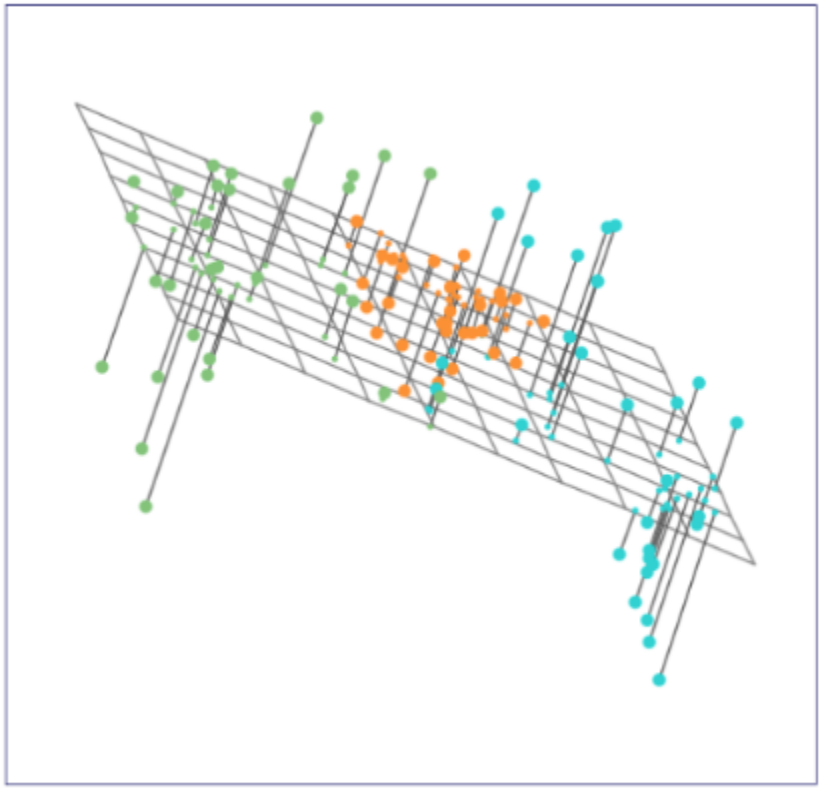
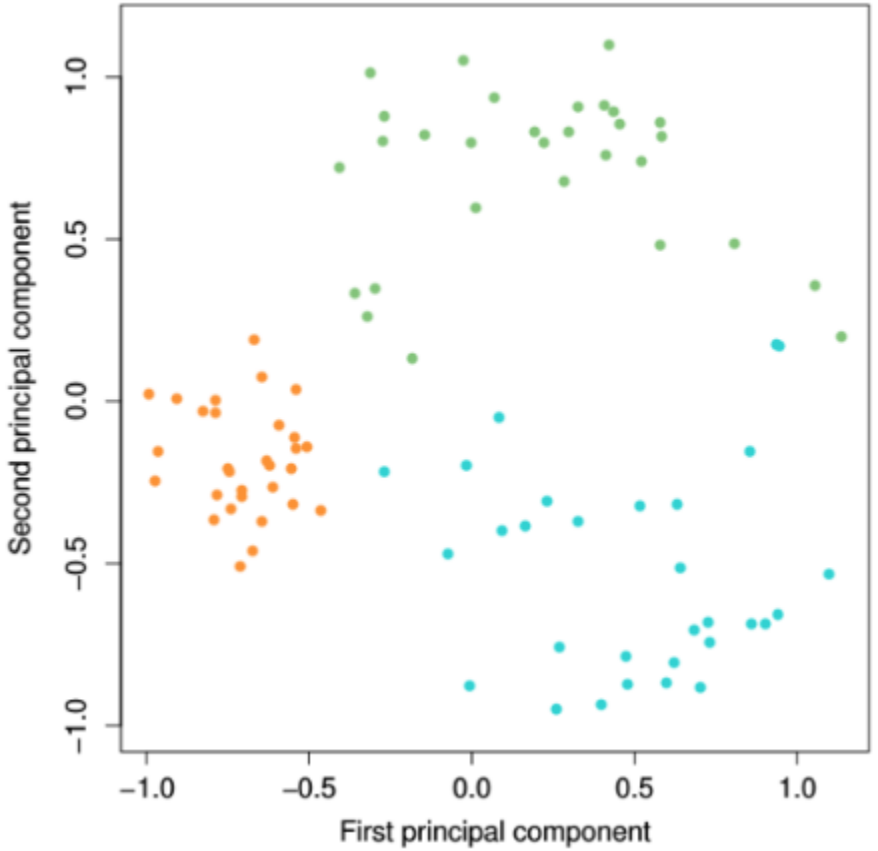




Redukcija dimenzionalnosti

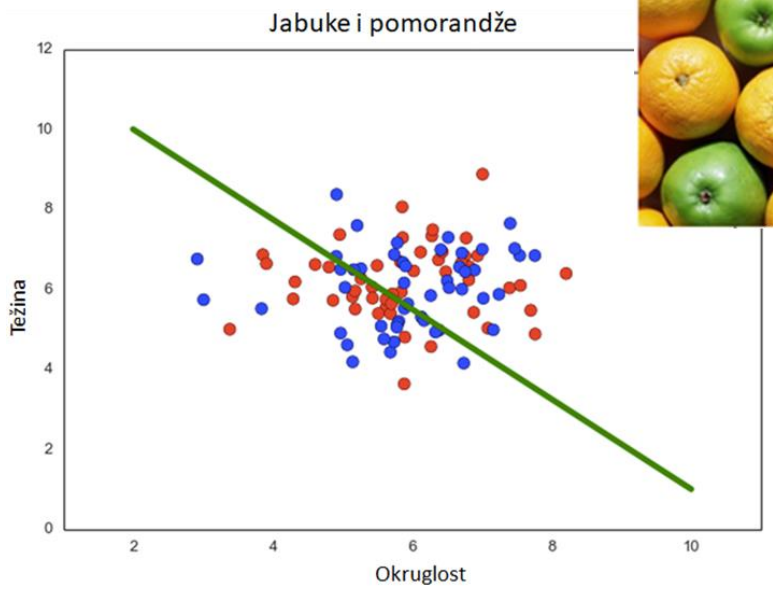
PCA (*Principal Component Analysis*)

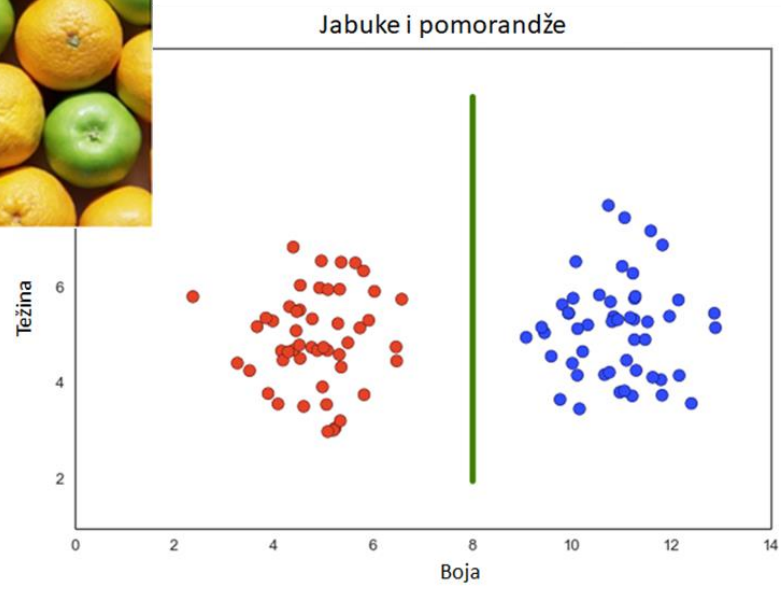
Kako da smanjimo broj dimenzija?

- Ideja: kreirati novi podskup obeležja koji dobro sumarizuje polazna obeležja

- Dobar podskup obeležja je onaj koji je *relevantan* za ciljnu funkciju f

- Na primer, onaj koji ima veliki kapacitet da napravi razliku između primera različitih klasa





Zbog čega želimo manje dimenzija?

1. Kompresija
   - manje zauzeće memorije i diska
   - (važnije) značajno ubrzanje obučavajućih algoritama

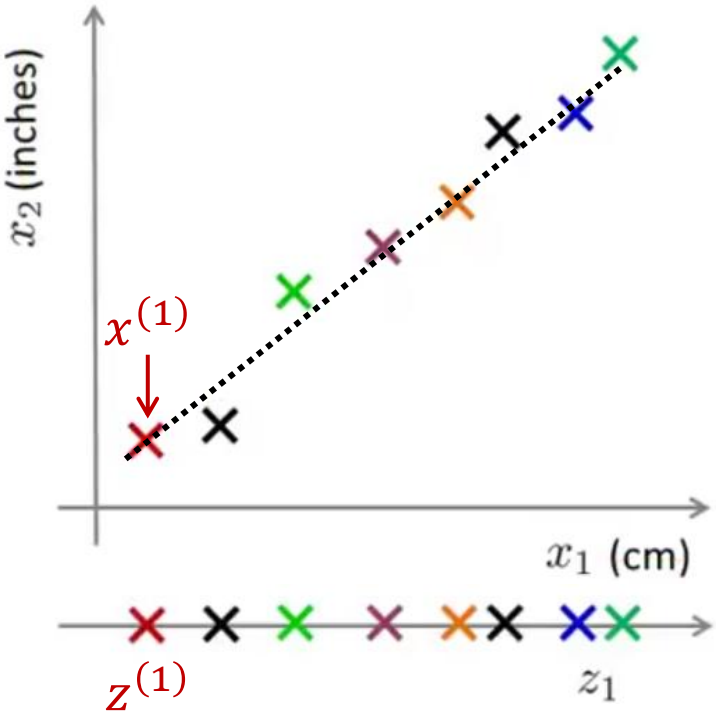
2. Uklanjanje šuma
   - Previše (irelevantnih) obeležja može da degradira performanse

3. Vizuelizacija
   - Bolje razumevanje podataka što može da omogući izgradnju boljih modela

Kompresija

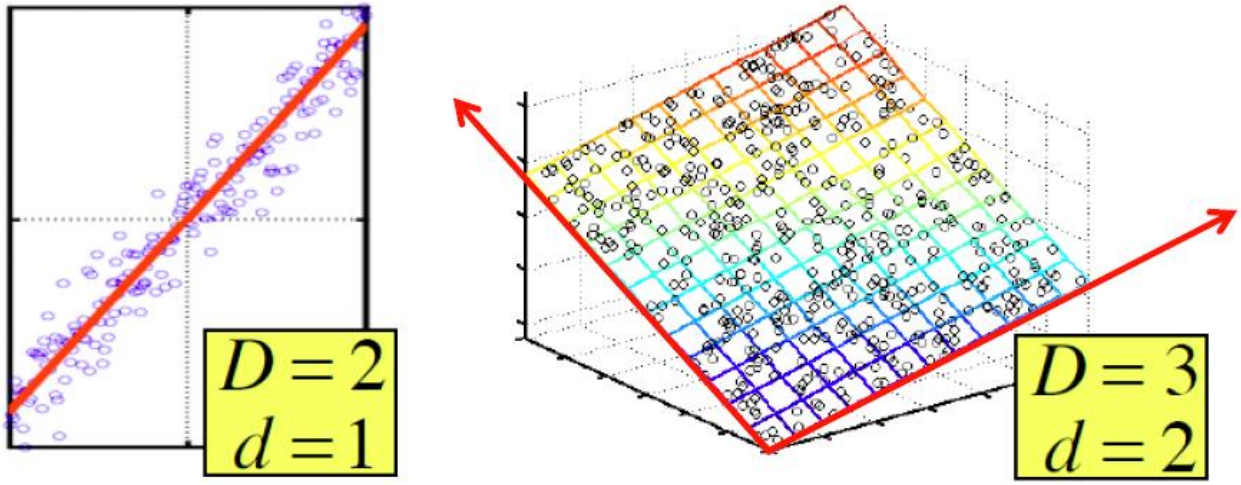
- Recimo da smo sakupili skup podataka sa veoma mnogo obeležja
- Ovde su grafički predstavljena samo dva obeležja:

  x_1 – dužina u cm, x_2 – ista dužinu u inčima



- Umesto da imamo dva odvojena (redudantna) obeležja, bolje bi bilo da redukujemo informaciju u jedno obeležje (jednu dimenziju)
  - $x^{(i)} \in \mathbb{R}^2 \rightarrow z^{(i)} \in \mathbb{R}$
  - Izvršili smo određenu aproksimaciju skupa podataka, ali smo prepolovili broj obeležja

Kompresija



- Pretpostavka: podaci leže tačno na ili blizu *d*-dimenzionog potprostora
- Ose ovog potprostora predstavljaju efektivnu reprezentaciju podataka
- U tipičnom zadatku redukcije dimenzionalnosti možemo imati više hiljada obeležja koja želimo da projektujemo u 100-dimenzioni prostor

Uklanjanje šuma

- Još jedan primer bi bilo automatsko prepoznavanje osobe koja se nalazi na slici
  - Interesuju nas sistematične varijacije koje zaista reprezentuju kako osoba izgleda
  - Ali na slikama možemo imati „šum“ poput promena u osvetljenju i drugih uslova pod kojim je snimak napravljen

- Prilikom automatskog klasifikovanja rukom pisanih cifara:
  - Pretvaranje slike u binarne
  - Skaliranje na istu dimenziju, npr. 16×16
  - Umesto 256 parametara možda možemo koristiti svega dva relevantna obeležja – prosečan intenzitet i simetrija

  …uklanjamo fluktuacije koje nisu relevantne za prepoznavanje o kojoj je cifri reč

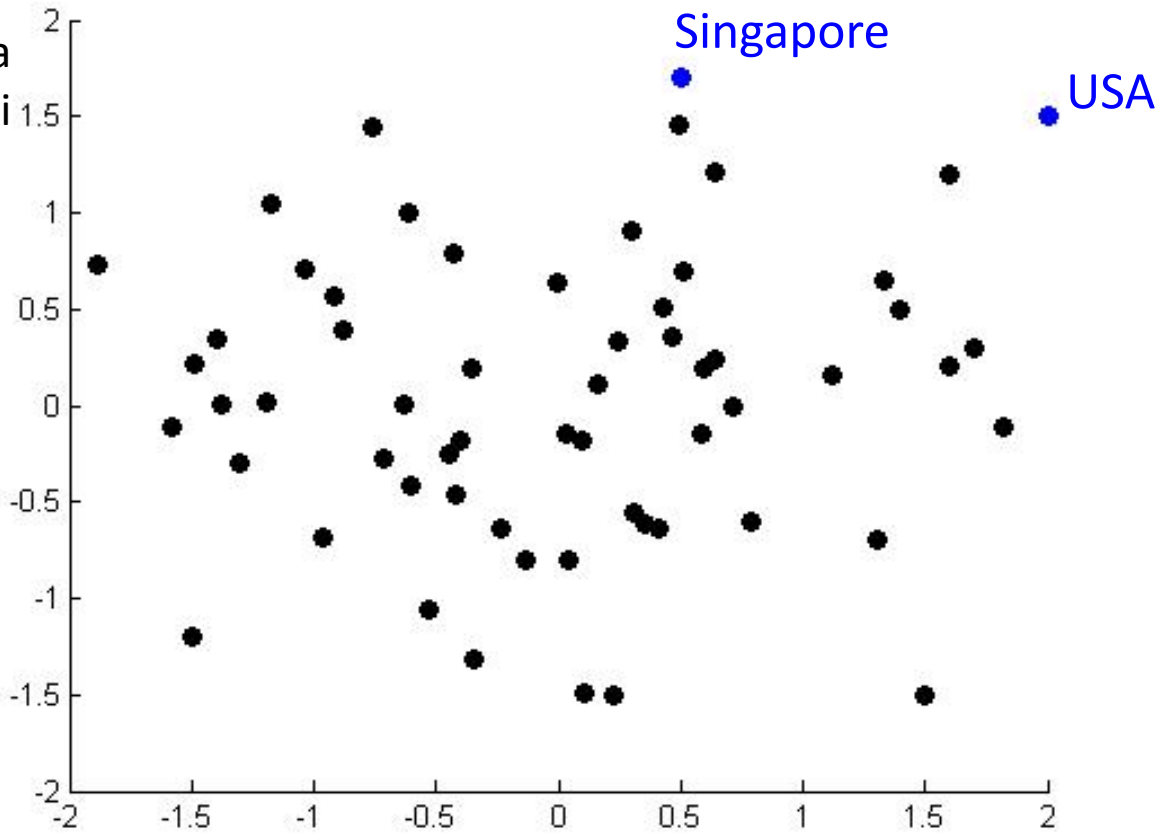
Vizuelizacija

Country	GDP (trillions of US$)	Per capita GDP (thousands of intl. $)	Human Develop-ment Index	Life expectancy	Poverty Index (Gini as percentage)	Mean household income (thousands of US$)	...
Canada	1.577	39.17	0.908	80.7	32.6	67.293	...
China	5.878	7.54	0.687	73	46.9	10.22	...
India	1.632	3.41	0.547	64.7	36.8	0.735	...
Russia	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapore	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...
...	...	...	...	...	...	...	

Vizuelizacija

~ GDP per capita/ekonomska aktivnost po osobi

Country		
Canada	1.6	1.2
China	1.7	0.3
India	1.6	0.2
Russia	1.4	0.5
Singapore	0.5	1.7
USA	2	1.5
…	…	…



~ ekonomska aktivnost zemlje (zavisi od veličine zemlje)

Oprez!

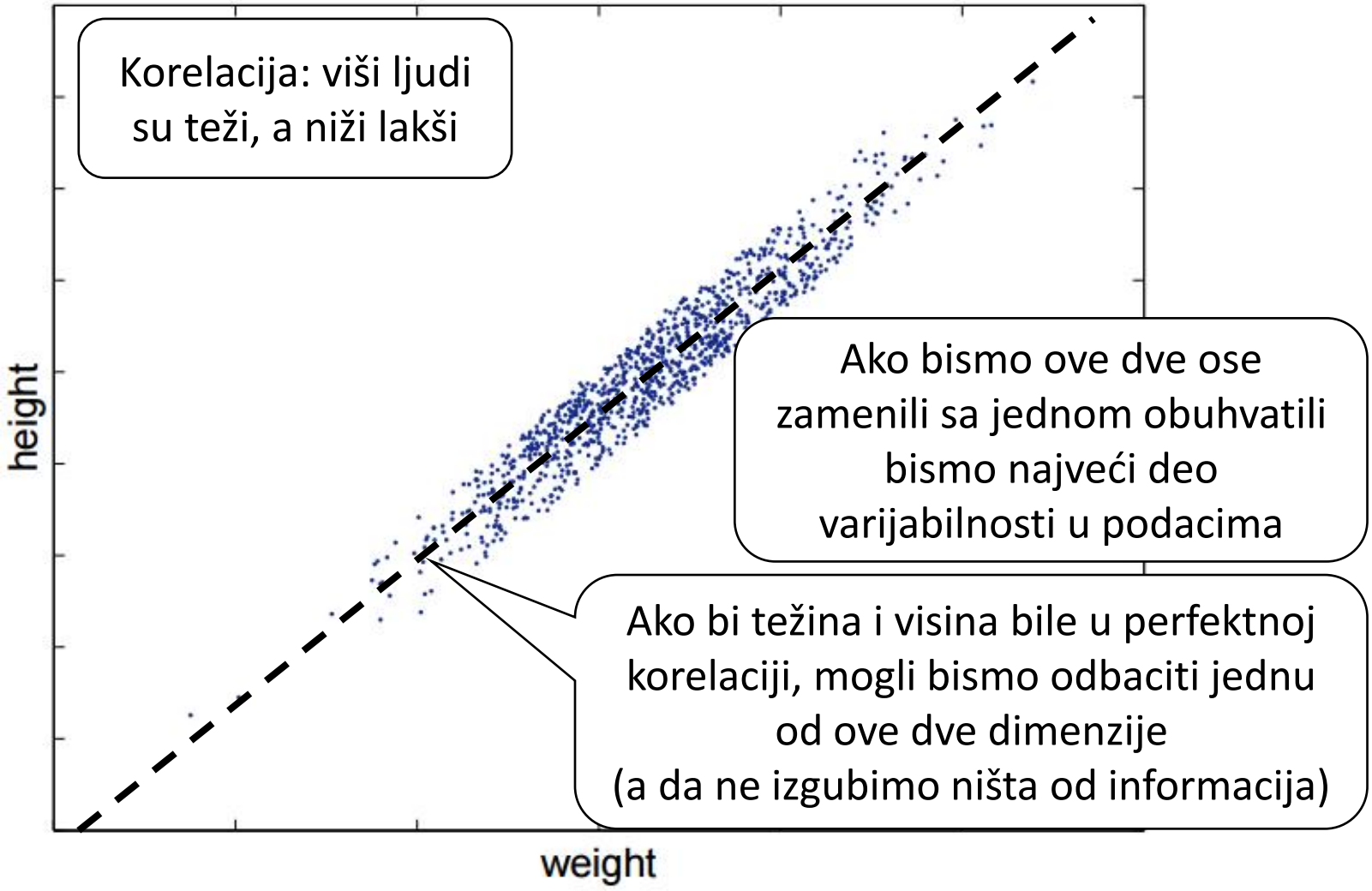
- Važno je redukciju dimenzionalnosti sprovesti na principijelan način
- Odbacujemo informacije – možemo da izgubimo one koje su ključne za obučavanje
- Važno je da algoritam sačuva koristan deo informacija, a odbaci šum

Kako da smanjimo broj dimenzija?

