Imputation with data depth

Pavlo Mozharovskyi[1,2,3]
joint work with
Julie Josse[2,4] and François Husson[2,3]

[1] Centre Henri Lebesgue, Rennes
[2] Agrocampus Ouest, Rennes
[3] Institut de Recherche Mathématique de Rennes (IRMAR)
[4] Institut National de Recherche en Informatique et en Automatique (INRIA), Orsay

48ème Journées de Statistique de la SFdS
Montpellier, May 31, 2016

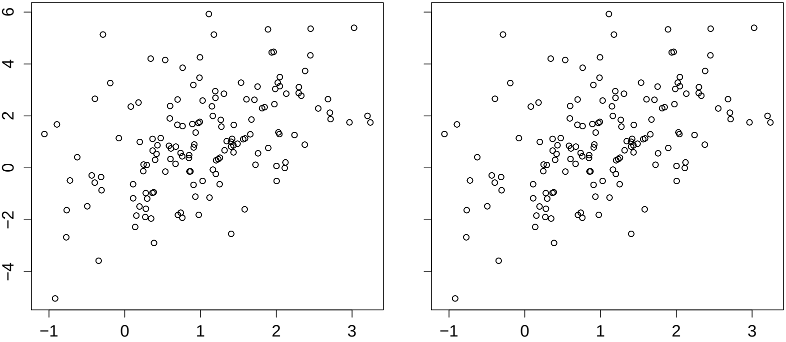
Contents

Contents

Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

Missing completely at random   Missing (not) at random

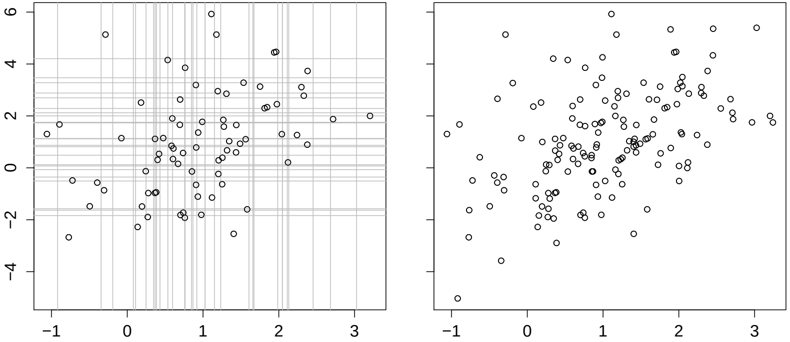


◆ – regression ● – zonoid depth ▲ – random forest

Imputation close to data and extrapolation

150 observasions of $X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$

Missing completely at random　　Missing (not) at random

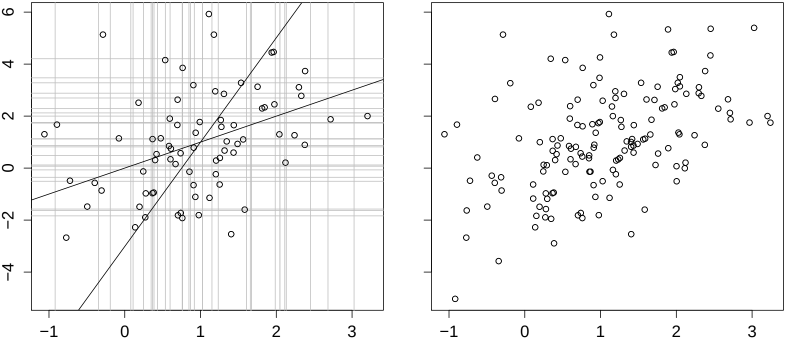


◆ – regression ● – zonoid depth ▲ – random forest

