Matematičke metode u marketingu
Multidimenzionalno skaliranje

Lavoslav Čaklović
PMF-MO

2016

MDS

Čemu služi:
– za redukciju dimenzije

Bazirano na:
– udaljenosti (sličnosti) među objektima

Problem:
Traži se projekcija najmanjeg ranga koja čuva zadane udaljenosti.

Obradit ćemo:

1. klasično MDS,
2. metrijsko MDS,
3. ne metrijsko MDS.

Objekti mogu imati metrijske karakteristike, ordinalne ili kategorijske.

Sličnost (bliskost[1]) udaljenost[2],

Fukciju $d : S \to \mathbb{R}$ nazivamo *metrikom* ako zadovoljava aksiome:

1. $\forall x, y \in S,\ d(x, y) \geq 0$ (pozitivnost)
2. $d(x, y) = 0 \iff x = y$ (definitnost)
3. $\forall x, y \in S,\ d(x, y) = d(y, x)$ (simetričnost)
4. $\forall x, y, z \in S,\ d(x, z) \leq d(x, y) + d(y, z)$. (tranzitivnost)

Primjer: euklidska metrika $\|x - y\|$, metrika Minkowskog (izvedena iz p-norme $\|x - y\|_p, p \geq 1$)

Problem. Zadan je konačan skup S i funkcija $d : S \to \mathbb{R}$ koja zadovoljava pozitivnost i simetričnost (obično kao simetrična matrica $D_{n\times n}$). Naći $x_1, \ldots x_n \in \mathbb{R}^p$ tako da je $d_{ij} \approx \|x_i - x_j\|$.

[1] similarity, proximity

[2] distance

Doprinosi

- U kontekstu klasične diferencijalne geometrije:
  Cayley (1841), Menger (1928), Fréchet (1935), Schoenberg (1935), Young-Householder (1938), Blumenthal (1938, 1953).
- Nemetričko skaliranje:
  Stumpf (1880) promatra mjeru razlikovanja (dissimilarity), Thurstonova škola – Messik & Abelson(1956), Coombs (1950), Kruskal (1973, 74) nerazumljiv za tadašnje psihometričare (STRESS), Tversky& Kranz (1968, 1970).
- Problem MDS sa šumom (probabilistic approach):
  Borg & Groenen (1997), MacKay (1989), Steyvers & Busey (2000).
- Primjena:
  Klasifikacija, strojno učenje, genetika, psihometrika, neuroznanost (prepoznavanje: zvuka, slike)...

Izometričko smještavanje[3]

Teorem (Keller, Torgensen (1958))

Zadan je konačni skup točaka $S \subset \mathbb{R}^k$, $n := \#S$ s ℓ_p metrikom. Tada je S izometrički smjestiv u ℓ_p^N za $N = \binom{n}{2}$.

Problem je kako smanjiti dimenziju N.

U praktičnim situacijama opravdano je zahtijevati da smještavanje dozvoljava određenu ϵ-distorziju metrike. To omogućava niže dimenzije smještavanja.

Za općenite metričke prostore (S, d) to nije uvijek moguće, napr. ako je S jedinična sfera u $\mathbb{R}^2$, a d lučna metrika.

[3] embedding

KLASIČNO[4] MDS

$$X_{m\times p} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \vdots \\ x_{m1} & \cdots & x_{mp} \end{bmatrix} \xrightarrow{PCA} Y = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ y_{21} & \cdots & y_{2p} \\ \vdots & & \vdots \\ y_{m1} & \cdots & y_{mp} \end{bmatrix}$$

D (matrica udaljenosti između redaka)

- Što ako je poznato samo D? Može li se rekonstruirati X ili Y?
- Nap. matrica D je invarijantna na translaciju i rotaciju pa X nije nužno jedinstvena.

U **klasičnom MDS** pretpostavlja se da je D matrica euklidskih udaljenosti. Definirajmo: $B = XX^{\tau}$. Tada su B i D vezane, pa je ideja izraziti B preko D i zatim faktorizacijom dobiti X (Keller, 1958).

