Odjel za matematiku
Sveučilište J. J. Strossmayera u Osijeku

Matematički praktikum

R. Scitovski, K. Sabo, I. Vazler

1 Reprezentant podataka iz $\mathbb{R}$

Zadani su podaci $y_1, y_2, \ldots, y_m \in \mathbb{R}$.
Treba odrediti realni broj $c^\star \in \mathbb{R}$ (reprezentant podataka) koji će što bolje reprezentirati podatke.

Definicija 1. Funkciju $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, koja ima svojstvo *pozitivne definitnosti*

$$\forall x, y \in \mathbb{R} \quad d(x,y) \geq 0 \quad \& \quad d(x,y) = 0 \quad \Leftrightarrow \quad x = y,$$

zovemo *kvazimetrička funkcija* (funkcija sličnosti, funkcija različitosti)

Primjer 1. *Dva najčešća primjera:*
(a) $d_{LS}(x,y) = (x-y)^2$ – *Least Squares (LS) kvazimetrička funkcija*
(b) $d_1(x,y) = |x-y|$ – l_1 *metrička funkcija (Manhattan metrika)*
(c) Primijetite da u $\mathbb{R}$ *vrijedi* $d_1(x,y) = d_2(x,y) = d_\infty(x,y) = d_p(x,y), \quad p \geq 1$

Zadatak 1. *Pokažite da funkcija* d_{LS} *iz prethodnog primjera nije metrika na* $\mathbb{R}$*, a da su funkcije* d_1, d_2, d_∞ *metrike na* $\mathbb{R}$*.*

Definicija 2. Neka je $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ kvazimetrička funkcija. Kažemo da je $c^\star \in \mathbb{R}$ *najbolji reprezentant* podataka $y_1, y_2, \ldots, y_m \in \mathbb{R}$ u odnosu na kvazimetričku funkciju d onda ako je

$$c^\star = \operatorname*{argmin}_{c \in \mathbb{R}} \sum_{i=1}^{m} d(c, y_i), \tag{1}$$

tj ako je $c^\star \in \mathbb{R}$ točka globalnog minimuma funkcionala $F\colon \mathbb{R} \to \mathbb{R}_+$

$$F(c) = \sum_{i=1}^{m} d(c, y_i). \tag{2}$$

Primjer 2. *Za LS-kvazimetričku funkciju najbolji reprezentant podataka* $y_1, y_2, \ldots, y_m \in \mathbb{R}$ *je obična* aritmetička sredina

$$c^\star_{LS} = \operatorname*{argmin}_{c \in \mathbb{R}} \sum_{i=1}^{m} d_{LS}(c, y_i) = \tfrac{1}{m} \sum_{i=1}^{m} y_i,$$

a odgovarajući funkcional glasi

$$F_{LS}(c) = \sum_{i=1}^{m} (y_i - c)^2.$$

Za l_1-kvazimetričku funkciju najbolji reprezentant podataka $y_1, y_2, \ldots, y_m \in \mathbb{R}$ je običnI medijan

$$c_1^\star = \operatorname*{argmin}_{c\in\mathbb{R}} \sum_{i=1}^{m} d_1(c, y_i) = \operatorname*{med}_i y_i,$$

a odgovarajući funkcional glasi

$$F_1(c) = \sum_{i=1}^{m} |y_i - c|.$$

Zadatak 2. *Dokažite tvrdnje iz Primjera 2 i nacrtajte grafove funkcija F_{LS} i F_1.*
Uputa: vidi (Sabo and Scitovski, 2008; Scitovski, 2004)

Definicija 3. Neka je $d\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ kvazimetrička funkcija. Kažemo da je $c^\star \in \mathbb{R}$ *najbolji reprezentant* podataka $y_1, y_2, \ldots, y_m \in \mathbb{R}$ s težinama $w_1, \ldots, w_m > 0$ u odnosu na kvazimetričku funkciju d onda ako je

$$c^\star = \operatorname*{argmin}_{c\in\mathbb{R}} \sum_{i=1}^{m} w_i d(c, y_i), \tag{3}$$

tj ako je $c^\star \in \mathbb{R}$ točka globalnog minimuma funkcionala $F\colon \mathbb{R} \to \mathbb{R}_+$

$$F(c) = \sum_{i=1}^{m} w_i d(c, y_i). \tag{4}$$

Primjer 3. *Za LS-kvazimetričku funkciju najbolji reprezentant podataka $y_1, y_2, \ldots, y_m \in \mathbb{R}$ s težinama $w_1, \ldots, w_m > 0$ je* težinska aritmetička sredina

$$c_{LS}^\star = \operatorname*{argmin}_{c\in\mathbb{R}} \sum_{i=1}^{m} w_i d_{LS}(c, y_i) = \tfrac{1}{W} \sum_{i=1}^{m} w_i y_i, \quad W = \sum_{i=1}^{m} w_i$$

a odgovarajući funkcional glasi

$$F_{LS}(c) = \sum_{i=1}^{m} w_i (y_i - c)^2.$$

Za l_1-kvazimetričku funkciju najbolji reprezentant podataka $y_1, y_2, \ldots, y_m \in \mathbb{R}$ s težinama $w_1, \ldots, w_m > 0$ je težinski medijan

$$c_1^\star = \operatorname*{argmin}_{c\in\mathbb{R}} \sum_{i=1}^{m} w_i d_1(c, y_i) = \operatorname*{med}_i(w_i, y_i),$$

a odgovarajući funkcional glasi

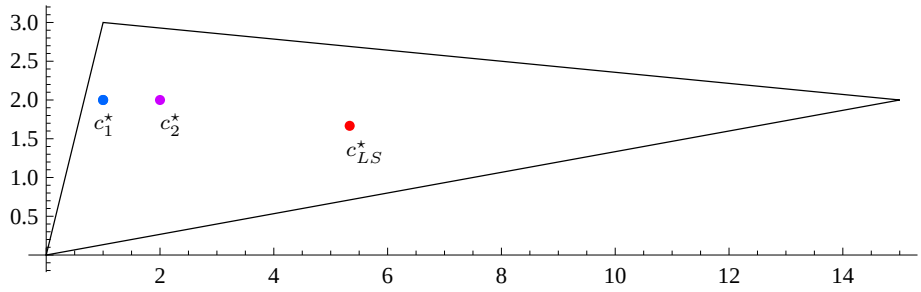
$$F_1(c) = \sum_{i=1}^{m} w_i |y_i - c|.$$

Zadatak 3. *Dokažite tvrdnje iz Primjera 3 za neki slučajno izabran skup podataka* $y_1, y_2, \ldots, y_m \in \mathbb{R}$ *s težinama* $w_1, \ldots, w_m > 0$. *Nacrtajte grafove funkcija* F_{LS} *i* F_1.
Uputa: vidi Sabo and Scitovski (2008)

2 Reprezentant podataka iz $\mathbb{R}^2$

Neka su $A_1, A_2, A_3 \in \mathbb{R}^2$ tri nekolinearne točke u ravnini. Točka $c^\star_{LS} \in \mathbb{R}^2$, za koju je suma kvadrata euklidskih udaljenosti do vrhova trokuta minimalna zove se **centroid** ili **Steinerova točka** (povezano s pojmom centra masa u fizici). Naka je

$$A_1 = (x_1, y_1), \quad A_2 = (x_2, y_2), \quad A_3 = (x_3, y_3).$$



Slika 1: Centroid ($c^\star_{LS}$), Medijan ($c^\star_1$) i Geometrijski medijan ($c^\star_2$)

Tada je centroid $c^\star_{LS} = (x_c, y_c)$ točka za koju se postiže minimum

$$\sum_{i=1}^{3} d_2^2(T, A_i) \to \min_{T \in \mathbb{R}^2}, \qquad tj. \quad \sum_{i=1}^{3} [(x - x_i)^2 + (y - y_i)^2] \to \min_{(x,y) \in \mathbb{R}^2}$$

$$\text{odakle dobivamo:} \quad x_c = \tfrac{1}{3} \sum_{i=1}^{3} x_i, \qquad y_c = \tfrac{1}{3} \sum_{i=1}^{3} y_i.$$

