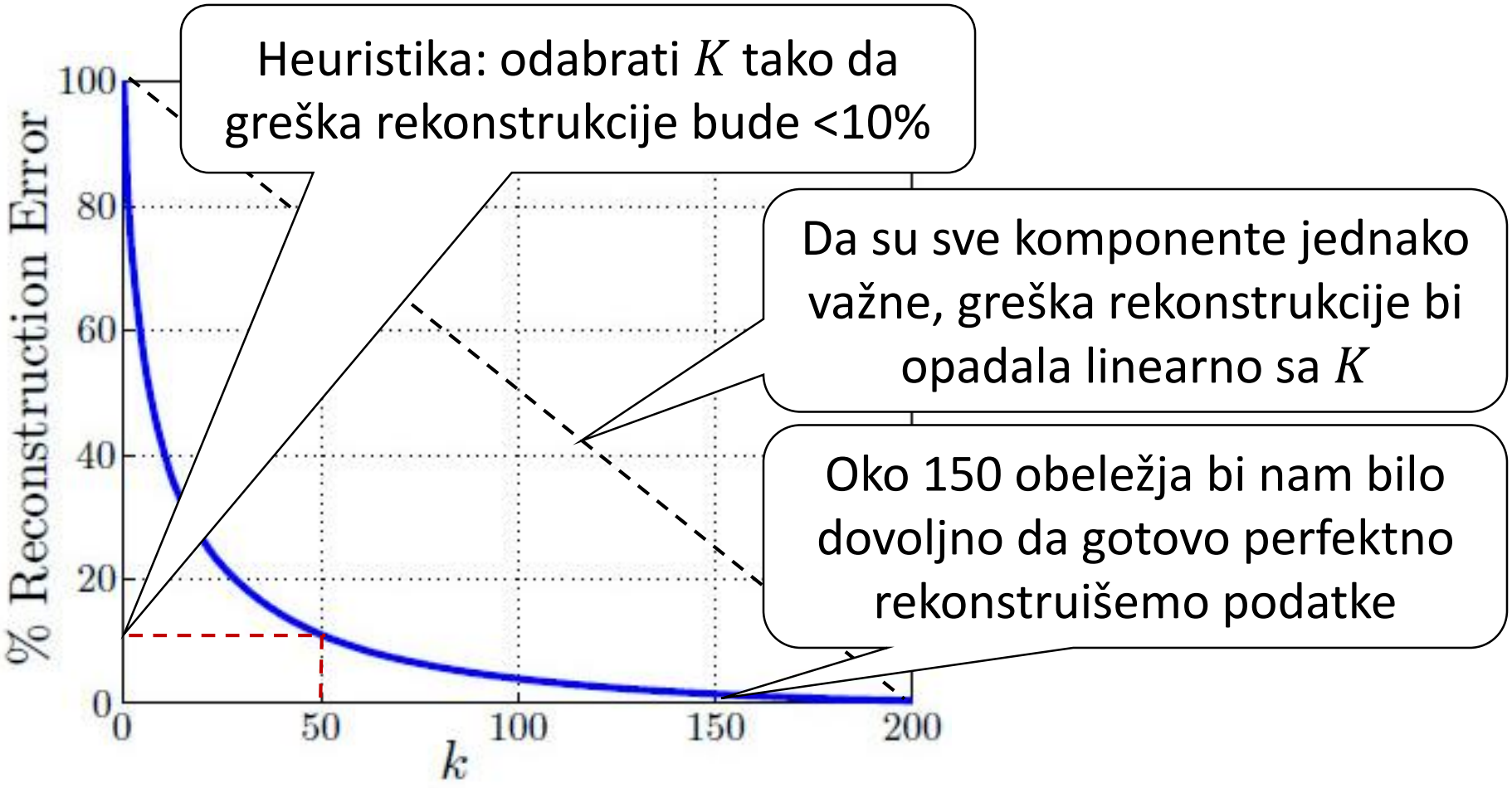
