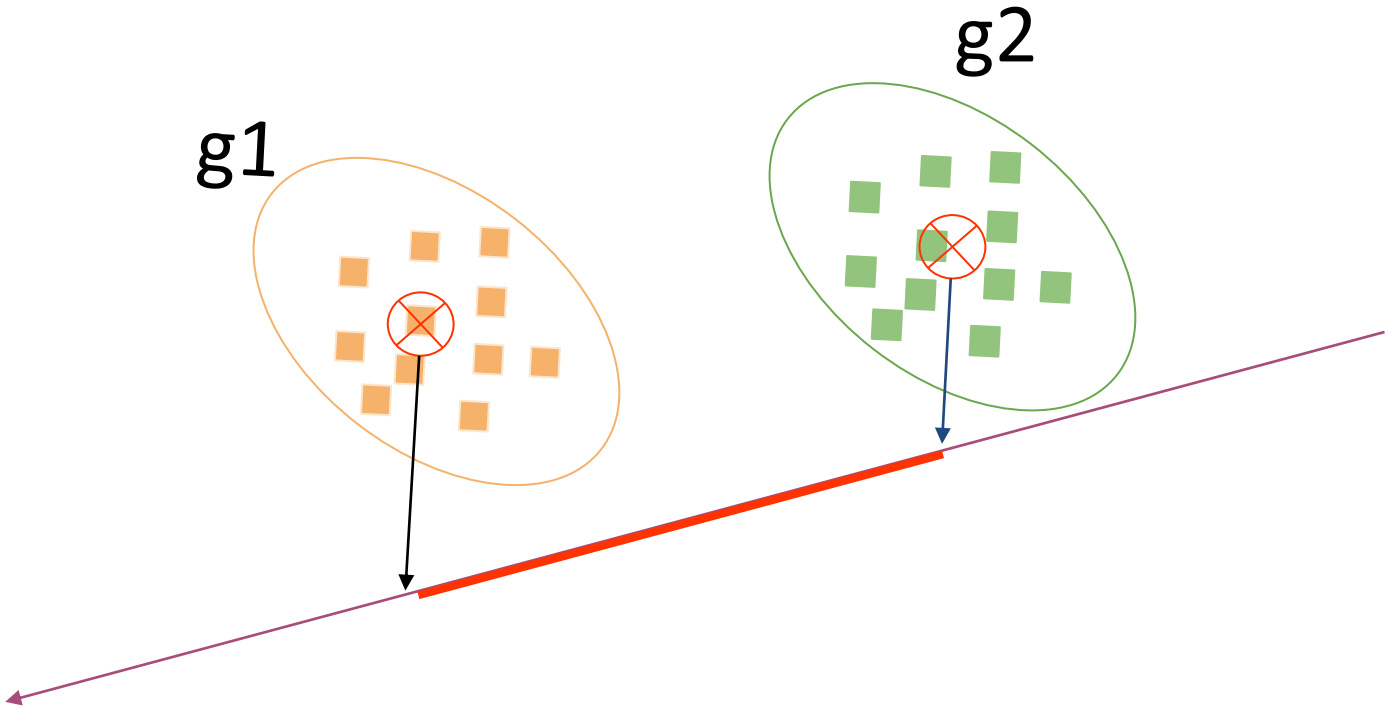

x2
g1
x1
g2

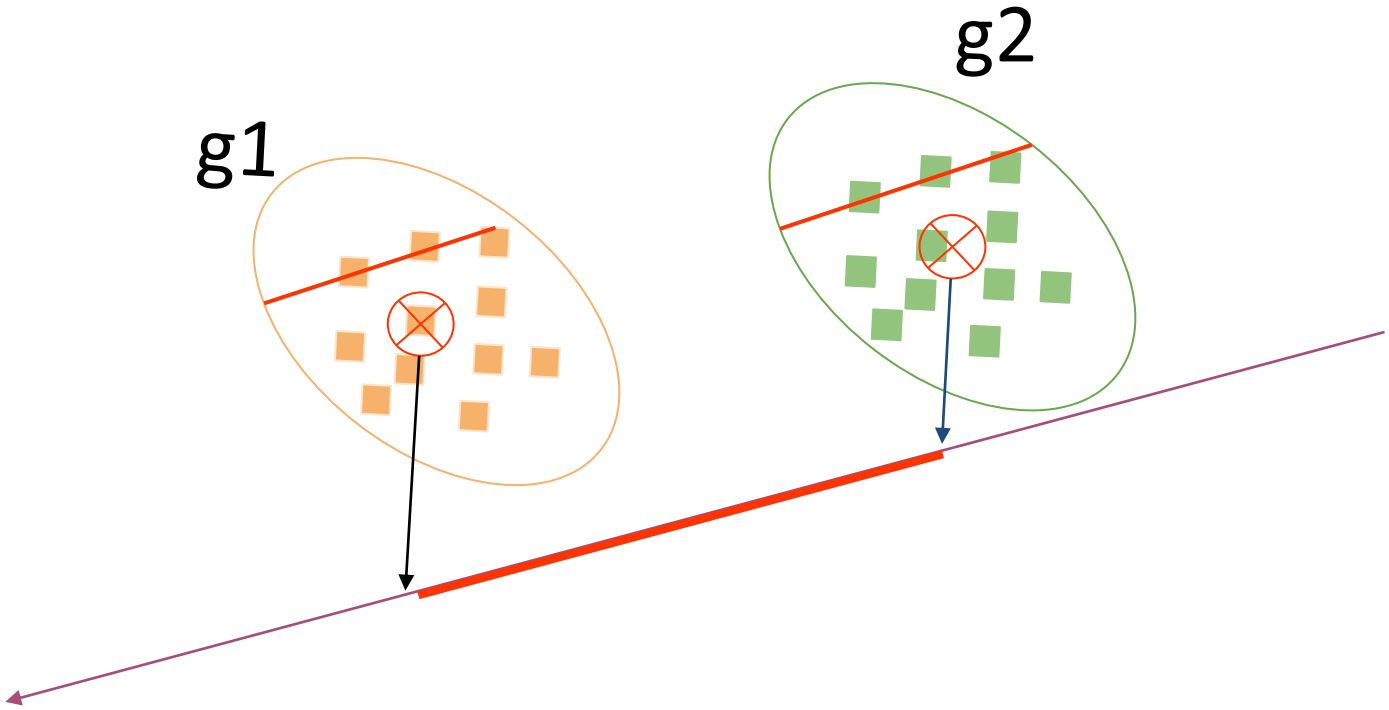
Šta maksimizujemo u KDA da bismo to postigli?

- Maksimizuje se odnos varijanse između grupa i varijanse unutar grupa

$$\lambda_p = \frac{v_p^t \mathrm{B} v_p}{v_p^t W v_p}$$


g1
g2


g1
g2


g1
g2

Kanonička korelacija u KDA

Kada imamo dve grupe ispitanika tada je

- Rho - kanonička korelacija između prediktorskog skupa i kriterijumske varijable
- Rho^2 – Kvadrat koeficijenta multiple(kanoničke) korelacije
  - Govori o proporciji varijanse razlike objašnjene pomoću skupa prediktorskih varijabli

$$\rho^2 = \frac{SSb}{SSw} \qquad \rho = \sqrt{\frac{SSb}{SSw}}$$

SSb - suma kvadrata odstupanja između grupa
SSw - suma kvadrata odstupanja unutar grupa

Koeficijenti u KDA

Standardizovani koeficijenti kanoničke diskriminacione funkcije

- Za koliko se SD promeni predviđena vrednost kriterijuma, ako se skor na prediktoru promeni za 1SD
- Koliko svaki prediktor doprinosi građenju kanoničke funkcije
- Analogni kanoničkim koeficijentima u KKA i beta ponderima u multiploj regresiji

Koeficijenti strukture

- Koliko svaki od prediktora korelira sa kanoničkom funkcijom
- Analogni koeficijentima strukture u KKA i regresionim faktorima u multiploj regresiji