Imputation close to data and extrapolation

150 observasions of $X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$

Missing completely at random   Missing (not) at random

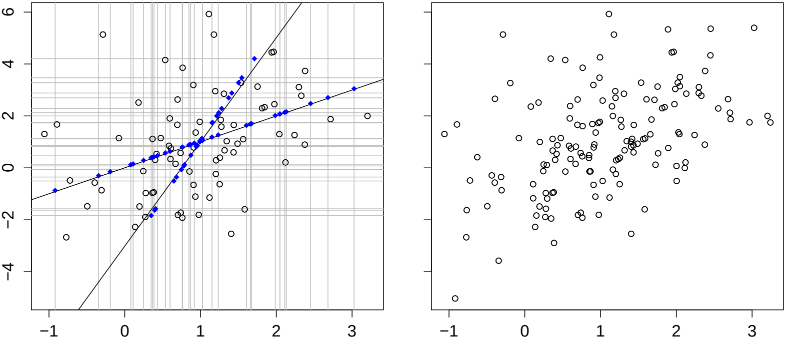


♦ – regression ● – zoonoid depth ▲ – random forest

Imputation close to data and extrapolation

150 observasions of $X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$

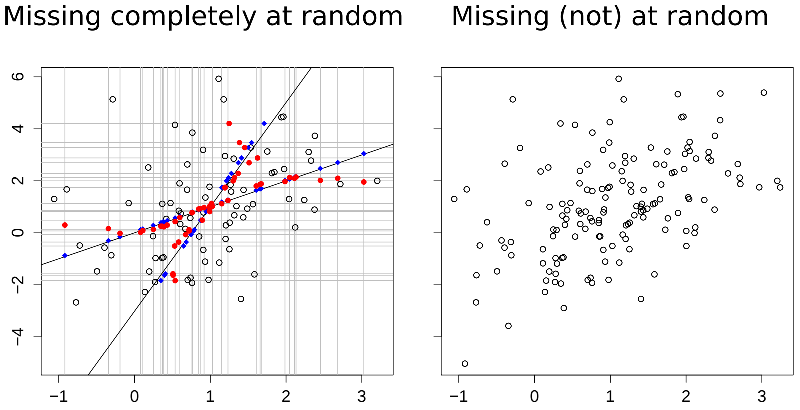
Missing completely at random Missing (not) at random



◆ – regression ● – zonoid depth ▲ – random forest

Imputation close to data and extrapolation

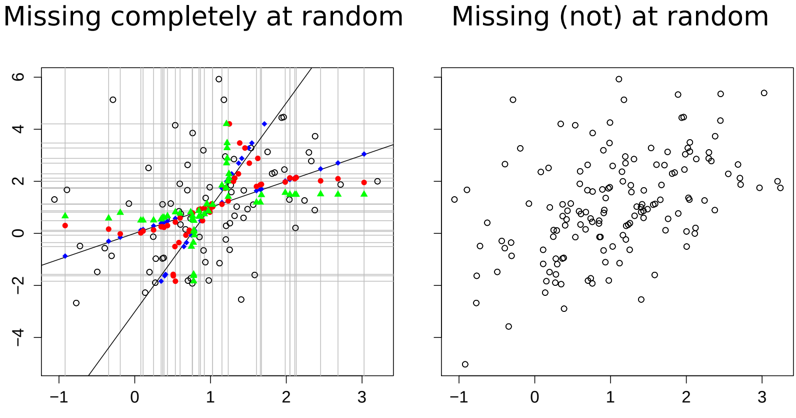
150 observasions of $X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$



◆ – regression ● – zonoid depth ▲ – random forest

Imputation close to data and extrapolation

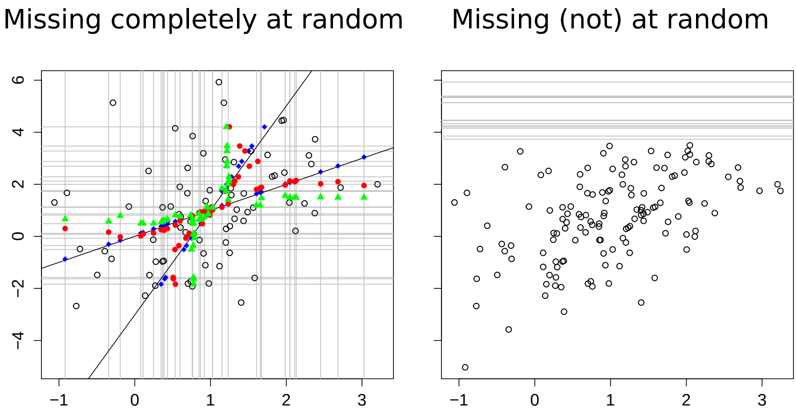
150 observasions of $X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$



