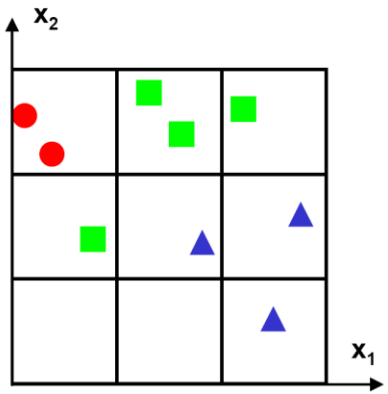
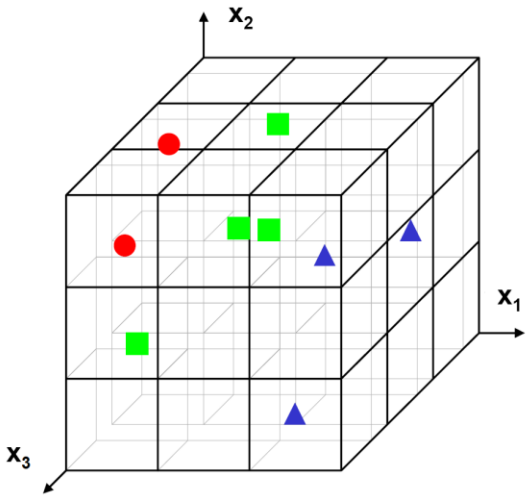
Razlaganje na glavne komponente (PCA) i linearna diskriminantna analiza (LDA)

- Razlaganje na glavne komponente (eng. *principal component analysis – PCA*)
  - Problem dimenzionalnosti
  - Selekcija i izdvajanje obeležja
  - Reprezentacija i klasifikacija signala
  - Razlaganje na glavne komponente (PCA)
- Linearna diskriminantna analiza (eng. *linear discriminant analysis – LDA*)
  - Slučajevi dveju i više klasa
  - Ograničenja i varijante LDA
  - Ostale metode za redukciju dimenzionalnosti

Problem dimenzionalnosti

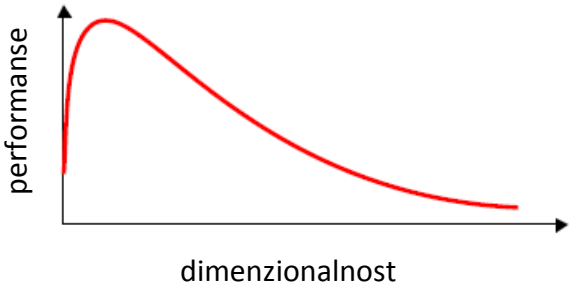
- Problem analize multivarijabilnih podataka kada broj dimenzija prostora raste
- Primer: problem klasifikacije uzoraka u tri klase (*N* uzoraka)
  - Rešenje:
    - podela prostora obeležja na ćelije jednakih dimenzija
    - dodela nepoznatog uzorka klasi čiji su uzorci najzastupljeniji u ćeliji u kojoj se on nalazi
  - Ako imamo samo jedno obeležje, uzorci klasa se po pravilu preklapaju i logično je uvesti i drugo obeležje da bi se popravila separabilnost klasa
  - Ako se uvede drugo obeležje, ili treba obezbediti N^2 uzoraka (za istu gustinu uzoraka) ili će gustina uzoraka drastično opasti (a u tri dimenzije će situacija biti još mnogo gora)
  - Podela na jednake ćelije nije najbolje rešenje, ali taj faktor je ovde od manjeg značaja





Problem dimenzionalnosti

- Za zadatu veličinu uzorka postoji maksimalni broj obeležja koji, kada se prekorači, rezultuje opadanjem (a ne porastom) performansi klasifikatora
  - Informacija koja se izgubi odbacivanjem nekih obeležja je često (više nego) kompenzovana preciznijom estimacijom GRV u niždimenzionalnom prostoru
- Različite implikacije problema dimenzionalnosti
  - Za D dimenzija potrebno je N^D uzoraka
  - Eksponencijalni rast kompleksnosti GRV sa porastom broja dimenzija
    - „Funkcija definisana u višedimenzionalnom prostoru često je mnogo kompleksnija nego funkcija definisana u niždimenzionalnom prostoru, a tu komplikovanu strukturu je znatno teže odrediti.“ (Friedman)



    - Da bi se bolje opisala komplikovana struktura potrebno je raspolagati gušćim skupom uzoraka
  - Šta raditi ako raspodela nije Gaussova?
    - Za jednodimenzionalnu slučajnu promenljivu postoji veliki broj različitih tipova GRV koji odgovaraju različitim problemima, ali je kod više dimenzija na raspolaganju jedino Gaussova
- Ljudi imaju izvanrednu sposobnost da razlikuju oblike i klastere u 1, 2 ili 3 dimenzije, ali ova sposobnost drastično opada za 4 ili više dimenzija
- Problem se može rešavati uzimanjem u obzir apriornog znanja, definisanjem glađe GRV ili nekim od postupaka *smanjenja dimenzionalnosti*

Smanjenje dimenzionalnosti

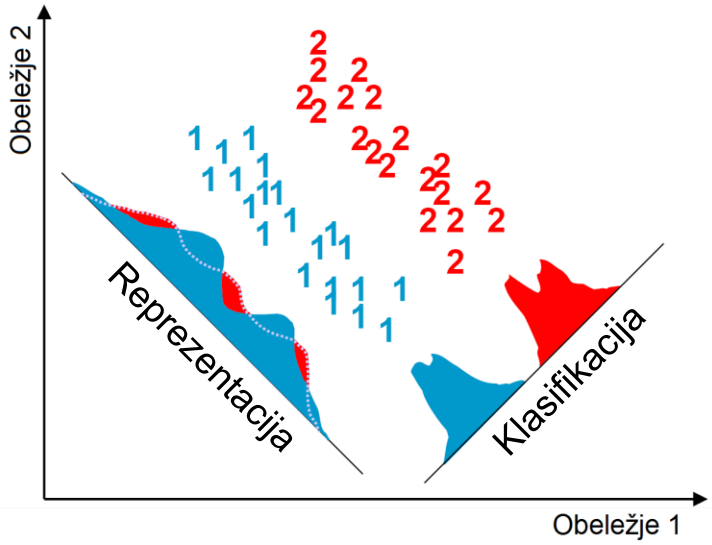
- Dva su osnovna pristupa smanjenju dimenzionalnosti
  - **Odabir (selekcija) obeležja**: biranje (bitnijih) obeležja iz skupa svih obeležja
  - **Izdvajanje obeležja**: formiranje (manjeg) skupa novih obeležja kombinovanjem postojećih

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{odabir obeležja}} \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_M} \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{izdvajanje obeležja}} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} \right)$$

- Problem izdvajanja obeležja se može predstaviti na sledeći način
  - Za dati prostor obeležja $\mathbb{R}^N$ naći preslikavanje $\mathbf{y} = f(\mathbf{x})$: $\mathbb{R}^N \rightarrow \mathbb{R}^M$, $M < N$, tako da transformisani vektor obeležja $\mathbf{y} \in \mathbb{R}^M$ u najvećoj meri očuva informacije ili strukturu prisutne u $\mathbb{R}^N$
- *Optimalno* preslikavanje $\mathbf{y} = f(\mathbf{x})$ je ono koje ne izaziva povećanje verovatnoće greške
  - Ne postoji opšti način da se ovo preslikavanje nađe jer optimalnost zavisi od problema
    - Pod pretpostavkom da je ovo preslikavanje linearno ($\mathbf{y} = \mathbf{W}\mathbf{x}$) problem je jednostavniji, ali postoje tehnike koje se bave i nelinearnim slučajem
  - Drugim rečima, Bayesovo pravilo odlučivanja primenjeno na inicijalni prostor $\mathbb{R}^N$ i na redukovani prostor $\mathbb{R}^M$ treba da dâ istu tačnost klasifikacije

Reprezentacija i klasifikacija

- Izbor preslikavanja **y** = f(**x**) za izdvajanje obeležja, diktiran je funkcijom cilja koju treba maksimizovati (ili minimizovati)
- U zavisnosti od kriterijuma koji koristi funkcija cilja, tehnike izdvajanja obeležja se dele u dve kategorije:
  - **Reprezentacija**: Cilj preslikavanja je da predstavi skup uzoraka što tačnije u prostoru sa manjim brojem dimenzija
  - **Klasifikacija**: Cilj preslikavanja je da pojača diskriminaciju između klasa u prostoru s manjim brojem dimenzija
- U okviru linearnog izdvajanja obeležja, najčešće se koriste dve tehnike:
  - Razlaganje na glavne komponente (PCA)
    - Koristi kriterijum **reprezentacije**
  - Linearna diskriminantna analiza (LDA)
    - Koristi kriterijum **klasifikacije**



Razlaganje na glavne komponente (PCA)

- Cilj je smanjenje dimenzionalnosti prostora uz očuvanje rasutosti (varijanse) podataka u višedimenzionom prostoru
  - Varijansa je nosilac informacije
- Neka je **x** *N*-dimenzionalni slučajni vektor koji se može izraziti kao linearna kombinacija ortonormalnih vektora $\boldsymbol{\Phi}_i$ iz baze:

$$\mathbf{x} = \sum_{i=1}^{N} x_i \boldsymbol{\Phi}_i, \quad \left\langle \boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j \right\rangle = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