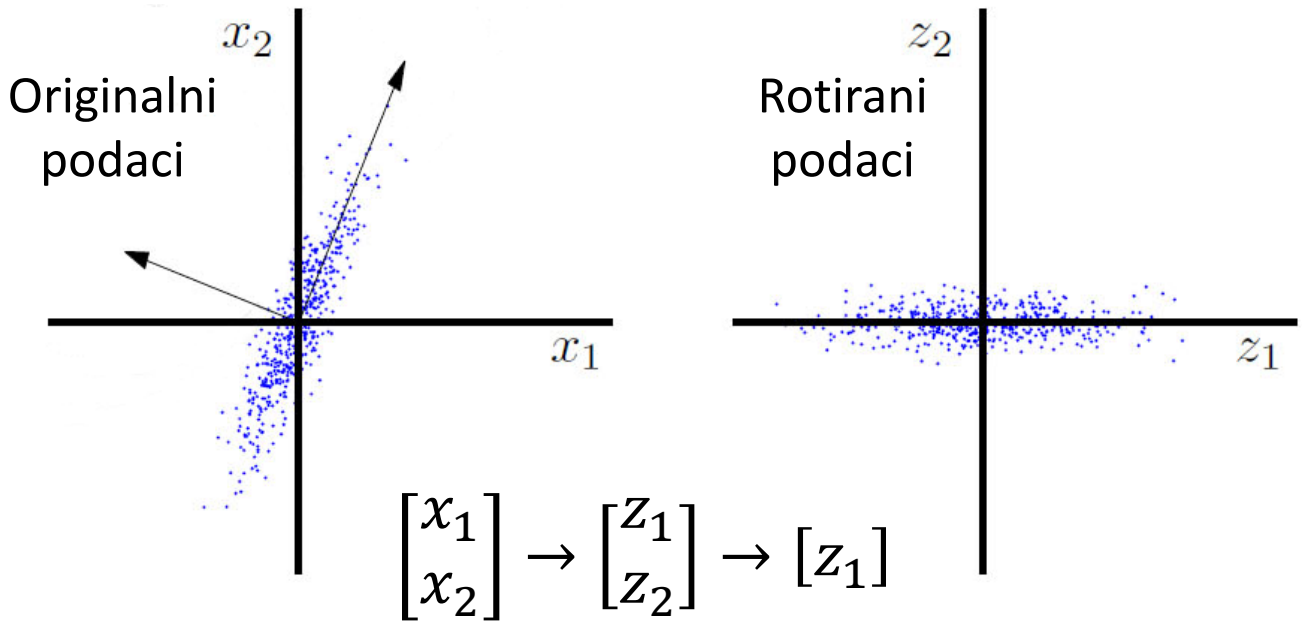
- Selekcija obeležja
  - Pronaći minimalan podskup obeležja koji nam može pomoći da razlikujemo klase
- Redukcija dimenzionalnosti
  - Kreirati nova obeležja koja će predstavljati neku kombinaciju starih obeležja

Principal Component Analysis (PCA)

- Recimo da smo sproveli anketu i zabeležili visinu i težinu grupe ljudi



Principal Component Analysis (PCA)

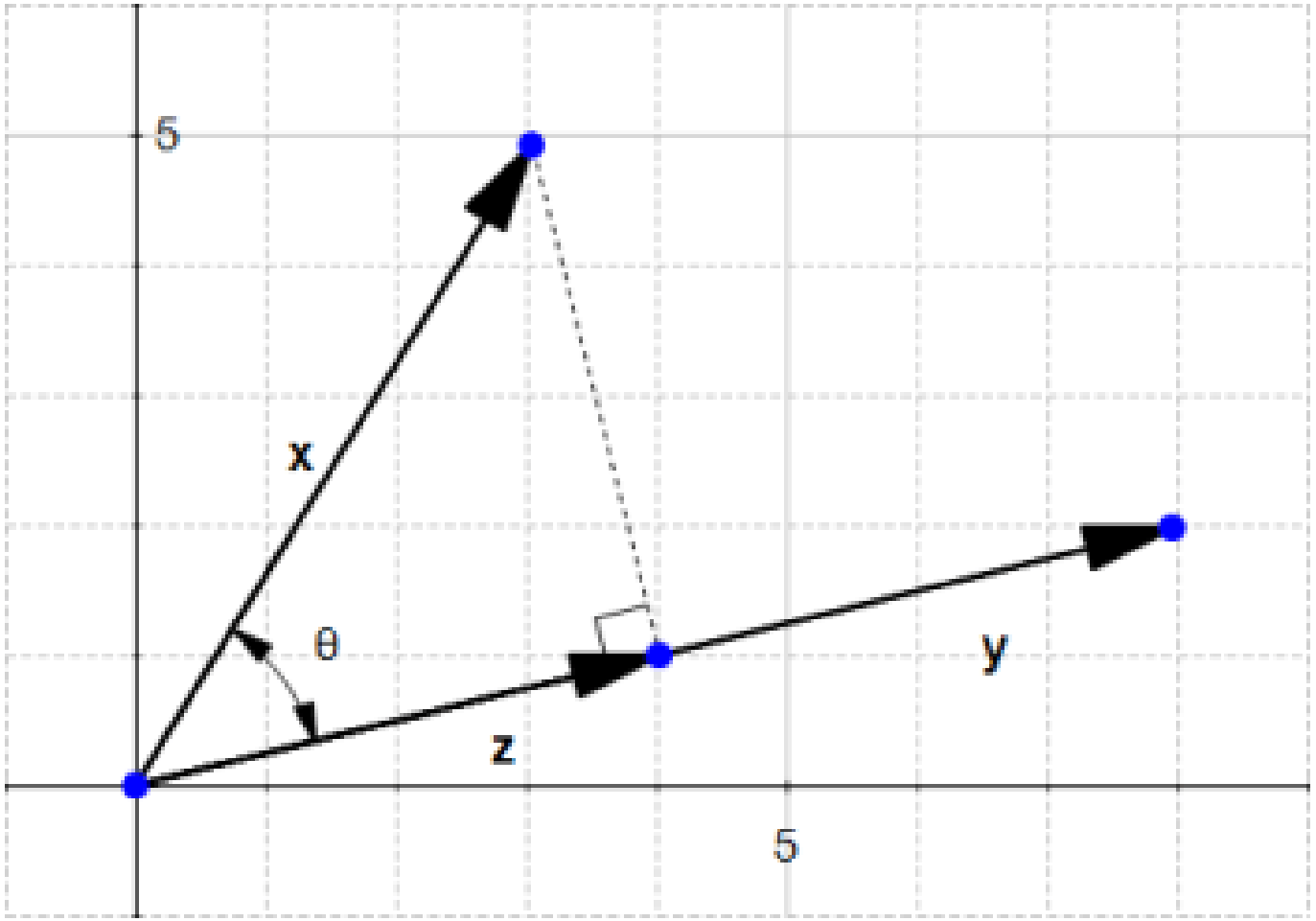


- PCA konstruiše mali broj *linearnih* obeležja koja sumarizuju ulazne podatke
- Ideja je da se rotiraju ose (linearna transformacija koja definiše novi koordinatni sistem), tako da u ovom sistemu
  - Identifikujemo dominantne dimenzije (informacije)
  - Odbacimo manje dimenzije (šum)

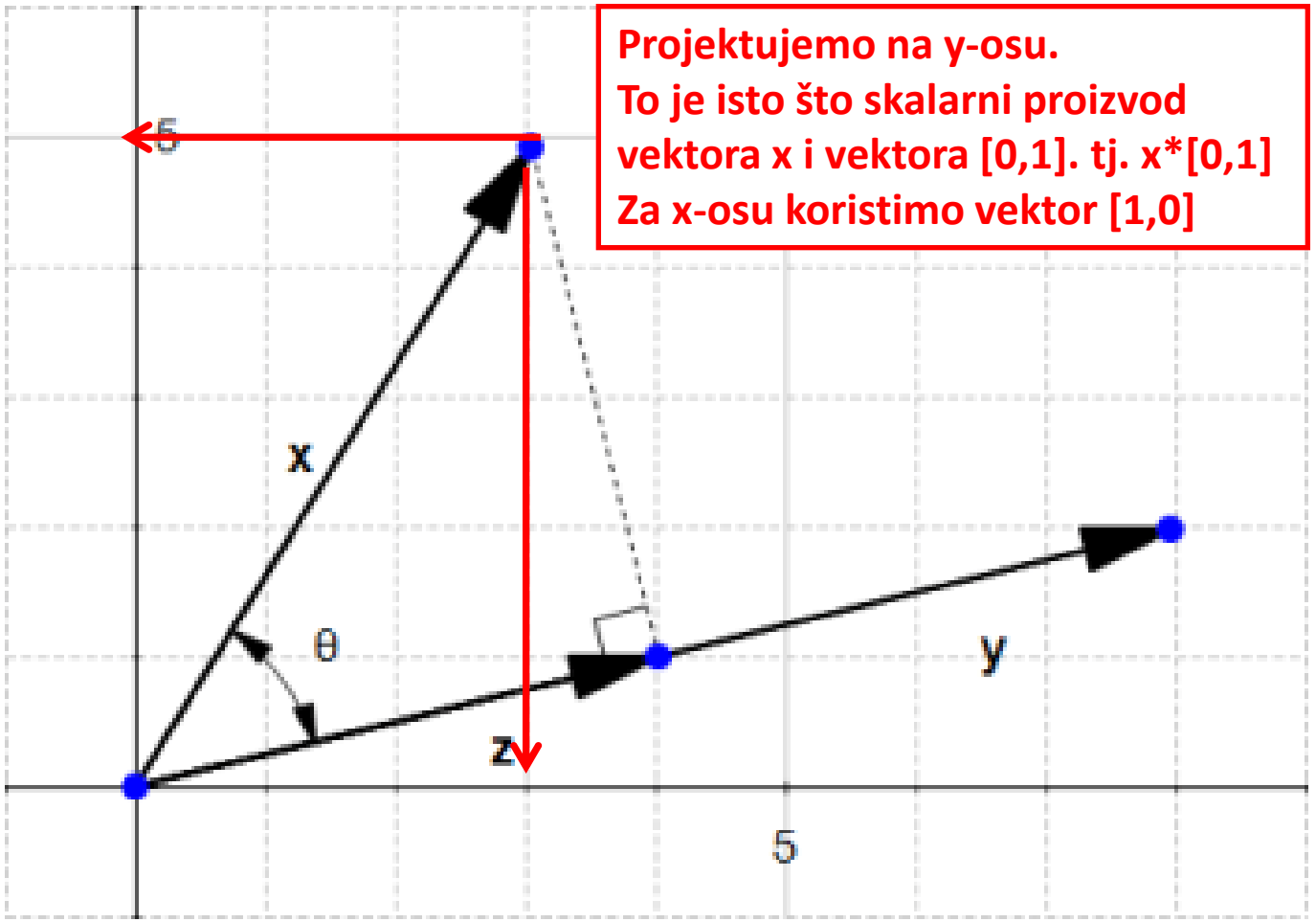
Koordinatni sistem

- Koordinatni sistem je definisan skupom ortonormalnih vektora (međusobno ortogonalni jedinični vektori)
- Dužina projekcije tačke x na *jedinični vektor* v je $x^T v$

Osnovni principi Linearne Projekcije

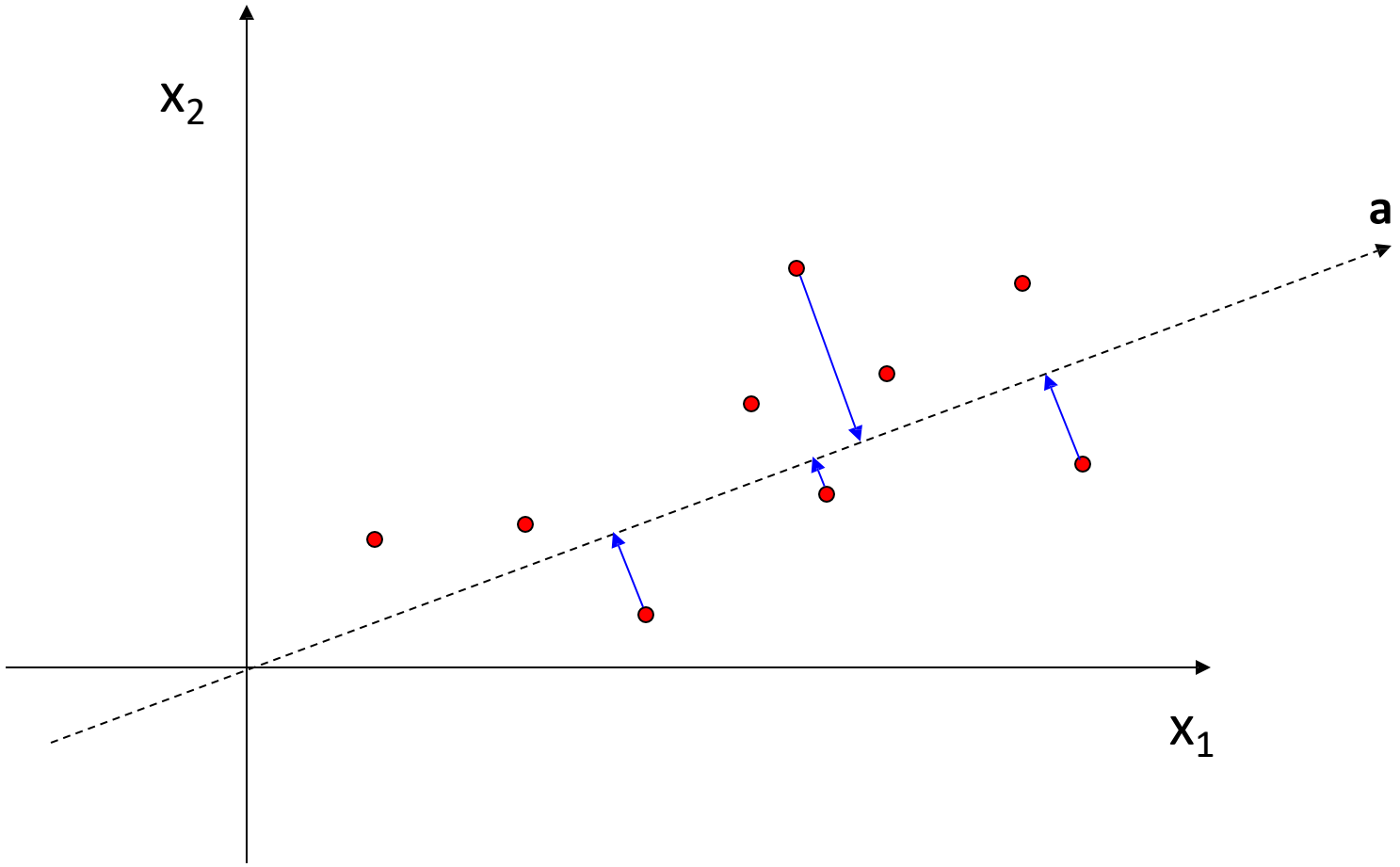


Osnovni principi Linearne Projekcije

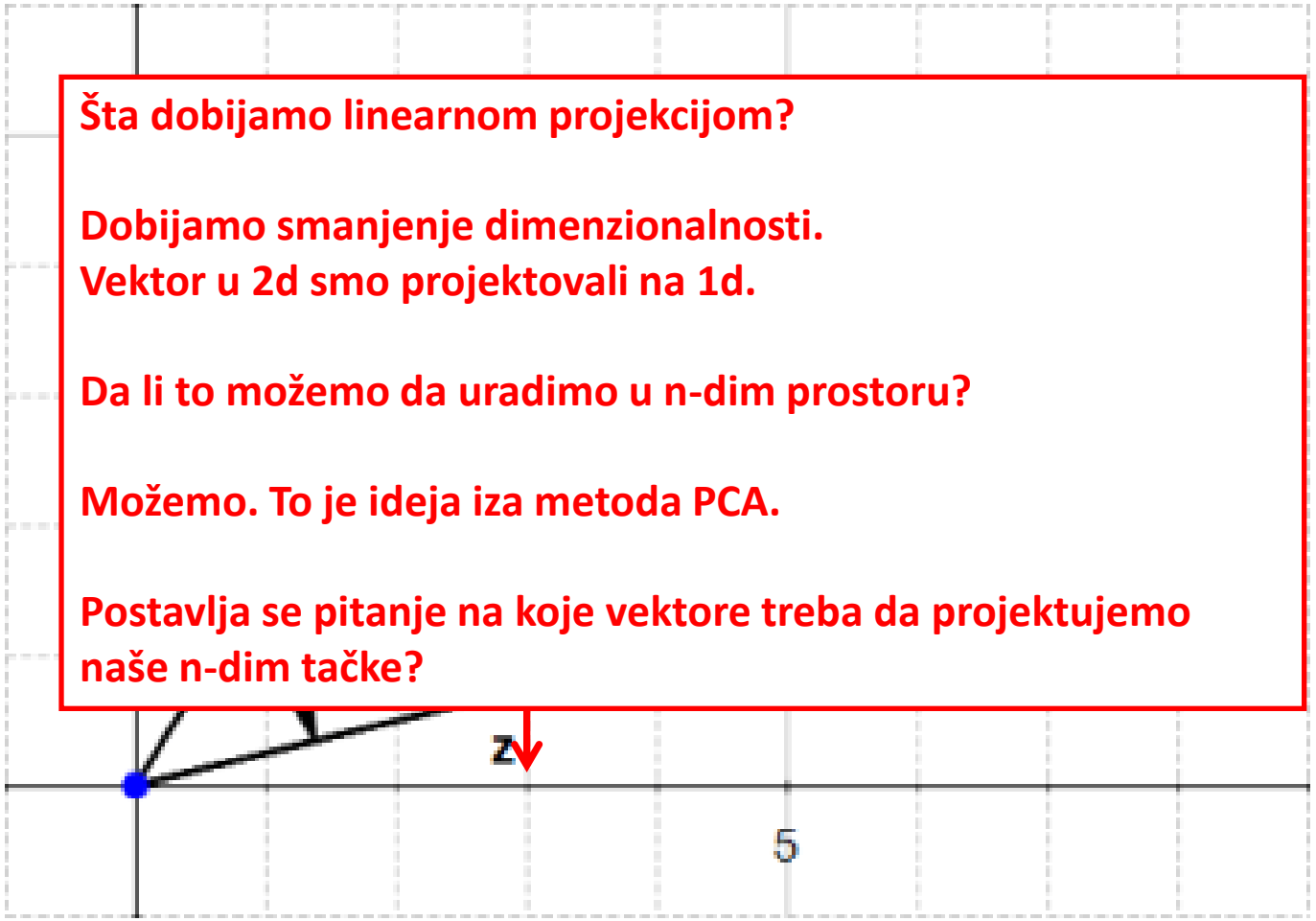


Projektujemo na y-osu.
To je isto što skalarni proizvod vektora x i vektora [0,1]. tj. x*[0,1]
Za x-osu koristimo vektor [1,0]

Primer projekcije iz 2d u 1d



Osnovni principi Linearne Projekcije



Šta dobijamo linearnom projekcijom?

