Metódy odhadovania kovariančnej matice priestorového mediánu

Katarína Burclová a Ján Somorčík

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky
Katedra aplikovanej matematiky a štatistiky

15.9.2016

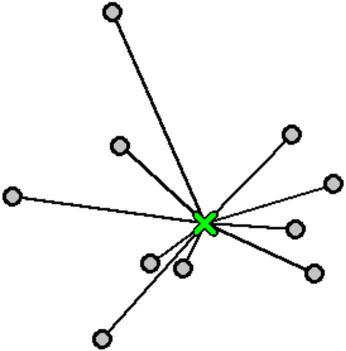
1. Priestorový medián

Definícia

$X_1, \ldots, X_n$ - náhodný výber z p-rozmerného rozdelenia s hustotou f.
Priestorový medián:

$$\hat{\theta} = \arg\min_{\phi \in \mathbb{R}^p} \sum_{i=1}^{n} \|X_i - \phi\|$$

Výpočet $\hat{\theta}$ - iteračný algoritmus z článku Vardi a Zhang (2000)



$n = 10,\ p = 2$

Asymptotická kovariančná matica priestorového mediánu

$X \in \mathbb{R}^p$, označme:

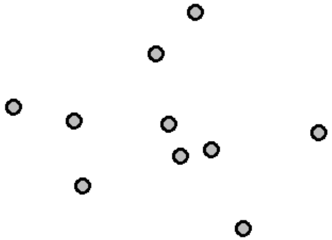
$$
\begin{aligned}
\theta - &\text{teoretický priestorový medián } X, \\
U(X) &= \|X\|^{-1} X, \\
Q(X) &= \|X\|^{-1} (I_p - \|X\|^{-2} X X^T), \\
A &= E[Q(X-\theta)], \\
B &= E[U(X-\theta)\, U^T (X-\theta)].
\end{aligned}
$$

Veta

$$n^{1/2}(\hat{\theta} - \theta) \dot{\sim} N_p(0, \underbrace{A^{-1} B A^{-1}}_{\mathcal{D}})$$

Ako odhadnúť $\mathcal{D}$ - Bose a Chaudhuri (1993)

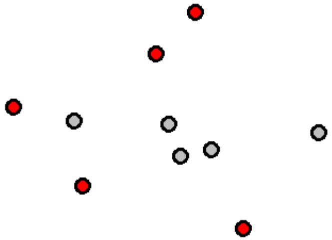
- dáta: $X_1, \ldots, X_n$



$n = 10,\ p = 2$

Ako odhadnúť $\mathcal{D}$ - Bose a Chaudhuri (1993)

- dáta: $X_1, \ldots, X_n$
- dve nezávislé skupiny $\{X_i, i \in S_n\}$ a $\{X_i, i \in S_n^c\}$
- $\#(S_n) = k_n$ a $\#(S_n^c) = n - k_n$

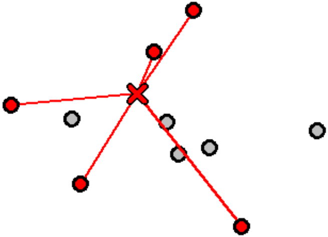


$n = 10$, $p = 2$, $k_n = 5$

Ako odhadnúť $\mathcal{D}$ - Bose a Chaudhuri (1993)

- dáta: $X_1, \ldots, X_n$
- dve nezávislé skupiny $\{X_i, i \in S_n\}$ a $\{X_i, i \in S_n^c\}$
- $\#(S_n) = k_n$ a $\#(S_n^c) = n - k_n$
- prvá časť - výberový medián:

$$\hat{\theta}_{k_n} = \arg \min_{\phi \in \mathbb{R}^p} \sum_{i \in S_n} \|X_i - \phi\|$$



$n = 10,\ p = 2,\ k_n = 5$

Ako odhadnúť $\mathcal{D}$ - Bose a Chaudhuri (1993)