- Neka je potrebno aproksimirati **x** sa samo *M* baznih vektora ($M < N$), što znači da koordinate $x_{M+1}, \ldots x_N$ treba zameniti nekim unapred odabranim konstantama b_i:

$$\hat{\mathbf{x}}(M) = \sum_{i=1}^{M} x_i \boldsymbol{\Phi}_i + \sum_{i=M+1}^{N} b_i \boldsymbol{\Phi}_i,$$

i greška takve reprezentacije iznosi:

$$\Delta\mathbf{x}(M) = \mathbf{x} - \hat{\mathbf{x}}(M) = \sum_{i=1}^{N} x_i \boldsymbol{\Phi}_i - \left(\sum_{i=1}^{M} x_i \boldsymbol{\Phi}_i + \sum_{i=M+1}^{N} b_i \boldsymbol{\Phi}_i \right) = \sum_{i=M+1}^{N} (x_i - b_i) \boldsymbol{\Phi}_i,$$

- Mera greške reprezentacije je srednja kvadratna vrednost norme vektora $\Delta\mathbf{x}(M)$:

$$\overline{\varepsilon^2}(M) = E\{\|\Delta\mathbf{x}(M)\|^2\} = \sum_{i=M+1}^{N} E\{(x_i - b_i)^2\}$$

Razlaganje na glavne komponente (PCA)

- Optimalne vrednosti za b_i mogu se naći izjednačavanjem izvoda srednje kvadratne greške po b_i sa nulom:

$$\frac{\partial}{\partial b_i} E\{(x_i - b_i)^2\} = -2(E\{x_i\} - b_i) = 0 \quad \Rightarrow \quad b_i = E\{x_i\},$$

  - Izostavljena obeležja treba zameniti njihovim očekivanim vrednostima (što je i logično)

- Srednja kvadratna greška se zatim može napisati u obliku:

$$\overline{\varepsilon^2}(M) = \sum_{i=M+1}^{N} E\{(x_i - E\{x_i\})^2\} = \sum_{i=M+1}^{N} E\{(\mathbf{x}^\mathrm{T}\boldsymbol{\phi}_i - E\{\mathbf{x}^\mathrm{T}\boldsymbol{\phi}_i\})^2\}$$

$$= \sum_{i=M+1}^{N} E\{(\mathbf{x}^\mathrm{T}\boldsymbol{\phi}_i - E\{\mathbf{x}^\mathrm{T}\boldsymbol{\phi}_i\})^\mathrm{T}(\mathbf{x}^\mathrm{T}\boldsymbol{\phi}_i - E\{\mathbf{x}^\mathrm{T}\boldsymbol{\phi}_i\})\}$$

$$= \sum_{i=M+1}^{N} \boldsymbol{\phi}_i^\mathrm{T} E\{(\mathbf{x} - E\{\mathbf{x}\})(\mathbf{x} - E\{\mathbf{x}\})^\mathrm{T}\}\boldsymbol{\phi}_i = \sum_{i=M+1}^{N} \boldsymbol{\phi}_i^\mathrm{T}\boldsymbol{\Sigma}_\mathbf{x}\boldsymbol{\phi}_i$$

- Prethodni izraz treba minimizovati po $\boldsymbol{\phi}_i$ uz ograničenje da $\boldsymbol{\phi}_i$ moraju biti ortonormalni
  - To se rešava metodom Lagrangeovih multiplikatora, odnosno, formiranjem izraza:

$$\overline{\varepsilon^2}(M) = \sum_{i=M+1}^{N} \boldsymbol{\phi}_i^\mathrm{T}\boldsymbol{\Sigma}_\mathbf{x}\boldsymbol{\phi}_i + \sum_{i=M+1}^{N} \lambda_i(1 - \boldsymbol{\phi}_i^\mathrm{T}\boldsymbol{\phi}_i),$$

  i izjednačavanjem njegovih izvoda po $\boldsymbol{\phi}_i$ sa nulom

Razlaganje na glavne komponente (PCA)

- Izjednačavanjem izvoda srednje kvadratne greške po $\boldsymbol{\phi}_i$ sa nulom dobija se:

$$\frac{\partial}{\partial \boldsymbol{\phi}_i}\overline{\varepsilon^2}(M) = \frac{\partial}{\partial \boldsymbol{\phi}_i}\left[\sum_{i=M+1}^{N}\boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\phi}_i + \sum_{i=M+1}^{N}\lambda_i(1-\boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{\phi}_i)\right]$$
$$= 2(\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\phi}_i - \lambda_i\boldsymbol{\phi}_i) = 0,$$

  odakle sledi da su $\boldsymbol{\phi}_i$ karakteristični vektori kovarijansne matrice $\boldsymbol{\Sigma}_{\mathbf{x}}$ a λ_i odgovarajuće karakteristične vrednosti

- Srednja kvadratna greška može se sada zapisati u obliku:

$$\overline{\varepsilon^2}(M) = \sum_{i=M+1}^{N}\boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\phi}_i = \sum_{i=M+1}^{N}\boldsymbol{\phi}_i^{\mathrm{T}}\Lambda_i\boldsymbol{\phi}_i = \sum_{i=M+1}^{N}\Lambda_i,$$

  i da bi se ovaj izraz minimizovao, Λ_i treba da budu upravo *najmanje karakteristične vrednosti* $\boldsymbol{\Sigma}_{\mathbf{x}}$

- Stoga, da bi slučajni vektor **x** bio predstavljen sa minimalnom srednjom kvadratnom greškom, treba ga predstaviti preko vektora $\boldsymbol{\phi}_i$ koji predstavljaju karakteristične vektore kovarijansne matrice $\boldsymbol{\Sigma}_{\mathbf{x}}$, i to one kojima odgovaraju *najveće karakteristične vrednosti* λ_i

Rezime

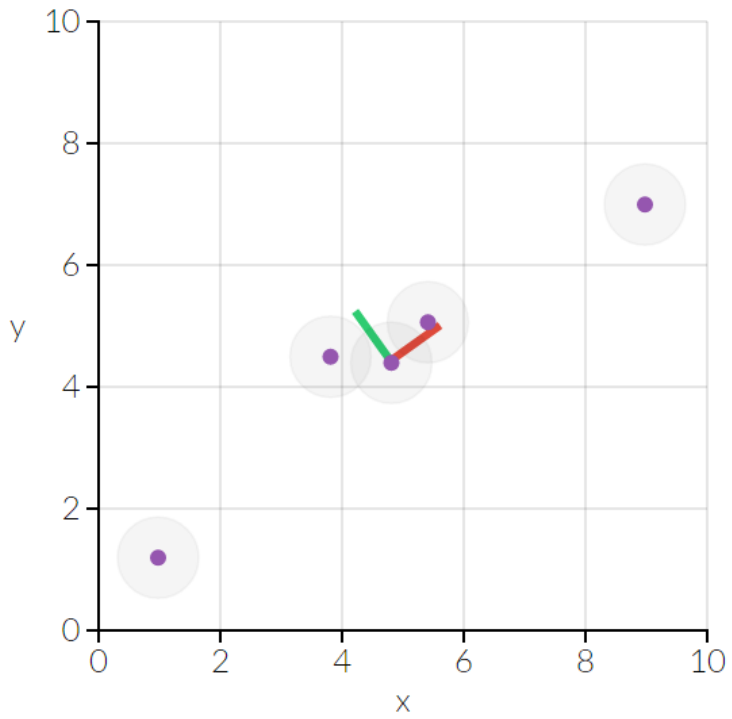
Smanjenje dimenzionalnosti pomoću PCA

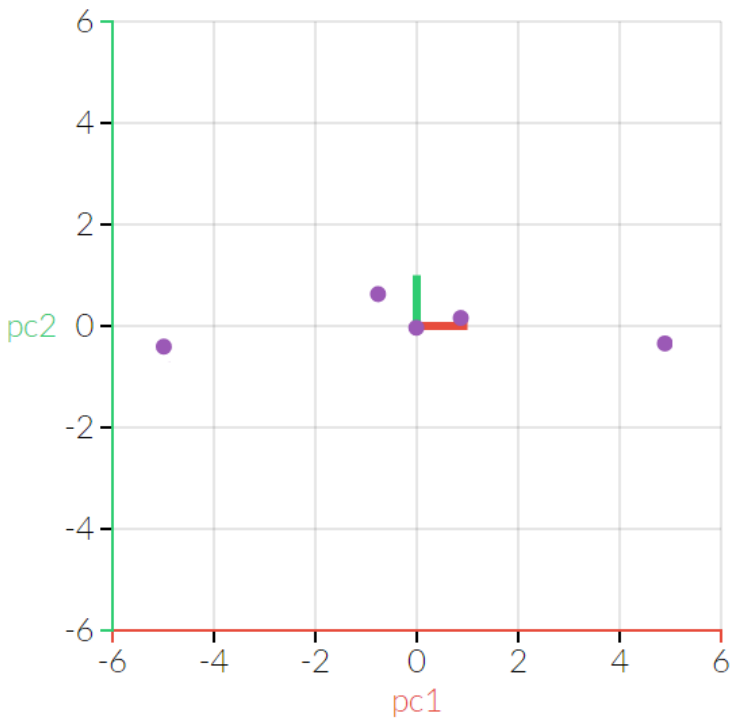
Optimalna* aproksimacija slučajnog vektora $\mathbf{x} \in \mathbb{R}^N$ na osnovu linearne kombinacije M ($M < N$) ortonormalnih vektora dobija se projekcijom vektora $\mathbf{x}$ na karakteristične vektore $\boldsymbol{\phi}_i$ kovarijansne matrice $\boldsymbol{\Sigma}_\mathbf{x}$, i to one kojima odgovara M najvećih karakterističnih vrednosti λ_i.