Točka $c^\star_1$ za koju se postiže minimum

$$\sum_{i=1}^{3} d_1(T, A_i) \to \min_{T \in \mathbb{R}^2}, \qquad tj. \quad \sum_{i=1}^{3} (|x - x_i| + |y - y_i|) \to \min_{(x,y) \in \mathbb{R}^2}$$

naziva se **medijan** točaka $A_1, A_2, A_3 \in \mathbb{R}^2$. Lako dobivamo

$$c^\star_1 = (\operatorname{med} x_i, \operatorname{med} y_i)\,.$$

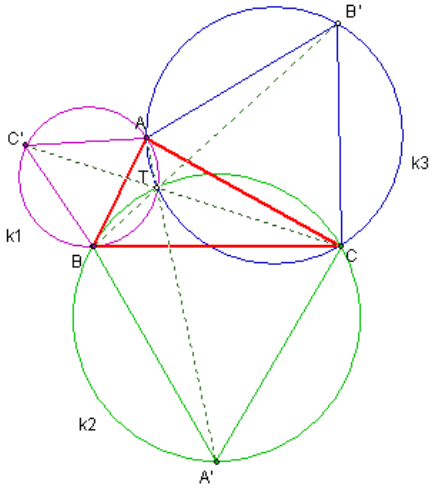
Točka $c^\star_2$ za koju se postiže minimum

$$\sum_{i=1}^{3} d_2(T, A_i) \to \min_{T \in \mathbb{R}^2}, \qquad tj. \quad \sum_{i=1}^{3} \sqrt{(x - x_i)^2 + (y - y_i)^2} \to \min_{(x,y) \in \mathbb{R}^2}$$

naziva se **geometrijski medijan** točaka $A_1, A_2, A_3 \in \mathbb{R}^2$ i ne može se eksplicitno izraziti.

Geometrijski medijan $c^\star_2$ može se dobiti geometrijskim konstrukcijama na dva načina:

- kao sjecište Torricellijevih kružnica;
- kao sjecište Simpsonovih pravaca.



Slika 2: Sliku izradile: D. Jankov, S. Sušić

2.1 Kvazimetričke funkcije i reprezentanti

Zadan je skup točaka $A = \{a^i = (x_i, y_i) \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$, odnosno
vektora $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$.
Treba odrediti točku $C^\star$ (odnosno vektor $c^\star$) koja će što bolje reprezentirati skup točaka A (odnosno skup vektora $\mathcal{A}$).

Definicija 4. Funkciju $d \colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$, koja ima svojstvo *pozitivne definitnosti*

$$\forall x, y \in \mathbb{R}^2 \quad d(x,y) \geq 0 \quad \& \quad d(x,y) = 0 \quad \Leftrightarrow \quad x = y,$$

zovemo *kvazimetrička funkcija* (funkcija sličnosti, funkcija različitosti) na $\mathbb{R}^2$.

Primjer 4. *Najčešći primjeri (Gan et al., 2007; Kogan, 2007; Späth, 1983):*
(a) $d_{LS}(x,y) = \|x-y\|_2^2 = (x-y)^T(x-y)$ – *Least Squares (LS) kvazimetrička funkcija*
(b) $d_2(x,y) = \|x-y\|_2 = \sqrt{(x-y)^T(x-y)}$ – l_2 *euklidska metrička funkcija*
(c) $d_1(x,y) = \|x-y\|_1$ – l_1 *metrička funkcija (Manhattan metrika)*
(d) $d_\infty(x,y) = \|x-y\|_\infty$ – l_∞ *Čebiševljeva metrička funkcija*
(e) $d_M(x,y) = (x-y)^T S^{-1}(x-y)$ – *Mahalanobis kvazimetrička funkcija*
($S \in \mathbb{R}^{2\times 2}$ *je simetrična pozitivno definitna matrica)*

Zadatak 4. *Pokažite da su sve funkcije iz prethodnog primjera kvazimetričke funkcije na* $\mathbb{R}^2$*, da funkcije* d_{LS} *i* d_M *iz prethodnog primjera nisu metrike na* $\mathbb{R}^2$*, a da su funkcije* d_1, d_2, d_∞ *metrike na* $\mathbb{R}^2$*.*

Zadatak 5. *Neka je* $S \in \mathbb{R}^{2\times 2}$ *je simetrična pozitivno definitna matrica. Pokažite da je* $d_\mu : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$*,*

$$d_\mu(x, y) = \sqrt{d_M(x, y)} = \sqrt{(x - y)^T S^{-1}(x - y)}$$

metrika na $\mathbb{R}^2$*. Udaljenost* $d_\mu(x, y)$ *u literaturi se može naći pod nazivom* Mahalanobis udaljenost.

Definicija 5. Neka je $d\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$ kvazimetrička funkcija. Kažemo da je $c^\star \in \mathbb{R}^2$ *najbolji reprezentant* (centar) skupa $\mathcal{A}$ u odnosu na kvazimetričku funkciju d onda ako je

$$c^\star = \operatorname*{argmin}_{c\in\mathbb{R}^2} \sum_{i=1}^{m} d(c, a^i), \tag{5}$$

tj. ako je $c^\star \in \mathbb{R}^2$ točka globalnog minimuma funkcionala $F\colon \mathbb{R}^2 \to \mathbb{R}_+$

$$F(c) = \sum_{i=1}^{m} d(c, a^i). \tag{6}$$

Zadatak 6. *[20 bodova]*

(a) Ako su $f_1, \ldots, f_m\colon \mathbb{R} \to \mathbb{R}$ *konveksne funkcije i* $w_1, \ldots, w_m > 0$*, pokažite da je tada* $\sum_{i=1}^{m} w_i f_i\colon \mathbb{R} \to \mathbb{R}$ *konveksna funkcija.*

(b) Vrijedi li obrat prethodne tvrdnje ? Ako je $f = f_1 + f_2\colon \mathbb{R} \to \mathbb{R}$ *konveksna funkcija, moraju li onda i funkcije* f_1, f_2 *biti konveksne ?*

(c) Neka su $y_1, \ldots, y_m \in \mathbb{R}$ *podaci s težinama* $w_1, \ldots, w_m > 0$*. Pokažite da su tada sljedeće funkcije konveksne*

$$(i)\, f(x) = \sum_{i=1}^{m} w_i (y_i - x)^2; \quad (ii)\, g(x) = \sum_{i=1}^{m} w_i |x - y_i|; \quad (iii)\, h(x) = \max_{i=1,\ldots,m} w_i |x - y_i|.$$

Koristći prethodni zadataka može se pokazati da vrijedi:

(a) Za LS-kvazimetričku funkciju najbolji reprezentant skupa $\mathcal{A}$ je *centroid (težište) skupa*

$$c^\star_{LS} = \operatorname*{argmin}_{c\in\mathbb{R}^2} \sum_{i=1}^{m} d_{LS}(c, a^i) = \tfrac{1}{m} \sum_{i=1}^{m} a^i = \left(\tfrac{1}{m} \sum_{i=1}^{m} x_i, \tfrac{1}{m} \sum_{i=1}^{m} y_i \right),$$

a odgovarajući funkcional glasi

$$F_{LS}(c) = \sum_{i=1}^{m} \|c - a^i\|_2^2.$$

(b) Za l_1-kvazimetričku funkciju najbolji reprezentant skupa $\mathcal{A}$ je *medijan skupa*

$$c_1^\star = \operatorname*{argmin}_{c\in\mathbb{R}^2} \sum_{i=1}^m d_1(c, a^i) = \operatorname*{med}_i a^i = \left(\operatorname*{med}_i x_i, \operatorname*{med}_i y_i\right),$$

a odgovarajući funkcional glasi

$$F_1(c) = \sum_{i=1}^m \|c - a^i\|_1.$$

Zadatak 7. *Dokažite prethodno navedene tvrdnje, nacrtajte* `ContourPlot` *funkcija F_{LS} i F_1 i napišite minimizirajuće funkcionale za preostale kvazimetričke/metričke funkcije.*