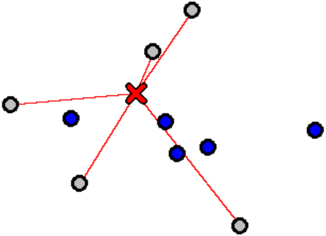
- druhá časť dát - matice A a B:

$$\hat{A}_{k_n} = \frac{1}{n-k_n} \sum_{i \in S_n^c} Q(X_i - \hat{\theta}_{k_n})$$

$$\hat{B}_{k_n} = \frac{1}{n-k_n} \sum_{i \in S_n^c} U(X_i - \hat{\theta}_{k_n}) U^T(X_i - \hat{\theta}_{k_n})$$

- asymptotická kovariančná matica:

$$\hat{\mathcal{D}}_{k_n} = \hat{A}_{k_n}^{-1} \hat{B}_{k_n} \hat{A}_{k_n}^{-1}$$



$n = 10,\ p = 2,\ k_n = 5$

Cieľ

Bose a Chaudhuri (1993): „*... This leaves us with a wide range of choices for k_n. ... However, we have not tried to dig deeper into this matter...*"

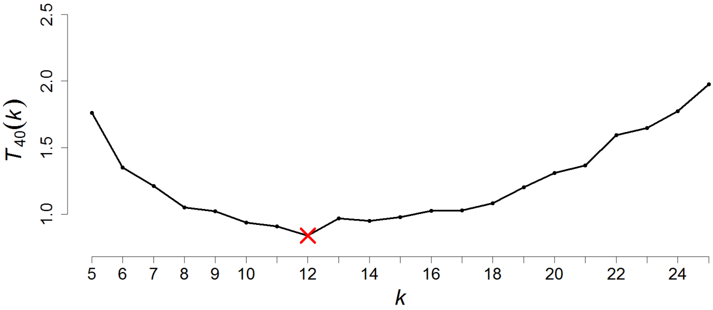
Podmienky na k_n:

- $k_n \in \{1, \ldots, n-1\}$
- $\lim_{n\to\infty} \frac{k_n}{n} \neq 0$ a $\lim_{n\to\infty}(1 - \frac{k_n}{n}) \neq 0$

Cieľ

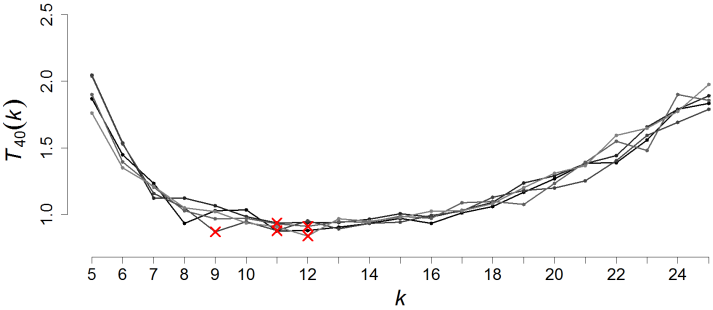
nájsť k_n^* - t. j. k_n minimalizujúce vzdialenosť matíc $\mathcal{D}$ a $\hat{\mathcal{D}}_{k_n}$

Štandardizované normálne rozdelenie, $p = 3$, $n = 40$



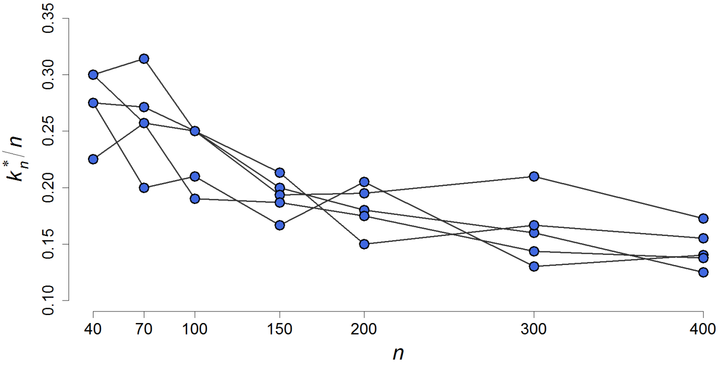
$T_{40}(k)$ odhad pre normovanú strednú vzdialenosť medzi $\mathcal{D}$ a jej odhadom $\mathcal{D}_k$ pri $n = 40$ vypočítaný zo simulácii