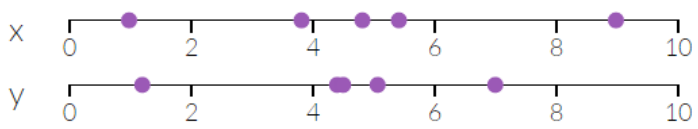
- Za svaki skup podataka dimenzionalnosti N postoji *unutrašnja (stvarna) dimenzionalnost* M, a to je minimalni broj slobodnih parametara preko kojih se on može opisati
  - Ako je unutrašnja dimenzionalnost manja od polazne, to znači da su podaci redundantni
  - Ako je $M < N$, to i dalje ne znači da će PCA moći da otkrije unutrašnju dimenzionalnost
- U opštem slučaju PCA dekoreliše ose u prostoru obeležja
  - Ako je u pitanju unimodalna Gaussova raspodela, PCA samim tim otkriva nezavisne ose
- Glavno ograničenje PCA je da ne razmatra separabilnost između klasa jer ne uzima u obzir oznake klasa pojedinih uzoraka
  - PCA vrši samo rotaciju koordinatnog prostora tako što poravnava ose rotiranog prostora sa pravcima maksimalne varijanse
  - Nema garancije da će pravci sa maksimalnom varijansom biti dobri za diskriminaciju!

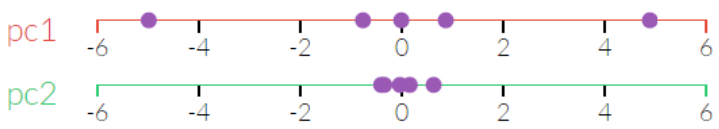
* Optimalnost je definisana kroz minimizaciju srednje kvadratne greške aproksimacije