Koeficijenti u KDA

	Multipla regresija	**Kanonička korelaciona analiza (KKA)**	**Kanonička diskriminaciona analiza (KDA)**
Doprinos izgradnji linearnog kompozita	Standardizovani regresioni koeficijent (beta ponder)	Kanonički koeficijent / koeficijent kanoničke funkcije	Standardizovani kanonički koeficijent / standardizovani koeficijent diskriminacione funkcije
Korelacija sa linearnim kompozitom	Regresioni faktori	Koeficijenti strukture / kanonički faktori	Koeficijenti strukture / diskriminacioni faktori

Centroidi

- Multivarijatne aritmetičke sredine tj. aritmetičke sredine linearne kombinacije
    - Tako svaka grupa ima svoj centroid, odnosno svoju aritmetičku sredinu na linearnoj kombinaciji prediktora (odnosno kanoničkoj funkciji)
- Imaju isto značenje kao aritmetičke sredine predviđenih vrednosti.
- Izražene su u standardizovanim skorovima tako da na osnovu njih zaključujemo za koliko se standardnih devijacija grupe razlikuju na diskriminacionoj funkciji.

Dva aspekta KDA

- Deskriptivni
  - Koja linearna kombinacija prediktora najbolje razlikuje (dve) grupe?
  - Odnosno koja je priroda razlika između grupa?
- Prediktivni
  - Koliko dobro možemo da predvidimo pripadnost grupi na osnovu kanoničkih funkcija

Predviđanje pripadnosti grupi

- Diskriminaciona funkcija je standardizovana kontinuirana varijabla
- Verovatnoća da ispitanik pripada svakoj od grupa može se odrediti na osnovu njegove udaljenosti od centroida grupe
  - U slučaju dve grupe, ako je verovatnoća da neki ispitanik pripada nekoj od grupa veća od 0.5 svrstavamo ga u tu grupu

Predviđanje pripadnosti grupi

Zatim je moguće uporediti predviđene i opažene grupne pripadnosti, te na osnovu preklapanja odrediti efikasnost klasifikacije

Što je udaljenost između centroida veća, manja je greška odnosno bolje je predviđanje

- Udaljenost od centroida direktno zavisi od veličine kanoničke korelacije
- Veća kanonička korelacija → bolje predviđanje

Nešto dobro poznato

Što je razlika između grupa u odnosu na varijansu unutar grupa (*F* količnik, r_{pb} ,η^2) <u>veća</u>, varijabla je potencijalno <u>značajnija za razlikovanje grupa</u>

- Drugim rečima, prediktorska varijabla na kojoj je varijansa između grupa velika, a varijansa unutar grupa mala, imaće visok kanonički koeficijent

Istovremeno, što je korelacija date prediktorske varijable sa ostalim prediktorima <u>veća</u> to je njen parcijalni doprinos (kanonički koeficijent) <u>manji</u>

- Odnosno, biće redundantna

Nešto dobro poznato

Dobar prediktor – visoki kanonički koeficijent i koeficijent strukture (visok doprinos KDF i visoka korelacija sa njom)

Loš prediktor – niski kanonički koeficijent i koeficijent strukture (nizak doprinos KDF i niska korelacija sa njom)

Redundantna varijabla – visoka korelacija sa KDF (koeficijent strukture), ali nizak doprinos izgradnji KDF (kanonički koeficijent); nije neophodna u modelu

Supresor - niska korelacija sa KDF (koeficijent strukture), ali visok doprinos izgradnji KDF (kanonički koeficijent), ili suprotan smer korelacije i doprinosa (oba moraju biti različita od 0); može ukazivati na posredan uticaj varijable ili biti statistički artefakt

PRIMER

Prediktorske varijable

emocionalnost - racionalnost

rigidnost

mizantropija

moralni relativizam

lokus kontrole

globalno samopostovanje (kompetentnosti)

polna privlacnost

evaluacija od strane drugih

fizicke sposobnosti

intelektualne sposobnosti

Pitamo se (deskriptivni aspekt KDA)

- Koliki procenat varijanse razlika među polovima možemo objasniti pomoću varijabli samopoimanja?

- Koliki je specifični doprinos pojedinih prediktora u objašnjenju razlika?
  - Kakva je struktura latentne dimenzije (kanoničke diskriminacione funkcije)?

- Kolika je udaljenost grupa na kanoničkoj diskriminacionoj funkciji?