Rozptyl optimálneho k_n^*



Simulácia zopakovaná 5-krát

Vplyv rozsahu súboru n



$N_3(0, I_3)$, pomer k_n^*/n klesá s n

Zhrnutie 1. Priestorový medián

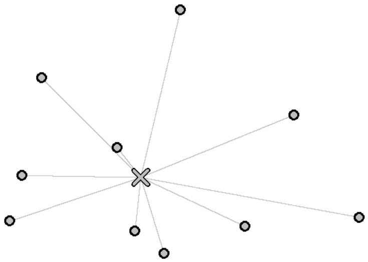
- Optimálna voľba parametra k_n je približne **15-30%** z celkového počtu dát
- k_n^*/n klesá s rastúcim rozsahom súboru n
- Podobné výsledky aj pre:
  - normálne rozdelenie s inou kovariančnou maticou (diagonálna, všeobecná)
  - viacrozmerné t-rozdelenie
- pri Cauchyho rozdelení $k_n^* > 30\% \cdot n$

2. Hodges-Lehmannov priestorový medián

Definícia

$X_1, \ldots, X_n$ - náhodný výber z p-rozmerného rozdelenia s hustotou f.
Hodges-Lehmannov priestorový medián:

$$\hat{\psi} = \arg \min_{\phi \in \mathbb{R}^p} \sum_{i \neq j}^{n} \left\| \frac{X_i + X_j}{2} - \phi \right\|$$



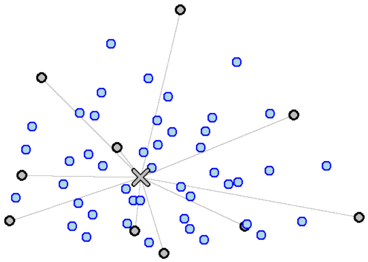
$n = 10,\ p = 2$

2. Hodges-Lehmannov priestorový medián

Definícia

$X_1, \ldots, X_n$ - náhodný výber z p-rozmerného rozdelenia s hustotou f.
Hodges-Lehmannov priestorový medián:

$$\hat{\psi} = \arg\min_{\phi \in \mathbb{R}^p} \sum_{i \neq j}^{n} \left\| \frac{X_i + X_j}{2} - \phi \right\|$$



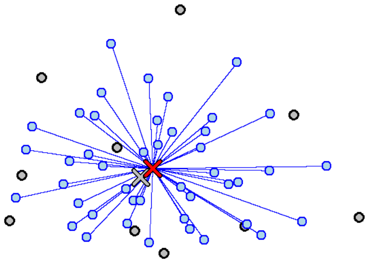
$n = 10,\ p = 2$

2. Hodges-Lehmannov priestorový medián

Definícia

$X_1, \ldots, X_n$ - náhodný výber z p-rozmerného rozdelenia s hustotou f.
Hodges-Lehmannov priestorový medián:

$$\hat{\psi} = \arg \min_{\phi \in \mathbb{R}^p} \sum_{i \neq j}^{n} \left\| \frac{X_i + X_j}{2} - \phi \right\|$$



$n = 10,\ p = 2$

Asymptotická kovariančná matica Hodges-Lehmannovho priestorového mediánu

$X^1, X^2, X^3 \in \mathbb{R}^p$, označme:

$\psi -$ teoretický Hodges-Lehmannov priestorový medián X,

$$C = E\left[Q\left(\frac{X^1+X^2}{2}-\psi\right)\right],$$

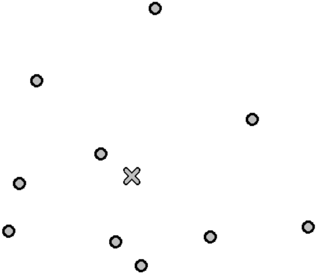
$$D = E\left[U\left(\frac{X^1+X^2}{2}-\psi\right)U^T\left(\frac{X^2+X^3}{2}-\psi\right)\right].$$