Primer 1

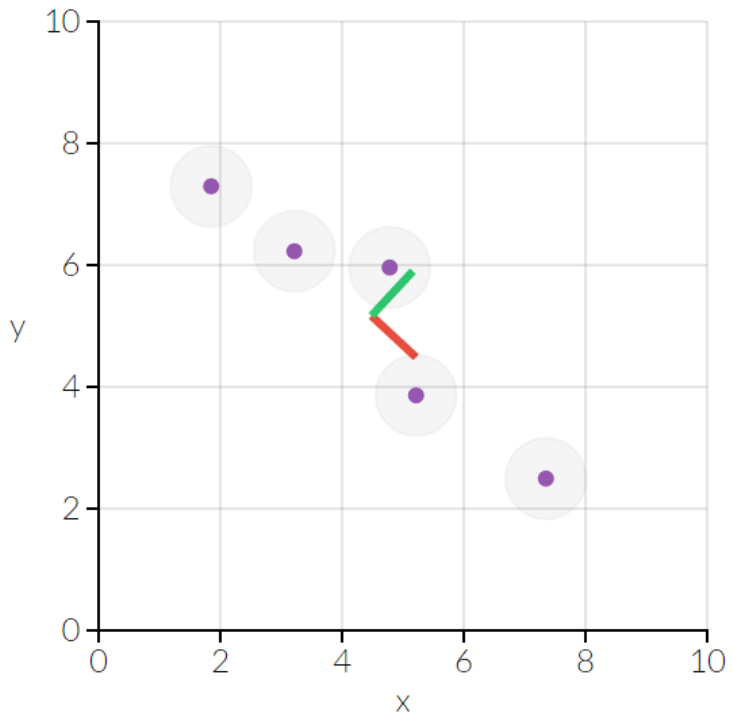


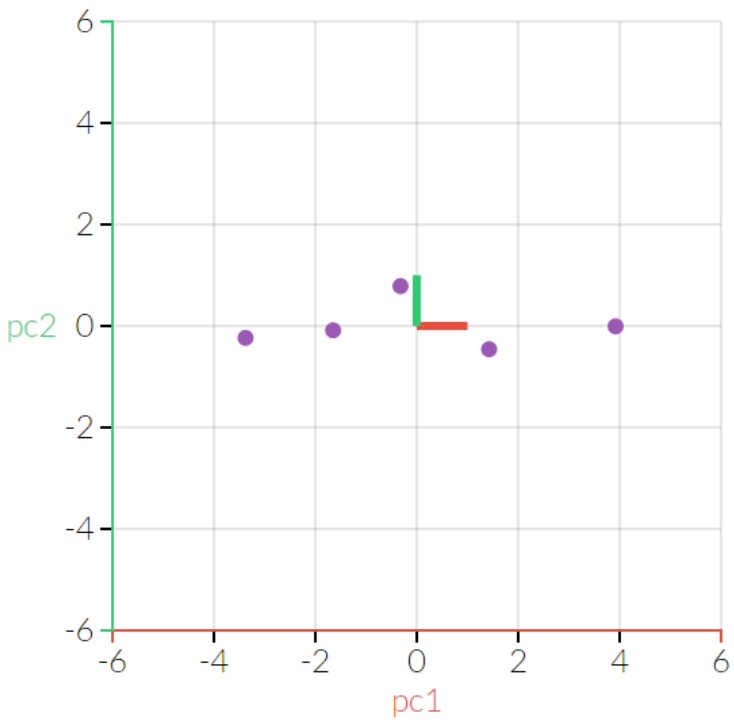


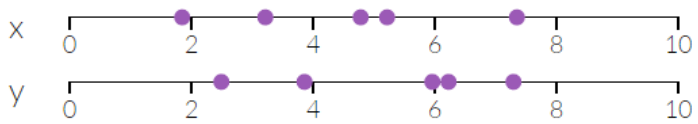


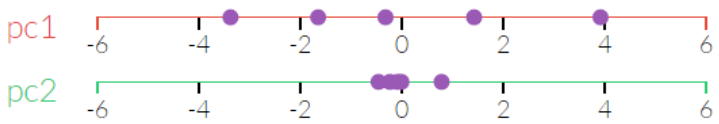


Primer 1

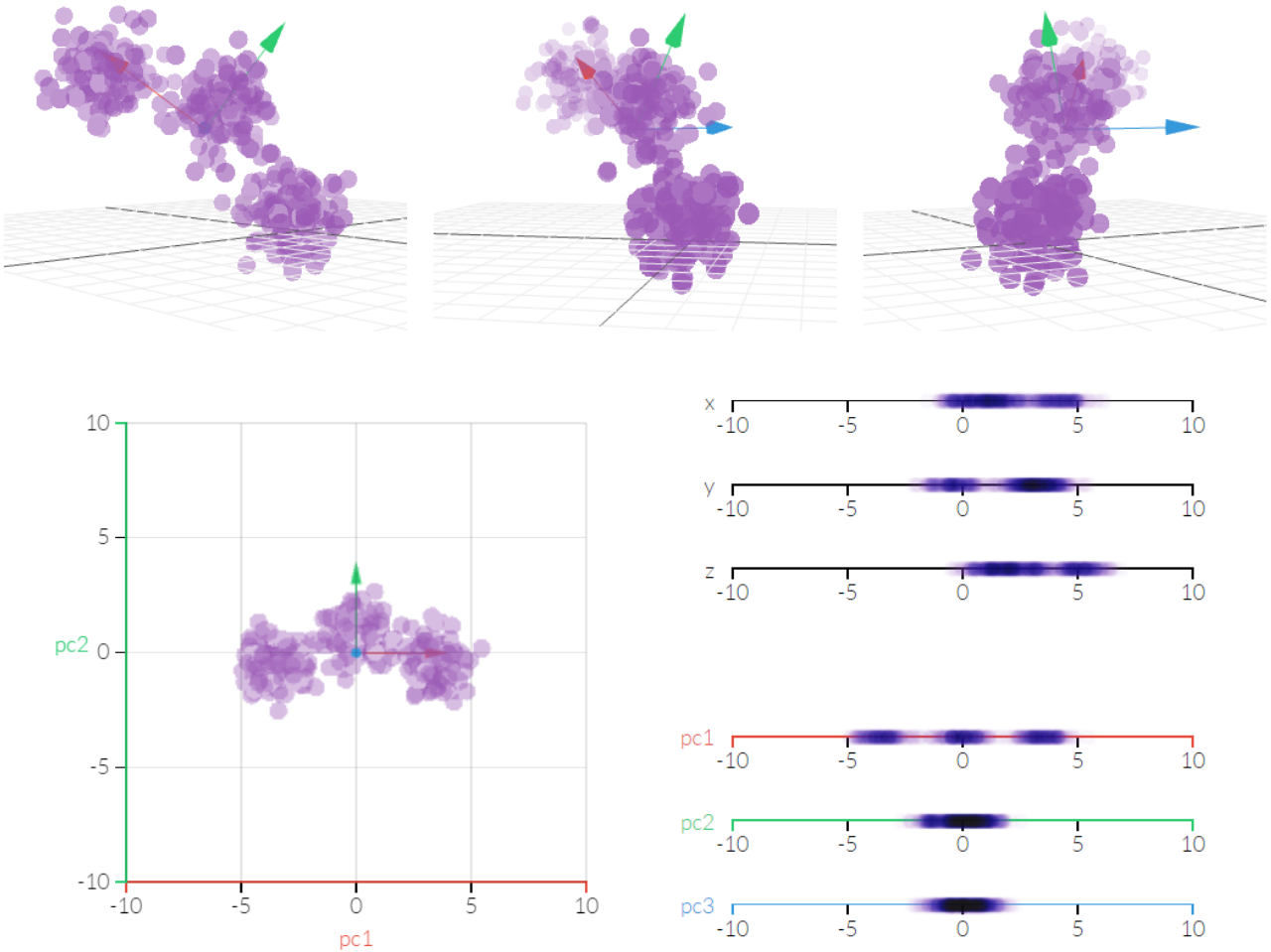

10
8
6
4
2
0
y
0
2
4
6
8
10
x


6
4
2
0
-2
-4
-6
pc2
-6
-4
-2
0
2
4
6
pc1


x
y
0
2
4
6
8
10

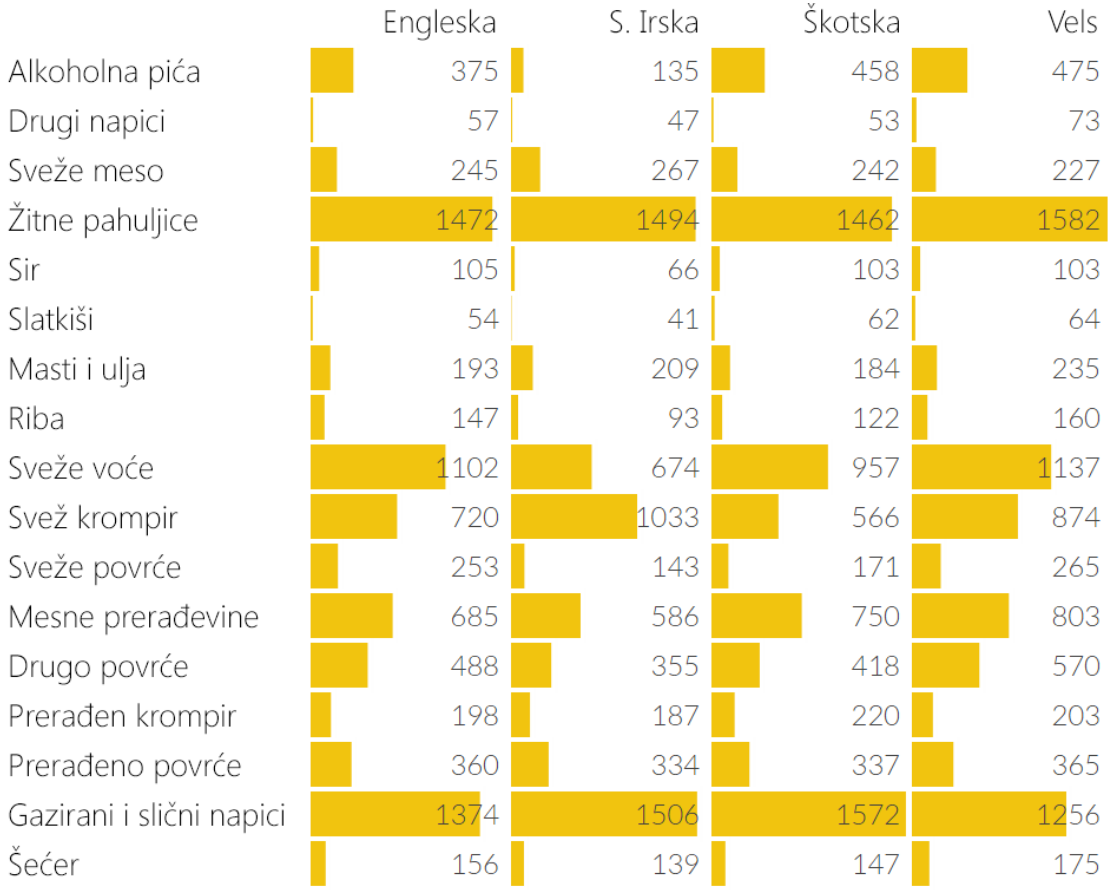

pc1
pc2
-6
-4
-2
0
2
4
6

Primer 2



Primer 3

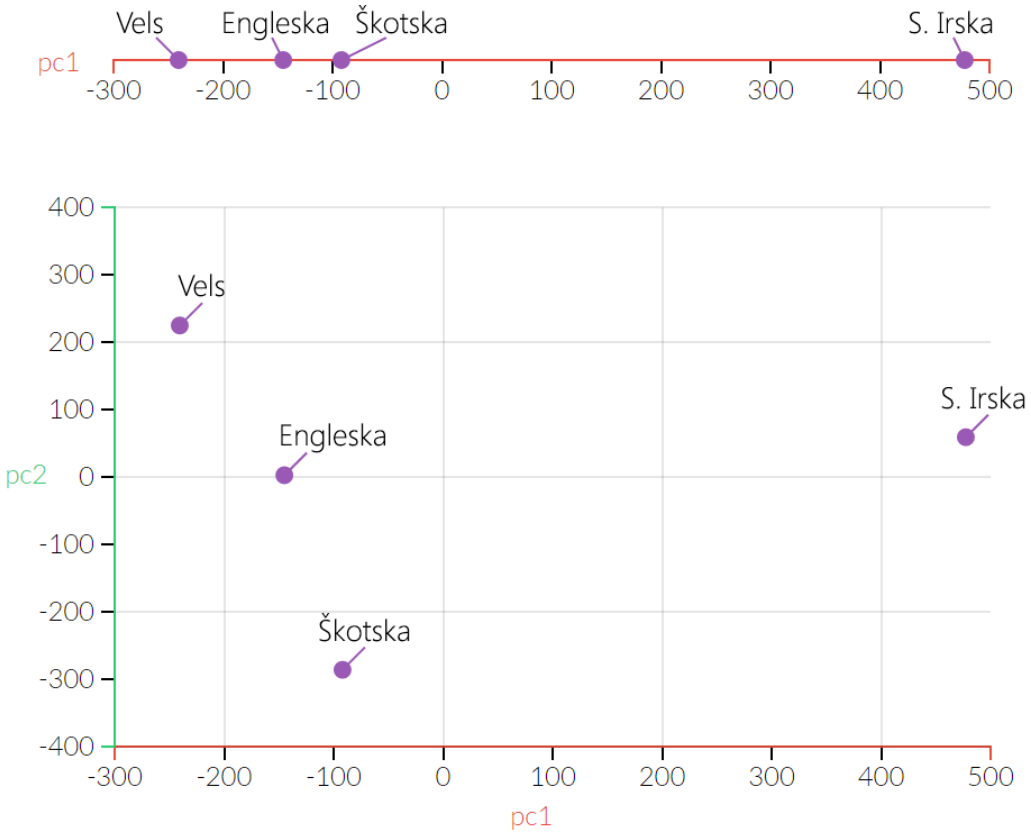
U tabeli je prikazana prosečna potrošnja određenih prehrambenih proizvoda u gramima dnevno po stanovniku u pojedinim regionima Ujedinjenog Kraljevstva



	Engleska	S. Irska	Škotska	Vels
Alkoholna pića	375	135	458	475
Drugi napici	57	47	53	73
Sveže meso	245	267	242	227
Žitne pahuljice	1472	1494	1462	1582
Sir	105	66	103	103
Slatkiši	54	41	62	64
Masti i ulja	193	209	184	235
Riba	147	93	122	160
Sveže voće	1102	674	957	1137
Svež krompir	720	1033	566	874
Sveže povrće	253	143	171	265
Mesne prerađevine	685	586	750	803
Drugo povrće	488	355	418	570
Prerađen krompir	198	187	220	203
Prerađeno povrće	360	334	337	365
Gazirani i slični napici	1374	1506	1572	1256
Šećer	156	139	147	175

Primer 3

Nakon PCA mnogo je jasnije koji se region izdvaja u odnosu na ostale, a to je ujedno i region koji je geografski odvojen od njih

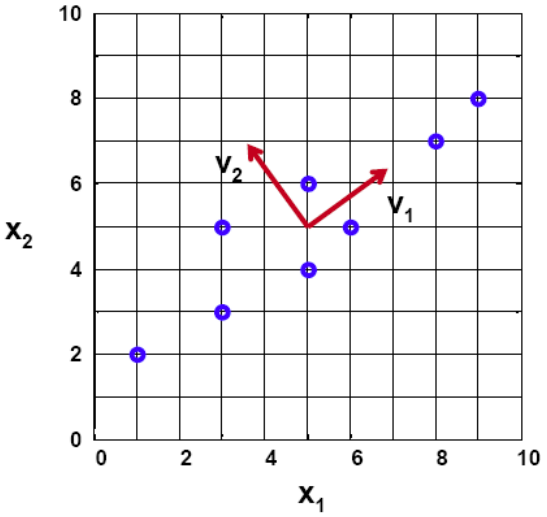


Primer 4

- Izračunati glavne komponente za sledeći 2-d skup podataka: X = {(1,2), (3,3), (3,5), (5,4), (5,6), (6,5), (8,7), (9,8)}

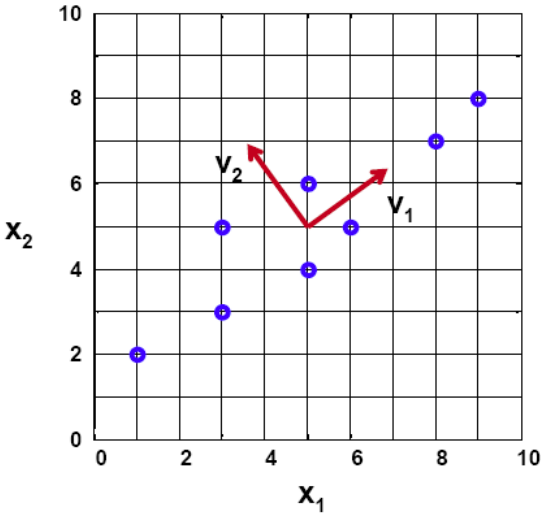
Primer 4

- Izračunati glavne komponente za sledeći 2-d skup podataka: X = {(1,2), (3,3), (3,5), (5,4), (5,6), (6,5), (8,7), (9,8)}



Primer 4

- Izračunati glavne komponente za sledeći 2-d skup podataka: X = {(1,2), (3,3), (3,5), (5,4), (5,6), (6,5), (8,7), (9,8)}