Dobijamo smanjenje dimenzionalnosti.
Vektor u 2d smo projektovali na 1d.

Da li to možemo da uradimo u n-dim prostoru?

Možemo. To je ideja iza metoda PCA.

Postavlja se pitanje na koje vektore treba da projektujemo naše n-dim tačke?

Primer: Euklidski koordinatni sistem

- Definisan je vektorima $v_1, \ldots, v_D$, gde vektor v_i ima *i*-tu koordinatu 1, a sve ostale koordinate 0

- Ulazni vektor x ima komponente $x_i = x^T v_i$ i možemo pisati

$$x = \sum_{i=1}^{D} x_i v_i = \sum_{i=1}^{D} (x^T v_i) v_i$$

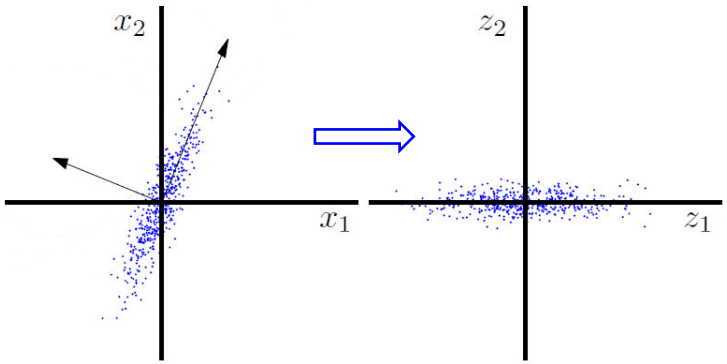
Koordinatni sistem

- Isto se može uraditi sa bilo kojom ortonormalnom bazom $v_1, \dots, v_D$:

$$x = \sum_{i=1}^{D} z_i v_i = \sum_{i=1}^{D} (x^T v_i) v_i$$

gde su *koordinate* u bazi $v_1, \dots, v_D$ date sa $z_i = (x^T v_i)$

- Cilj PCA je da konstruiše intuitivniju bazu gde je većina koordinati mala



- Male koordinate tretiramo kao slučajne fluktuacije i postavljamo ih na 0
- Nadamo se da smo ovim smanjili dimenzionalnost, a sačuvali većinu važnih informacija

Zadatak

- Projektovati *D*-dimenzioni prostor u *K*-dimenzioni prostor $x^{(i)} \in \mathbb{R}^D \rightarrow z^{(i)} \in \mathbb{R}^K \ (K \leq D)$

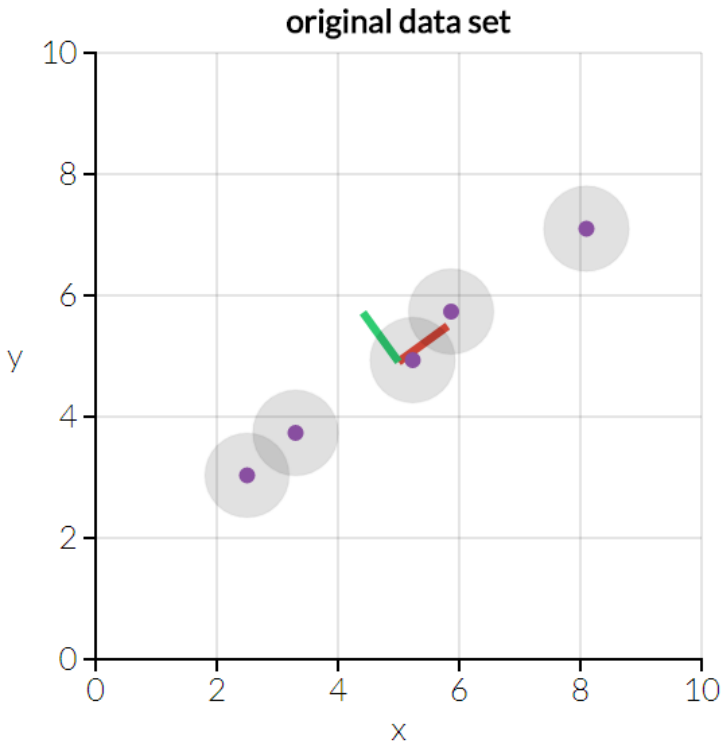
1. Pronaći ose novog koordinatnog sistema: $v_1, v_2, \ldots, v_D$
2. Transformisati x u novi prostor (koordinate transformisanog vektora su $z_1, z_2, \ldots, z_D$)
3. Recimo da su prvih $K \leq D$ koordinati informativne. Odbacićemo preostale koordinate da bismo dobili ulazni vektor redukovane dimenzionalnosti:

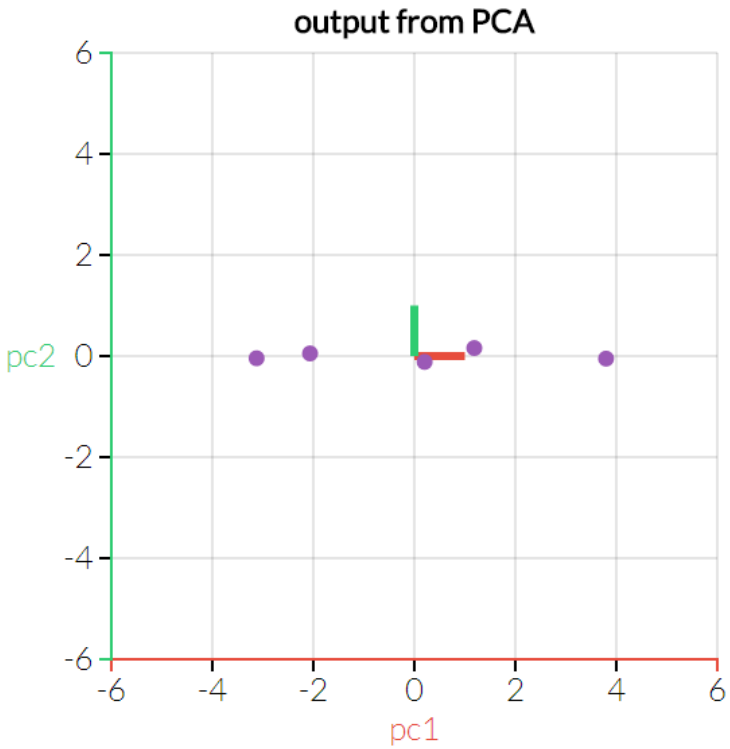
$$z = \begin{bmatrix} z_1 \\ z_2 \\ \ldots \\ z_K \end{bmatrix} = \begin{bmatrix} x^T v_1 \\ x^T v_2 \\ \ldots \\ x^T v_K \end{bmatrix} = \Phi(x)$$

PCA – linearne projekcije

Postavlja se pitanje na koje vektore treba da projektujemo naše n-dim tačke?

PCA bira vektore koji imaju pravac koji sadrži najviše varijabilnosti naših podataka

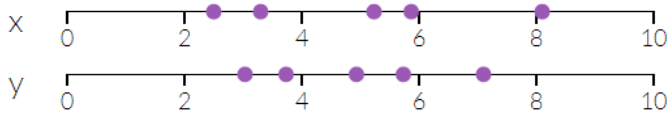




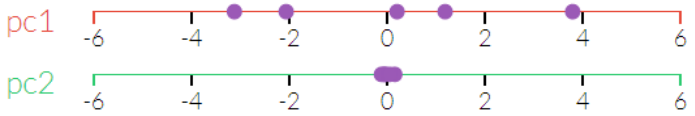
PCA – linearne projekcije

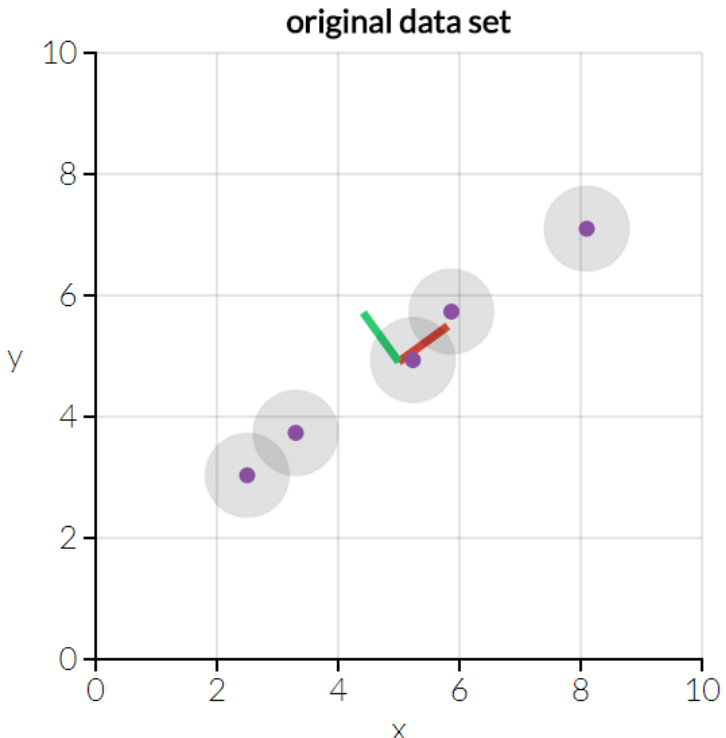
PCA bira vektore koji imaju pravac koji sadrži najviše varijabilnosti naših podataka

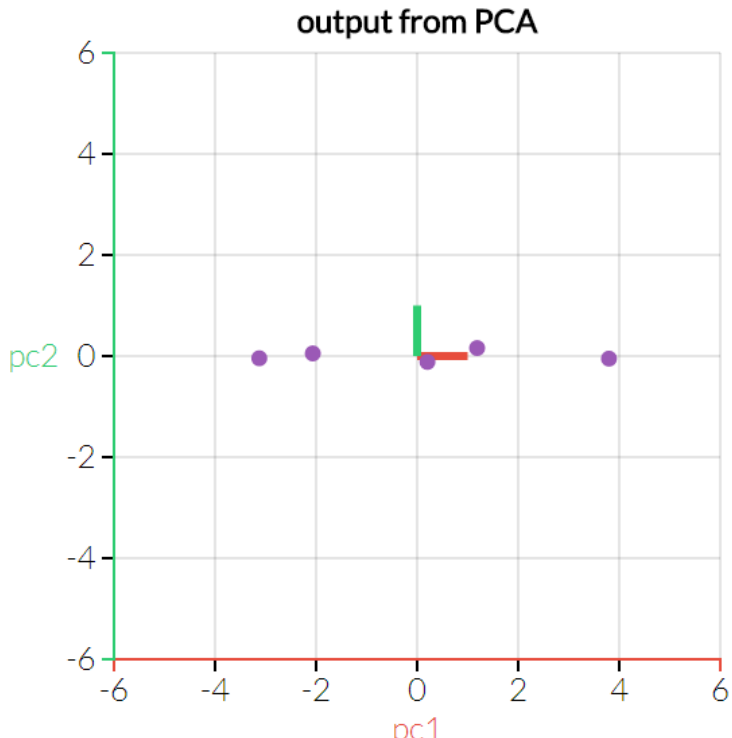
Varijabilost po x, y



Varijabilost po PCA osama



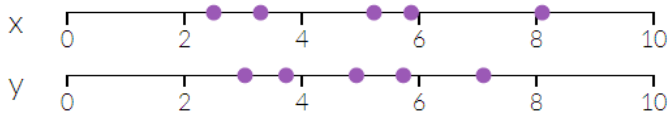




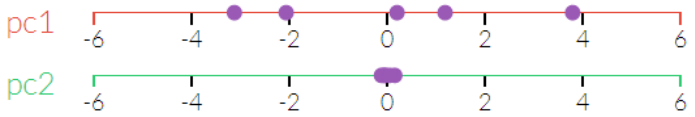
PCA – linearne projekcije

PCA bira vektore koji imaju pravac koji sadrži najviše varijabilnosti naših podataka

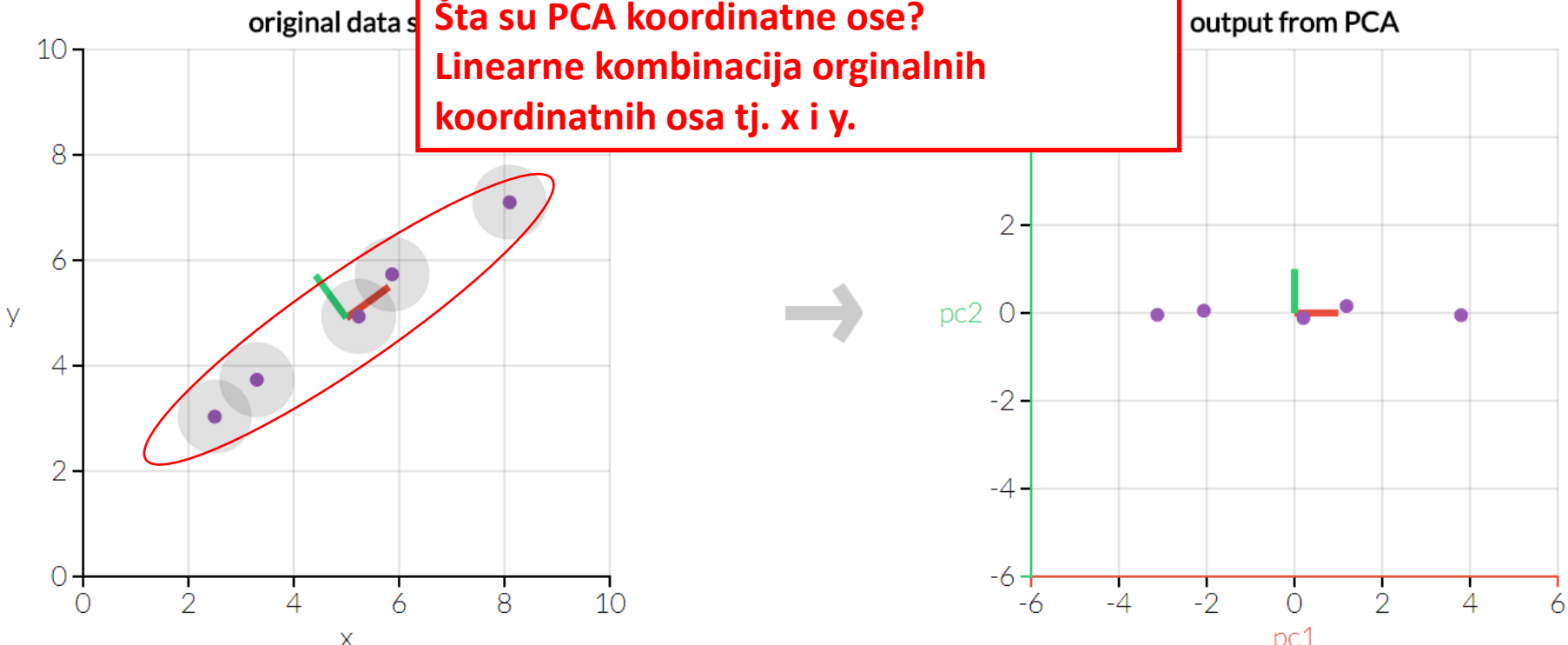
Varijabilost po x, y



Varijabilost po PCA osama



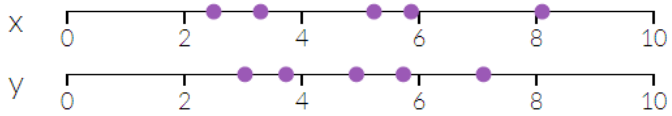
Šta su PCA koordinatne ose?
Linearne kombinacija orginalnih koordinatnih osa tj. x i y.



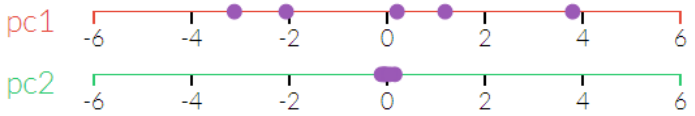
PCA – linearne projekcije

PCA bira vektore koji imaju pravac koji sadrži najviše varijabilnosti naših podataka

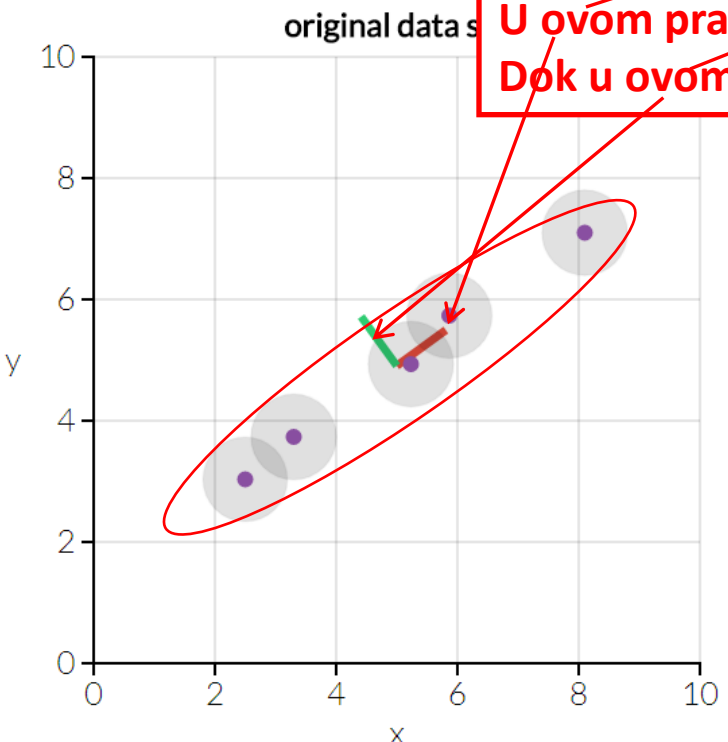
Varijabilost po x, y

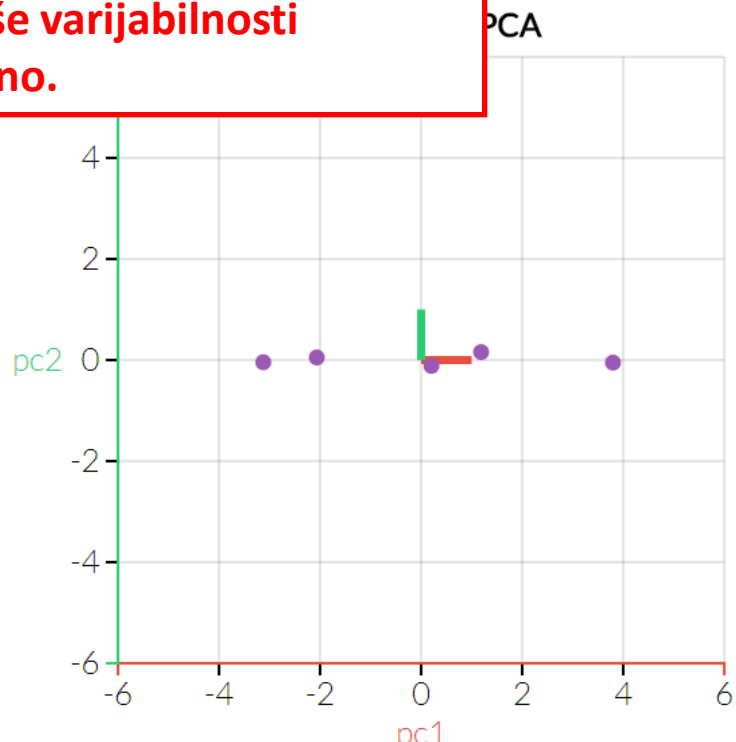


Varijabilost po PCA osama



U ovom pravcu imamo najviše varijabilnosti
Dok u ovom imamo minimalno.