Veta

$$n^{1/2}(\hat{\psi}-\psi)\dot{\sim}N_p(0,\underbrace{4C^{-1}DC^{-1}}_{\mathcal{D}})$$

Ako odhadnúť $\mathcal{D}$ - Bose a Chaudhuri (1993)

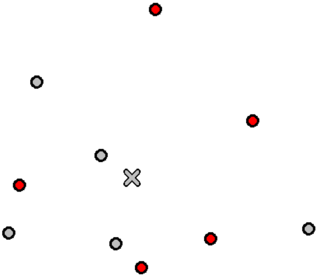
- dáta: $X_1, \ldots, X_n$



$n = 10$, $p = 2$

Ako odhadnúť $\mathcal{D}$ - Bose a Chaudhuri (1993)

- dáta: $X_1, \ldots, X_n$
- dve nezávislé skupiny
  $\{X_i, i \in S_n\}$ a $\{X_i, i \in S_n^c\}$
- $\#(S_n) = k_n$ a $\#(S_n^c) = n - k_n$

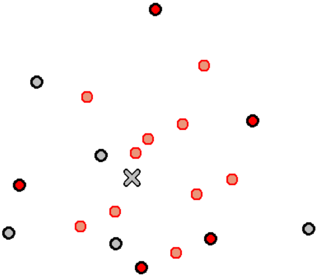


$n = 10$, $p = 2$, $k_n = 5$

Ako odhadnúť $\mathcal{D}$ - Bose a Chaudhuri (1993)

- dáta: $X_1, \ldots, X_n$
- dve nezávislé skupiny $\{X_i, i \in S_n\}$ a $\{X_i, i \in S_n^c\}$
- $\#(S_n) = k_n$ a $\#(S_n^c) = n - k_n$
- prvá časť - výberový medián:

$$\hat{\psi}_{k_n} = \arg\min_{\phi \in \mathbb{R}^p} \sum_{\substack{i,j \in S_n \\ i \neq j}} \left\| \frac{X_i + X_j}{2} - \phi \right\|$$

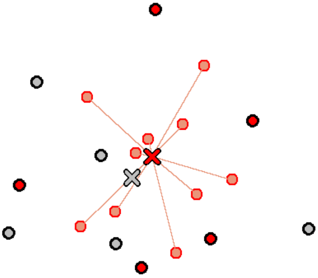


$n = 10,\ p = 2,\ k_n = 5$

Ako odhadnúť $\mathcal{D}$ - Bose a Chaudhuri (1993)

- dáta: $X_1, \ldots, X_n$
- dve nezávislé skupiny
  $\{X_i, i \in S_n\}$ a $\{X_i, i \in S_n^c\}$
- $\#(S_n) = k_n$ a $\#(S_n^c) = n - k_n$
- prvá časť - výberový medián:

$$\hat{\psi}_{k_n} = \arg\min_{\phi \in \mathbb{R}^p} \sum_{\substack{i,j \in S_n \\ i \neq j}} \left\| \frac{X_i + X_j}{2} - \phi \right\|$$



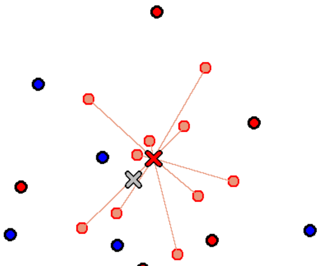
$n = 10,\ p = 2,\ k_n = 5$

Ako odhadnúť $\mathcal{D}$ - Bose a Chaudhuri (1993)

- druhá časť dát - matice C a D:

$$\hat{C}_{k_n} = \sum_{\substack{i,j \in S_n^c \\ i \neq j}} \frac{Q\left(\frac{X_i+X_j}{2} - \hat{\psi}_{k_n}\right)}{(n-k_n)(n-k_n-1)}$$

$$\hat{D}_{k_n} = \sum_{\substack{i,j,l \in S_n^c \\ i \neq j \neq l \neq i}} \frac{U\left(\frac{X_i+X_j}{2} - \hat{\psi}_{k_n}\right) U^T\left(\frac{X_j+X_l}{2} - \hat{\psi}_{k_n}\right)}{(n-k_n)(n-k_n-1)(n-k_n-2)}$$