◆ – regression ● – zoonoid depth ▲ – random forest

Imputation close to data and extrapolation

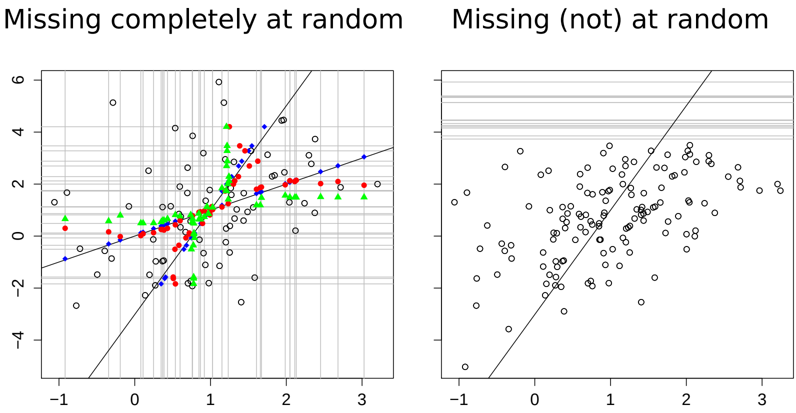
150 observasions of $X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$



◆ – regression ● – zonoid depth ▲ – random forest

Imputation close to data and extrapolation

$$150 \text{ observasions of } X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$

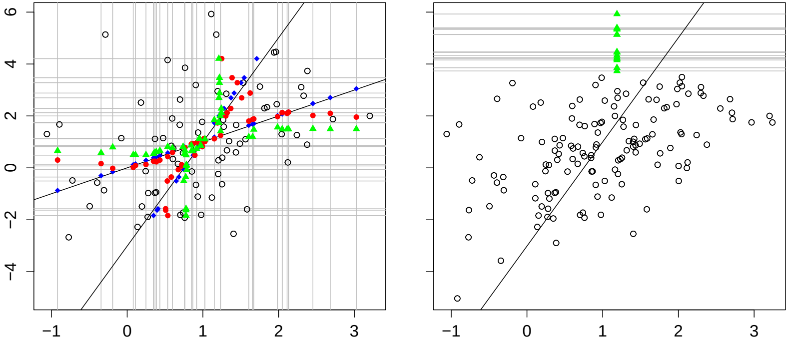


♦ – regression ● – zonoid depth ▲ – random forest

Imputation close to data and extrapolation

150 observasions of $X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$

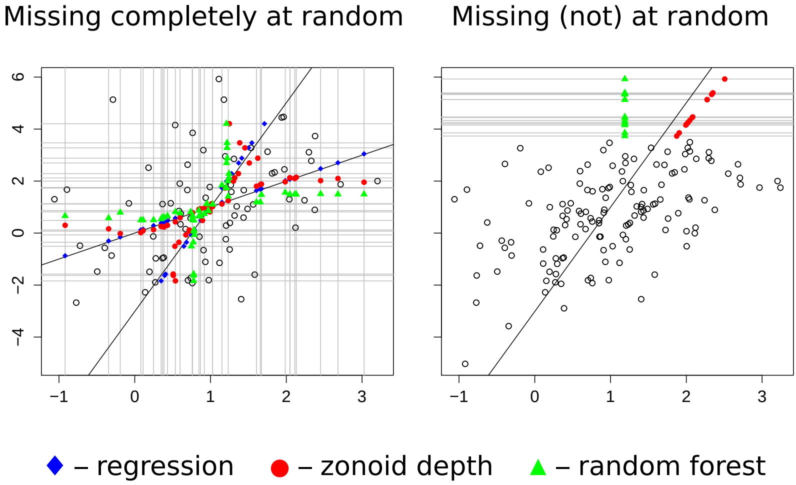
Missing completely at random　　　Missing (not) at random