- Rešenje:
  - Necentrirana estimacija kovarijansne matrice skupa uzoraka je:

$$\boldsymbol{\Sigma}_x = \begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}$$

  - Karakteristične vrednosti kovarijansne matrice su nule odgovarajuće karakteristične jednačine:

$$\boldsymbol{\Sigma}_x \mathbf{v} = \lambda \mathbf{v} \;\Rightarrow\; \left|\boldsymbol{\Sigma}_x - \lambda \mathbf{I}\right| = \begin{vmatrix} 6.25-\lambda & 4.25 \\ 4.25 & 3.5-\lambda \end{vmatrix} = 0 \;\Rightarrow\; \lambda_1 = 9{,}34;\ \lambda_2 = 0{,}41$$

  - Karakteristični vektori su rešenja sledećih (linearnih) sistema jednačina:

$$\begin{bmatrix} 6{,}25 & 4{,}25 \\ 4{,}25 & 3{,}5 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} \lambda_1 v_{11} \\ \lambda_1 v_{12} \end{bmatrix} \;\Rightarrow\; \mathbf{v}_1 = \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0{,}81 \\ 0{,}59 \end{bmatrix}$$

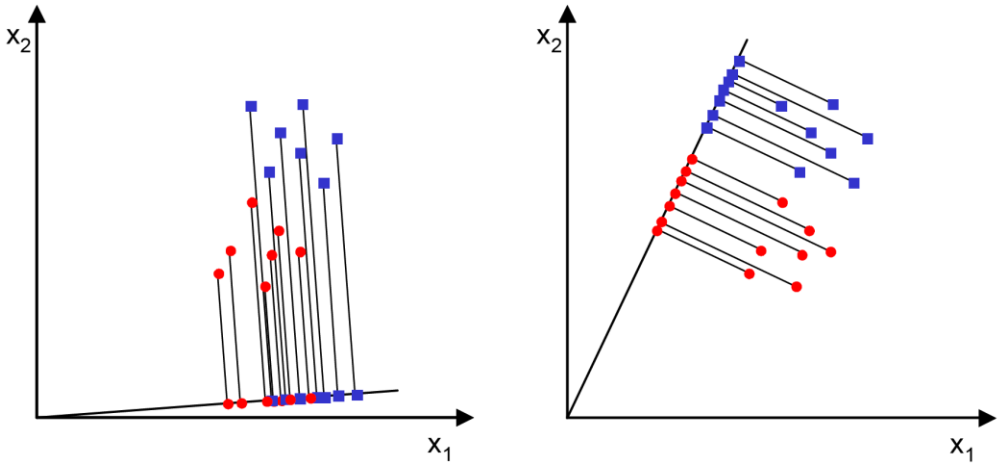
$$\begin{bmatrix} 6{,}25 & 4{,}25 \\ 4{,}25 & 3{,}5 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} \lambda_2 v_{21} \\ \lambda_2 v_{22} \end{bmatrix} \;\Rightarrow\; \mathbf{v}_2 = \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} -0{,}59 \\ 0{,}81 \end{bmatrix}$$

Linearna diskriminantna analiza (2 klase)

- Cilj LDA je da smanji dimenzionalnost, a da pritom sačuva što je moguće više diskriminatornih informacija
- Ako je dat skup D-dimenzionalnih uzoraka $\{\mathbf{x}^{(1)},\mathbf{x}^{(2)},\ldots,\mathbf{x}^{(N)}\}$, od kojih N_1 pripadaju klasi ω_1 i N_2 koji pripadaju klasi ω_2, cilj je naći pravac na koji ove uzorke treba projektovati tako da se maksimizuje separabilnost između klasa
  - Matematički gledano, projektovanjem uzoraka $\mathbf{x}$ na pravac $\mathbf{w}$ dobijaju se skalari y:

$$y = \mathbf{w}^T\mathbf{x},$$

tako da se dimenzionalnost smanjuje na 1, i od svih mogućih pravaca treba odabrati onaj koji maksimizuje separabilnost skalarâ koji se odnose na različite klase



Linearna diskriminantna analiza (2 klase)

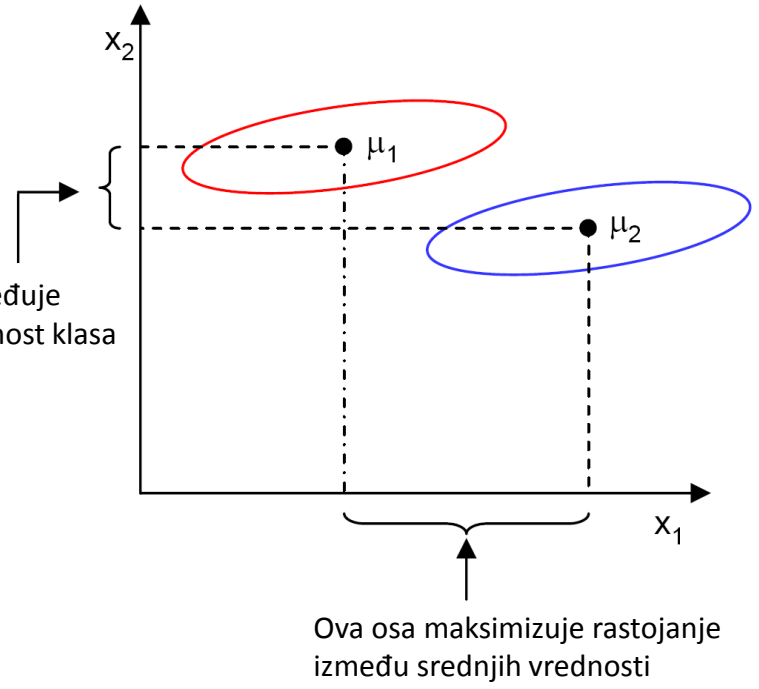
- Da bi se odredio pravac **w** potrebno je odrediti meru separabilnosti između projektovanih klasa
  - Vektor srednje vrednosti klasa u polaznom prostoru i odgovarajući skalar u potprostoru projekcija su:

$$\boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{\mathbf{x}\in\omega_i}\mathbf{x} \quad \text{i} \quad \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i}\mathbf{w}^{\mathrm{T}}\mathbf{x} = \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_i$$

  - Kao funkcija cilja može se izabrati rastojanje između projekcija vektora srednjih vrednosti:

$$J(\mathbf{w}) = \left|\tilde{\mu}_1 - \tilde{\mu}_2\right| = \left|\mathbf{w}^{\mathrm{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right|,$$

    ali rastojanje između projekcija vektora srednjih vrednosti nije najbolja mera pošto ne uzima u obzir varijansu uzoraka u okviru klasa



Fisherova linearna diskriminanta

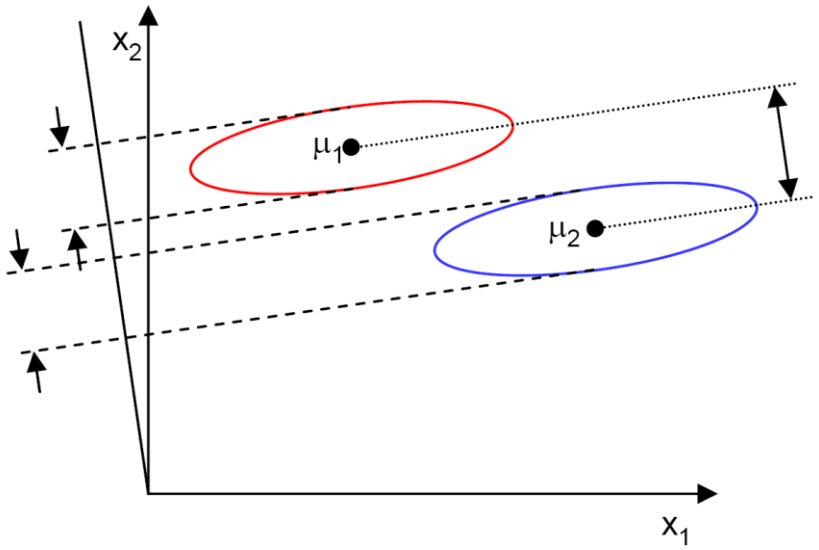
- Rešenje je u minimizaciji funkcije razlike projektovanih srednjih vrednosti normalizovane merom rasipanja uzoraka unutar projektovanih klasa
  - Za svaku od klasa definiše se *rasipanje* projekcija (koje je ekvivalent varijanse):

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$$

  - Veličina $\tilde{s}_1^2 + \tilde{s}_2^2$ predstavlja *unutarklasno rasipanje* projektovanih uzoraka
  - *Fisherova linearna diskriminanta* predstavlja linearnu funkciju $y = \mathbf{w}^T\mathbf{x}$ koja maksimizuje funkciju cilja:

$$J(\mathbf{w}) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

  - Rezultat je projekcija kod koje su uzorci iz iste klase projektovani veoma blizu jedni drugima, dok su centroidi, istovremeno, što je moguće udaljeniji jedan od drugog



Fisherova linearna diskriminanta

- Za nalaženje optimalne projekcije $\mathbf{w}^*$ treba izraziti $J(\mathbf{w})$ kao eksplicitnu funkciju $\mathbf{w}$
- Mera rasipanja u višedimenzionalnom prostoru obeležja $\mathbf{x}$ definiše se *matricom rasipanja*:

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\omega_i} (\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}},$$

odnosno, matricom unutarklasnog rasipanja (eng. *within-class scatter matrix*):