- asymptotická kovariančná matica:

$$\hat{\mathcal{D}}_{k_n} = 4\hat{C}_{k_n}^{-1}\hat{D}_{k_n}\hat{C}_{k_n}^{-1}$$

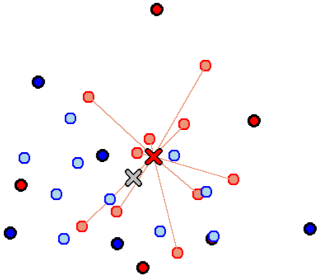
$n = 10,\ p = 2,\ k_n = 5$

Ako odhadnúť $\mathcal{D}$ - Bose a Chaudhuri (1993)

- druhá časť dát - matice C a D:

$$\hat{C}_{k_n} = \sum_{\substack{i,j \in S_n^c \\ i \neq j}} \frac{Q\left(\frac{X_i + X_j}{2} - \hat{\psi}_{k_n}\right)}{(n-k_n)(n-k_n-1)}$$

$$\hat{D}_{k_n} = \sum_{\substack{i,j,l \in S_n^c \\ i \neq j \neq l \neq i}} \frac{U\left(\frac{X_i + X_j}{2} - \hat{\psi}_{k_n}\right) U^T \left(\frac{X_j + X_l}{2} - \hat{\psi}_{k_n}\right)}{(n-k_n)(n-k_n-1)(n-k_n-2)}$$

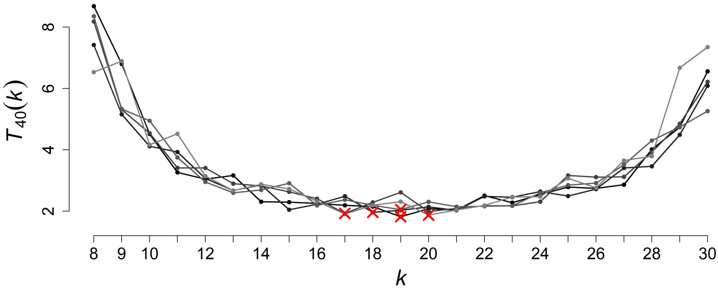


$n = 10,\ p = 2,\ k_n = 5$

- asymptotická kovariančná matica:

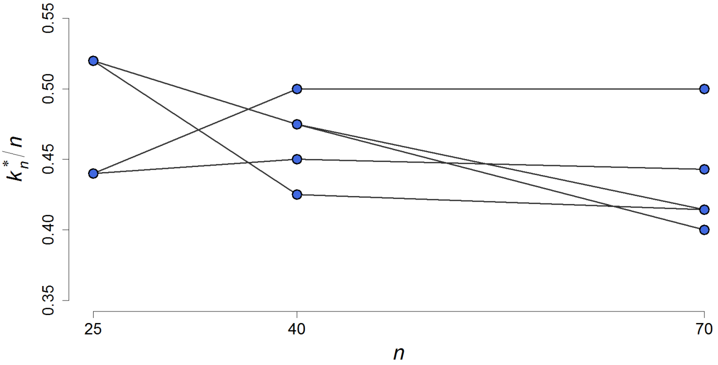
$$\hat{\mathcal{D}}_{k_n} = 4 \hat{C}_{k_n}^{-1} \hat{D}_{k_n} \hat{C}_{k_n}^{-1}$$

Rozptyl optimálneho k_n^*



$N_3(0, I_3),\ T_{40}(k)$ konvexné $\Rightarrow$ existujú k_{40}^*

Vplyv rozsahu súboru n



$N_3(0, I_3)$, k_n^* nezávisí od n

Zhrnutie 2. Hodges-Lehmannov priestorový medián

- Optimálna voľba parametra k_n je približne **40-50%** z celkového počtu dát
- Podobné výsledky aj pre:
  - normálne rozdelenie s inou kovariančnou maticou (diagonálna, všeobecná)
  - viacrozmerné t-rozdelenie
- pri Cauchyho rozdelení $k_n^* > 50\% \cdot n$