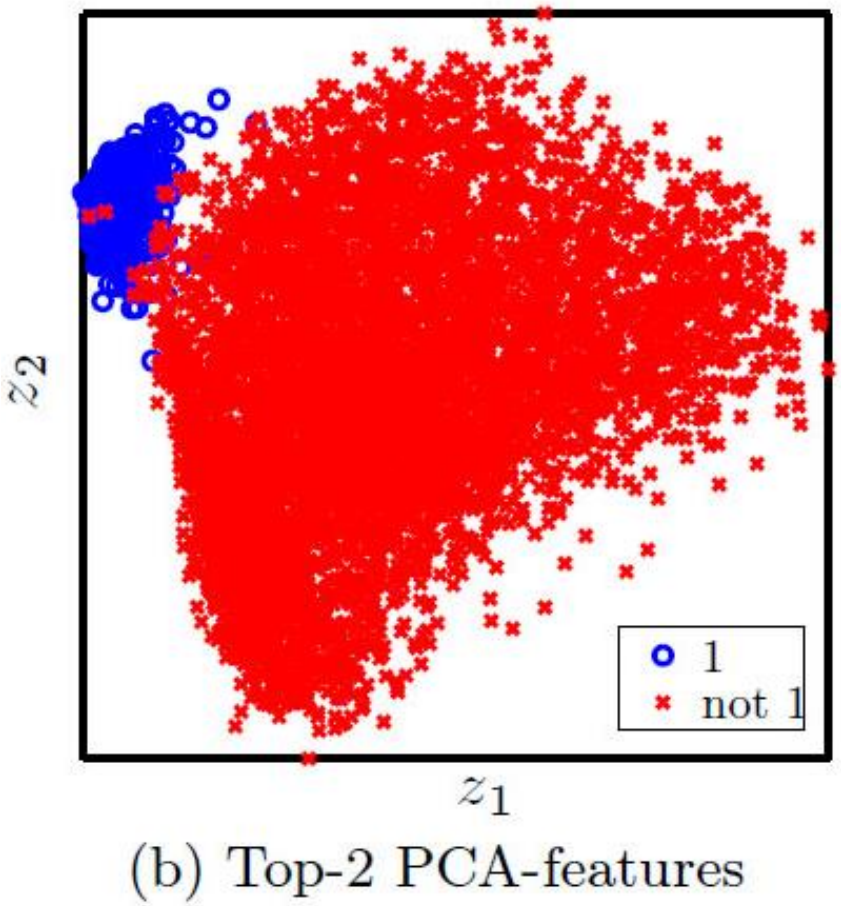
Primer: OCR