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

- Rasipanje projekcije y može se izraziti kao funkcija matrice rasipanja u prostoru obeležja $\mathbf{x}$:

$$\tilde{s}_i^2 = \sum_{y\in\omega_i} (y-\tilde{\mu}_i)^2 = \sum_{x\in\omega_i} (\mathbf{w}^{\mathrm{T}}\mathbf{x}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_i)^2 = \sum_{x\in\omega_i} \mathbf{w}^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\mathbf{w} = \mathbf{w}^{\mathrm{T}}\mathbf{S}_i\mathbf{w},$$

što znači da je unutarklasno rasipanje projekcija jednako:

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^{\mathrm{T}}\mathbf{S}_1\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{S}_2\mathbf{w} = \mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}$$

- Slično, razlika projekcija srednjih vrednosti može se izraziti preko vektora srednjih vrednosti u polaznom prostoru obeležja:

$$(\mu_1-\mu_2)^2 = (\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_1 - \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_2)^2 = \mathbf{w}^{\mathrm{T}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\mathbf{w} = \mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w},$$

gde je $\mathbf{S}_B$ matrica međuklasnog rasipanja (eng. *between-class scatter matrix*)

  - rang matrice $\mathbf{S}_B$ može biti najviše 1

- Fisherov kriterijum može se sada izraziti preko matrica $\mathbf{S}_B$ i $\mathbf{S}_W$:

$$J(\mathbf{w}) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}}$$

Fisherova linearna diskriminanta

- Maksimum $J(\mathbf{w})$ nalazi se izjednačavanjem izvoda po $\mathbf{w}$ sa nulom:

$$\frac{\mathrm{d}J(\mathbf{w})}{\mathrm{d}\mathbf{w}} = \frac{\mathrm{d}}{\mathrm{d}\mathbf{w}}\left[\frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}}\right] = 0 \quad \Rightarrow \quad \mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}\cdot 2\mathbf{S}_B\mathbf{w} - \mathbf{w}^{\mathrm{T}}\mathbf{S}_B\mathbf{w}\cdot 2\mathbf{S}_W\mathbf{w} = 0$$

- Deljenjem sa $2\mathbf{w}^{\mathrm{T}}\mathbf{S}_W\mathbf{w}$ dobija se:

$$\mathbf{S}_B\mathbf{w} - J(\mathbf{w})\mathbf{S}_W\mathbf{w} = 0 \quad \Rightarrow \quad \mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} - J(\mathbf{w})\mathbf{w} = 0$$

- Poslednja jednačina predstavlja karakteristični problem za matricu $\mathbf{S}_W^{-1}\mathbf{S}_B$ i može se rešavati kao takva, ali može i jednostavnije:

$$J(\mathbf{w})\mathbf{S}_W\mathbf{w} = \mathbf{S}_B\mathbf{w} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\mathbf{w} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\cdot\alpha,$$

  gde su α i $J(\mathbf{w})$ neki skalari (uopšte nije bitno koji jer je od interesa samo *pravac* vektora $\mathbf{w}$!)

- Odatle se, kao optimalni vektor $\mathbf{w}^*$, koji definiše linearnu Fisherovu diskriminantu, dobija:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} J(\mathbf{w}) = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

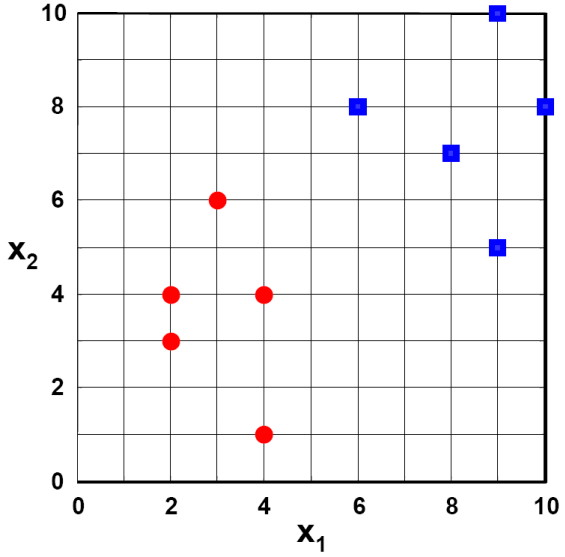
  - Strogo uzevši, $\mathbf{w}$ ne predstavlja diskriminantnu funkciju već samo izbor pravca na koji se uzorci projektuju

Fisherova linearna diskriminanta (primer)

- Naći LDA projekciju za dvodimenzionalni skup podataka:
  - X_1 = {(4,1), (2,4), (2,3), (3,6), (4,4)}
  - X_2 = {(9,10), (6,8), (9,5), (8,7), (10,8)}

Fisherova linearna diskriminanta (primer)

- Naći LDA projekciju za dvodimenzionalni skup podataka:
  - X_1 = {(4,1), (2,4), (2,3), (3,6), (4,4)}
  - X_2 = {(9,10), (6,8), (9,5), (8,7), (10,8)}

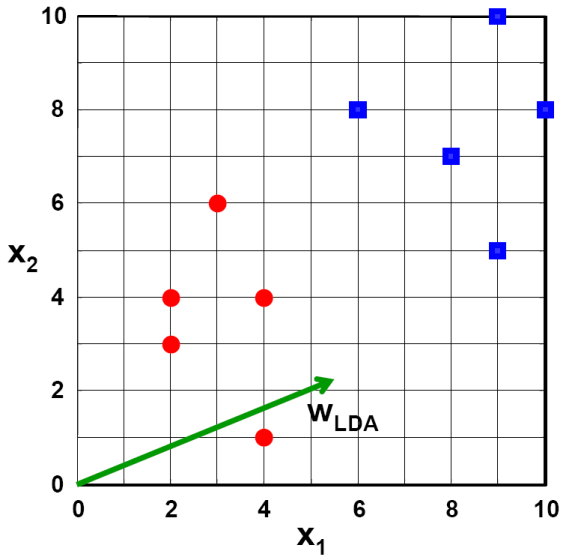


Fisherova linearna diskriminanta (primer)

- Naći LDA projekciju za dvodimenzionalni skup podataka:
  - X_1 = {(4,1), (2,4), (2,3), (3,6), (4,4)}
  - X_2 = {(9,10), (6,8), (9,5), (8,7), (10,8)}
- Rešenje:
  - Srednje vrednosti i kovarijanse u okviru klasa jednake su:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3.00 \\ 3.60 \end{bmatrix};\ \mathbf{S}_1 = \begin{bmatrix} 0{,}80 & -0{,}40 \\ -0{,}40 & 2{,}60 \end{bmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 8.40 \\ 7.60 \end{bmatrix};\ \mathbf{S}_2 = \begin{bmatrix} 1{,}84 & -0{,}04 \\ -0{,}04 & 2{,}64 \end{bmatrix}$$



  - Međuklasno i unutarklasno rasipanje iznosi:

$$\mathbf{S}_B = \begin{bmatrix} 29{,}16 & 21{,}60 \\ 21{,}60 & 16{,}00 \end{bmatrix};\ \mathbf{S}_W = \begin{bmatrix} 2{,}64 & -0{,}44 \\ -0{,}44 & 5{,}28 \end{bmatrix}$$

  - LDA projekcija se dobija kao rešenje karakteristične jednačine:

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{v} = \lambda\mathbf{v} \quad \Rightarrow \quad \left|\mathbf{S}_W^{-1}\mathbf{S}_B - \lambda\mathbf{I}\right| = \begin{vmatrix} 11{,}89-\lambda & 8{,}81 \\ 5{,}08 & 3{,}76-\lambda \end{vmatrix} = 0 \quad \Rightarrow \quad \lambda = 15{,}65$$

$$\begin{bmatrix} 11{,}89 & 8{,}81 \\ 5{,}08 & 3{,}76 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 15{,}65\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0{,}91 \\ 0{,}39 \end{bmatrix}$$

ili direktno:

$$\mathbf{w}^* = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{bmatrix} -0{,}91 \\ -0{,}39 \end{bmatrix}$$

Linearna diskriminantna analiza (K klasa, $K > 2$)

- Fisherova diskriminanta može se iskoristiti i za klasifikaciju u K klasa ($K > 2$):
  - Višestrukom primenom binarne klasifikacije („jedan protiv svih“ ili „svako protiv svakog“)
  - Direktnom generalizacijom u više dimenzija
- Generalizacija Fisherove diskriminante:
  - Umesto jedne projekcije y, sada se traži m projekcija $[y_1, y_2, \ldots, y_m]$, određenih sa m vektora pravaca $\mathbf{w}_i$ (po pravilu je $m \leq K - 1$)
    - Ovi vektori mogu biti poslagani u kolone projekcione matrice $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2\ \ldots\ \mathbf{w}_m]$:

$$y_i = \mathbf{w}_i^{\mathrm{T}}\mathbf{x} \quad \Rightarrow \quad \mathbf{y} = \mathbf{W}^{\mathrm{T}}\mathbf{x}$$

    - Unutarklasno i međuklasno rasipanje u polaznom prostoru obeležja generalizuju se kao:

$$\mathbf{S}_W = \sum_{i=1}^{K}\mathbf{S}_i = \sum_{i=1}^{K}\sum_{\mathbf{x}\in\omega_i}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}, \qquad \boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{\mathbf{x}\in\omega_i}\mathbf{x}$$