Definicija 6. Neka je $d\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$ kvazimetrička funkcija. Kažemo da je $c^\star \in \mathbb{R}^2$ *najbolji reprezentant* skupa $\mathcal{A}$ s težinama $w_1, \ldots, w_m > 0$ u odnosu na kvazimetričku funkciju d onda ako je

$$c^\star = \operatorname*{argmin}_{c\in\mathbb{R}^2} \sum_{i=1}^m w_i d(c, a^i), \tag{7}$$

tj. ako je $c^\star \in \mathbb{R}^2$ točka globalnog minimuma funkcionala $F\colon \mathbb{R}^2 \to \mathbb{R}_+$

$$F(c) = \sum_{i=1}^m w_i d(c, a^i). \tag{8}$$

Vrijedi:

(a) Za LS-kvazimetričku funkciju najbolji reprezentant skupa $\mathcal{A}$ s težinama $w_1, \ldots, w_m > 0$ je *centroid (težište) skupa*

$$c_{LS}^\star = \operatorname*{argmin}_{c\in\mathbb{R}^2} \sum_{i=1}^m w_i d_{LS}(c, a^i) = \tfrac{1}{W} \sum_{i=1}^m w_i a^i, \quad W = \sum_{i=1}^m w_i, \quad \text{tj.}$$
$$c_{LS}^\star = \left(\tfrac{1}{W} \sum_{i=1}^m w_i x_i, \tfrac{1}{W} \sum_{i=1}^m w_i y_i\right),$$

a odgovarajući funkcional glasi

$$F_{LS}(c) = \sum_{i=1}^m w_i \|c - a^i\|_2^2.$$

(b) Za l_1-kvazimetričku funkciju najbolji reprezentant skupa $\mathcal{A}$ je *medijan skupa*

$$c_1^\star = \operatorname*{argmin}_{c\in\mathbb{R}^2} \sum_{i=1}^m w_i d_1(c, a^i) = \operatorname*{med}_i(w_i, a^i) = \left(\operatorname*{med}_i(w_i, x_i), \operatorname*{med}_i(w_i, y_i)\right),$$

a odgovarajući funkcional glasi

$$F_1(c) = \sum_{i=1}^m w_i \|c - a^i\|_1.$$

2.2 Reprezentant podataka iz $\mathbb{R}^n$

Zadan je skup točaka $A = \{a^i \in \mathbb{R}^n \colon i = 1, \ldots, m\}$, odnosno
vektora $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i)^T \in \mathbb{R}^n \colon i = 1, \ldots, m\}$ s težinama $w_1, \ldots, w_m > 0$.

Najbolji LS-reprezentant skupa $\mathcal{A}$ je **težinski centroid** skupa vektora $\mathcal{A} \subset \mathbb{R}^n$

$$c_{LS}^\star = \operatorname*{argmin}_{c \in \mathbb{R}^n} \sum_{i=1}^m w_i \|c - a^i\|_2^2 = \left(\tfrac{1}{m} \sum_{i=1}^m w_i a_1^i, \ldots, \tfrac{1}{m} \sum_{i=1}^m w_i a_n^i \right),$$

jer se na njemu postiže globalni minimum funkcionala

$$F_{LS}(c) = \sum_{i=1}^m w_i \|c - a^i\|_2^2.$$

Najbolji l_1-reprezentant skupa $\mathcal{A}$ je **težinski median** skupa vektora $\mathcal{A} \subset \mathbb{R}^n$

$$c_1^\star = \operatorname*{argmin}_{c \in \mathbb{R}^n} \sum_{i=1}^m w_i \|c - a^i\|_1 = \left(\operatorname*{med}_{j=1,\ldots,m} (w_j, a_1^j), \ldots, \operatorname*{med}_{j=1,\ldots,m} (w_j, a_n^j) \right)^T \tag{9}$$

jer se na njemu postiže globalni minimum funkcionala

$$F_1(c) = \sum_{i=1}^m w_i \|c - a^i\|_1.$$

Naime, vrijedi

$$\begin{aligned}
\min_{x \in \mathbb{R}^n} \sum_{i=1}^m w_i \|x - a^i\|_1 &= \min_{x \in \mathbb{R}^n} \sum_{i=1}^m w_i \left(\sum_{k=1}^n |x_k - a_k^i| \right) \\
&= \min_{x \in \mathbb{R}^n} \sum_{k=1}^n \left(\sum_{i=1}^m w_i |x_k - a_k^i| \right) = \sum_{k=1}^n \min_{x_k \in \mathbb{R}} \sum_{i=1}^m w_i |x_k - a_k^i| \\
&= \sum_{k=1}^n \sum_{i=1}^m w_i | \operatorname*{med}_{j=1,\ldots,m} (w_j, a_k^j) - a_k^i| = \sum_{i=1}^m \sum_{k=1}^n w_i | \operatorname*{med}_{j=1,\ldots,m} (w_j, a_k^j) - a_k^i| \\
&= \sum_{i=1}^m w_i \|c^* - a^i\|_1.
\end{aligned}$$

3 Mahalanobis kvazimetrička funkcija

Primjenu Mahalanobis kvazimetričke funkcije razmotrit ćemo najprije u ravni. Zadan je skup skup točaka $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^n \colon i = 1, \ldots, m\}$, odnosno skup vektora $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$, s odgovarajućim težinama $w_1, \ldots, w_m > 0$. Možemo shvaćati da su x_i vrijednosti nezavisne varijable, a y_i odgovarajuće vrijednosti zavisne varijable.

3.1 Pravac u ravnini kao reprezentant podataka

Za dani skup podataka $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 \colon i = 1, \dots m\}$ s težinama $w_1, \dots, w_m > 0$ treba odrediti afinu funkciju

$$\mathcal{L}(x) = \alpha x + \beta$$

čiji graf prolazi što bliže točkama $a^i = (x_i, y_i)$, $i = 1, \dots, m$. Traženje ovog pravca možemo shvatiti kao traženje najboljeg reprezentanta podataka u formi pravca.

Ako su pogreške u izmjerenim vrijednostima nezavisne varijable zanemarive, a pogreške u izmjerenim vrijednostima zavisne varijable normalno distribuirane nezavisne slučajne varijable s očekivanjem 0, onda parametre α, β afine funkcije možemo odrediti u smislu običnik najmanjih kvadrata (OLS) minimizirajući funkcional

$$S(\alpha, \beta) = \sum_{i=1}^{m} w_i (y_i - \alpha x_i - \beta)^2.$$

Minimizacija funkcionala S je jednostavni linearni LS problem.

Ako se značajne pogreške mogu očekivati u izmjerenim vrijednostima nezavisne varijable i u izmjerenim vrijednostima zavisne varijable, onda parametre α, β afine funkcije određujemo u smislu potpunih najmanjih kvadrata (total least squares – TLS) minimizirajući funkcional

$$T(\alpha, \beta, \delta) = \sum_{i=1}^{m} w_i \left[(y_i - \alpha(x_i + \delta_i) - \beta)^2 + \delta_i^2 \right], \quad \delta = (\delta_1, \dots, \delta_m)^T \in \mathbb{R}^m,$$

(vidi (Jukić et al., 1999, KOI 1998)). Minimizacija funkcionala T je nelinearni minimizacijski problem s $m + 2$ varijable.

U geometrijskom smislu kod TLS pristupa minimizira se suma kvadrata ortogonalnih udaljenosti točaka podataka do pravca. Budući da se u našem slučaju radi o procjeni parametara najjednostavnije linearne model-funkcije, traženje odgovarajućeg najboljeg TLS-pravca može se značajno pojednostaviti.