◆ – regression　● – zonoid depth　▲ – random forest

Imputation close to data and extrapolation

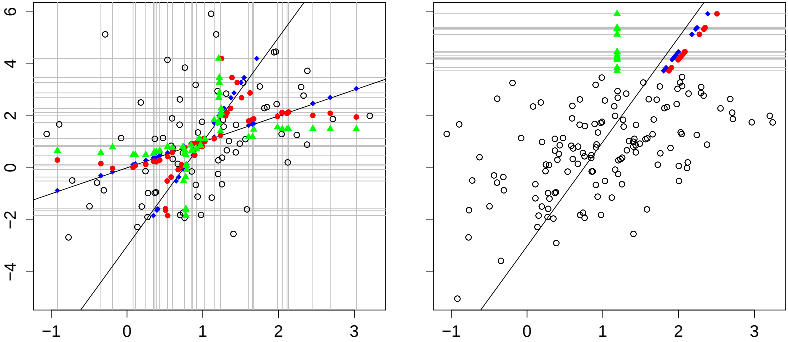
$$150 \text{ observasions of } X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$$



Imputation close to data and extrapolation

150 observasions of $X \sim Normal\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$

Missing completely at random Missing (not) at random

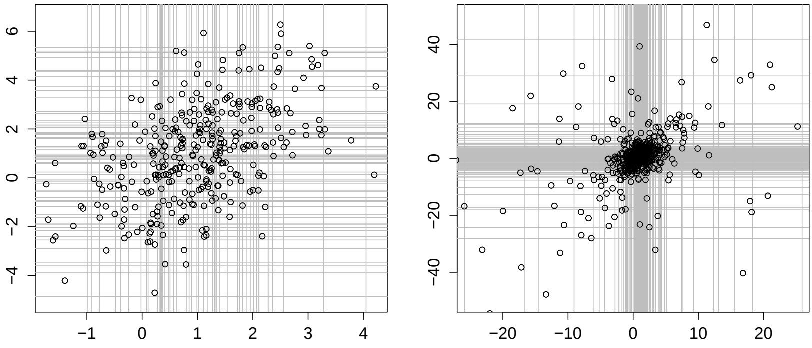


◆ – regression ● – zonoid depth ▲ – random forest

Robust imputation

15% (left) and 100% (right) from $Cauchy\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$

Robust to outliers (MCAR) Robust to disribution (MCAR)

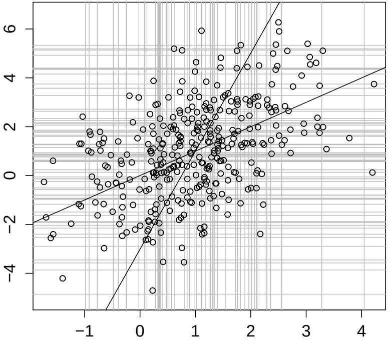


◆ – regression ● – Tukey depth

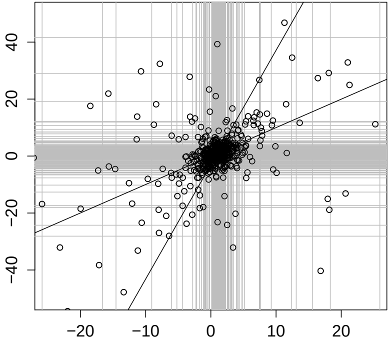
Robust imputation

15% (left) and 100% (right) from $Cauchy\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$

Robust to outliers (MCAR)



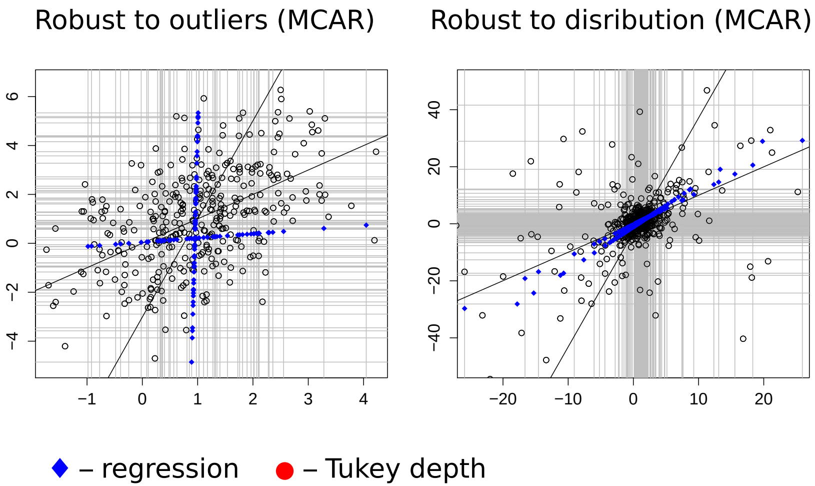
Robust to disribution (MCAR)