[4] Gower, 1966

Dakle:

$$b_{ij} = \sum_{k=1}^{m} x_{ik}x_{jk}, \text{ a } d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}. \tag{1}$$

Zbog invarijantnosti na translacije, možemo zamišljati da je X centrirana, tj.

$$\forall k, \sum_{i=1}^{m} x_{ik} = 0.$$

To ima za posljedicu

$$\sum_{j=1}^{m} b_{ij} = \sum_{j=1}^{m}\sum_{k=1}^{p} x_{ik}x_{jk} = \sum_{k=1}^{p} x_{ik} \sum_{j=1}^{m} x_{jk} = 0. \tag{2}$$

Stoga ćemo tražiti B tako da zadovoljava (2).

Zbog (1) i (2) je

$$\bullet \sum_{i=1}^{m} d_{ij}^2 = tr(B) + m * b_{jj} \qquad \bullet \sum_{j=1}^{m} d_{ij}^2 = m * b_{ii} + tr(B)$$

odnosno $$\sum_{j=1}^{m}\sum_{i=1}^{m} d_{ij}^2 = 2m * tr(B).$$

Nadalje, zbog (1)

$$\begin{aligned} b_{ij} &= \frac{1}{2}(b_{ii} + b_{jj} - d_{ij}^2) \\ &= \frac{1}{2}\left(\frac{1}{m}\sum_{i=1}^{m} d_{ij}^2 + \frac{1}{m}\sum_{j=1}^{m} d_{ij}^2 - \frac{1}{m^2}\sum_{j=1}^{m}\sum_{i=1}^{m} d_{ij}^2 - d_{ij}^2\right) \end{aligned}$$

Time smo dobili B. Ostaje izračunati X.

Faktorizacija:

$$B = V\Lambda V^\tau,$$

V ort. matrica svojstvenih vektora, $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ matrica pripadnih svojstvenih vrijednosti, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

Kriterij odabira q-dim aproksimacije (reprezentacije) ovisi o:

$$c_q{}^5 := \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Rješenje jednadžbe $XX^\tau = B$ ja sada $\boldsymbol{X = V\Lambda^{1/2}}$.

Napomena: Ako je polazna matrica X **ranga** $q < p$, onda je i $r(B) = q$, tj. $\lambda_k = 0,\ k > q$, pa je umjesto V dovoljno uzeti **prvih** q stupaca $[v_1\ v_2 \ldots v_q]$, a u matrici Λ **prvih** q sv. vrijednosti.

[5] $c_q \geq 0.8$

NE METRIČKO SKALIRANJE

Što ako B nije simetrična pozitivno definitna? Napr. ako je $D = (d_{ij})$ i d_{ij} nije euklidska udaljenost[6]. Osim toga, D može biti izvedena iz matrice sličnosti $S = (s_{ij})$ na neki od sljedećih načina:

$$d_{ij} = \text{const.} - s_{ij} \qquad d_{ij} = 1/s_{ij}$$
$$d_{ij} = s_{ii} + s_{jj} - 2s_{ij}$$