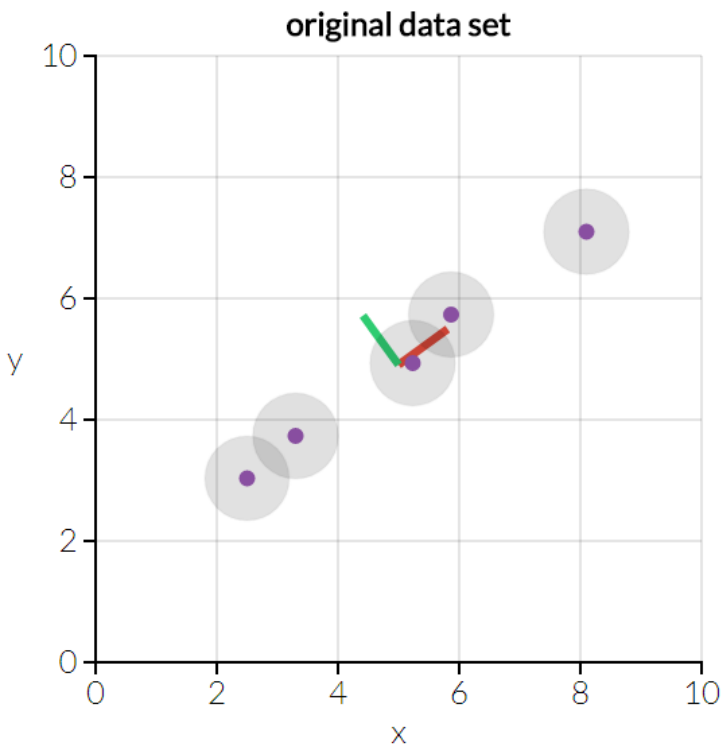


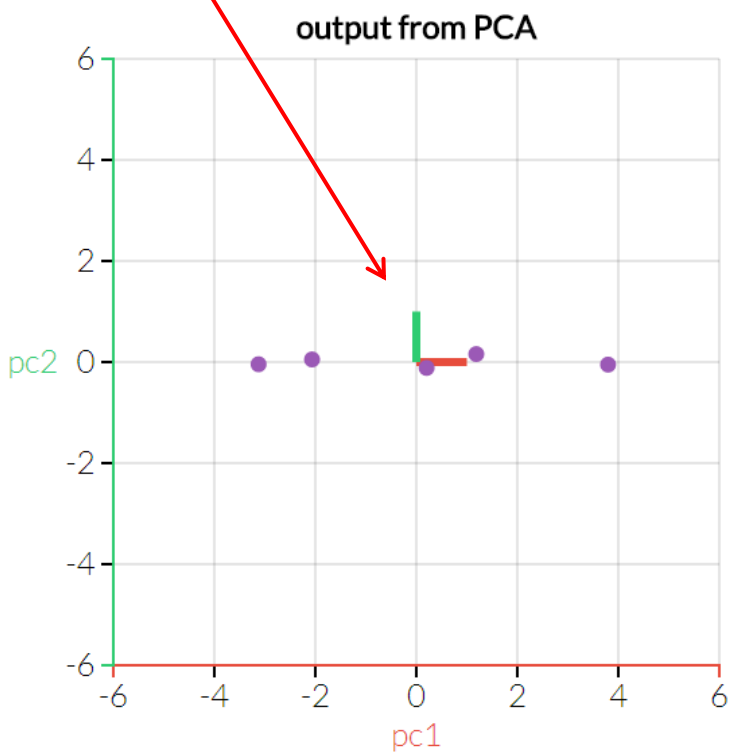


PCA – linearne projekcije

PCA bira vektore koji imaju pravac koji sadrži najviše varijabilnosti naših podataka.

Dobijeni vektori čine novi koordinatni sistem.





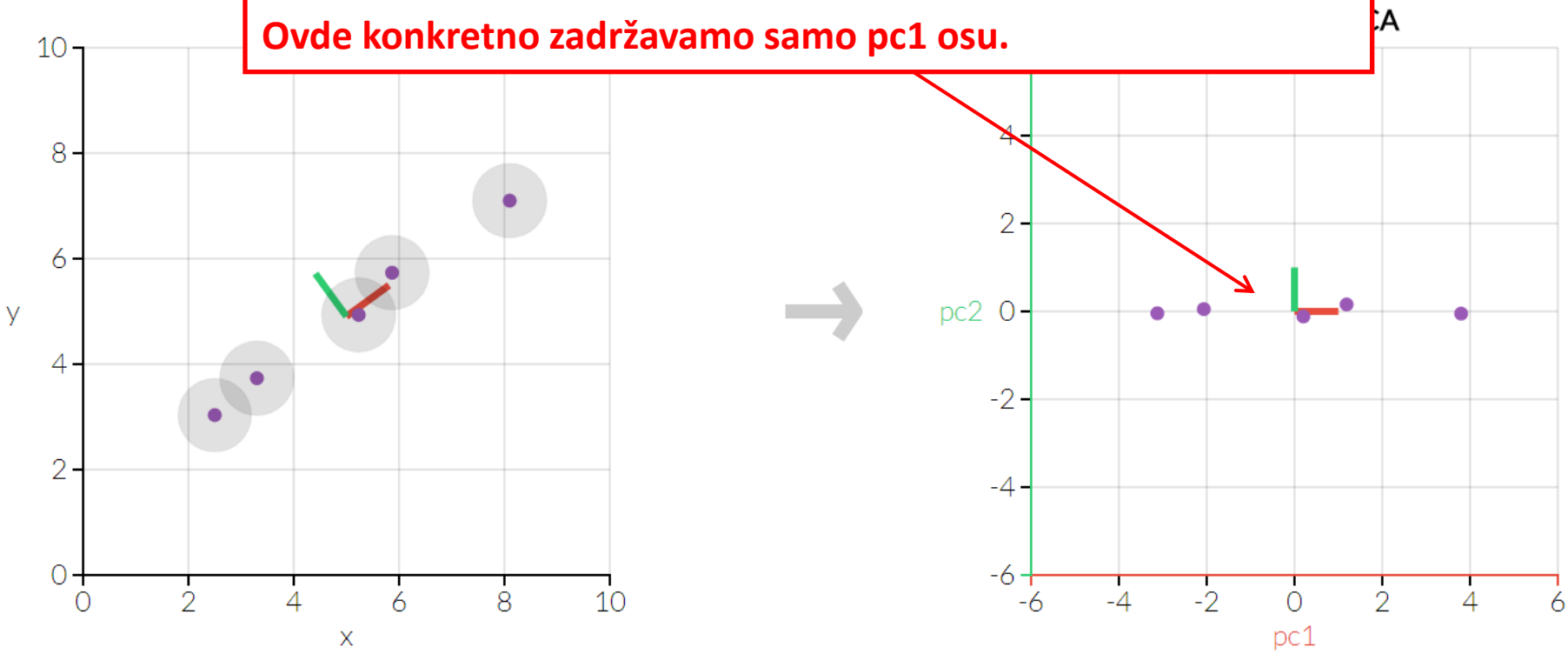
PCA – linearne projekcije

Dobijeni vektori čine novi koordinatni sistem.

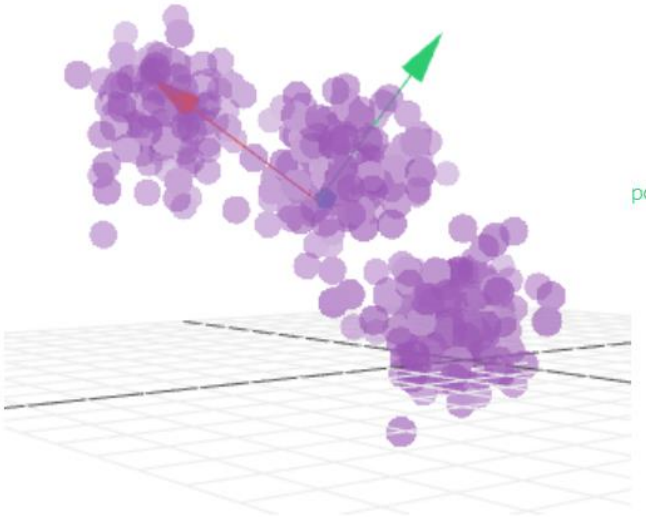
Kako radimo redukciju dimenzionalnosti?

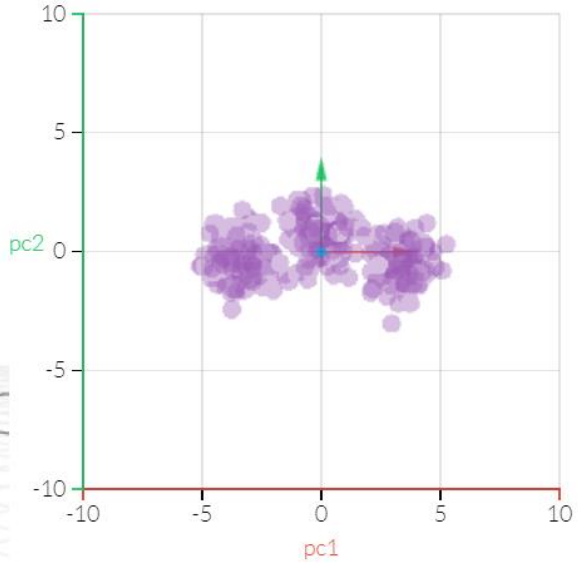
Odbacimo pravce koji imaju malu varijabilnost.

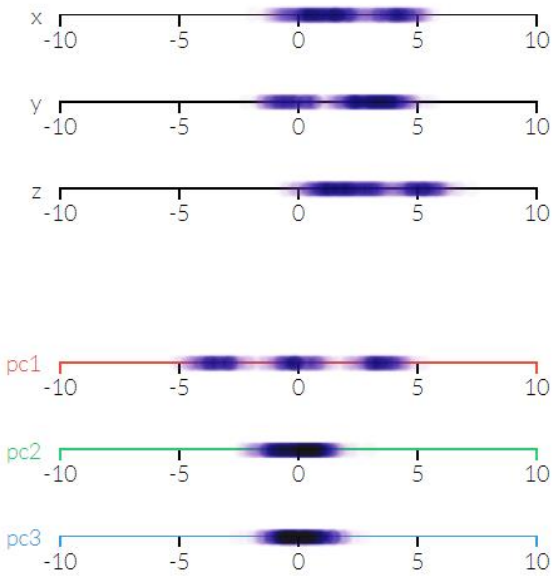
Ovde konkretno zadržavamo samo pc1 osu.