Općenito, traženi pravac zadan je u implicitnom obliku

$$ux + vy + c = 0. \tag{10}$$

Kvadrat udaljenosti točke $T_i = (x_i, y_i)$ do pravca (10) zadana je formulom[1]

$$d_i^2 = \frac{(ux_i + vy_i + c)^2}{u^2 + v^2}, \tag{11}$$

[1] Ako je jednadzba pravca dana u eksplicitnom obliku $y = kx + l$, onda je udaljenost tocke $a^i = (x_i, y_i)$ do pravca dana formulom:

$$d_i = \frac{|y_i - \alpha x_i - \beta|}{\sqrt{\alpha^2 + 1}}$$

Primjedba 1. Uvjek mozemo pretpostaviti da za parametre u, v pravca (10) vrijedi $u^2 + v^2 = 1$. Naime, kako u i v ne mogu istovremeno iščezavati, množeći jednadžbu (10) s $1/\sqrt{u^2+v^2}$, dobivamo

$$\hat{u}x + \hat{v}y + \hat{c} = 0, \qquad \text{gdje je } \hat{u} = \tfrac{u}{\sqrt{u^2+v^2}}, \quad \hat{v} = \tfrac{v}{\sqrt{u^2+v^2}}, \quad \hat{c} = \tfrac{c}{\sqrt{u^2+v^2}}, \qquad \hat{u}^2 + \hat{v}^2 = 1.$$

Zato formula za kvadrat udaljenosti tocke T_i do pravca (10) s uvjetom $u^2 + v^2 = 1$ postaje jednostavnija

$$d_i^2 = (ux_i + vy_i + c)^2, \tag{12}$$

a umjesto minimizacije funkcionala T možemo minimizirati funkcional

$$G(u, v, c) = \sum_{i=1}^{m} w_i d_i^2 = \sum_{i=1}^{m} w_i (ux_i + vy_i + c)^2 \quad \text{uz uvjet} \quad u^2 + v^2 = 1. \tag{13}$$

Lema 1. *Zadani su podaci $a^i = (x_i, y_i)$, $i = 1, \dots m$ s težinama $w_i > 0$. Najbolji TLS pravac prolazi centroidom podataka $\bar{c} = (\bar{x}, \bar{y})$, gdje je*

$$\bar{x} = \frac{1}{W}\sum_{i=1}^{m} w_i x_i, \qquad \bar{y} = \frac{1}{W}\sum_{i=1}^{m} w_i y_i, \qquad W = \sum_{i=1}^{m} w_i.$$

Dokaz. Uočimo najprije da pravac zadan jednadžbom

$$ux + vy + c = 0, \qquad u^2 + v^2 = 1,$$

prolazi centroidom podataka $\bar{c} = (\bar{x}, \bar{y})$, onda ako njegova jednadžba glasi

$$u(x - \bar{x}) + v(y - \bar{y}) = 0, \qquad u^2 + v^2 = 1. \tag{14}$$

Kako je

$$\begin{aligned} G(u, v, c) &= \sum_{i=1}^{m} w_i (ux_i + vy_i - (-c))^2 \\ &\geq \sum_{i=1}^{m} w_i (ux_i + vy_i - (u\bar{x} + v\bar{y}))^2 = \sum_{i=1}^{m} w_i \left(u(x_i - \bar{x}) + v(y_i - \bar{y})\right)^2, \end{aligned}$$

što znači da između svih pravaca $ux + vy + c = 0, \quad u^2 + v^2 = 1$, pravac koji prolazi centroidom podataka ima najmanju moguću težinsku sumu kvadrata ortogonalnih odstupanja. □

Sukladno *Lemi 1* najbolji TLS-pravac tražit ćemo u obliku

$$u(x - \bar{x}) + v(y - \bar{y}) = 0, \qquad u^2 + v^2 = 1. \tag{15}$$

minimizirajući funkcional

$$F(u,v) = \sum_{i=1}^{m} w_i \left[u(x_i - \bar{x}) + v(y_i - \bar{y})\right]^2, \quad \text{uz uvjet} \quad u^2 + v^2 = 1. \tag{16}$$

Uz oznake

$$B := \begin{bmatrix} (x_1 - \bar{x}) & (y_1 - \bar{y}) \\ \vdots & \vdots \\ (x_m - \bar{x}) & (y_m - \bar{y}) \end{bmatrix}, \quad D = \operatorname{diag}(w_1, \ldots, w_m), \quad t = \begin{bmatrix} u \\ v \end{bmatrix}.$$

funkcional F može se zapisati u obliku

$$F(u,v) = \|\sqrt{D}\, Bt\|_2^2, \quad \|t\| = 1.$$

Sukladno Nievergelt (1994), vrijedi sljedeći teorem. Potrebni korišteni pojmovi mogu se pronaći kod Truhar (2010).

Teorem 1. *Funkcional F zadan s (16) postiže svoj minimum na svakom jediničnom vektoru $t = (u,v)^T$ koji odgovara najmanjoj (manjoj) svojstvenoj vrijednosti simetrične pozitivno definitne matrice*

$$B^T DB = \begin{bmatrix} \sum_{i=1}^{m} w_i (x_i - \bar{x})^2 & \sum_{i=1}^{m} w_i (x_i - \bar{x})(y_i - \bar{y}) \\ \sum_{i=1}^{m} w_i (x_i - \bar{x})(y_i - \bar{y}) & \sum_{i=1}^{m} w_i (y_i - \bar{y})^2 \end{bmatrix} \tag{17}$$

Zadatak 8. *Dokažite Teorem 1.*

Primjer 5. *Zadani su podaci*

w_i	1	1	1	1	1
x_i	1	2	3	4	5
y_i	1	3	4	2	3

Centroid podataka je u točki $\bar{c} = (3, \frac{13}{5})$, a matrice B, D i $B^T DB$ u ovom slučaju su

$$B = \begin{bmatrix} -2 & -8/5 \\ -1 & 2/5 \\ 0 & 7/5 \\ 1 & -3/5 \\ 2 & 2/5 \end{bmatrix}, \qquad D = \operatorname{diag}(1,1,1,1,1), \qquad B^T DB = \begin{bmatrix} 10 & 3 \\ 3 & 26/5 \end{bmatrix}.$$

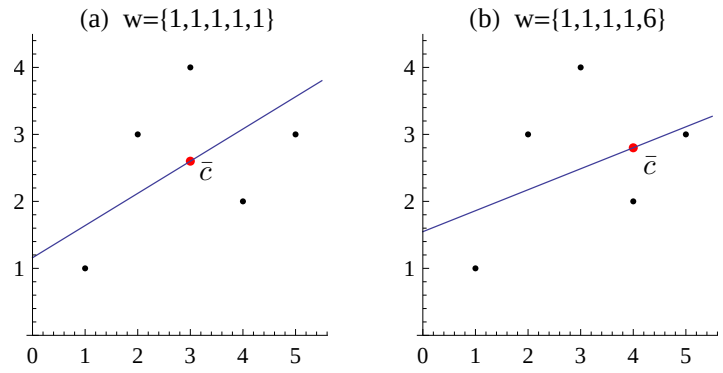
Svojstvene vrijednosti matrice $B^T DB$ su $\lambda_1 = 11.4419$, $\lambda_2 = 3.75813$, a jedinični svojstveni vektor koji odgovara manjoj svojstvenoj vrijednosti je $t = (-0.433189, 0.901303)^T$. Zato jednadžba najboljeg TLS-pravca glasi (vidi Sliku 3a)

$$-0.433189\,(x-3) + 0.901303\,(y - \frac{13}{5}) = 0, \quad \text{odnosno} \quad y = 0.480625x + 1.15813.$$