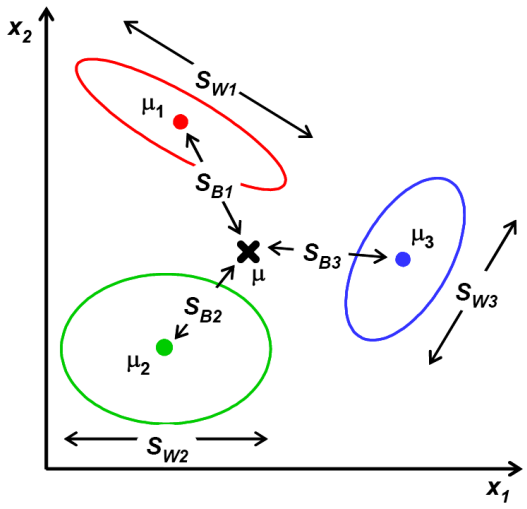
$$\mathbf{S}_B = \sum_{i=1}^{K}N_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^{\mathrm{T}}, \qquad \boldsymbol{\mu} = \frac{1}{N}\sum_{\forall\mathbf{x}}\mathbf{x}$$

    gde je N_i broj uzoraka klase ω_i a N broj uzoraka svih klasa

    - Slično se generalizuju i matrice rasipanja nakon projekcije:

$$\tilde{\mathbf{S}}_W = \sum_{i=1}^{K}\tilde{\mathbf{S}}_i = \sum_{i=1}^{K}\sum_{\mathbf{y}\in\omega_i}(\mathbf{y}-\tilde{\boldsymbol{\mu}}_i)(\mathbf{y}-\tilde{\boldsymbol{\mu}}_i)^{\mathrm{T}}, \qquad \tilde{\boldsymbol{\mu}}_i = \frac{1}{N_i}\sum_{\mathbf{y}\in\omega_i}\mathbf{y}$$

$$\tilde{\mathbf{S}}_B = \sum_{i=1}^{K}N_i(\tilde{\boldsymbol{\mu}}_i-\tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}_i-\tilde{\boldsymbol{\mu}})^{\mathrm{T}}, \qquad \tilde{\boldsymbol{\mu}} = \frac{1}{N}\sum_{\forall\mathbf{y}}\mathbf{y}$$



Linearna diskriminantna analiza (K klasa, $K > 2$)

- Na osnovu prethodnog izvođenja za slučaj dveju klasa važi:

$$\tilde{\mathbf{S}}_W = \mathbf{W}^{\mathrm{T}} \mathbf{S}_W \mathbf{W}$$

$$\tilde{\mathbf{S}}_B = \mathbf{W}^{\mathrm{T}} \mathbf{S}_B \mathbf{W}$$

- I u ovom slučaju se traži projekcija koja maksimizuje odnos međuklasnog i unutarklasnog rasipanja, a pošto projekcija sada više nije skalar (ima m dimenzija), skalarna funkcija cilja dobija se kao odnos odgovarajućih determinanata:

$$J(\mathbf{W}) = \frac{|\tilde{\mathbf{S}}_B|}{|\tilde{\mathbf{S}}_W|} = \frac{|\mathbf{W}^{\mathrm{T}} \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^{\mathrm{T}} \mathbf{S}_W \mathbf{W}|}$$

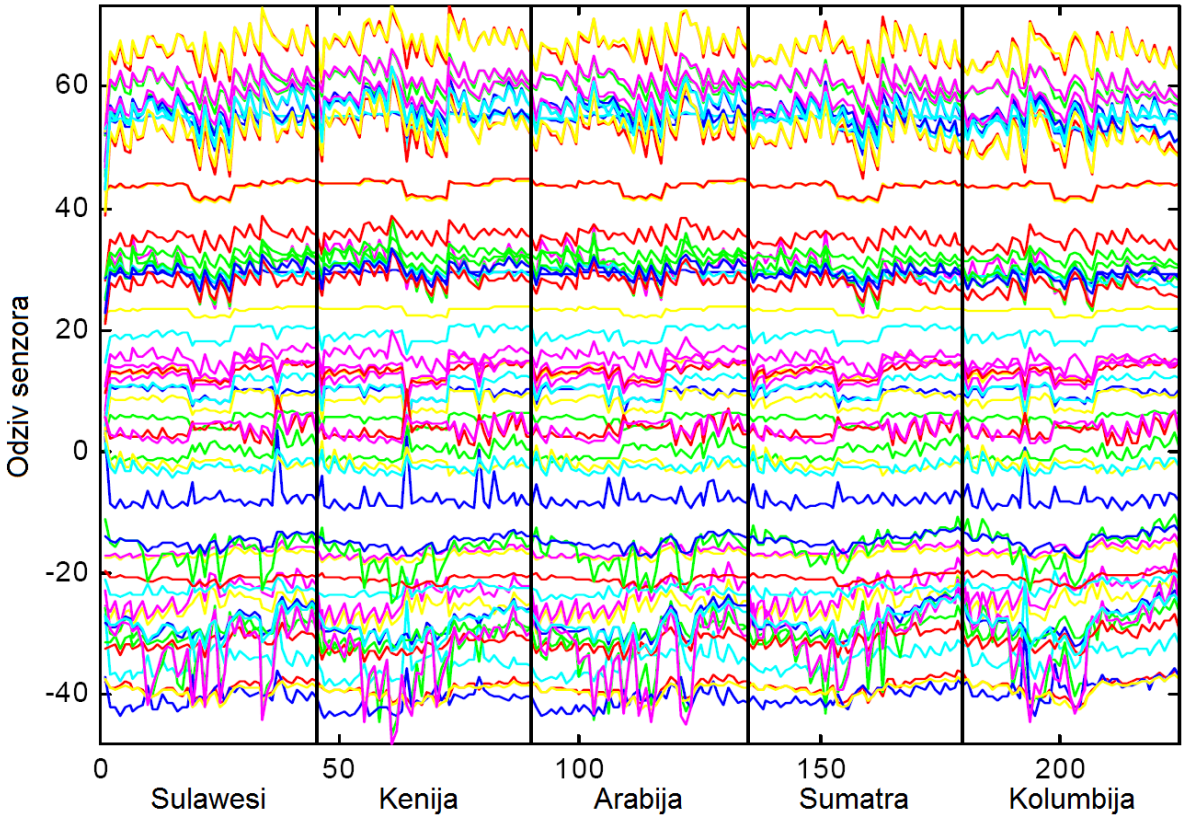
- Može se pokazati da je matrica projekcije $\mathbf{W}^*$ koja maksimizuje ovaj odnos ona čije su kolone karakteristični vektori koji odgovaraju najvećim karakterističnim vrednostima $\mathbf{S}_W^{-1}\mathbf{S}_B$

$$\mathbf{W}^* = \left[\mathbf{w}_1^* \;\; \mathbf{w}_2^* \;\; \ldots \;\; \mathbf{w}_m^*\right] = \arg\max_{\mathbf{w}} \left\{ \frac{\mathbf{W}^T \mathbf{S}_B \mathbf{W}}{\mathbf{W}^T \mathbf{S}_W \mathbf{W}} \right\} \;\Rightarrow\; (\mathbf{S}_B - \lambda_i \mathbf{S}_W)\mathbf{w}_i^* = 0$$

  - Pošto je $\mathbf{S}_B$ zbir K matrica ranga najviše 1, a pri tom postoji i linearna zavisnost između $\boldsymbol{\mu}$ i pojedinih $\boldsymbol{\mu}_i$, $\mathbf{S}_B$ je ranga najviše $K - 1$, pa isto važi i za $\mathbf{S}_W^{-1}\mathbf{S}_B$
  - Zbog toga postoji najviše $K - 1$ karakterističnih vrednosti $\mathbf{S}_W^{-1}\mathbf{S}_B$ različitih od nule, što je i razlog zašto je kod ove metode $m \leq K - 1$

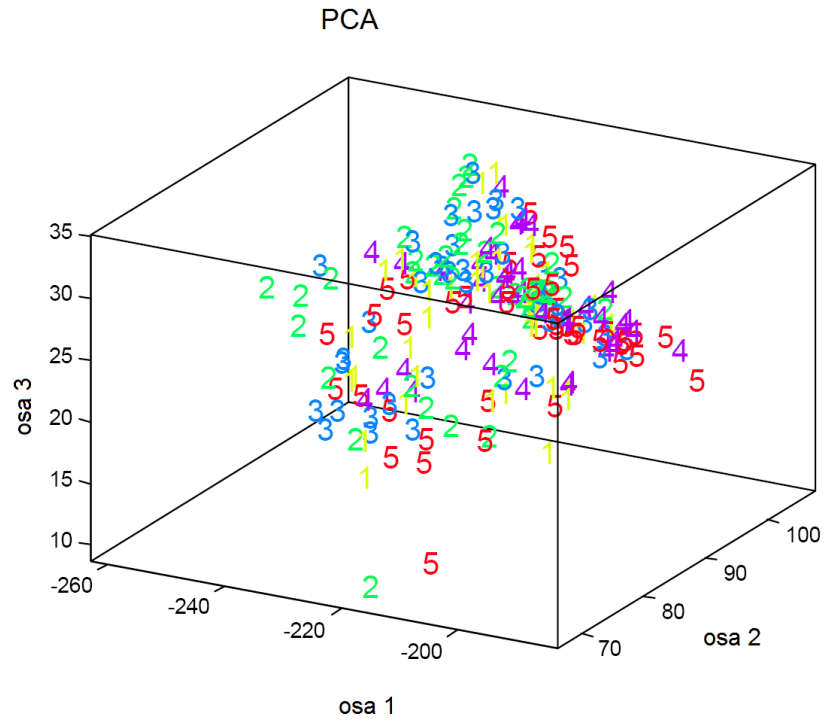
Primer 1 (klasifikacija kafe)