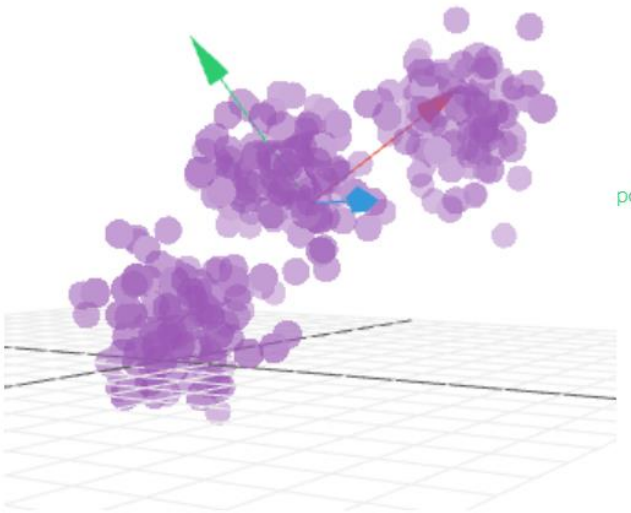
PCA – 3d primer

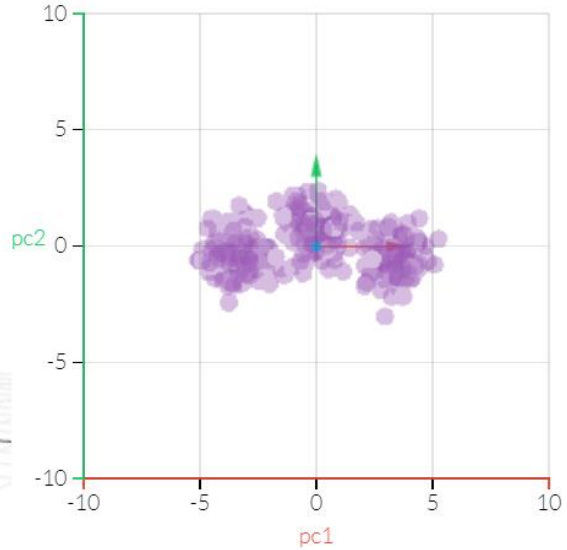


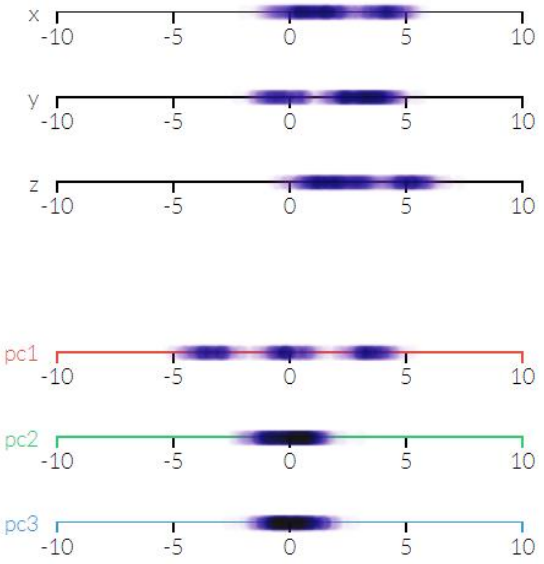




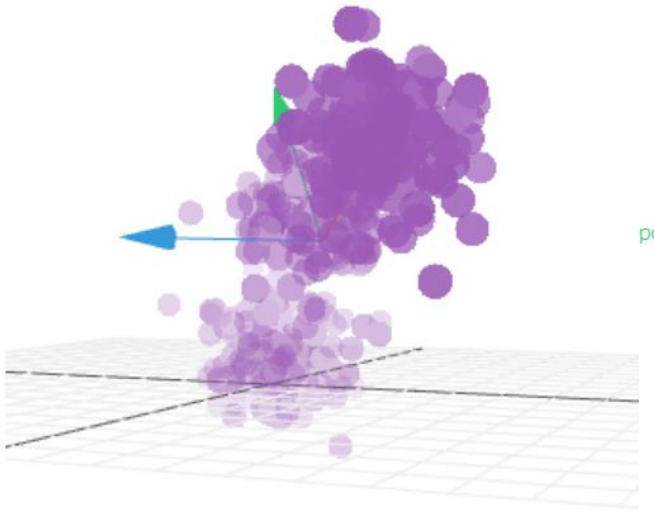
PCA – 3d primer – drugi ugao

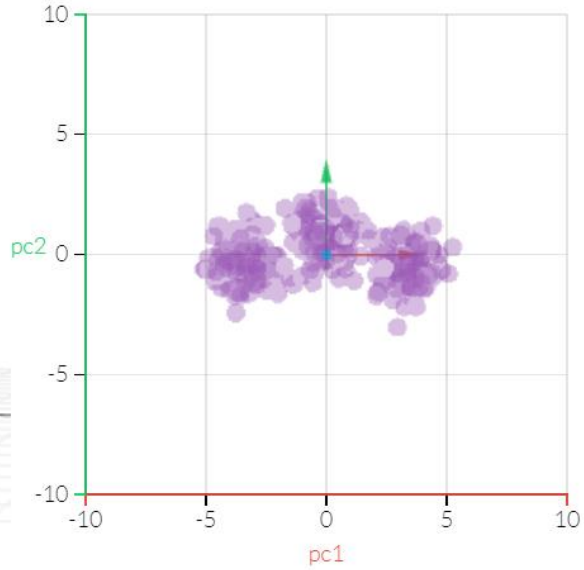


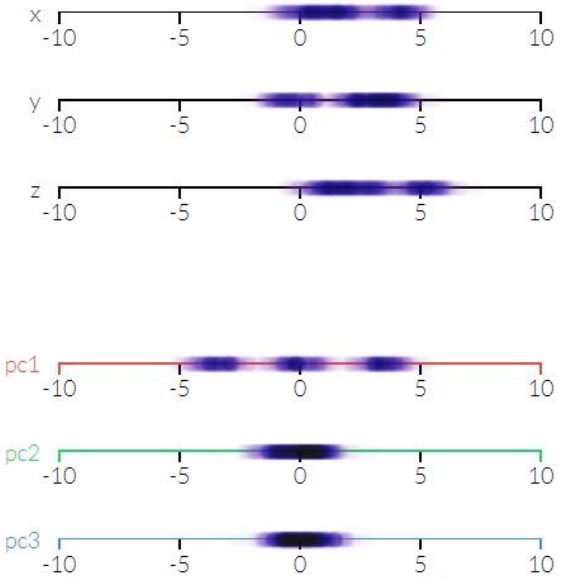




PCA – 3d primer – treći ugao

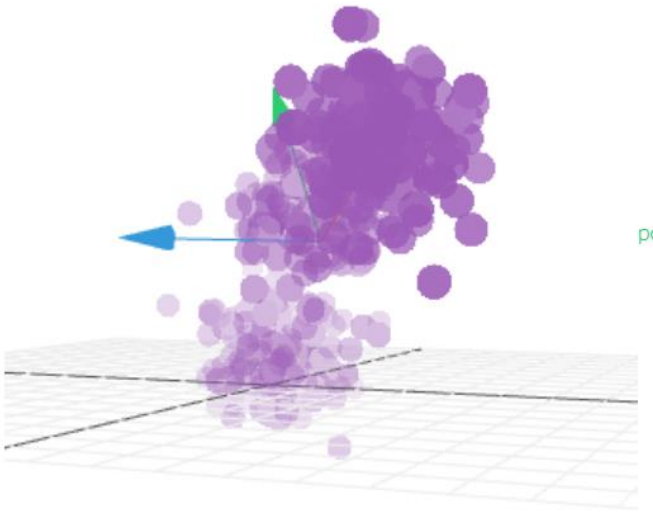


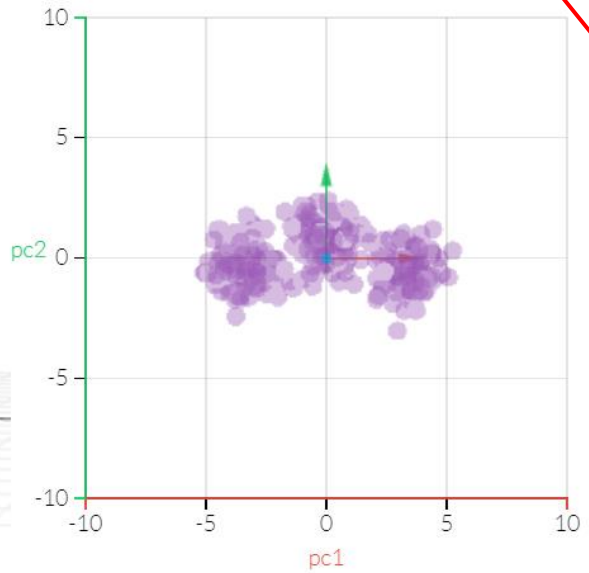


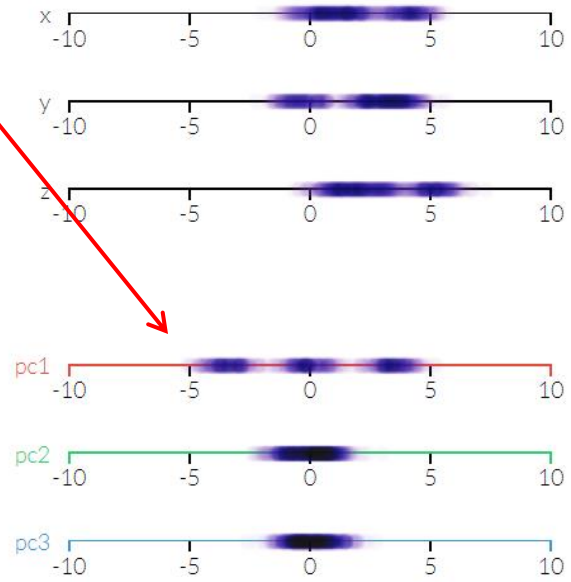


PCA – 3d primer – treći ugao

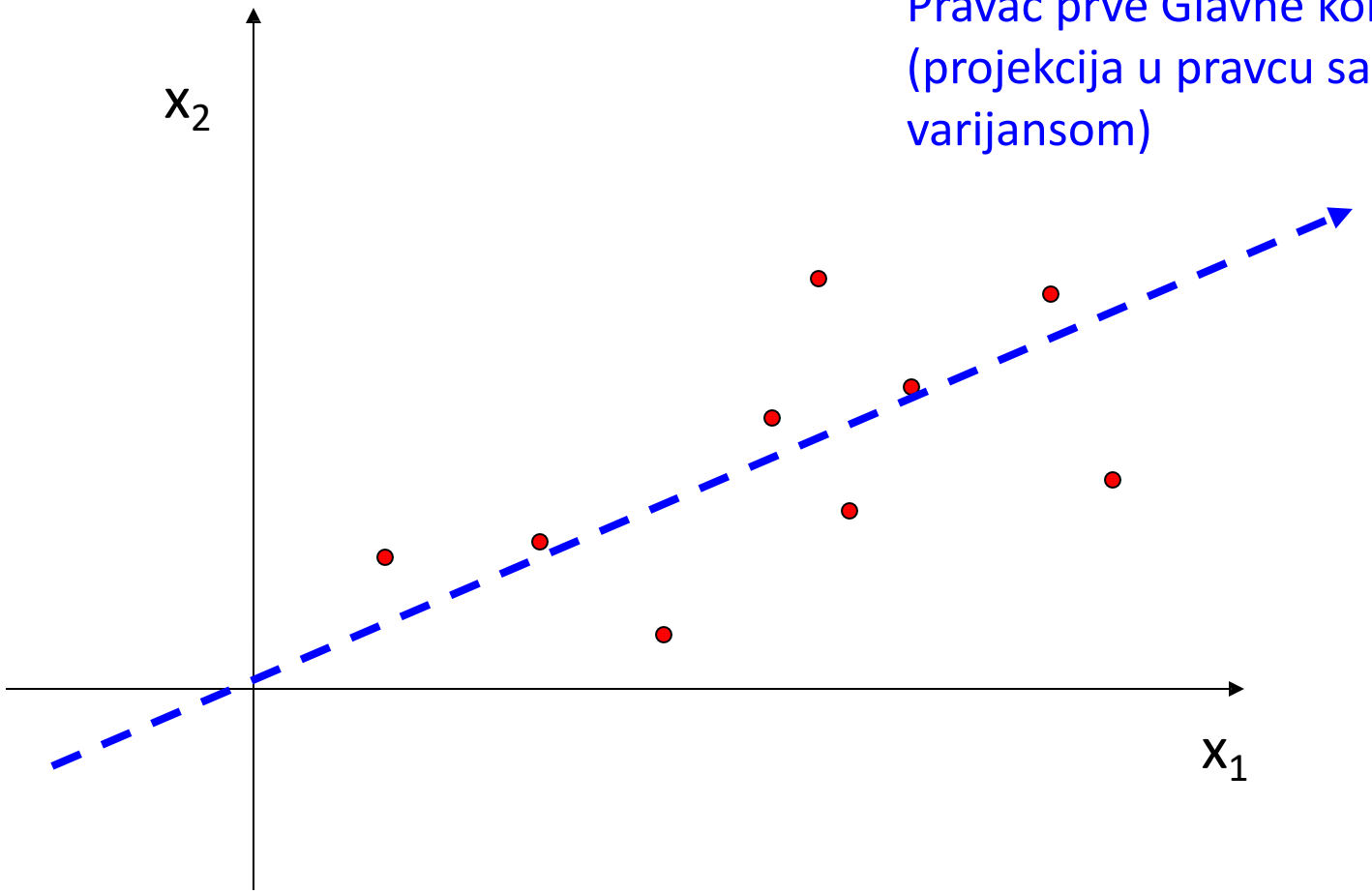
3d prostor možemo svesti na jednu dimenziju



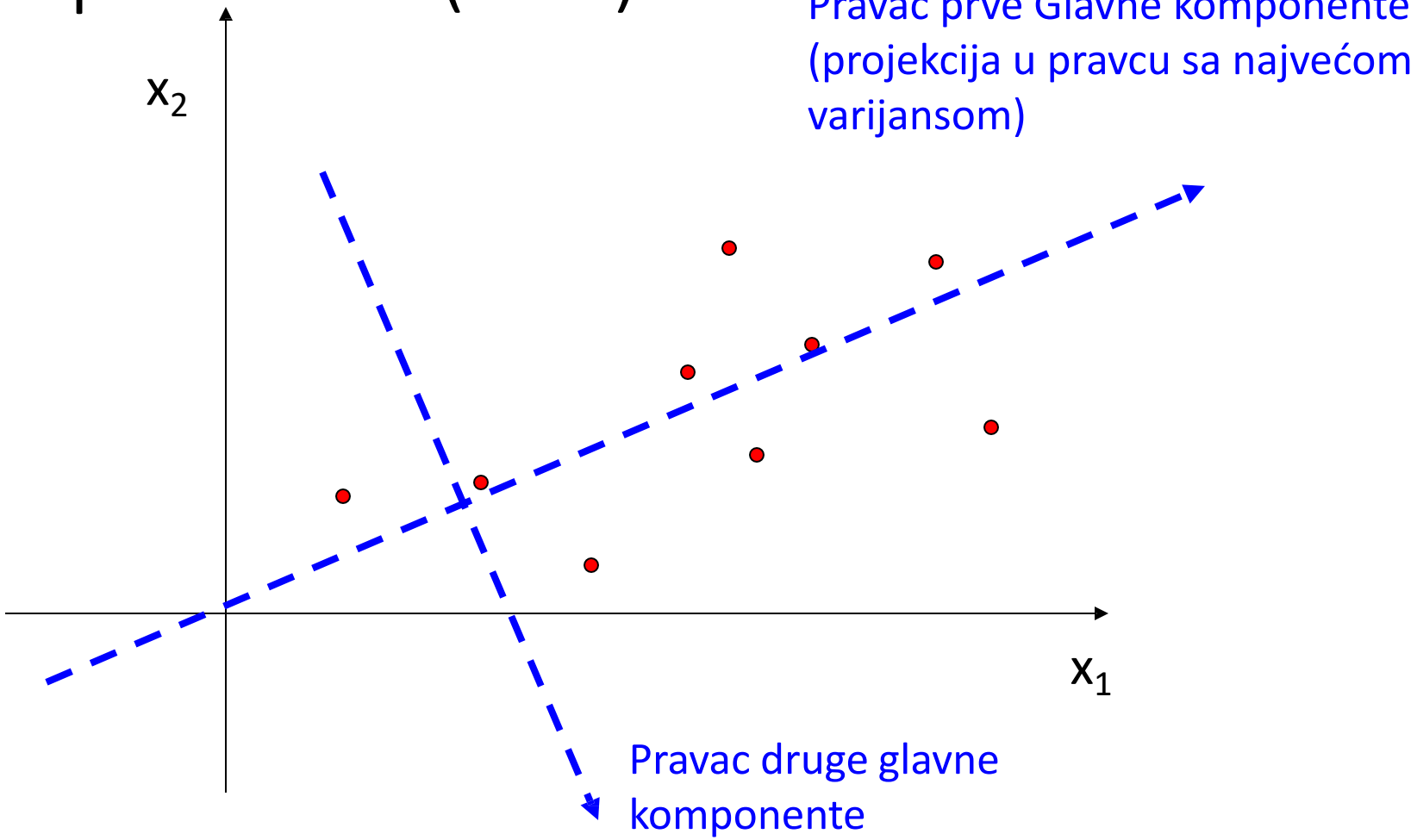




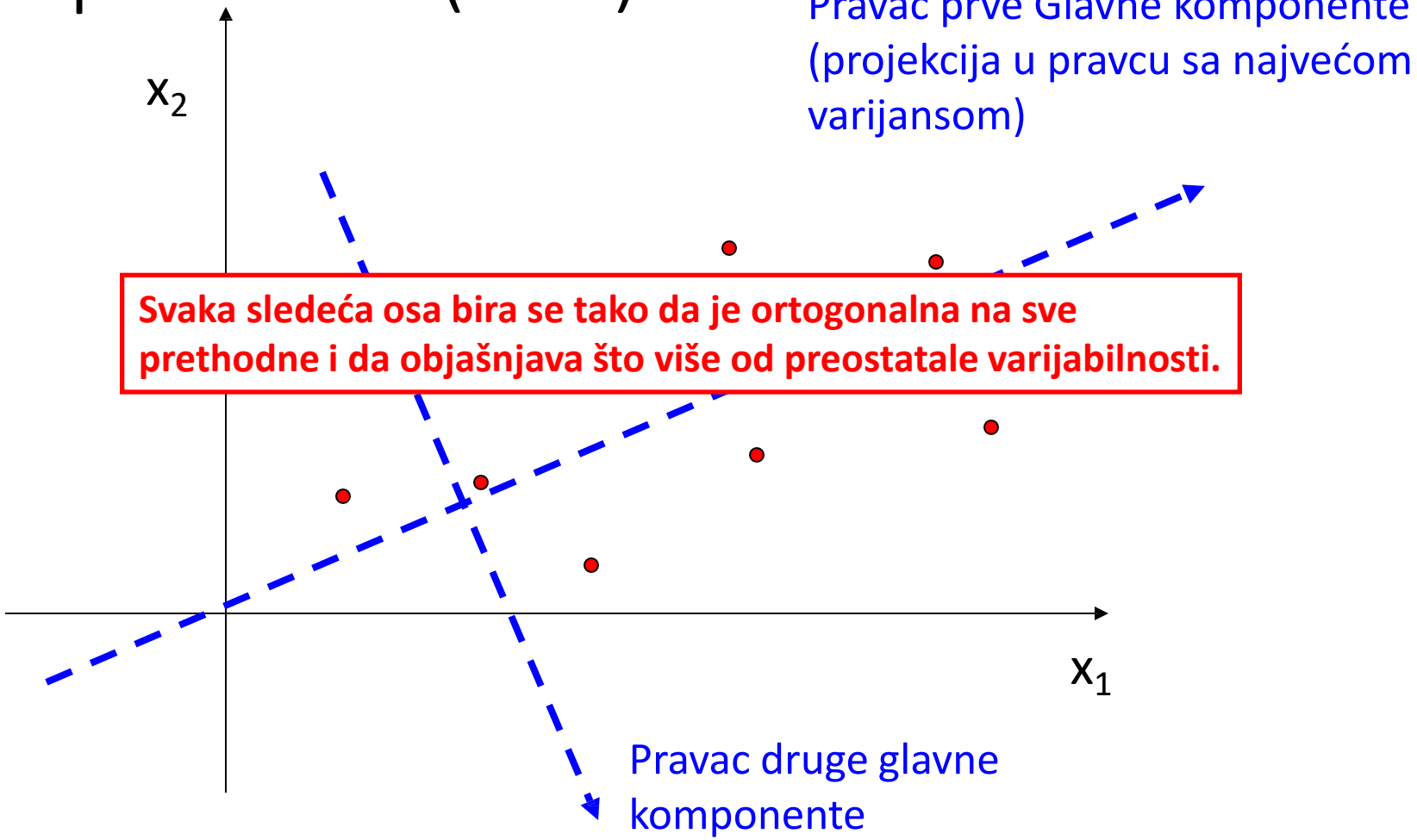
PCA Kako odabrati sledeću komponentu (osu)?



PCA Kako odabrati sledeću komponentu (osu)?



PCA Kako odabrati sledeću komponentu (osu)?



PCA – matrica podataka

X = d x N matrica podataka: kolone = d-dimenzioni vektori podataka

X=

	Ex 1	Ex 2	Ex 3
Refund	Yes	No	No
Mar. status	Single	Married	Single
Tax. income	125k	100k	70k
Evade	No	No	No

N x d

Kada primenjujemo PCA na skup podataka radimo sa matricom podataka.

Atributi su koordinatne ose u orginalnom prostoru.

PCA komponente su linearne kombinacije atributa.

Na primer pc1=3*Refund-0.8*Tax.income+2.1*Evade-0.4*Mar.status

Naravno, nominalne atribute moramo da konvertujemo u numeričke.

Kako izračunavamo glavne komponente?

Postoji mnogo metoda da se dobiju glavne komponente.

Jedan od najčešćih načina je određivanje sopstvenih vektora matrice kovarijansi.

$$BB^T = \begin{vmatrix} (a_1-\mu_1)^2+(b_1-\mu_1)^2+(c_1-\mu_1)^2 & s_{12} & s_{13} \\ s_{21} & (a_2-\mu_2)^2+(b_2-\mu_2)^2+(c_2-\mu_2)^2 & s_{23} \\ s_{31} & s_{32} & (a_3-\mu_3)^2+(b_3-\mu_3)^2+(c_3-\mu_3)^2 \end{vmatrix}$$

Zbir dijagonalnih elemenata je ukupna varijansa skupa podataka

Varijansa atributa 1:

$$S_{11} = \frac{1}{3-1}\left((a_1-\mu_1)^2+(b_1-\mu_1)^2+(c_1-\mu_1)^2\right)$$

Ko-varijansa atributa 1 i 2:

$$S_{21} = \frac{1}{3-1}\left((a_1-\mu_1)(a_2-\mu_2)+(b_1-\mu_1)(b_2-\mu_2)+(c_1-\mu_1)(c_2-\mu_2)\right)$$

Kako izračunavamo glavne komponente?

Hajde da analiziramo matricu kovarijansi.
Prvo na glavnoj dijagonali je varijansa svakog atributa redom.
Kad ih saberemo dobijamo ukupnu varijansu celog skupa podataka.

$$S_{11} = \frac{1}{3-1}\left((a_1-\mu_1)^2 + (b_1-\mu_1)^2 + (c_1-\mu_1)^2\right)$$

$$BB^T = \begin{vmatrix} (a_1-\mu_1)^2 + (b_1-\mu_1)^2 + (c_1-\mu_1)^2 & s_{12} & s_{13} \\ s_{21} & (a_2-\mu_2)^2 + (b_2-\mu_2)^2 + (c_2-\mu_2)^2 & s_{23} \\ s_{31} & s_{32} & (a_3-\mu_3)^2 + (b_3-\mu_3)^2 + (c_3-\mu_3)^2 \end{vmatrix}$$

Zbir dijagonalnih elemenata je ukupna varijansa skupa podataka

Kako izračunavamo glavne komponente?

Šta su elementi koji nisu na glavnoj dijagonali?
To su kovarijanse – govore nam o linearnom odnosu atributa po parovima.

Ko-varijansa atributa 1 i 2:

$$S_{21} = \frac{1}{3-1}\left((a_1-\mu_1)(a_2-\mu_2)+(b_1-\mu_1)(b_2-\mu_2)+(c_1-\mu_1)(c_2-\mu_2)\right)$$

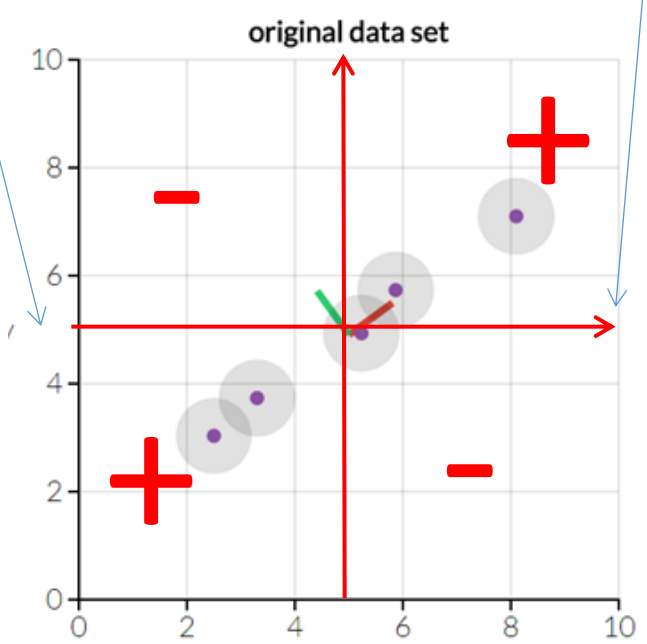
$$BB^T = \begin{vmatrix} (a_1-\mu_1)^2+(b_1-\mu_1)^2+(c_1-\mu_1)^2 & s_{12} & s_{13} \\ s_{21} & (a_2-\mu_2)^2+(b_2-\mu_2)^2+(c_2-\mu_2)^2 & s_{23} \\ s_{31} & s_{32} & (a_3-\mu_3)^2+(b_3-\mu_3)^2+(c_3-\mu_3)^2 \end{vmatrix}$$

Kako izračunavamo glavne komponente?

Šta su elementi koji nisu na glavnoj dijagonali?
To su kovarijanse – govore nam o linearnom odnosu atributa po parovima.

Ko-varijansa atributa 1 i 2:

$$S_{21} = \frac{1}{3-1}\left((a_1 - \mu_1)(a_2 - \mu_2) + (b_1 - \mu_1)(b_2 - \mu_2) + (c_1 - \mu_1)(c_2 - \mu_2)\right)$$



Matematika PCA

Theorem 1. *If A is symmetric (meaning $A^T = A$), then A is orthogonally diagonalizable and has only real eigenvalues. In other words, there exist real numbers $\lambda_1, \ldots, \lambda_n$ (the eigenvalues) and orthogonal, non-zero real vectors $\vec{v}_1, \ldots, \vec{v}_n$ (the eigenvectors) such that for each $i = 1, 2, \ldots, n$:*

$$A\vec{v}_i = \lambda_i \vec{v}_i.$$

This is a very powerful result (often called the Spectral Theorem), but it is limited by the fact that it applies only to symmetric matrices. Nevertheless, we can still get some use out of the theorem in general with the following observation:

Exercise 1. *If A is any $m \times n$ matrix of real numbers, then the $m \times m$ matrix AA^T and the $n \times n$ matrix A^TA are both symmetric.*

Thus, we can apply the theorem to the matrices AA^T and A^TA. It is natural to ask how the eigenvalues and eigenvectors of these matrices are related.

Proposition 2. *The eigenvalues of AA^T and A^TA are nonnegative numbers.*

Matematika PCA

Theorem 1. *If A is symmetric (meaning $A^T = A$), then A is orthogonally diagonalizable and has only real eigenvalues. In other words, there exist real numbers $\lambda_1, \ldots, \lambda_n$ (the eigenvalues) and orthogonal, non-zero real vectors $\vec{v}_1, \ldots, \vec{v}_n$ (the eigenvectors) such that fo*

Drugim rečima,

Ako je matrica simetirčna onda ima sopstvene vektore koji su ortogonalni jedan na drugog – to nam treba za PCA!

Sopstvenim vektorima odgovaraju sopstvene vrednosti.

This is a … it is limited by the fact … can still get some use ou…

Exercise 1 … *matrix AA^T and the $n \times$* …

Thus, w… s natural to ask how the eigenvalues and eigenvectors of these matrices are related.

Proposition 2. *The eigenvalues of AA^T and A^TA are nonnegative numbers.*

Matematika PCA

Ako uzmemo neku matricu A i pomnožimo je sa A^T (transponovanom) onda su AA^T i A^TA simetrične.

Theorenally diagonaliz-
able and bers $\lambda_1, \ldots, \lambda_n$
(the eige eigenvectors)
such that for each $i = 1, 2, \ldots, n$*:*

$$A\vec{v}_i = \lambda_i \vec{v}_i.$$

This is a very powerful result (often called the Spectral Theorem), but it is limited by the fact that it applies only to symmetric matrices. Nevertheless, we can still get some use out of the theorem in general with the following observation:

Exercise 1. *If A is any $m \times n$ matrix of real numbers, then the $m \times m$ matrix AA^T and the $n \times n$ matrix A^TA are both symmetric.*

Thus, we can apply the theorem to the matrices AA^T and A^TA. It is natural to ask how the eigenvalues and eigenvectors of these matrices are related.

Proposition 2. *The eigenvalues of AA^T and A^TA are nonnegative numbers.*

Matematika PCA

Ako uzmemo neku matricu A i pomnožimo je sa A^T (transponovanom) onda su AA^T i A^TA simetrične.

Na koji način možemo to da iskoristimo?

Pa, matricu kovarijansi ćemo dobiti kao proizvod neke A i A^T.

Theoren…*lly diagonaliz-*
able and …*bers* $\lambda_1, \ldots, \lambda_n$
(the eige … *eigenvectors)*
such that

This i… …ut it is limited
by the fa… …e can still get
some use…

Exercise 1. *If A is any $m \times n$ matrix of real numbers, then the $m \times m$ matrix AA^T and the $n \times n$ matrix A^TA are both symmetric.*

Thus, we can apply the theorem to the matrices AA^T and A^TA. It is natural to ask how the eigenvalues and eigenvectors of these matrices are related.

Proposition 2. *The eigenvalues of AA^T and A^TA are nonnegative numbers.*

Matematika PCA

Matricu kovarijansi ćemo dobiti kao proizvod neke A i A^T.

Šta dobijamo time?

Dokaz da od matrice kovarijansi možemo da dobijemo sopstvene vektore i sopstvene vrednosti.