Ako tezinu $w_5 = 1$ promijenimo u $w_5 = 6$, centroid podataka postaje točka $\bar{c} = (4, \frac{14}{5})$, a matrice B, D i $B^T DB$ u ovom slučaju su

$$B = \begin{bmatrix} -3 & -9/5 \\ -2 & 1/5 \\ -1 & 6/5 \\ 0 & -4/5 \\ 1 & 1/5 \end{bmatrix}, \qquad D = \mathrm{diag}(1,1,1,1,6), \qquad B^T DB = \begin{bmatrix} 20 & 5 \\ 5 & 28/5 \end{bmatrix}.$$

Svojstvene vrijednosti matrice $B^T DB$ su $\lambda_1 = 21.5658, \quad \lambda_2 = 4.03416$, a jedinični svoj-



Slika 3: Najbolji TLS-pravac

stveni vektor koji odgovara manjoj svojstvenoj vrijednosti je $t = (-0.298856, 0.954298)^T$. Zato jednadžba najboljeg TLS-pravca glasi (vidi Sliku 3b)

$$-0.298856\,(x-4) + 0.954298\,(y - \frac{14}{5}) = 0, \quad \text{odnosno} \quad y = 0.313169x + 1.54733.$$

Primjedba 2. Matrica $\frac{1}{W} B^T DB$ je uz neke uvjete na podatke (vidi Lemu 2) pozitivno definitna simetrična matrica. Njezine svojstvene vrijednosti su realni brojevi, a odgovarajući svojstveni vektori međusobno okomiti. U smjeru svojstvenog vektora, koji odgovara većoj svojstvenoj vrijednosti usmjeren je i najbolji TLS-pravac. Pravac okomit na najbolji TLS-pravac ima smjer svojstvenog vektora, koji odgovara manjoj svojstvenoj vrijednosti ove matrice.

U statističkoj literaturi ova matrica naziva se *kovarijacijska matrica* (en.: covariance matrix) slučajnih varijabli x, y, a smjerovi svojstvenih vektora nazivaju se glavni smjerovi (en.: principal components). U smjeru svojstvenog vektora, koji odgovara većoj svojstvenoj vrijednosti varijanca podataka je veća, a u smjeru svojstvenog vektora, koji odgovara manjoj svojstvenoj vrijednosti varijanca podataka je manja.

3.2 Mahalanobis kvazimetrička funkcija u ravnini

Neka je $\mathcal{S}\colon X_0 \to X_0$ linearni operator kontrakcije/dilatacije kome u bazi $e = \{e_1, e_2\}$ pripada dijagonalna matrica

$$S(e) = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}, \quad \alpha, \beta > 0.$$

Operator $\mathcal{S}$ jediničnu kružnicu (uz primjenu l_2-udaljenosti) sa središtem u ishodištu

$$K = \{x \in \mathbb{R}^2 \colon \|x\|_2^2 = 1\} = \{(x_1, x_2) \in \mathbb{R}^2 \colon x_1^2 + x_2^2 = 1\},$$

preslikava na elipsu

$$\mathcal{S}(K) = \{\xi = (\xi_1, \xi_2) \in \mathbb{R}^2 \colon \tfrac{\xi_1^2}{\alpha^2} + \tfrac{\xi_2^2}{\beta^2} = 1\},$$

jer je

$$\mathcal{S}(x_1 e_1 + x_2 e_2) = \alpha x_1 e_1 + \beta x_2 e_2 =: \xi_1 e_1 + \xi_2 e_2.$$

Primjer 6. *Linearni operator $\mathcal{S}\colon X_0 \to X_0$ u bazi $e = \{e_1, e_2\}$ zadan je matricom $S(e) = \begin{bmatrix} 3 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$. On ravninu (x, y) u smjeru baznog vektora e_1 produljuje 3 puta, a u smjeru baznog vektora e_2 skraćuje na pola. Jediničnu kružnicu K sa središtem u ishodištu transformira u elipsu s glavnom poluosi duljine 3 u smjeru baznog vektora e_1 i sporedne poluosi duljine $\frac{1}{2}$ u smjeru baznog vektora e_2 (vidi Sliku 4).*

Treba definirati takvu kvazimetričku funkciju $d_M\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$ primjenom koje će točke na elipsi $\mathcal{S}(K)$ biti jednako udaljene od ishodišta O, odnosno primjenom koje će elpsa $\mathcal{S}(K)$ postati jedinična kružnica u prostoru snabdjevenom kvazimetričkom funkcijom d_M. To ćemo postići tako da udaljenosti u smjeru svojstvenih vektora uzimamo obrnuto proporcionalno veličini odgovarajuće svojstvene vrijednosti operatora $\mathcal{S}$. Upravo takvo djelovanje ima inverzni linearni operator $\mathcal{S}^{-1}$.

Zato možemo definirati novu normu $\|\cdot\|_\mu\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$,

$$\|x\|_\mu := \|S^{-1/2}x\|_2 = \sqrt{x^T S^{-1} x}. \tag{18}$$

u kojoj će elipsa $\mathcal{S}(K)$ postati kružnica

$$K_\mu = \{x \in \mathbb{R}^2 \colon \|x\|_\mu = 1\}.$$

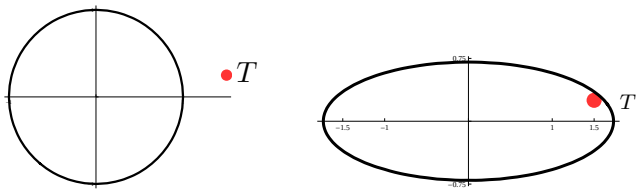
Odgovarajuću kvazimetričku funkciju

$$d_M(x, y) := \|x - y\|_\mu^2 = \|S^{-1/2}x\|_2^2 = (x - y)^T S^{-1}(x - y), \quad x, y \in \mathbb{R}^2, \tag{19}$$

nazivamo *Mahalanobis kvazimetrička funkcija*, a udaljenost $d_M(x, y)$ zvat ćemo M-udaljenost točaka $x, y \in \mathbb{R}^2$.

Primjer 7. *LS-udaljenost točke* $T = (1.5, 0.25)$ *do ishodišta je* $d_{LS}(T, O) = 2.3125$*, a M-udaljenost definirana s* (19) *i matricom* S *iz Primjera 6 iznosi* $d_M(T, O) = 0.875$*.*

(a) LS-udaljenost (b) M-udaljenost



Slika 4: Udaljenost točke $T = (1.5, 0.25)$ do ishodišta O

3.2.1 Mahalanobis udaljenost inducirana skupom točaka podataka u $\mathbb{R}^2$

Zadan je skup $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$ podataka/točaka s težinama $w_1, \ldots, w_m > 0$. *Kako pronaći linearni operator kontrakcije/dilatacije koji će "otkriti" smjerove izduženja ?*

1. Odrediti centroid
$$\bar{c} = (\bar{x}, \bar{y}), \qquad \bar{x} = \tfrac{1}{W} \sum_{i=1}^{m} w_i x_i, \quad \bar{y} = \tfrac{1}{W} \sum_{i=1}^{m} w_i y_i, \quad W = \sum_{i=1}^{m} w_i.$$

2. Prema Teoremu 1 smjer TLS-pravca, a onda i glavnog smjera podataka (first principal component) je u smjeru svojstvenog vektora koji pripada većoj svojstvene vrijednosti matrice (17). Sporedni smjer (second principal component) uzima se okomito na prvi, dakle, u smjeru svojstvenog vektora koji odgovara manjoj svojstvenoj vrijednosti iste matrice. Zbog toga glavne smjerove tražit ćemo u smjeru svojstvenih vektora *kovarijacijske matrice*
$$S = \begin{bmatrix} \frac{1}{W} \sum_{i=1}^{m} w_i (x_i - \bar{x})^2 & \frac{1}{W} \sum_{i=1}^{m} w_i (x_i - \bar{x})(y_i - \bar{y}) \\ \frac{1}{W} \sum_{i=1}^{m} w_i (x_i - \bar{x})(y_i - \bar{y}) & \frac{1}{W} \sum_{i=1}^{m} w_i (y_i - \bar{y})^2 \end{bmatrix} \tag{20}$$
Prema Lemi **??** matrica S je pozitivno definitna pa možemo definirati Mahalanobis kvazimetričku funkciju $d_M \colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$,
$$d_M(x, y) = \|x - y\|_\mu^2 = (x - y)^T S^{-1} (x - y). \tag{21}$$

Lema 2. *Neka je* $\mathcal{A} = \{a^i = (x_i, y_i) \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$ *skup vektora s centroidom* $\bar{c} = (\bar{x}, \bar{y})$*, pri čemu su vektori* $(x_1 - \bar{x}, \ldots, x_m - \bar{x})^T$*,* $(y_1 - \bar{y}, \ldots, y_m - \bar{y})^T$ *linearno nezavisni. Tada je kovarijacijska matrica* S *zadana s* (20) *simetrična pozitivno definitna matrica.*

Simetričnost je očigledna, a pozitivna definitnost slijedi iz Cauchy-Schwarz-Buniakowsky nejednakosti.