Pitamo se (prediktivni aspekt KDA)

- Kakva je preciznost pogađanja pripadnosti grupi u odnosu na slučajno pogađanje (osetljivost i specificitet)
- U koju grupu bi bio svrstan svaki ispitanik
- Kolika je verovatnoća da svaki od ispitanika pripada svakoj od grupa
- Kolika je osetljivost i specificitet nekog skupa indikatora (baterije testova)

Interpretacija kanoničke diskriminacione funkcije

Za svaku pojedinačnu varijablu je prilikom interpretacije potrebno uzeti u obzir:

- razliku između grupa na toj varijabli
- njenu korelaciju sa kanoničkom diskriminacionom funkcijom (koeficijent strukture)
- njen koeficijent kanoničke diskriminativne funkcije
- njene izvorne korelacije sa ostalim prediktorima

Proseci grupa

	muški			ženski			total		
	mean	SD	N	mean	SD	N	mean	SD	N
emoc	2.99	0.62	150	3.12	0.6	104	3.04	0.61	254
rigid	3.19	0.62	150	3.22	0.65	104	3.2	0.63	254
miza	3.12	0.67	150	3.12	0.64	104	3.12	0.66	254
moral	3.13	0.64	150	3.16	0.55	104	3.14	0.61	254
ekste	3.24	0.54	150	3.27	0.52	104	3.25	0.53	254
genko	3.15	0.67	150	3.14	0.65	104	3.15	0.66	254
izgled	3.42	0.68	150	3.44	0.69	104	3.46	0.68	254
social	3.93	0.63	150	3.9	0.57	104	3.92	0.6	254
fizicki	3.98	0.71	150	3.51	0.77	104	3.79	0.77	254
intelek	3.76	0.63	150	3.69	0.52	104	3.73	0.58	254

Proseci grupa

	muški			ženski			total		
	mean	SD	N	mean	SD	N	mean	SD	N
emoc	**2.99**	**0.62**	**150**	**3.12**	**0.6**	**104**	**3.04**	**0.61**	**254**
rigid	3.19	0.62	150	3.22	0.65	104	3.2	0.63	254
miza	3.12	0.67	150	3.12	0.64	104	3.12	0.66	254
moral	3.13	0.64	150	3.16	0.55	104	3.14	0.61	254
ekste	3.24	0.54	150	3.27	0.52	104	3.25	0.53	254
genko	3.15	0.67	150	3.14	0.65	104	3.15	0.66	254
izgled	3.42	0.68	150	3.44	0.69	104	3.46	0.68	254
social	3.93	0.63	150	3.9	0.57	104	3.92	0.6	254
fizicki	**3.98**	**0.71**	**150**	**3.51**	**0.77**	**104**	**3.79**	**0.77**	**254**
intelek	3.76	0.63	150	3.69	0.52	104	3.73	0.58	254

Univarijatna analiza varijanse

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
emoc	,989	2,846	1	252	,093
rigid	,999	,131	1	252	,718
miza	1,000	,004	1	252	,947
moral	,999	,150	1	252	,698
ekste	,999	,234	1	252	,629
genkom	1,000	,006	1	252	,940
izgled	1,000	,120	1	252	,729
social	,999	,136	1	252	,713
fizicki	,910	25,057	1	252	,000
intelekt	,996	,966	1	252	,327

Homogenost kovarijansi

Test Results

Box's M		62,244
F	Approx.	1,082
	df1	55
	df2	159490,086
	Sig.	,314

Tests null hypothesis of equal population covaria

Box M parametar testira homogenost kovarijansi grupa

Odgovara na pitanje da li se odnosi između varijabli u dvema grupama razlikuju

svojstvena vrednost

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,139(a)	100,0	100,0	,350

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

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,878	32,187	10	,000

$$\lambda_p = \frac{v_p^t \mathrm{B} v_p}{v_p^t W v_p}$$

Kanonička korelacija reflektuje važnost latentne varijable za diskriminaciju

Kanonička korelacija

Kanonička korelacija (Rho) je mera povezanosti nominalne varijable i linearne kombinacije nezavisnih varijabli,