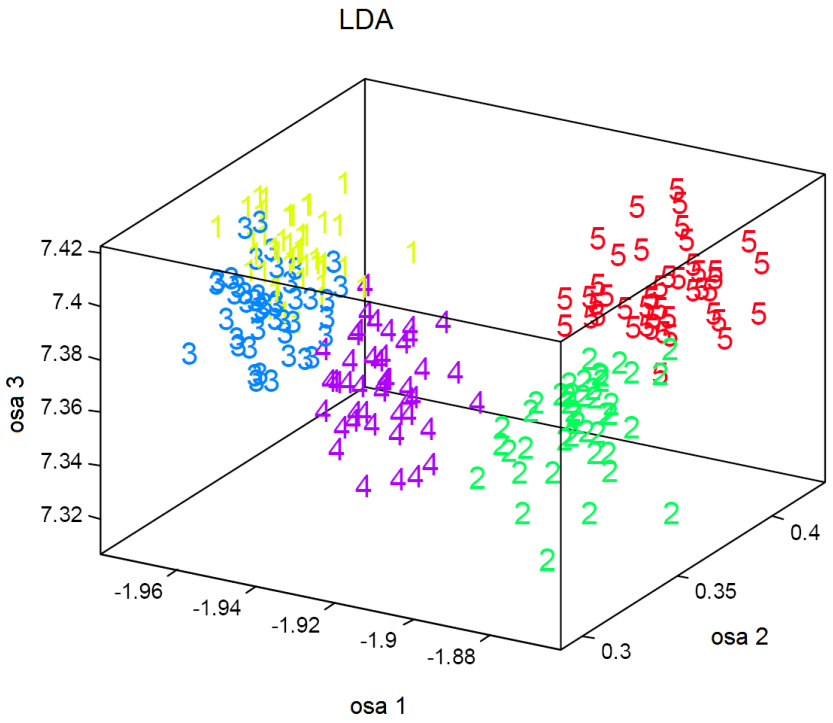
- Pet različitih vrsta kafe izloženo je nizu hemijskih gasnih senzora
- Svaka vrsta kafe „pomirisana“ je 45 puta i odziv gasnih senzora obrađen je da bi se dobio 60-dimenzioni vektor obeležja



Primer 1 (klasifikacija kafe)

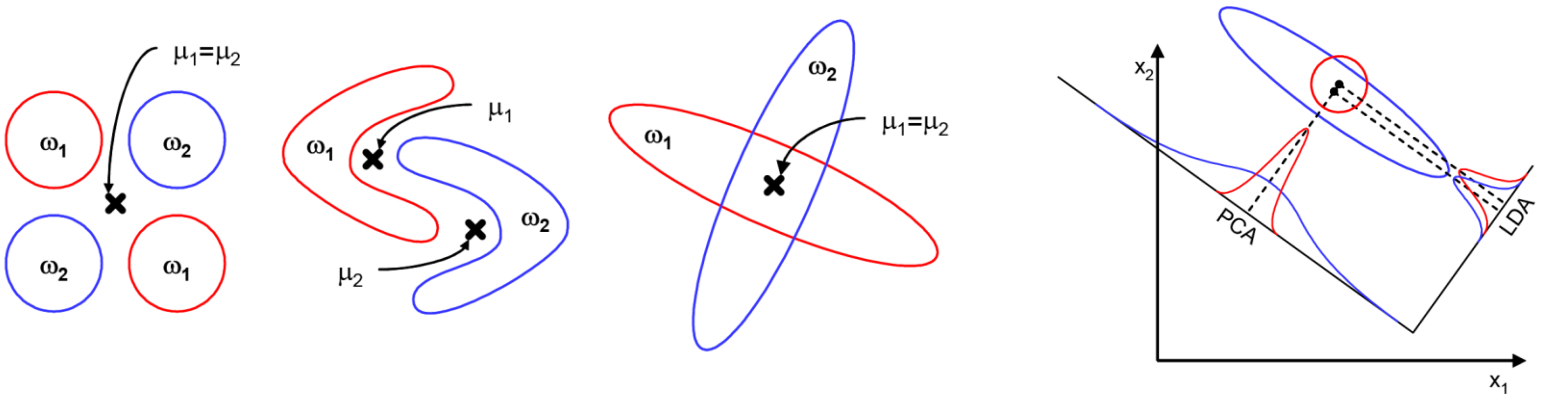
- Sa 3-d dijagrama rasejanja jasno se vidi da LDA u ovom slučaju nadmašuje PCA u pogledu klasne diskriminacije
  - Ovo je još jedan primer u kom se pravac sa najvećom diskriminatornom informacijom ne poklapa sa pravcem maksimalne varijanse





Ograničenja LDA

- LDA daje maksimalno $K - 1$ projekcija vektora obeležja
  - Ako se na osnovu procene verovatnoće greške klasifikacije ustanovi da je potrebno više obeležja, neke druge metode se moraju upotrebiti da bi se obezbedila ta dodatna obeležja
- LDA je parametarski metod jer podrazumeva da su obeležja unutar klasa raspodeljena po Gaussovoj (unimodalnoj) raspodeli
  - Ako raspodele značajno odstupaju od Gaussove, LDA projekcije neće moći da očuvaju kompleksnu strukturu podataka, koja može biti neophodna za klasifikaciju (slike levo)



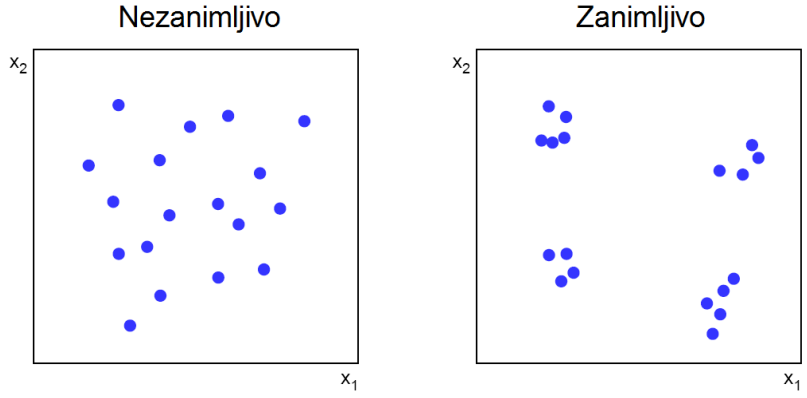
- LDA je neuspešna i kada se diskriminatorna informacija ne nalazi u srednjoj vrednosti, već u varijansi uzorka (slika desno)

Varijante LDA

- Neparametarska LDA (Fukunaga)
  - Ne podrazumeva Gaussovu raspodelu, već izračunava matricu međuklasnog rasipanja $\mathbf{S}_B$ koristeći lokalnu informaciju i *kNN* pravilo:
    - Matrica $\mathbf{S}_B$ je punog ranga i omogućava dobijanje više od $K - 1$ obeležja
- Ortonormalna LDA (Okada & Tomita)
  - Određuju se projekcije koje maksimizuju Fisherov kriterijum i ujedno su ortonormalne
    - Koristi se metoda koja kombinuje rešavanje karakteristične jednačine za $\mathbf{S}_W^{-1}\mathbf{S}_B$ i Gram-Schmidtov postupak ortonormalizacije
    - Metoda sekvencijalno pronalazi ose koje maksimizuju Fisherov kriterijum u potprostoru koji je ortogonalan prostoru već izdvojenih obeležja
    - Metoda takođe omogućava pronalaženje više od $K - 1$ obeležja
- Generalizovana LDA (Lowe)
  - GLDA uopštava Fisherov kriterijum uvođenjem funkcije cene, slične onoj uvedenoj pri izračunavanju Bayesovog rizika
    - Parovi klasa s većom cenom C_{ij} biće više udaljeni u niždimenzionalnom prostoru projekcija
- Višeslojni perceptroni (Webb & Lowe)
  - Skriveni slojevi perceptrona vrše nelinearnu diskriminantnu analizu
    - Matrice rasipanja izmerene su na izlazu iz poslednjeg skrivenog sloja

Druge metode za smanjenje dimenzionalnosti

- Exploratory Projection Pursuit (Friedman & Tukey)
  - Traga se za *M*-dimenzionalnom linearnom projekcijom podataka koja maksimizuje meru "zanimljivosti" (obično *M* = 2 ili 3)
  - Zanimljivost se definiše kao odstupanje od višedimenzione normalne raspodele
    - Ova mera nije isto što i varijansa i obično je invarijantna na skaliranje podataka, a invarijantna je i na afine transformacije, tako da ne zavisi od korelacije između obeležja
  - EPP traži projekcije koje maksimalno razdvajaju klastere a pri tome ih održavaju kompaktnim
    - Metoda je slična Fisherovom kriterijumu ali ne koristi klasne labele
    - Kada se jedna interesantna projekcija pronađe, važno je ukloniti strukturu koju ta projekcija otkriva da bi se druge zanimljive projekcije lakše pronašle



Druge metode za smanjenje dimenzionalnosti

- Nelinearno preslikavanje (Sammon)
  - Traži se preslikavanje na *M*-dimenzionalni prostor koje bi očuvalo rastojanja između tačaka u originalnom *N*-dimenzionalnom prostoru
    - Ovo se postiže minimizovanjem sledeće funkcije cilja:

$$E(d,d') = \sum_{i<j} \frac{(d(P_i,P_j) - d'(P'_i,P'_j))^2}{d(P_i,P_j)}$$

    - Originalni metod nije eksplicitno određivao preslikavanje, već je samo tabelarno preslikavao elemente iz skupa za obuku
    - Novije realizacije zasnovane na primeni neuralne mreže određuju eksplicitno preslikavanje i takođe uzimaju u obzir funkcije cene
    - Sammonovo preslikavanje povezano je s multidimenzionalnim skaliranjem podataka (MDS), multivarijabilnom statističkom tehnikom koja se koristi u društvenim naukama