Zadatak 9. *Neka su $u, v \in \mathbb{R}$ proizvoljni brojevi i $a = (x_1, \ldots, x_m)$, $b = (y_1, \ldots, y_m) \in \mathbb{R}^m$ vektori.*

(a) Ako su vektori $a, b \in \mathbb{R}^m$ linearno nezavisni (zavisni), moraju li i vektori $(x_1 - u, \ldots, x_m - u)$, $(y_1 - v, \ldots, y_m - v) \in \mathbb{R}^m$ biti linearno nezavisni (zavisni) ?

(b) Ako su vektori $(x_1 - u, \ldots, x_m - u)$, $(y_1 - v, \ldots, y_m - v) \in \mathbb{R}^m$ linearno nezavisni (zavisni), moraju li i vektori $a, b \in \mathbb{R}^m$ biti linearno nezavisni (zavisni) ?

Primjer 8. *Svojstvene vrijednosti simetrične matrice $S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ su $\lambda_1 = 3$, $\lambda_2 = 1$, a odgovarajući jedinični svojstveni vektori $u_1 = \frac{\sqrt{2}}{2}(1, 1)$, $u_2 = \frac{\sqrt{2}}{2}(-1, 1)$. Pomoću ove matrice u okolini ishodišta $O = (0, 0)$ generirali smo $m = 500$ slučajnih točaka naredbom (Slika 5)*

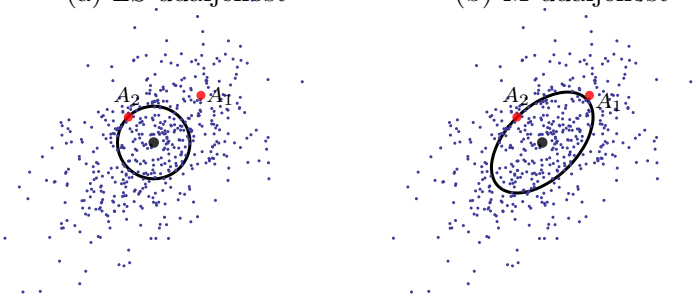
```
SeedRandom[23]
RandomReal[MultinormalDistribution[0, S],m]
```

LS-centroid $\bar{c} = (\bar{x}, \bar{y})$ i kovarijacijska matrica prema (20) je

$$\bar{c} = (0.0154519, -0.0510907), \qquad \text{cov} = \begin{bmatrix} 1.95989 & 0.97467 \\ 0.97467 & 1.93348 \end{bmatrix}.$$

Svojstvene vrijednosti su $\lambda_1 = 2.92145$, $\lambda_2 = 0.971928$. Odgovarajući svojstveni vektori su $u_1 = (-0.71188, -0.702301)$ (glavni smjer - first principal component), $u_2 = (0.702301, -0.71188)$ (sporedni smjer - second principal component).

(a) LS-udaljenost (b) M-udaljenost



Slika 5: Skup točaka generiran kovarijacijskom matricom S

Izaberimo točku $A_1 = (1.3, 1.3)$ na glavnom smeru. Njena LS-udaljenost do ishodišta $O = (0, 0)$ je $d_{LS}(A_1, O) = 3.38$, a njena M-udaljenost do ishodišta je $d_M(A_1, O) = 1.15707$

(Slika 5a). Točku $A_2 = (-.7, .7)$ izaberimo na sporednom smjeru. Njena LS-udaljenost do ishodišta je $d_{LS}(A_2, O) = 0.98$, a njena M-udaljenost do ishodišta je $d_M(A_2, O) = 1.00827$ (Slika 5b).

Kao što se može vidjeti u ravnini snabdjevenoj LS-kvazimetričkom funkcijom točka A_1 je daleko od jedinične kružnice, a točka A_2 blizu jedinične kružnice. U ravnini snabdjevenoj M-kvazimetričkom funkcijom obje točke A_1, A_2 su blizu jedinične kružnice.

Teorem 2. *Neka je* $A = \{T_i = (x_i, y_i) \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$ *skup točaka, a* $\mathcal{A} = \{a^i = (x_i, y_i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$ *odgovarajući skup vektora s odgovarajućim težinama* $w_i > 0$ *i* $S \in \mathbb{R}^{2\times 2}$, $S > 0$ *simetrična pozitivno definitna matrica. Tada se LS-centroid podudara s M-centoidom.*

Dokaz. U skladu s Definicijom 5 M-centroid skupa $\mathcal{A}$ je vektor

$$c_M^\star = \operatorname*{argmin}_{c\in\mathbb{R}^2} \sum_{i=1}^m w_i d_M(c, a^i) = \operatorname*{argmin}_{c\in\mathbb{R}^2} \sum_{i=1}^m w_i (c - a^i)^T S^{-1} (c - a^i),$$

na kome se postiže minimum funkcionala

$$F_M(c) = \sum_{i=1}^m w_i d_M(c, a^i) = \sum_{i=1}^m w_i (c - a^i)^T S^{-1} (c - a^i).$$

Stacionarna točka $c_M^\star$ funkcionala F_M određena je s

$$\sum_{i=1}^m w_i S^{-1} (c - a^i) = 0.$$

Kako je Hessian funkcionala F_M pozitivno definitan ($H_{F_M} = 2S^{-1} \sum w_i$), na vektoru

$$c_M^\star = \tfrac{1}{\sum w_i} \sum_{i=1}^m w_i a^i = c_{LS}^\star,$$

postiže se globalni minimum funkcionala F_M. Dakle, M-centroid i LS-centroid skupa $\mathcal{A}$ se podudaraju. □

3.2.2 Mahalanobis udaljenost inducirana skupom točaka podataka u $\mathbb{R}^n$

Općenito neka je $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i) \in \mathbb{R}^n \colon i = 1, \ldots, m\}$ skup točaka, a $\mathcal{A} = \{a^i = (a_1^i, \ldots, a_n^i)^T \in \mathbb{R}^2 \colon i = 1, \ldots, m\}$ odgovarajući skup vektora s težinama $w_1, \ldots, w_m > 0$.