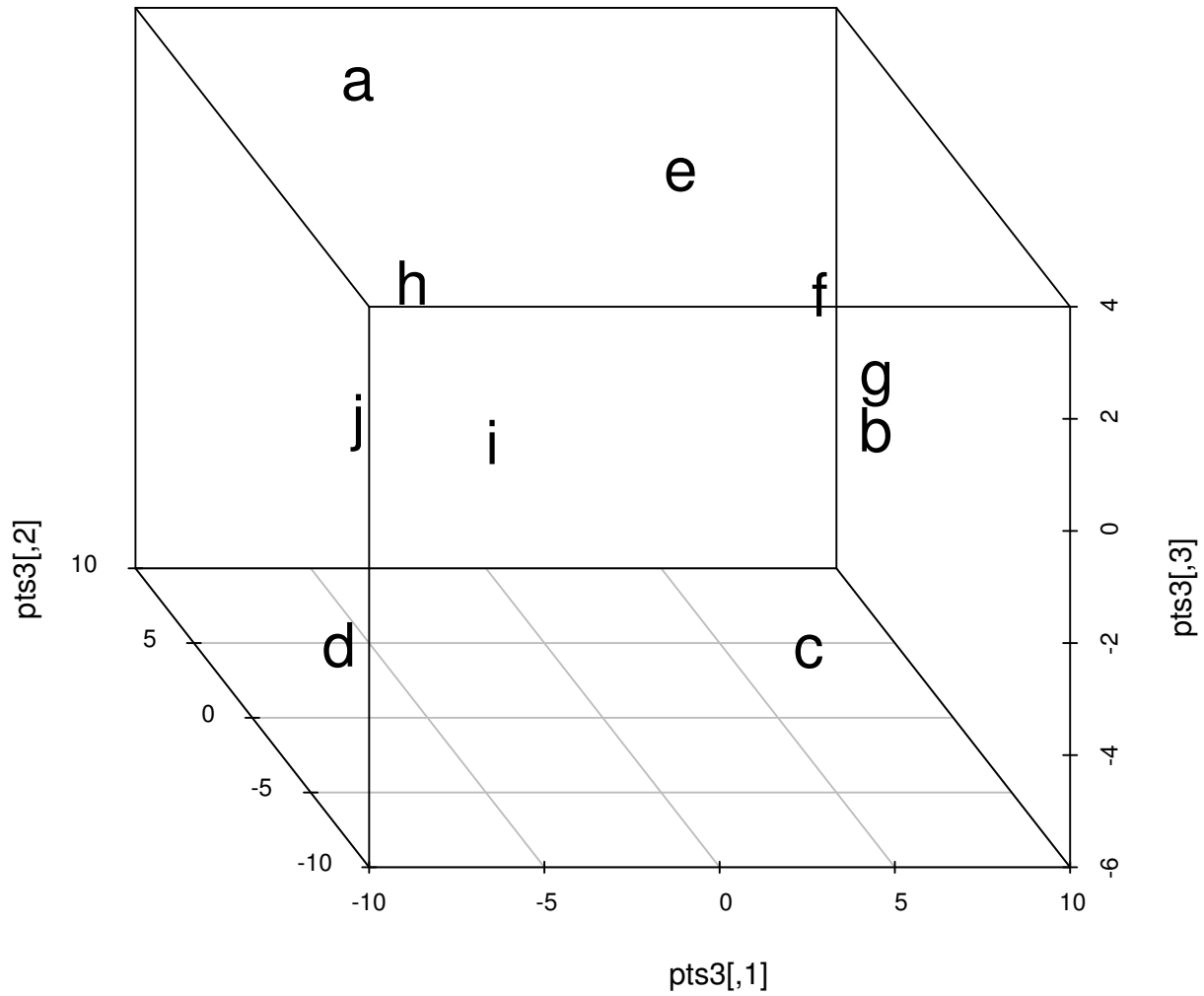
Štoviše, postoje realne situacije kad $d_{ij} \neq d_{ji}$.
Zadana je matrica udaljenosti $D_{m\times m}$. Traži se reprezentacija $X \in \mathbb{R}^{m\times q}$ ($\|x^i - x^j\|^2 =: \hat{d}_{ij}^2$) tako da je

$$\textbf{STRESS} \to S_X := \frac{\sum_{i<j}(\hat{d}_{ij} - d_{ij})^2}{\sum_{i<j} d_{ij}^2}$$