- 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

PCA obeležja i ručna konstrukcija obeležja

- Ručno konstruisana obeležja: simetriju i intenzitet
- Slično ovim obeležjima, dva obeležja dobijena pomoću PCA omogućavaju prilično jasno razlikovanje klasa
- Glavna prednost PCA obeležja nad ručno konstruisanim obeležjima jeste što su ova automatski generisana – ne moramo znati ništa o problemu prepoznavanja cifara kako bismo ih dobili
- Ovo je takođe njihov najveći nedostatak – dobijena obeležja su dokazano dobra da rekonstruišu originalne ulazne podatke, ali, nema garancije da će biti korisna prilikom rešavanja problema koji pokušavamo da rešimo
- U praksi, korišćenje domenskog znanja za konstrukciju obeležja, zajedno sa automatskim metodama za redukciju dimenzionalnosti obično ima najbolje performanse

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

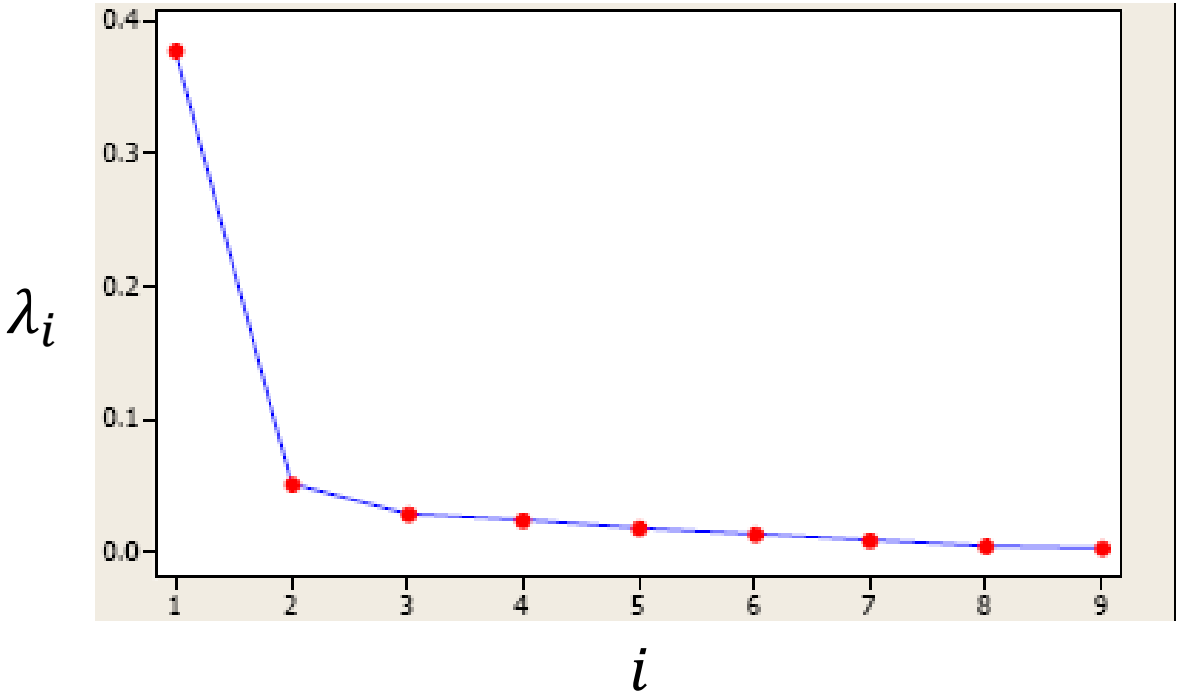
Proportion:

- 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

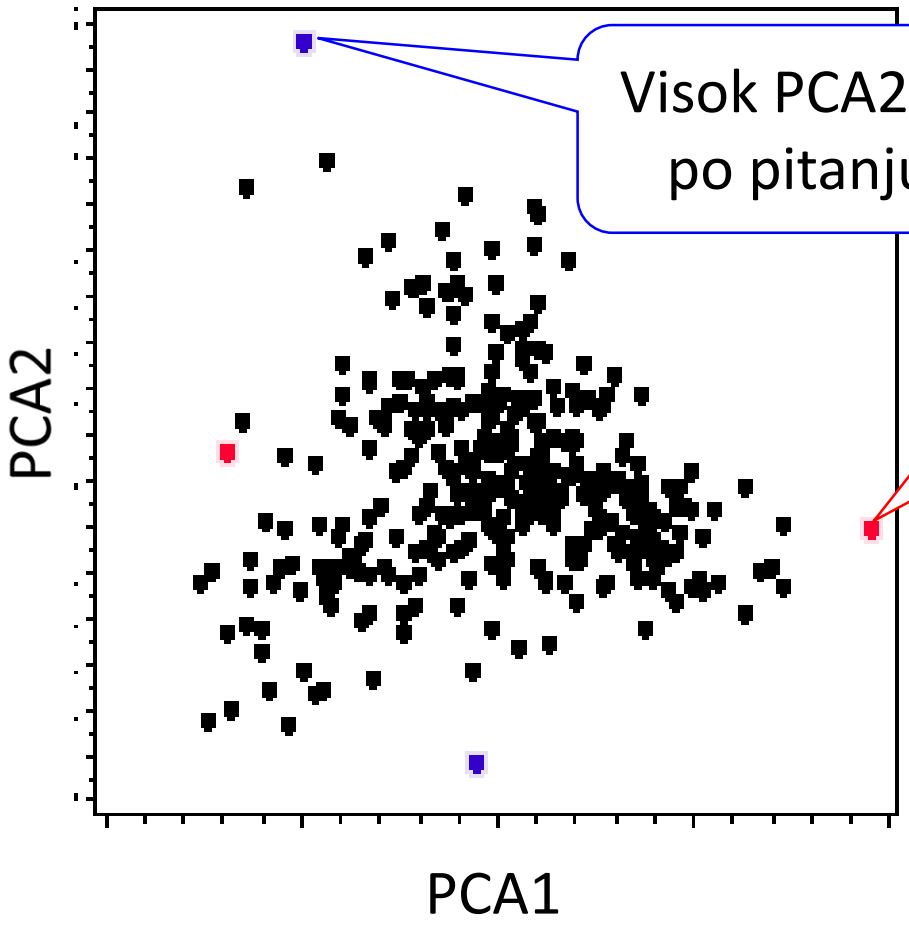
PCA2:

- 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