(i) Centroid skupa A zadan je s

$$c^\star = \operatorname*{argmin}_{c\in\mathbb{R}^n} \sum_{i=1}^m w_i d_M(c, a^i) = \tfrac{1}{W} \sum_{i=1}^m w_i a^i, \qquad W = \sum_{i=1}^m w_i;$$

(ii) Neka je

$$B = \begin{bmatrix} c_1^\star - a_1^1 & \cdots & c_n^\star - a_n^1 \\ \vdots & \ddots & \vdots \\ c_1^\star - a_1^m & \cdots & c_n^\star - a_n^m \end{bmatrix}, \qquad D = \operatorname{diag}(w_1, \ldots, w_m);$$

(iii) Tada je kovarijacijska matrica $S = \frac{1}{W} B^T D B$ zadana s

$$\frac{1}{W} \begin{bmatrix} \sum w_i (c_1^\star - a_1^i)^2 & \sum w_i (c_1^\star - a_1^i)(c_2^\star - a_2^i) & \cdots & \sum w_i (c_1^\star - a_1^1)(c_n^\star - a_n^i) \\ \sum w_i (c_2^\star - a_2^i)(c_1^\star - a_1^i) & \sum w_i (c_2^\star - a_2^i)^2 & \cdots & \sum w_i (c_2^\star - a_2^1)(c_n^\star - a_n^i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum w_i (c_n^\star - a_n^i)(c_1^\star - a_1^i) & \sum w_i (c_n^\star - a_n^i)(c_2^\star - a_2^i) & \cdots & \sum w_i (c_n^\star - a_n^i)^2 \end{bmatrix};$$

(iv) Mahalanobis udaljenost definira se s $d_M \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ (Durak, 2011),

$$d_M(x, y) = (x - y)^T S^{-1} (x - y).$$

4 Geometrijski medijan

Točku $c_2^\star \in \mathbb{R}^2$,

$$c_2^\star = \operatorname*{argmin}_{c \in \mathbb{R}^2} \sum_{i=1}^m w_i d_2(c, a^i), \tag{22}$$

u kojoj se postiže globalni minimum funkcionala

$$F(c) = \sum_{i=1}^m w_i d_2(c, a^i) = \sum_{i=1}^m w_i \sqrt{(x - x_i)^2 + (y - y_i)^2} = \sum_{i=1}^m w_i \|c - a^i\|, \tag{23}$$

zovemo **težinski geometrijski median** skupa vektora $\mathcal{A}$ (odnosno skupa točaka A) u ravnini i ne može se eksplicitno izraziti.

4.1 Weiszfeldov algoritam za traženje geometrijskog mediana

Stacionarne točke funkcionala F iz (23) određene su s $\frac{\partial F}{\partial x} = \frac{\partial F}{\partial y} = 0$, iz čega dobivamo sustav

$$\sum_{i=1}^m w_i \frac{x - x_i}{\sqrt{(x - x_i)^2 + (y - y_i)^2}} = 0, \quad \sum_{i=1}^m w_i \frac{y - y_i}{\sqrt{(x - x_i)^2 + (y - y_i)^2}} = 0, \tag{24}$$

odnosno u vektorskom obliku

$$\operatorname{grad} F(c) = \sum_{i=1}^m w_i \|c - a^i\|^{-1} (c - a^i) = 0. \tag{25}$$

Sustav (24) možemo zapisati u obliku

$$x = \frac{\sum_{i=1}^{m} w_i \frac{x_i}{\sqrt{(x-x_i)^2+(y-y_i)^2}}}{\sum_{s=1}^{m} w_s \frac{1}{\sqrt{(x-x_s)^2+(y-y_s)^2}}}, \qquad y = \frac{\sum_{i=1}^{m} w_i \frac{y_i}{\sqrt{(x-x_i)^2+(y-y_i)^2}}}{\sum_{s=1}^{m} w_s \frac{1}{\sqrt{(x-x_s)^2+(y-y_s)^2}}}, \tag{26}$$

odnosno uz oznake

$$\omega_i(x,y) = \frac{\frac{w_i}{\sqrt{(x-x_i)^2+(y-y_i)^2}}}{\sum_{s=1}^{m} w_s \frac{1}{\sqrt{(x-x_s)^2+(y-y_s)^2}}}, \quad i = 1, \ldots, m. \tag{27}$$

u obliku

$$x = \sum_{i=1}^{m} \omega_i(x,y) x_i, \qquad y = \sum_{i=1}^{m} \omega_i(x,y) y_i. \tag{28}$$

Slično u vektorskom obliku iz (25) dobivamo

$$c = \frac{\sum_{i=1}^{m} w_i \|c - a^i\|^{-1} a^i}{\sum_{s=1}^{m} w_s \|c - a^s\|^{-1}}. \tag{29}$$

što možemo zapisati kao

$$c = \sum_{i=1}^{m} \omega_i(c) a^i, \qquad \omega_i(c) = \frac{w_i \|c - a^i\|^{-1}}{\sum_{s=1}^{m} w_s \|c - a^s\|^{-1}}. \tag{30}$$

Uz povoljan izbor početne aproksimacije $(x_0, y_0) \in \mathbb{R}^2$ sustav nelinearnih jednadžbi (26), odnosno (28), rješava se metodom jednostavnih iteracija Scitovski (2004) i poznat je pod nazivom **Weiszfeldov algoritam** (1936)

$$x^{(k+1)} = \sum_{i=1}^{m} \omega_i(x^{(k)}, y^{(k)}) x_i, \quad y^{(k+1)} = \sum_{i=1}^{m} \omega_i(x^{(k)}, y^{(k)}) y_i, \qquad k = 0, 1, \ldots, \tag{31}$$

odnosno u vektorskom obliku

$$c^{(k+1)} = \sum_{i=1}^{m} \omega_i(c^{(k)}) a^i. \tag{32}$$

Zadatak 10. *Pokažite da za težinske funkcije* $(x,y) \mapsto \omega_i(x,y)$ *zadane s* (27) *vrijedi*

$$0 < \omega_i(x,y) < 1, \qquad \sum_{i=1}^{m} \omega_i(x,y) = 1.$$

Zadatak 11. *Konstruirajte Weiszfeldov algoritam za traženje najboljeg reprezentanta za Mahalanobis udaljenost* $d_\mu(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$, *gdje je* $S \in \mathbb{R}^{2\times 2}$ *simetrična pozitivno definitna matrica. Možete li algoritam generalizirati za* $\mathbb{R}^n$ *?*

5 Reprezentant podataka na jediničnoj kružnici

Zadan je skup realnih brojeva $t_1, \ldots, t_m \in [0, 2\pi]$, pomoću kojih gradimo skup točaka $\mathcal{A} = \{a^i(t_i) = (\cos t_i, \sin t_i) \in \mathbb{R}^2 \colon t_i \in [0, 2\pi]\}$, na jediničnoj kružnici $K = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 = 1\}$

Treba odrediti točku $c^\star(t^\star) \in K$ koja će što bolje reprezentirati skup točaka A.

Funkcija $d_K \colon \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$,

$$d_K(a(t_1), b(t_2)) = \begin{cases} |t_1 - t_2|, & \text{if } |t_1 - t_2| < \pi, \\ 2\pi - |t_1 - t_2|, & \text{if } |t_1 - t_2| > \pi. \end{cases} \tag{33}$$

je *metrička funkcija* na K.

Točka $c^\star(t^\star) \in K$ je *najbolji reprezentant* skupa $\mathcal{A}$ s težinama $w_1, \ldots, w_m > 0$ u odnosu na metričku funkciju d_K onda ako je

$$c^\star(t^\star) = \underset{t \in [0,2\pi]}{\operatorname{argmin}} \sum_{i=1}^{m} w_i d_K(c(t), a^i(t_i)), \tag{34}$$

tj. ako je $c^\star(t^\star) \in K$ točka globalnog minimuma funkcionala $F \colon [0, 2\pi] \to \mathbb{R}_+$

$$F(t) = \sum_{i=1}^{m} w_i d_K(c(t), a^i(t_i)). \tag{35}$$

6 Prepoznavanje riječi

$\mathcal{A} = \{a^i = (x_1, \ldots, x_n)^T \in \{0, 1\}^n \colon i = 1, \ldots, m\} \subset \mathbb{R}^n$

$d \colon \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$ – kvazimetrička funkcija, primjerice

d_{LS}, d_1, $d_c(x, y) = 1 - \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$ (kosinus).

U nekom tekst prisutnost neke riječi kodira se s 1, a odsutnost te riječi iz teksta s 0. Postavlja se pitanje o sličnosti/različitosti dva teksta obzirom na prisutnost/odsutnost promatranih riječi. Tekst u kome je prisutno/odsutno $n \geq 1$ izabranih riječi prikazat ćemo vektorom iz $\mathbb{R}^n$ s komponentama 0 ili 1.