3. Porovnanie s ďalšími metódami odhadu

- Resamplingové metódy:

  1. Bootstrap
  2. Jackknife

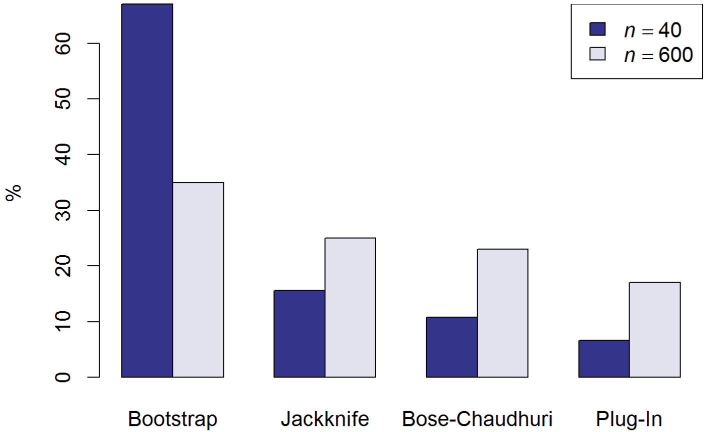
- Metódy založené na maticiach A a B (resp. C a D):

  3. Bose-Chaudhuri (1993)
  4. Plug-In odhad

Cieľ

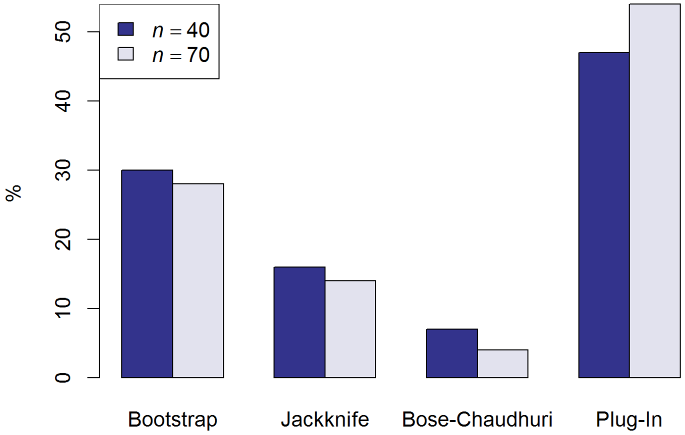
Porovnať metódy z hľadiska výpočtovej zložitosti a presnosti výsledného odhadu.

Priestorový medián



Úspešnosť jednotlivých metód.

Hodges-Lehmannov medián



Úspešnosť jednotlivých metód.

Zhrnutie 3. Porovnanie s ďalšími metódami odhadu

Na odhad asymptotickej kovariančnej matice:

1. **Priestorového mediánu** použiť
   - pri malom n **bootstrap**
   - pri veľkom n časovo nenáročný **Plug-In** alebo **Bose-Chaudhuri**
2. **Hodges-Lehmannovho priestorového mediánu** použiť
   - **Plug-In**

Ďakujem za pozornosť.

Literatúra:

- BOSE, A. – CHAUDHURI, P. 1993. On the dispersion of multivariate median. In: *Ann. Inst. Statist. Math.*, roč. 45, č. 3, s. 541 – 550.
- CHAUDHURI, P. 1992. Multivariate location estimation using extension of R-estimates through U-statistics type approach. In: *The Annals of Statistics*, roč. 20, s. 897 – 916.
- VARDI, Y. – ZHANG, C.-H. 2000. The multivariate L_1-median and associated data depth. In: *The Proceedings of the National Academy of Sciences USA (PNAS)*, roč. 97, č. 4, s. 1423 – 1426.
- R Development Core Team 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.