Sopstveni vektori su ustvari PCA komponente!

Theorenally diagonaliz-
able and bers $\lambda_1, \ldots, \lambda_n$
(the eige eigenvectors)
such that

This i ut it is limited
by the fa e can still get
some use

Exercise n matrix AA^T
and the r

Thus, we can apply the theorem to the matrices AA^T and A^TA. It is natural to ask how the eigenvalues and eigenvectors of these matrices are related.

Proposition 2. *The eigenvalues of AA^T and A^TA are nonnegative numbers.*

Mate

Prikazan je način formiranja matrice kovarijansi kao prozivod A i A^T.

Centralizacija matrice podataka, primer:

$$\vec{x}_1 = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}, \quad \vec{x}_2 = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}, \quad \vec{x}_3 = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}, \quad \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

$$B = \begin{bmatrix} a_1 - \mu_1 & b_1 - \mu_1 & c_1 - \mu_1 \\ a_2 - \mu_2 & b_2 - \mu_2 & c_2 - \mu_2 \\ a_3 - \mu_3 & b_3 - \mu_3 & c_3 - \mu_3 \\ a_4 - \mu_4 & b_4 - \mu_4 & c_4 - \mu_4 \end{bmatrix} \qquad S = \frac{1}{n-1} B B^T$$

Varijansa atributa 1:

$$S_{11} = \frac{1}{3-1}\left((a_1 - \mu_1)^2 + (b_1 - \mu_1)^2 + (c_1 - \mu_1)^2\right)$$

Ko-varijansa atributa 1 i 2:

$$S_{21} = \frac{1}{3-1}\left((a_1 - \mu_1)(a_2 - \mu_2) + (b_1 - \mu_1)(b_2 - \mu_2) + (c_1 - \mu_1)(c_2 - \mu_2)\right)$$

Matematika PCA

Centralizacija matrice podataka, primer:

$$\vec{x}_1 = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}, \quad \vec{x}_2 = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}, \quad \vec{x}_3 = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}, \quad \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

$$B = \begin{bmatrix} a_1 - \mu_1 & b_1 - \mu_1 & c_1 - \mu_1 \\ a_2 - \mu_2 & b_2 - \mu_2 & c_2 - \mu_2 \\ a_3 - \mu_3 & b_3 - \mu_3 & c_3 - \mu_3 \\ a_4 - \mu_4 & b_4 - \mu_4 & c_4 - \mu_4 \end{bmatrix} \qquad S = \frac{1}{n-1} B B^T$$

Varijansa

$$S_{11} = \frac{1}{3}$$

Ko-varijan

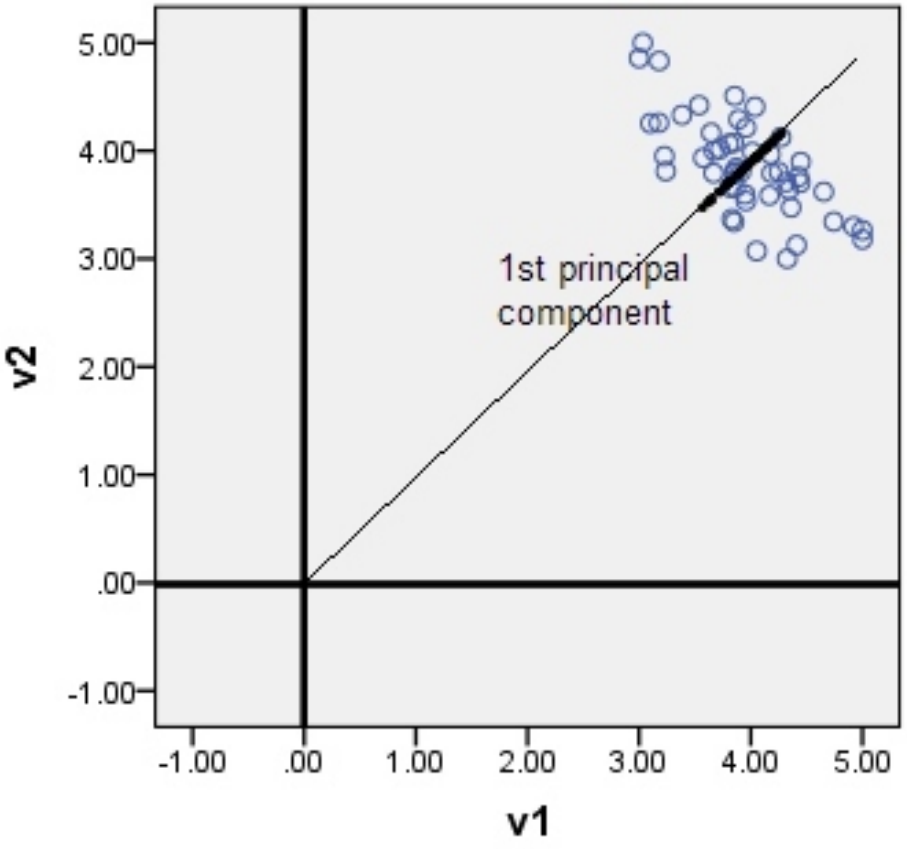
$$S_{21} = \frac{1}{3}$$

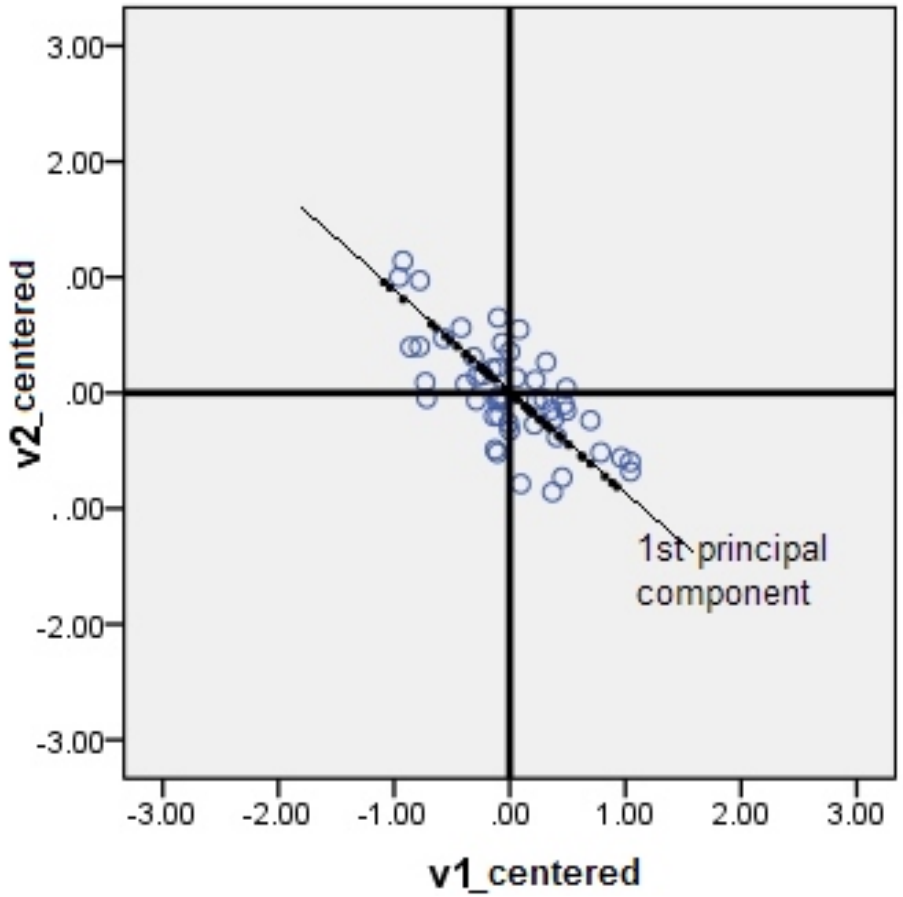
$c_2 - \mu_2))$

Prvo centriramo podatke. Redom za svaki atribut (a,b,c) idemo po vrednostima (a_1,a_2...) i od svake vrednosti oduzmemo srednju vrednost tog atributa (μ_1, μ_2, μ_3).
Tako dobijamo matricu B.

Centriranje podataka - Informativno

1. Podaci <u>moraju</u> biti centrirani





Centriranje podataka - Informativno

- Postupak centriranja podataka (*mean normalization*):
  a. Dat je trening skup $T = \left\{\left(x^{(i)}, y^{(i)}\right), i \in \{1, \ldots, N\}, x^{(i)} \in \mathbb{R}^D\right\}$
  b. Za svako obeležje $d \in \{1, \ldots, D\}$ izračunati srednju vrednost: $\mu_d = \frac{1}{N}\sum_{i=1}^{N} x_d^{(i)}$
  c. Za svako obeležje d: $x_d^{(i)} \leftarrow x_d^{(i)} - \mu_d$

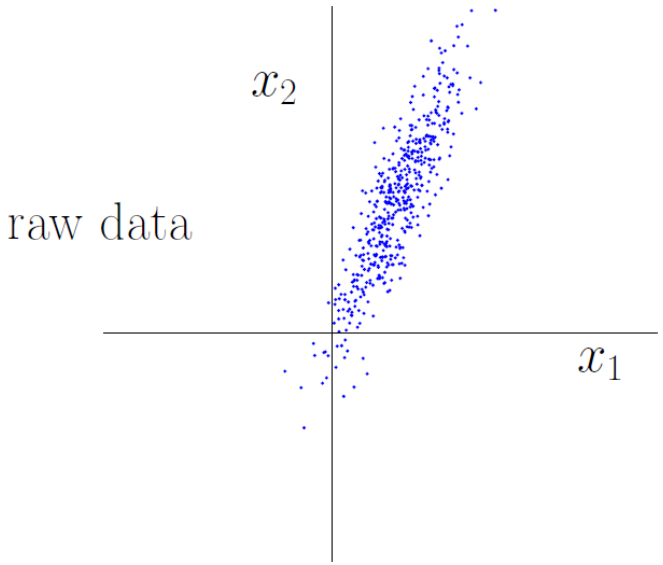
Centriranje podataka - Informativno

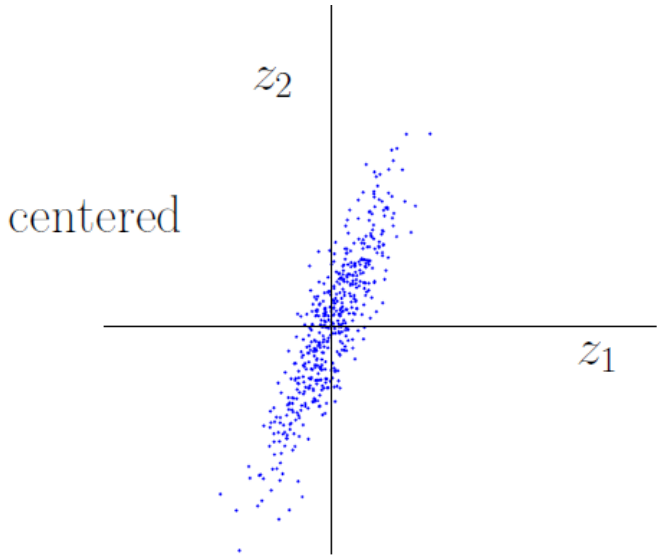
2. Normalizovati podatke (opciono)
   - Ako se opsezi vrednosti različitih obeležja veoma razlikuju, obeležja treba skalirati tako da se kreću u približno istom opsegu, npr.

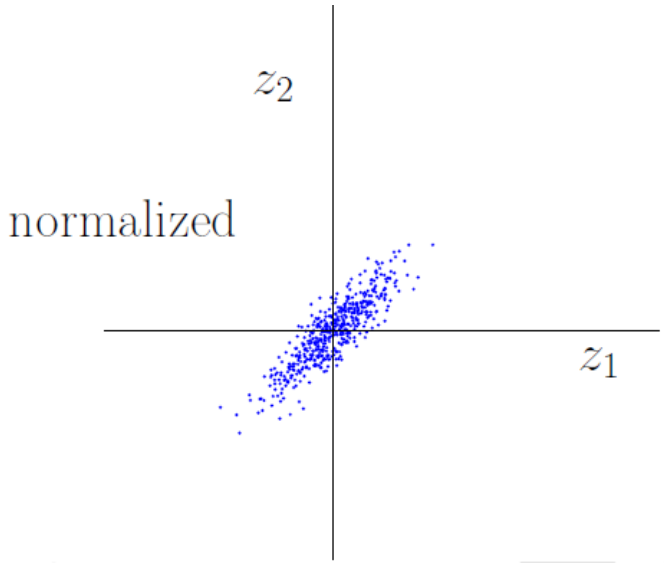
$$x_d^{(i)} \leftarrow \frac{x_d^{(i)}}{\sigma_d}$$

   - Velike razlike u opsezima varijabli koje potiču iz (proizvoljnog) odabira jedinice u kojima ih izražavamo su problem za PCA

Centriranje podataka - Informativno







Matematika PCA

Centralizacija matrice podataka, primer:

Šta su elementi BB^T pomnožene sa 1/n-1?

Varijanse atributa i kovarijanse.

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix}$$

$$B = \begin{bmatrix} a_1 - \mu_1 & b_1 - \mu_1 & c_1 - \mu_1 \\ a_2 - \mu_2 & b_2 - \mu_2 & c_2 - \mu_2 \\ a_3 - \mu_3 & b_3 - \mu_3 & c_3 - \mu_3 \\ a_4 - \mu_4 & b_4 - \mu_4 & c_4 - \mu_4 \end{bmatrix} \qquad S = \frac{1}{n-1} B B^T$$

Varijansa atributa 1:

$$S_{11} = \frac{1}{3-1}\left((a_1 - \mu_1)^2 + (b_1 - \mu_1)^2 + (c_1 - \mu_1)^2\right)$$

Ko-varijansa atributa 1 i 2:

$$S_{21} = \frac{1}{3-1}\left((a_1 - \mu_1)(a_2 - \mu_2) + (b_1 - \mu_1)(b_2 - \mu_2) + (c_1 - \mu_1)(c_2 - \mu_2)\right)$$

Matematika PCA

Centralizacija matrice podataka, primer:

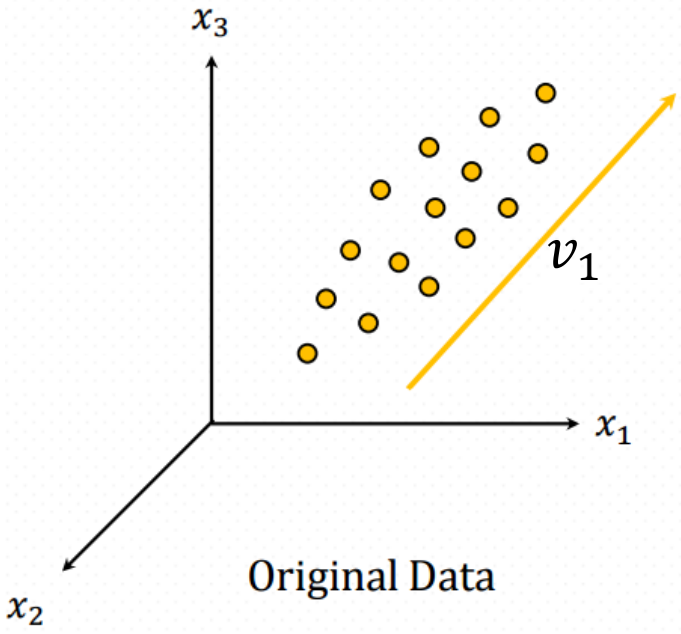
Pokazali smo da je matrica kovarijansi simetrična.

Sad iz nje možemo da dobijemo sopstvene vektore i sopstvene vrednosti.

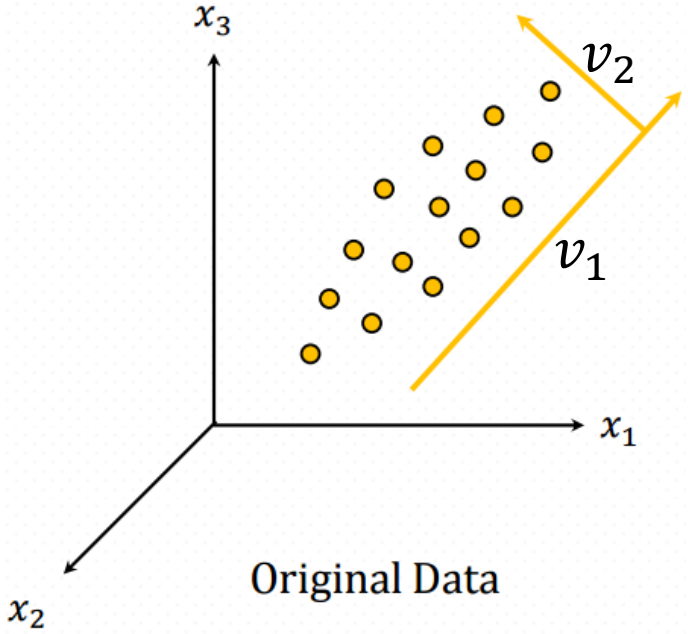
Tumačimo ih kao na sledećem slajdu.

$$BB^T = \begin{vmatrix} (a_1-\mu_1)^2+(b_1-\mu_1)^2+(c_1-\mu_1)^2 & s_{12} & s_{13} \\ s_{21} & (a_2-\mu_2)^2+(b_2-\mu_2)^2+(c_2-\mu_2)^2 & s_{23} \\ s_{31} & s_{32} & (a_3-\mu_3)^2+(b_3-\mu_3)^2+(c_3-\mu_3)^2 \end{vmatrix}$$

PCA: dekompozicija matrice Σ



Prvi sopstveni vektor v_1 (sa najvećim λ_1) je u pravcu najveće varijabilnosti podataka



Naredna projekcija v_2 (drugo po veličini λ_2) je u pravcu sledeće najveće varijabilnosti, pri čemu je ortogonalna na sve prethodne projekcije

(da nije ortogonalna, obuhvatala bi varijabilnost već obuhvaćenu prethodnim projekcijama)

Zadatak PCA – alternativna formulacija

- Do sada izložena fromulacija zadatka PCA bila je fokusirana na objašnjavanje varijabilnosti podataka.
- Postoji alternativna formulacija zadatka PCA koja se oslanja na grešku rekonstrukcije.
- Od formulacije zadatka zavisi i metodologija pomoću koje se rešava, pa tako dobijamo dve metodolgije za određivanje PCA komponenti.
- Obe formulacije zadatka PCA su međusobno jednake, odnosno njihovim rešavanjem dobijaju se isti rezultati.
- Detljanije u nastavku...