♦ – regression ● – Tukey depth

Robust imputation

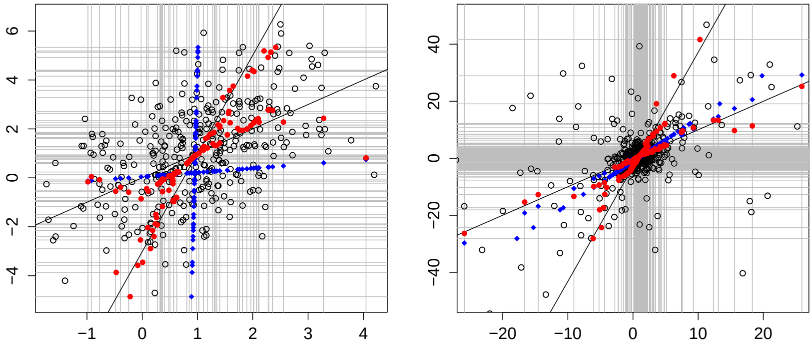
15% (left) and 100% (right) from $Cauchy\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$



Robust imputation

15% (left) and 100% (right) from $Cauchy\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 4\end{pmatrix}\right)$

Robust to outliers (MCAR) Robust to disribution (MCAR)



◆ – regression ● – Tukey depth

Contents

Statistical data depth

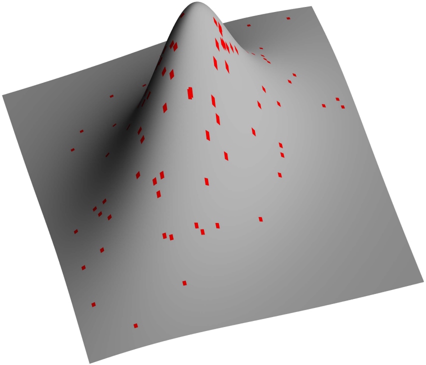
A **data depth** measures, how "close" a given point is located to the "center" of a distribution. For $\boldsymbol{x} \in \mathbb{R}^d$ and a d-variate random vector X distributed as $P \in \mathcal{M}$, a data depth is a function

$$D : \mathbb{R}^d \times \mathcal{M} \to [0, 1], (\boldsymbol{x}, P) \mapsto D(\boldsymbol{x}|P)$$

that is:

- **affine invariant**: $D(A\boldsymbol{x} + b|AX + b) = D(\boldsymbol{x}|X)$;
- **vanishing at infinity**: $\lim_{||\boldsymbol{x}|| \to \infty} D(\boldsymbol{x}|X) = 0$;
- **monotone w.r.t. the deepest point**: for any $\boldsymbol{x}^* \in \max_{\boldsymbol{x} \in \mathbb{R}^d} D(\boldsymbol{x}|X)$, any $\boldsymbol{x} \in \mathbb{R}^d$, and any $0 \leq \alpha \leq 1$ it holds: $D(\boldsymbol{x}|X) \leq D(\boldsymbol{x}^* + \alpha(\boldsymbol{x} - \boldsymbol{x}^*)|X)$;
- **upper semicontinuous in x**: the upper-level sets, **depth regions**, $D_\alpha(X) = \{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}|X) \geq \alpha\}$ are closed for all α;
- (**quasiconcave in x**): the upper-level sets are convex for all α.

Mahalanobis depth (Mahalanobis, 1936)

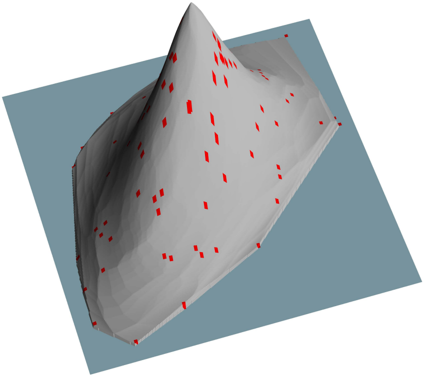


- Mahalanobis distance:
  $(d^M)^2(\boldsymbol{x}|X) =$
  $(\boldsymbol{x} - \boldsymbol{\mu}_X)'\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_X)$

- **Mahalanobis depth:**
  $D^M(\boldsymbol{x}|X) = \frac{1}{1+(d^M)^2(\boldsymbol{x}|X)}$

- \- *moment estimates* for $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$,
  or robust estimates:
  \- *minimum volume ellipsoid*,
  \- *minimum covariance determinant*.