- Ako je Rho = 0, aritmetičke sredine (centroidi) obe grupe su identične (raspodele se preklapaju)
- Ako je Rho = 1, varijansa između grupa jednaka je totalnoj varijansi

Koeficijenti strukture i koeficijenti kanoničke diskriminativne funkcije

Standardized Canonical Discriminant Function Coefficients

	Function
	1
emoc	,413
rigid	,003
miza	-,137
moral	,004
ekste	,171
genkom	,246
izgled	,233
social	,150
fizicki	-1,063
intelekt	,014

Structure Matrix

	Function
	1
fizicki	-,845
emoc	,285
intelekt	-,166
ekste	,082
moral	,065
social	-,062
rigid	,061
izgled	-,059
genkom	-,013
miza	,011

Centroidi – aritmetičke sredine grupa na KDF

Functions at Group Centroids

pol ispitanika	Function 1
muski	-,309
zenski	,446

Unstandardized canonical discriminant functions evaluated at group means

Centroidi su ključni za razumevanje razlika između grupa

- Iako znamo prirodu kanoničke funkcije, bez centroida ne znamo koji je smer tih razlika (koja grupa postiže više skorove na diskriminacionoj funkciji)

Diskriminantni skor

Skor ispitanika na diskriminacionoj funkciji

Kada su rezultati nestandardizovani, skor je rezultat primene obrasca

$DFL = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c$

Kada su rezultati standardizovani, skor računamo na sledeći način:

$DFL = \beta_1z_1 + \beta_2z_2 + \ldots + \beta_nz_n$

Fišerovi koeficijenti klasifikacije

- Služe da se izračuna skor pripadnosti za svaku grupu
- $Yg = f_1x_1 + f_2x_2 + ... + f_nx_n$
- Svaka grupa ima svoje koeficijente koji omogućavaju izračunavanje skora "pripadnosti" za svaku grupu

Prelomni (Cutoff) skor

- Kada su grupe jednake, prelomni skor je jednak tački koja je ekvidistantna obema aritmetičkim sredinama (centroida)
  - Odnosno, nalazi se tačno između dva centroida
- Kada grupe nisu jednake, onda se u obzir uzima veličina uzorka, tako da će biti bliži centroidu veće grupe
  - To znači da ćemo pouzdanije predviđati pripadnost većoj grupi

Tačnost klasifikacije

kojoj grupi ispitanik stvarno pripada

u koju smo grupu svrstali ispitanika na osnovu skora na KDF

			Membership		
		pol ispitanika	muski	zenski	Total
Original	Count	muski	99	51	150
		zenski	39	65	104
	%	muski	66,0	34,0	100,0
		zenski	37,5	62,5	100,0

a. 64,6% of original grouped cases correctly classified.

Što su veći elementi van glavne dijagonale to je specificitet testa manji

- Odnosno, test više greši prilikom klasifikacije

Uslovi za KDA

Stvarne kategorije - nikad ne treba raditi KDA na varijabli koja je veštački podeljena u kategorije

- Naravno, softver ne zna ništa o prirodi varijabli
- Da li su onda i dijagnostičke kategorije veštačke?
- Ovaj uslov zapravo znači da ne treba od kontinualne varijable da pravimo kategoričku, pa da na njoj radimo KDA

Linearna nezavisnost prediktora - ni jedna varijabla ne sme biti linearna kombinacija bilo koje dve ili više prediktorskih varijabli

Uslovi za KDA

Ne sme biti velika disproporcija između veličine grupa

Broj ispitanika bi trebalo da bude minimalno 5 puta veći od broja prediktorskih varijabli

Najmanje intervalni nivo merenja na prediktorskim varijablama

Uslovi za KDA

- Varijabilitet veći od 0 u svim grupama - unutargrupni varijabilitet > 0
- Slučajna raspodela greške modela (reziduala)
- Homogenost varijansi (homoscedasticity) i kovarijansi
  - Box M previše osetljiva mera
    - Posebno na velikim uzorcima

Uslovi za KDA

- Niska međuzavisnost prediktora
- Linearnost veza između prediktora
- Multinormalna raspodela

Hvala na pažnji!

Pitanja?