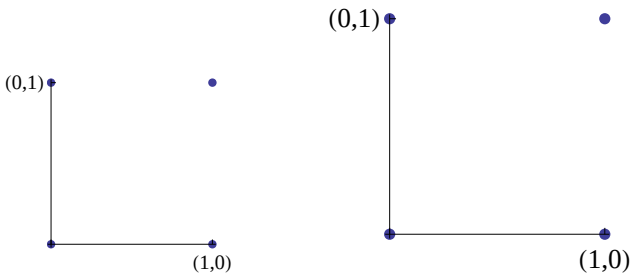
Primjer 9. *Promatramo tekstove u kojima se mogu pojaviti riječi: A,B,C. Neka je primjerice (vidi Sliku 6):*

$a^1 = (1, 1, 0)$*: tekst u kome se pojavljuju riječi A,B, a ne pojavljuje riječ C*

$a^2 = (1, 0, 0)$*: tekst u kome se pojavljuje riječ A, a ne pojavljuju riječi B,C*

$a^3 = (1, 0, 1)$*: tekst u kome se pojavljuju riječi A, C, a ne pojavljuje riječ B*

$a^4 = (0, 0, 1)$*: rečenica u kojoj se pojavljuje riječ C, a ne pojavljuju riječi A, B*



Slika 6: Skup $\mathcal{A}$ za $n = 2$ i $n = 3$

U svrhu ispitivanja sličnosti/različitosti tekstova obzirom na prisutnost/odsutnost nekih riječi možemo pokušati iskoristiti ranije spomenute metričke funkcije d_1, d_2, d_∞. U znanstvenoj literaturi (vidi primjerice (Berry and Kogan, 2010; Zhang, 2009)) u tu svrhu koriste se neke tzv. *kvazimetričke funkcije*, kao što su

$$d_{LS}(x,y) = \|x-y\|^2 \quad \text{– Least Squares (LS) kvazimetrička funkcija}$$
$$d_c(x,y) = 1 - \frac{\langle x,y\rangle}{\|x\|\cdot\|y\|} \quad \text{– kosinus kvazimetrička funkcija}$$

Za prethodno spomenuti primjer dobivamo

$$\begin{array}{lll} d_{LS}(a^1,a^2)=1, & d_{LS}(a^1,a^3)=2, & d_{LS}(a^1,a^4)=3 \\ d_1(a^1,a^2)=1, & d_1(a^1,a^3)=2, & d_1(a^1,a^4)=3 \\ d_c(a^1,a^2)=1-\frac{\sqrt{2}}{2}=0.29, & d_c(a^1,a^3)=\frac{1}{2}, & d_c(a^1,a^4)=0 \end{array}$$

Prema LS-kvazimetričkoj funkciji (a također i prema l_1-metričkoj funkciji) tekstovi a^1 i a^2 su najsličniji (najbliži), a tekstovi a^1 i a^4 najrazličitiji (najudaljeniji) obzirom na pojavu riječi A,B,C.

I prema kosinus-metričkoj funkciji d_c tekstovi a^1 i a^2 su najsličniji (najbliži), a tekstovi a^1 i a^4 potpuno različiti (maksimalno udaljeni) obzirom na pojavu riječi A,B,C.

Primjer 10. *Promatramo tekstove u kojima se mogu pojaviti riječi: A,B,C,D,E. Neka je primjerice:*

$a^1 = (1,0,0,0,1)$*: tekst u kome se pojavljuju riječi A, E, a ne pojavljuju riječi B,C,D*

$a^2 = (0,1,1,0,0)$*: tekst u kome se pojavljuju riječi B, C, a ne pojavljuju riječi A,D,E*

$a^3 = (1,0,0,0,0)$*: tekst u kome se pojavljuje riječe A, a ne pojavljuju riječi B,C,D, E*

$d_{LS}(a^i,a^j)$	a^1	a^2	a^3	$d_1(a^i,a^j)$	a^1	a^2	a^3	$d_c(a^i,a^j)$	a^1	a^2	a^3
a^1	0	4	1	a^1	0	4	1	a^1	0	1	0.29
a^2	4	0	3	a^2	4	0	3	a^2	1	0	1
a^3	1	3	0	a^3	1	3	0	a^3	0.29	1	0

Iz ovog primjera vidi se da kosinus-kvazimetrička funkcija puno bolje identificira sličnosti/različitosti tekstova obzirom na prisutnost/odsutnost riječi A,B,C,D,E (objasnite to na osnovi brojeva iz tablica!).

Literatura

M. W. Berry, J. Kogan, *Text Mining. Applications and Theory*, Wiley, 2010.

D. L. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

S. Butenko, W. A. Chaovalitwongse, P. M. Pardalos, *Clustering Challenges in Biological Networks*, World Scientific, 2009.

R. Cupec, R. Grbić, K. Sabo, R. Scitovski, *Three points method for searching the best least absolute deviations plane*, Applied Mathematics and Computation, **215**(2009), 983–994

B Durak, *A Classification Algorithm Using Mahalanobis Distances Clustering of Data with Applications on Biomedical Data Set*, The Graduate School of Natural and Applied Sciences of Middle East Technical University, 2011

C. A. Floudas, C. E. Gounaris, *A review of recent advances in global optimization*, J. Glob. Optim. **45**(2009), 3-–38

G. Gan, C Ma, J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 2007.

C. Iyigun, A. Ben-Israel, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters **38**(2010) 207–214

D. R. Jones, C. D. Perttunen, B. E. Stuckman, *Lipschitzian optimization without the Lipschitz constant*, JOTA **79**(1993), 157-181

D. Jukić, R. Scitovski and H. Späth, *Partial linearization of one class of the nonlinear total least squares problem by using the inverse model function*, Computing **62**(1999), 163-178.

D. Jukić, R. Scitovski and Š. Ungar, *The best total least squares line in* $\mathbb{R}^3$, Proceedings of the 7th International Conference on Operational Research KOI98, I. Aganović, T. Hunjak and R. Scitovski, Eds., Osijek, 1999, 311-316.

J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press, 2007.

F. Leisch, *A toolbox for K-centroids cluster analysis*, Computational Statistics & Data Analysis **51**(2006), 526-544

A. Neumaier, *Complete search in continuous global optimization and constraint satisfaction*, Acta Numerica (2006), 271-369.

Y. Neivergelt, *Total least squares: state-of-the-art regression in numerical analysis*, SIAM Review, **36**(1994), 258-264

P. M. Pardalos, P. Hansen, *Data Mining and Mathematical Programming*, American Mathematical Society, Providence, 2008.

J. Reese, *Solution methods for the p-median problem: an annotated bibliography*, Published online in Wiley InterScience, Wiley, 2006.

K. Sabo, R. Scitovski, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal **50**(2008), 185–198 doi:10.1017/S1446181108000345

K. Sabo, R. Scitovski, I.Vazler, *One-dimensional center-based l_1-clustering method*, Optimization Letters (accepted) DOI:10.1007/s11590-011-0389-9

A. Schöbel, D. Scholz, *The big cube small cube solution method for multidimensional facility location problems*, Computers & Operations Research **37**(2010), 115–122

R. Scitovski, *Numerička matematika*, Odjel za matematiku, Sveučilište u Osijeku, Osijek, 2004.

H. Späth, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.

M. Teboulle, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research **8**(2007), 65–102

N. Truhar, *Numerička linearna algebra*, Odjel za matematiku, Sveučilište u Osijeku, Osijek, 2010.

I. Vazler, K. Sabo, R. Scitovski, *Weighted median of the data in solving least absolute deviations problems*, Communications in Statistics - Theory and Methods, to appear in 2011

H. Zhang, Statistical Clustering Analysis: An Introduction, in S. Butenko, W. A. Chaovalitwongse, and P. M. Pardalos (eds.), *Clustering Challenges in Biological Networks*, World Scientific, 2009, pp 101–126