Odbačene informacije

- Da smo sačuvali sve komponente $z_1, \ldots, z_D$, mogli bismo rekonstruisati originalni vektor x:

$$x = \sum_{i=1}^{D} z_i v_i$$

- Sa prvih K komponenti, rekonstrukcija je

$$\hat{x} = \sum_{i=1}^{K} z_i v_i$$

- Magnituda odbačenih informacija (greška rekonstrukcije) je:

$$\|x - \hat{x}\|^2 = \left\| \sum_{i=K+1}^{D} z_i v_i \right\|^2 = \sum_{i=K+1}^{D} z_i^2$$

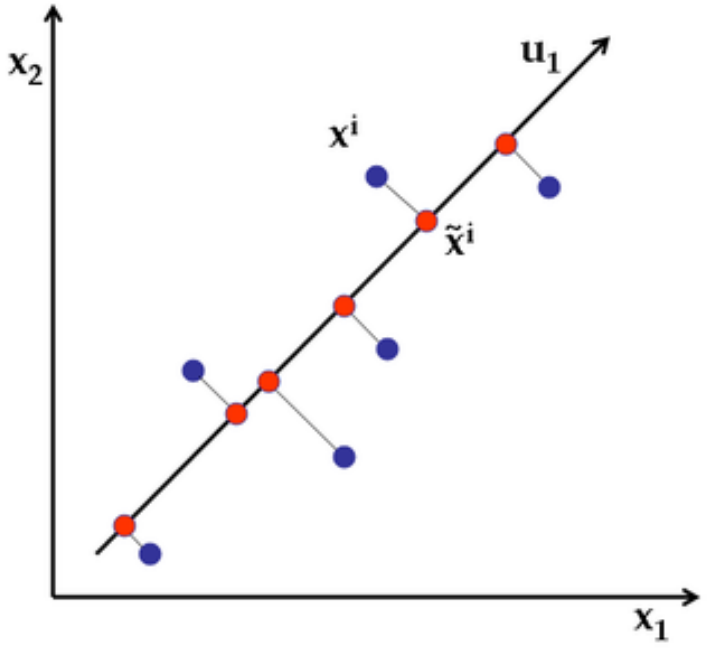
Odbačene informacije

- Novi koordinatni sistem je dobar ako je suma grešaka rekonstrukcije izračunata za sve primere skupa podataka mala ($\hat{x}_n \approx x_n$ za $n \in \{1, \ldots, N\}$), tj., ako je malo:

$$\sum_{n=1}^{N} \|x_n - \hat{x}_n\|^2$$

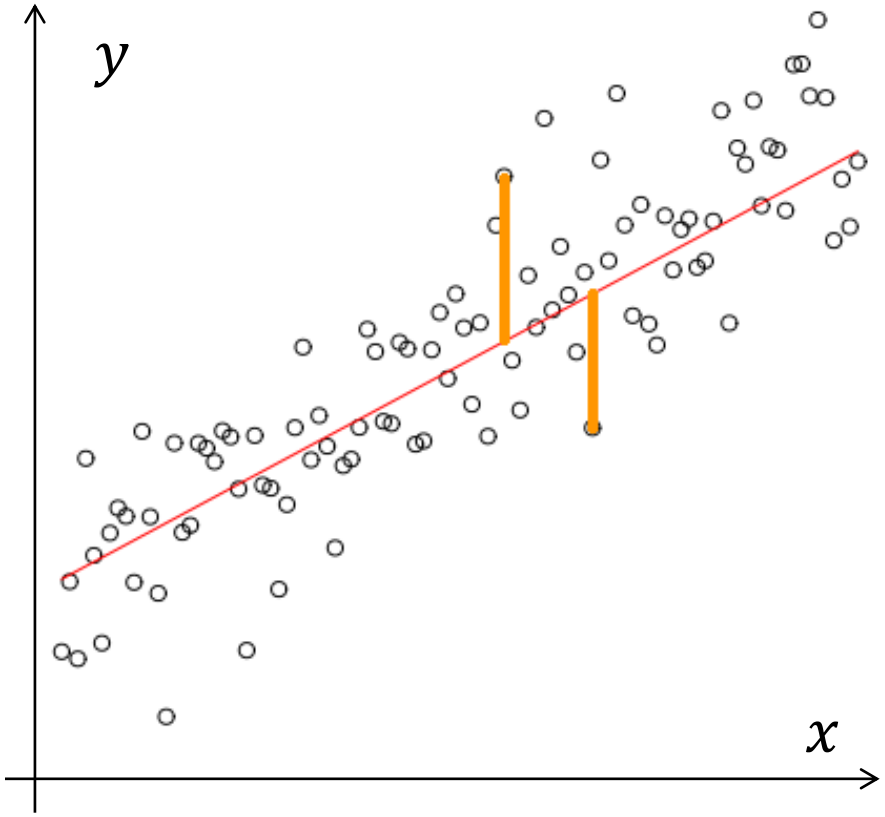
- PCA zato pronalazi koordinatni sistem koji mimimizuje ukupnu grešku rekonstrukcije

- Minimizujemo rastojanja tačaka od njihovih projekcija



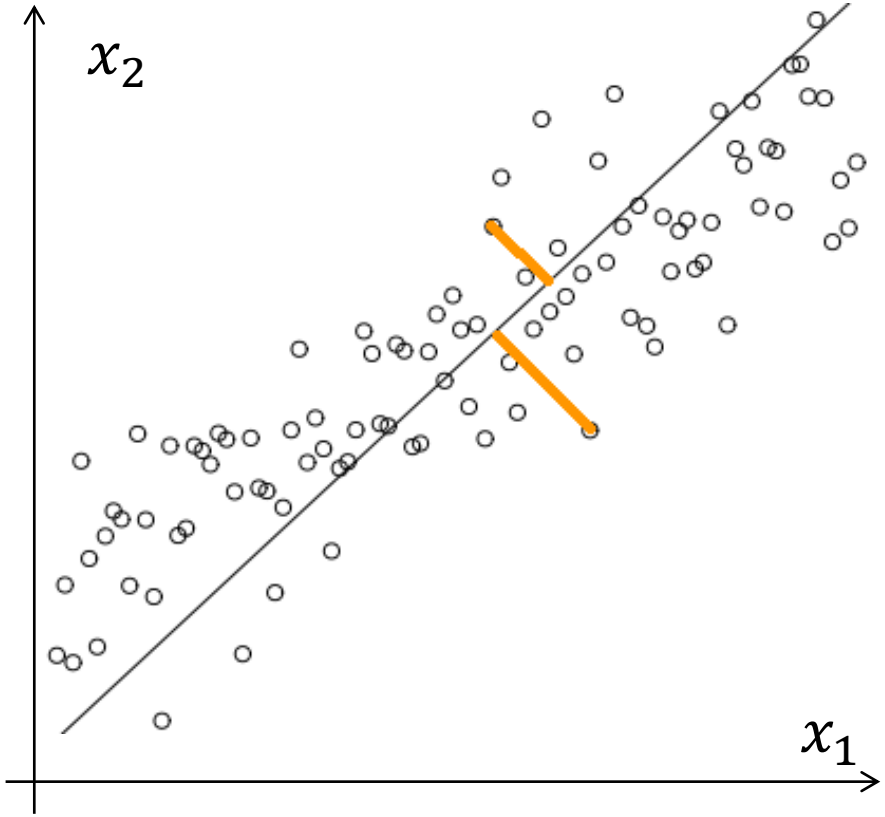
PCA *nije* linearna regresija

Linearna regresija



Specijalno obeležje y (ciljna varijabla) $y = f(x)$

PCA

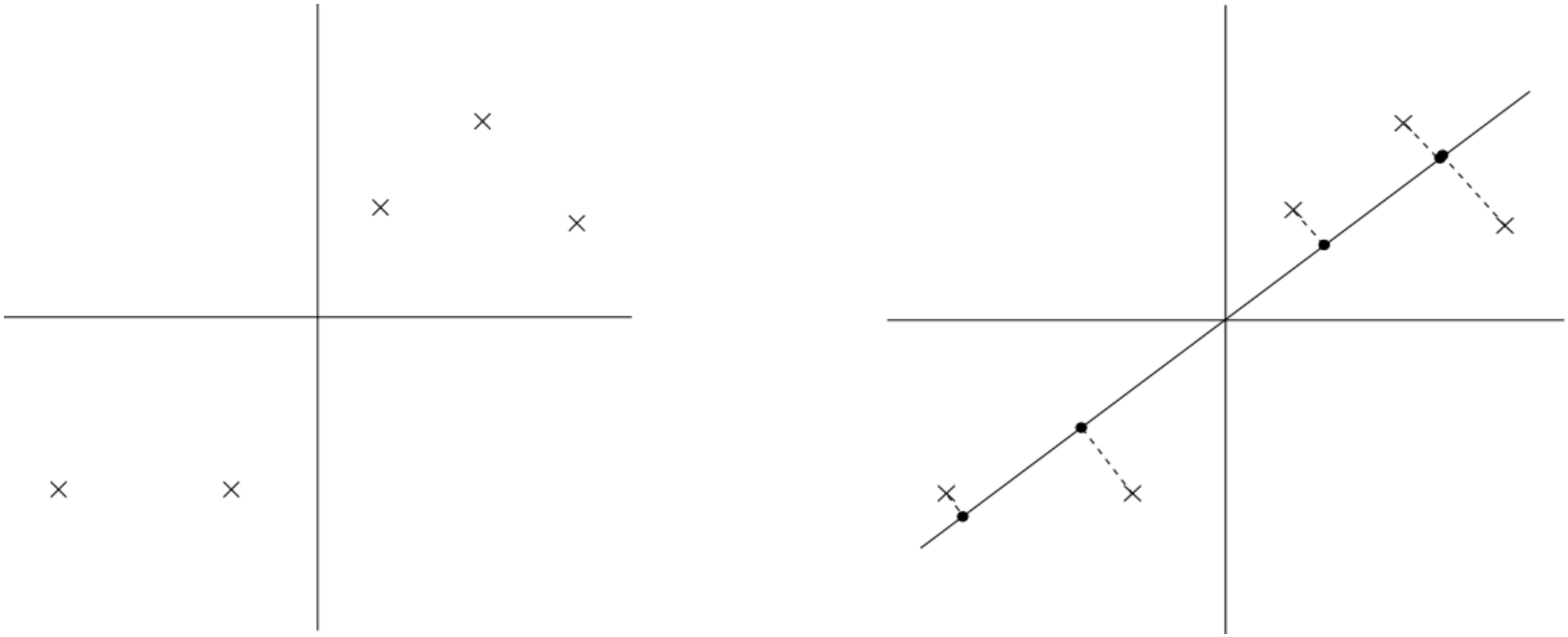


Sva obeležja su tretirana identično

PCA – druga (ekvivalentna) formulacija

- Cilj: zadržati što više informacija u podacima

  ⇒ novi koordinatni sistem određujemo tako da zadržimo što je moguće više *varijabilnosti* u dobijenoj projekciji

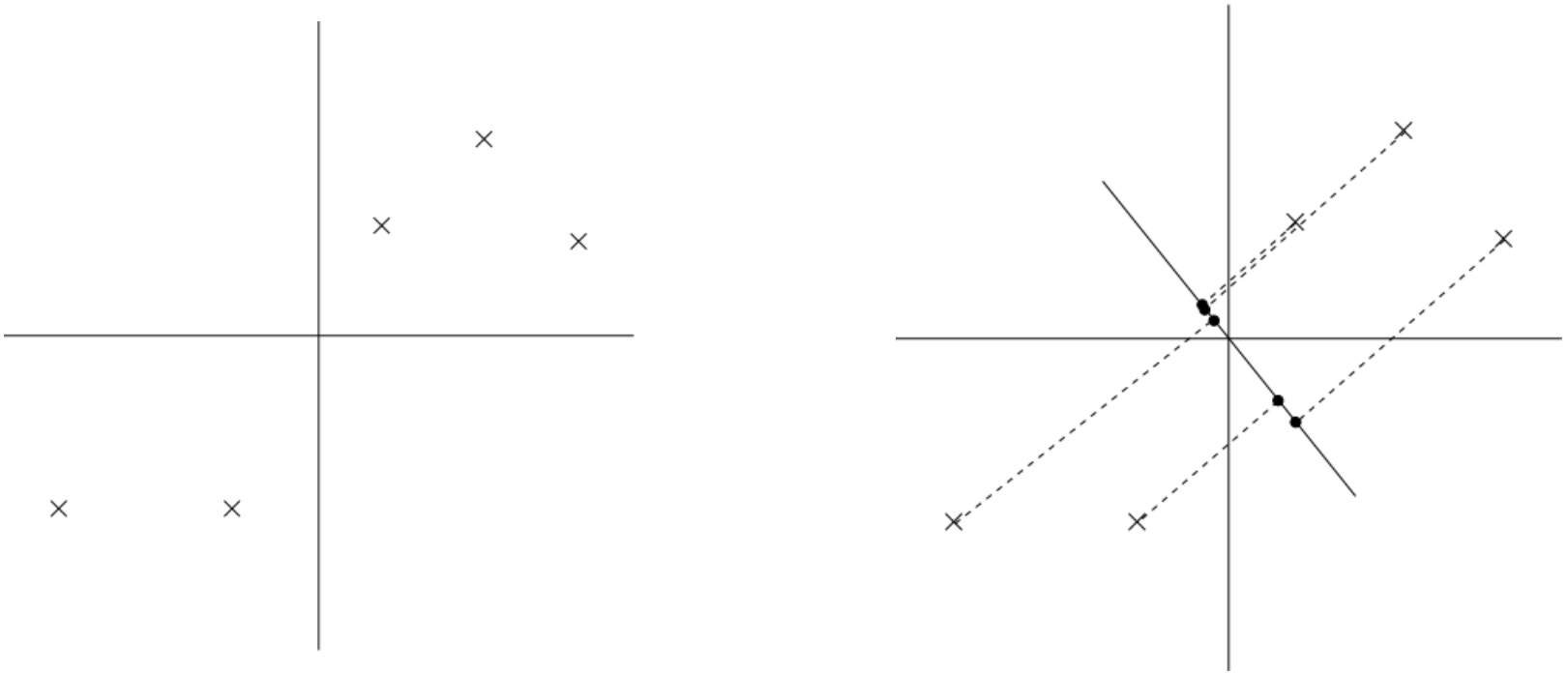


Projektovani podaci i dalje imaju dosta veliku varijansu. Podaci imaju tendenciju da budu daleko od nule

PCA – druga (ekvivalentna) formulacija

- Cilj: zadržati što više informacija u podacima

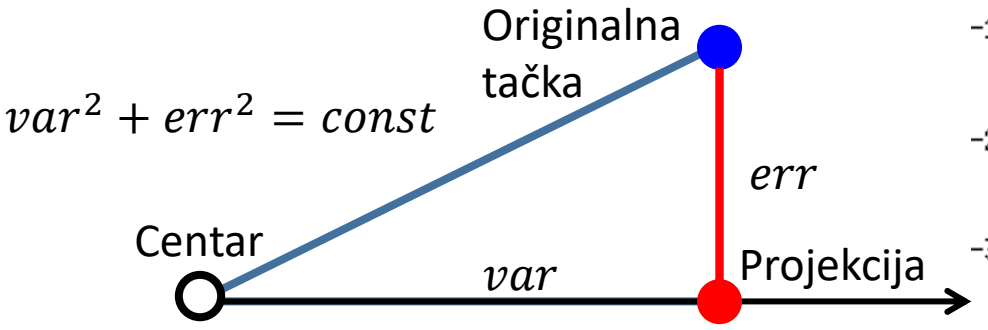
  ⇒ novi koordinatni sistem određujemo tako da zadržimo što je moguće više *varijabilnosti* u dobijenoj projekciji

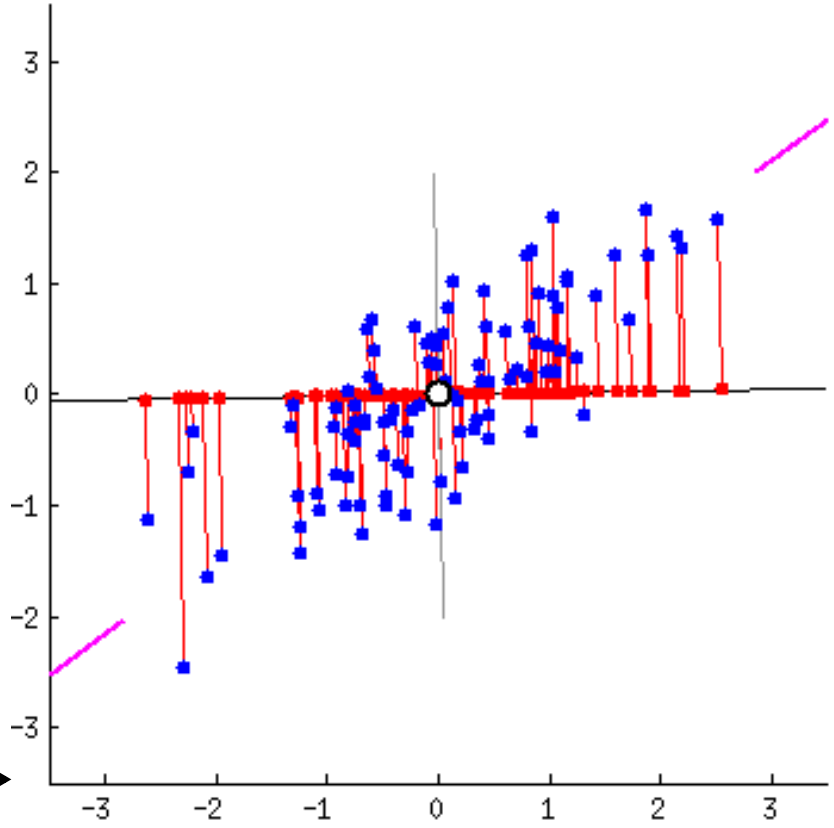


Projektovani podaci imaju znatno manju varijansu i mnogo su bliži nuli

PCA – dve ekvivalentne formulacije

- Ove dve formulacije PCA su ekvivalentne
  - Na slici možete primetiti da je zadržana varijansa projekcija („širina“ crvenih tačaka na novoj osi) najveća istovremeno kada je greška rekonstrukcije (suma crvenih linija) najmanja





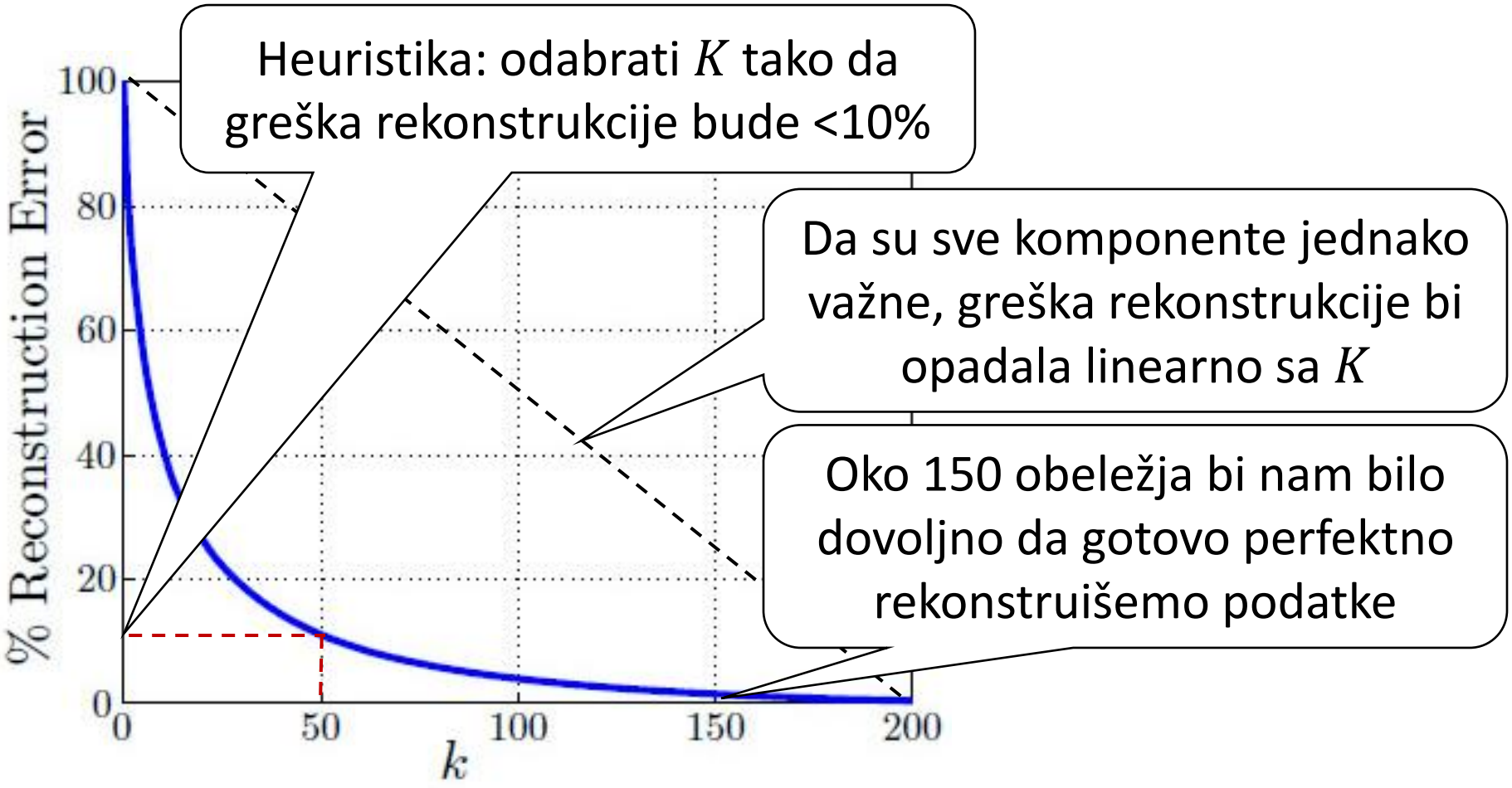
- PCA: Principal Component Analysis (analiza glavnih komponenti)
  - Nova osa je prva glavna komponenta

Kako odabrati broj novih obeležja *K*?