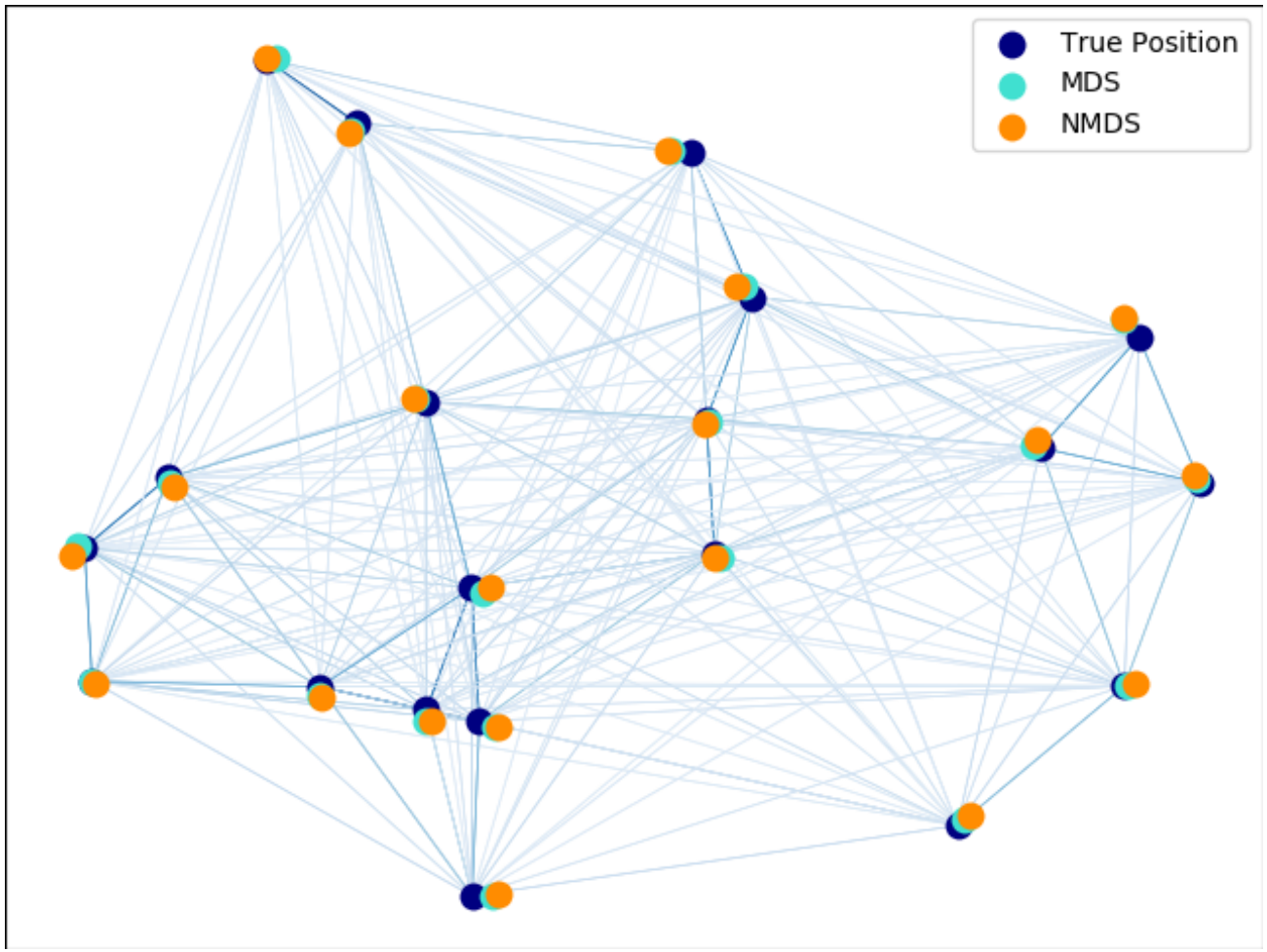
minimalno (alt. $d \leftarrow \varphi(d), \varphi : \mathbb{R} \to \mathbb{R}$ monotona).

[6] eng. *dissimilarity matrix*

Ne metričko skaliranje (rangirani brendovi): `isoMDS(·,k=3)`



Razlika: metrička MDS i nemetrička MDS.



Author: Nelle Varoquaux <nelle.varoquaux@gmail.com>