- Ako nam je cilj vizuelizacija, odabraćemo $K = 2$ ili $K = 3$
- Ako nam je cilj da ubrzamo obučavajući algoritam, obično ćemo *K* odabrati tako da zadržimo određeni procenat varijanse (tipično 99%, 95% ili 90%)

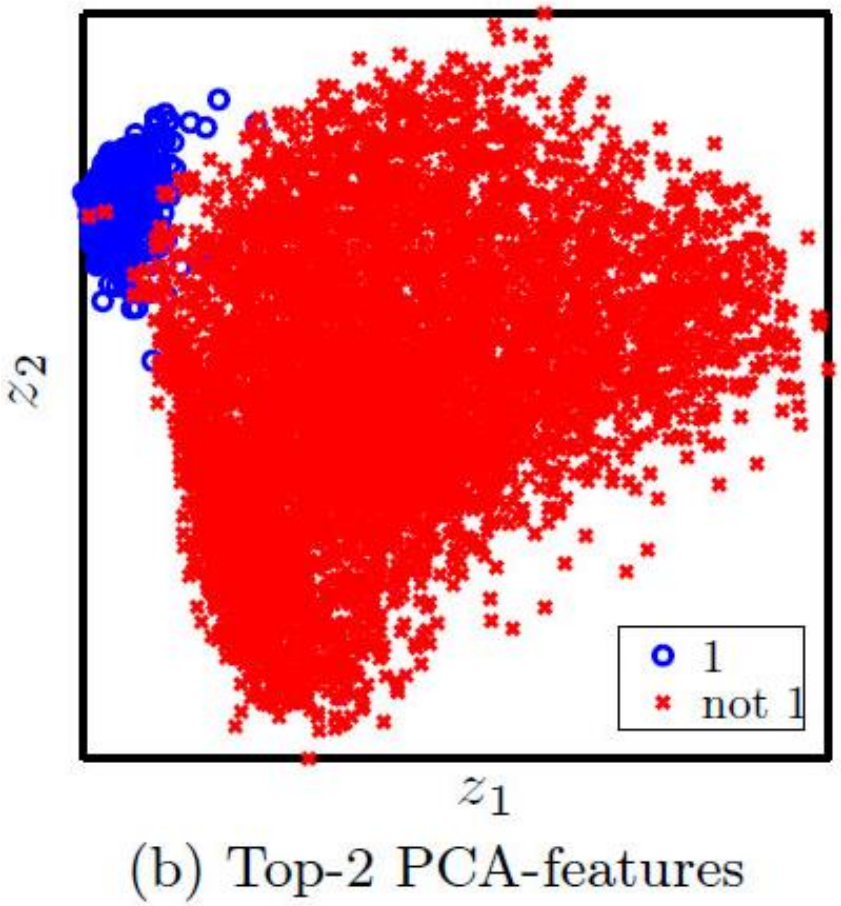
Primer: OCR

- prepoznavanje rukom pisanih cifara sa slika 16×16
- $X \in \mathbb{R}^{7291 \times 256}$ ćemo transformisati u prostor $\mathbb{R}^{7291 \times 2}$



Primer: OCR

- prepoznavanje rukom pisanih cifara sa slika 16×16
- $X \in \mathbb{R}^{7291 \times 256}$ ćemo transformisati u prostor $\mathbb{R}^{7291 \times 2}$



(b) Top-2 PCA-features

- Prikazane su dva najvažnija obeležja dobijena uz pomoć PCA
- Skup podataka je prikazan u novom prostoru i obeležene su instance anotirane kao 1 (plavo) i ostale (crveno)
- Vidimo da nova obeležja omogućavaju prilično jasno razlikovanje klasa

Primer PCA: ocene opština

- Places Rated Almanac (Boyer and Savageau)
- 329 opština ocenjeno na osnovu sledećih kriterijuma:
  1. Klima i zemljište
  2. Cena stambenog prostora
  3. Zdravstvo i životna sredina
  4. Kriminal
  5. Transport
  6. Obrazovanje
  7. Umetnost
  8. Rekreacija
  9. Ekonomija
- Problem: puno dimenzija – teško za interpretaciju podataka

Primena PCA na podatke

Component	Eigenvalue	Proportion	Cumulative
1	0.3775	0.7227	0.7227
2	0.0511	0.0977	0.8204
3	0.0279	0.0535	0.8739
4	0.0230	0.0440	0.9178
5	0.0168	0.0321	0.9500
6	0.0120	0.0229	0.9728
7	0.0085	0.0162	0.9890
8	0.0039	0.0075	0.9966
9	0.0018	0.0034	1.0000
Total	0.5225		

Ukupna varijansa ($\lambda_1 + \cdots + \lambda_9$)

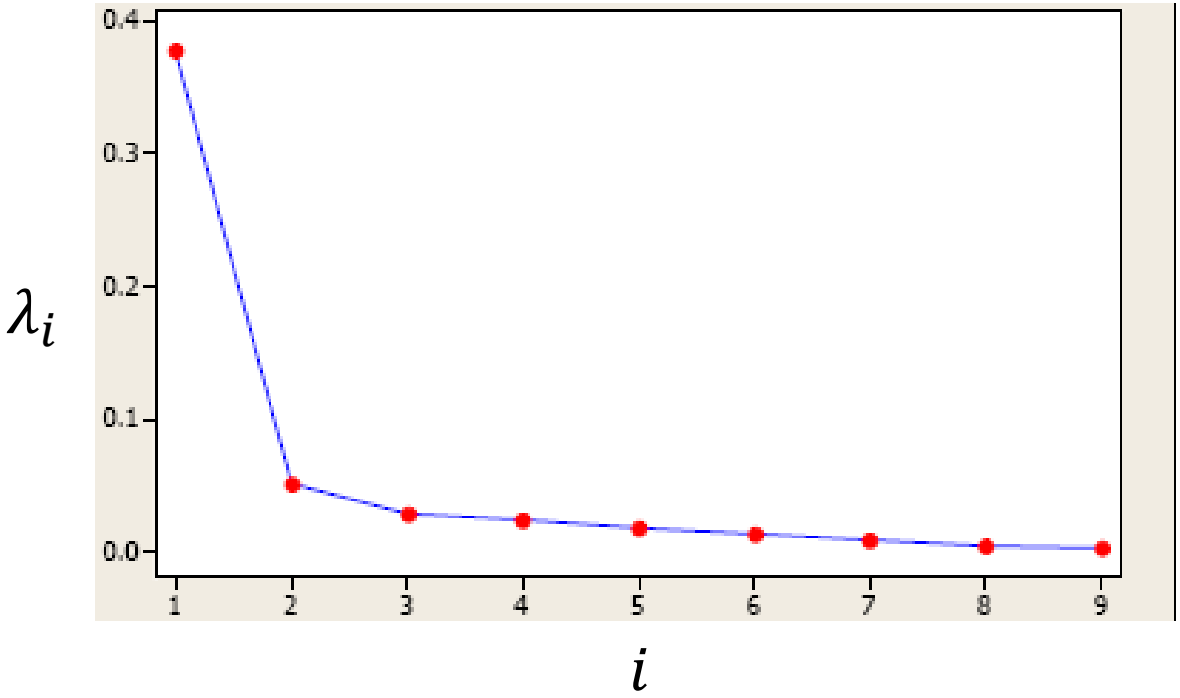
Proportion:

- deo varijanse objašnjen od strane glavne komponente
- Npr., za prvu komponentu $\frac{0.3775}{0.5223} = 0.7227$, odnosno oko 72% varijanse u podacima je objašnjeno prvom glavnom komponentom

Cumulative:

- Kumulativna varijansa dobijena dodavanjem uzastopnih udela u varijansi prvih *K* glavnih komponenti
- Npr. prve dve glavne komponente zajedno objašnjavaju $0.7227 + 0.0977 = 0.8204$

Odabir K



- Jedan način: odabrati prvih 5 komponenti jer je na taj način zadržano 95% varijanse u podacima. Ovo je razuman procenat ako je naš cilj prediktivno modelovanje
- Drugi način: pogledati grafik na slici – nakon 3. komponente, preostale sopstvene vrednosti su male i približno iste veličine. Prve 3 komponente objašnjavaju 87% varijanse. Ovo je razumno visok procenat ako je naš cilj interpretacija podataka

Prva glavna komponenta

$$Z = XV$$

$$z_1 = 0.0351 \cdot \text{climate} + 0.0993 \cdot \text{housing} + 0.4078 \cdot \text{health} + 0.1004 \cdot \text{crime} + 0.1501 \cdot \text{transportation} + 0.0321 \cdot \text{education} + 0.8743 \cdot \text{arts} + 0.1590 \cdot \text{recreation} + 0.0195 \cdot \text{economy}$$

- Magnitude koeficijenata predstavljaju udeo originalnih varijabli u datoj glavnoj komponenti
- Ali, imajte na umu da ove magnitude zavise i od varijanse datih varijabli

Interpretacija glavnih komponenti

	Principal Component		
Variable	**1**	**2**	**3**
Climate	0.190	0.017	0.207
Housing	**0.544**	0.020	0.204
Health	**0.782**	**-0.605**	0.144
Crime	0.365	0.294	**0.585**
Transportation	**0.585**	0.085	0.234
Education	0.394	-0.273	0.027
Arts	**0.985**	0.126	-0.111
Recreation	**0.520**	0.402	**0.519**
Economy	0.142	0.150	0.239

- U cilju interpretacije, izračunaćemo korelaciju originalnih varijabli sa glavnim komponentama
- Posmatraćemo najsnažnije korelacije po apsolutnoj vrednosti
- Šta se smatra „snažnom“ korelacijom je subjektivno. Ovde su uzete u obzir korelacije preko 0.5

Interpretacija glavnih komponenti

	Principal Component		
Variable	**1**	**2**	**3**
Climate	0.190	0.017	0.207
Housing	**0.544**	0.020	0.204
Health	**0.782**	**-0.605**	0.144
Crime	0.365	0.294	**0.585**
Transportation	**0.585**	0.085	0.234
Education	0.394	-0.273	0.027
Arts	**0.985**	0.126	-0.111
Recreation	**0.520**	0.402	**0.519**
Economy	0.142	0.150	0.239

PCA1:

- Uvećava se sa uvećanjem *housing*, *health, transportation, arts i recreation* – ovih 5 kriterijuma variraju zajedno, ako se jedna poveća, i ostale će
- Komponentu možemo videti kao meru kvaliteta umetnosti, zdravlja, transporta i rekreacije i viših cena nekretnina
- Najjača korelacija je sa umetnošću
- Ima smisla da su opštine sa puno umetnosti i najskuplje

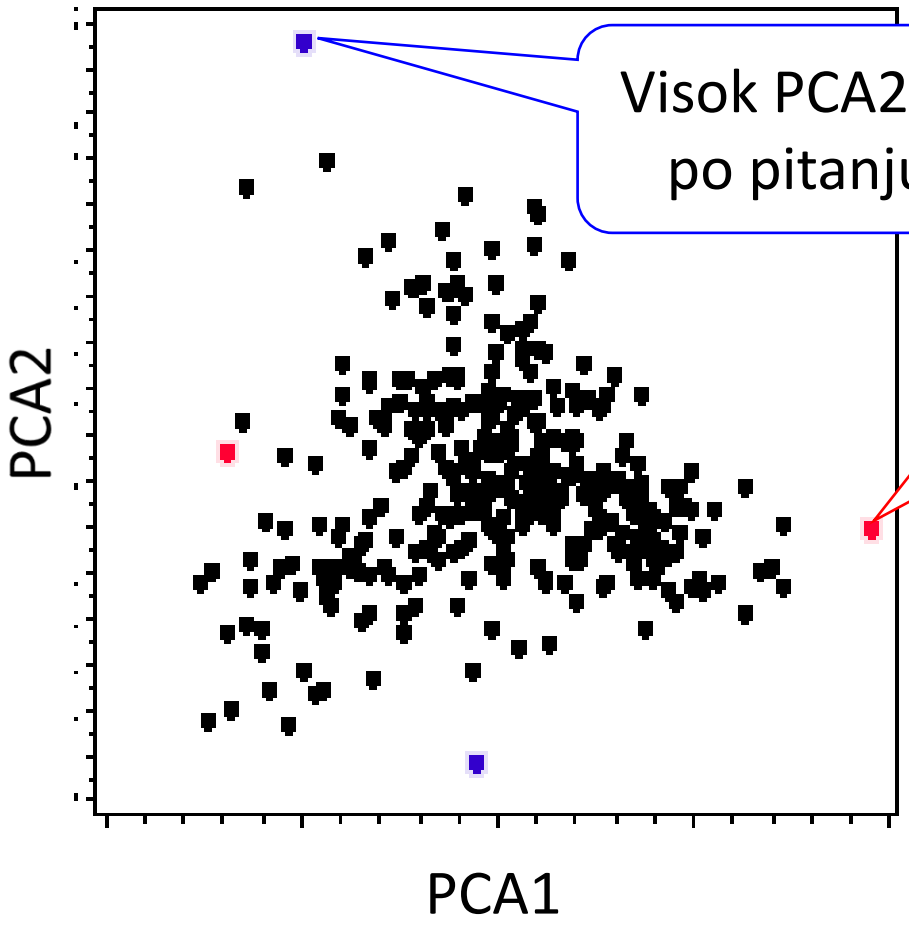
PCA2:

- Uvećava se sa opadanjem kvaliteta zdravstva

PCA3:

- Uvećava se sa porastom kriminala i rekreacije
- Ovo ukazuje da opštine sa većim kriminalom imaju i više rekreacionih ustanova

Interpretacija glavnih komponenti



Svaka tačka predstavlja opštinu

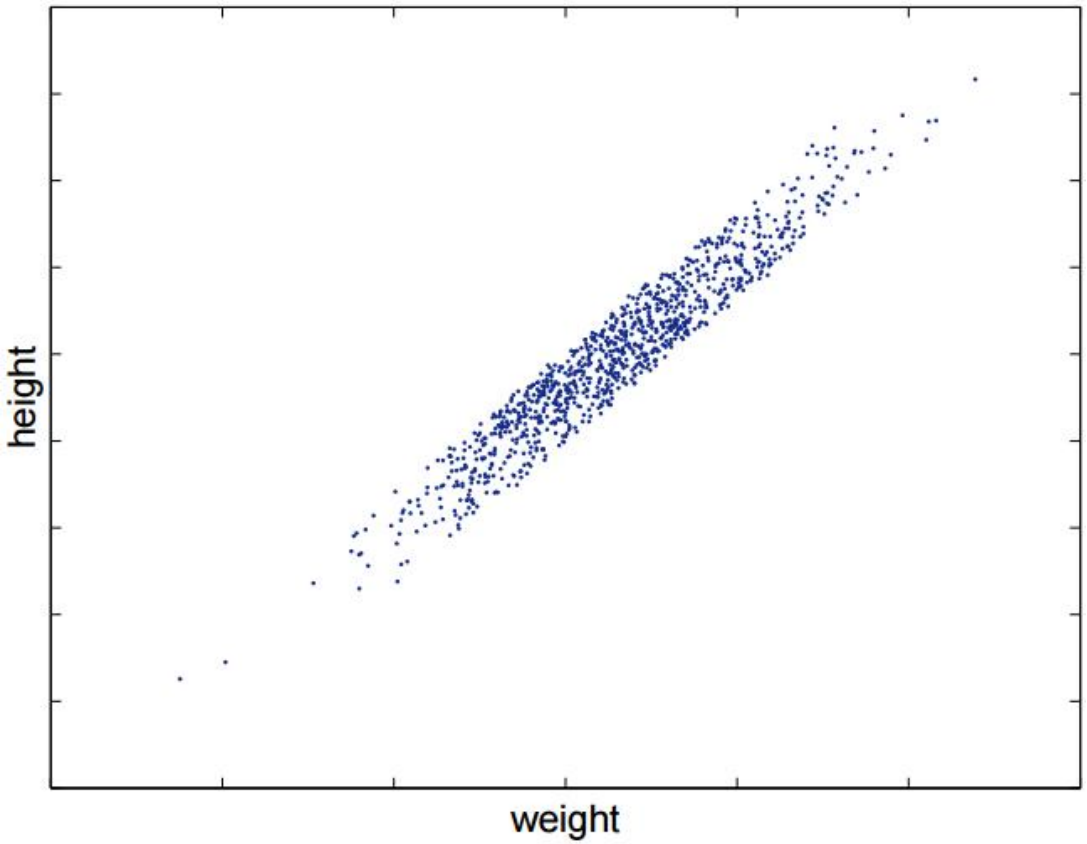
Visok PCA2 (loša opština po pitanju zdravstva)

Visok PCA1 (visoke vrednosti umetnosti, zdravstva,...)

Mogli bismo razmatrati lokacije opština da vidimo da li opštine za koje je neka komponenta visoka/niska pripadaju istoj regiji u zemlji

Nedostaci PCA: Interpretacija

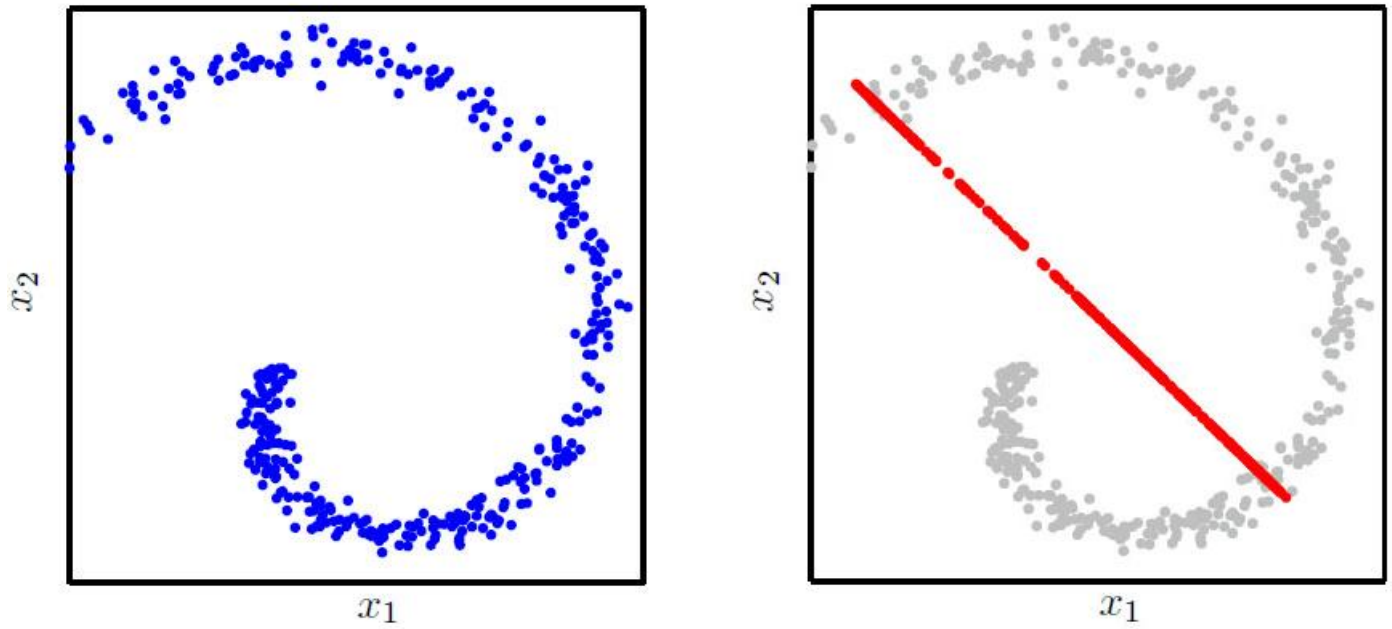
- Obeležja dobijena pomoću PCA metode su linearne kombinacije originalnih obeležja. Ovo je često teško interpretirati



Težinu i visinu menjamo samo jednom koordinatom.

Kako interpretirati ovu novu dimenziju, odnosno, pridodati joj neko fizičko značenje?

Nedostaci PCA: Linearnost



(a) Data in $\mathcal{X}$ space (b) Top-1 PCA reconstruction

- Podaci (a) približno leže na jednodimenzionoj površi (krivoj)
- Međutim, ako pokušamo da rekonstruišemo ove podatke pomoću PCA, rezultati su katastrofalni (b)
- Razlog je što kriva nije linearna, a PCA može da kreira samo linearna obeležja