Zonoid depth (Koshevoy, Mosler, 1997)



- Define the **zonoid trimmed region**:

  $$D^z_\alpha(X) = \Big\{ \int_{\mathbb{R}^d} \boldsymbol{x} g(\boldsymbol{x}) dP : \quad g : \mathbb{R}^d \mapsto [0, \tfrac{1}{\alpha}] \text{ measurable and } \int_{\mathbb{R}^d} g(\boldsymbol{x}) dP = 1 \Big\}$$

  for $\alpha \in (0, 1]$,

  $$D^z_0(X) = \mathrm{cl}\big(\cup_{\alpha \in (0,1]} D^z_\alpha(X)\big)$$

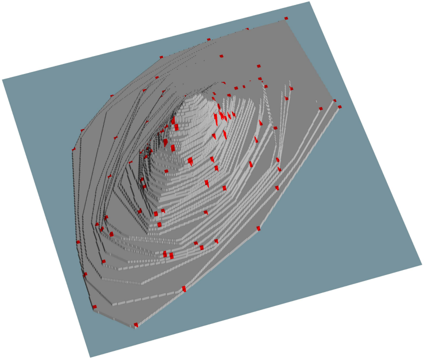
  for $\alpha = 0$.

- **Zonoid depth** is then:

  $$D^z(\boldsymbol{x}|X) = \sup\big\{\alpha : \boldsymbol{x} \in D^z_\alpha(X)\big\},$$

  or 0 if no such $D^z_\alpha(X)$ exists for $\alpha \in [0, 1]$.

Tukey depth (Tukey, 1975)



- The smallest probability of X in a closed halfspace containing $\boldsymbol{x}$:

  $$D^T(\boldsymbol{x}|X) = \inf\{P_X(H) : \boldsymbol{x} \in H, H \text{ a closed halfspace}\}$$

- Empirical version of the **Tukey depth**:

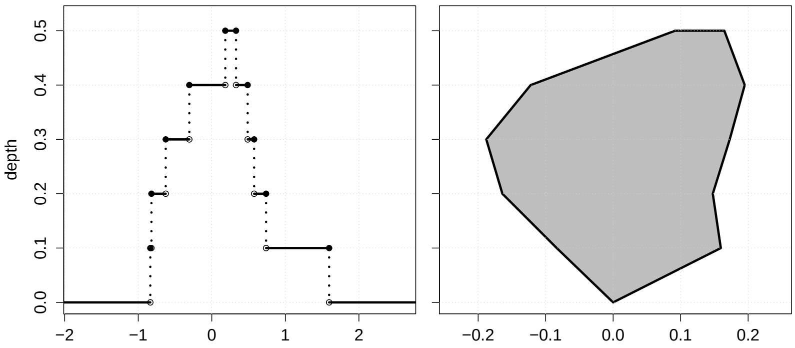
  $$D^{T(n)}(\boldsymbol{x}|\boldsymbol{X}) = \frac{1}{n} \min_{\boldsymbol{r} \in S^{d-1}} \sharp\{i : \boldsymbol{x}_i' \boldsymbol{r} \geq \boldsymbol{x}' \boldsymbol{r}\}$$

Depth lift

By adding a real dimension to the trimmed regions $D_\alpha(X)$, $\alpha \in [0, \alpha_{max}]$, define the **depth lift**:

$$D(X) = \{(\alpha, \boldsymbol{y}) \in [0, \alpha_{max}] \times \mathbb{R}^d : \boldsymbol{y} = \alpha \boldsymbol{x}, \boldsymbol{x} \in D_\alpha(X), \alpha \in [0, \alpha_{max}]\}$$

Example: Tukey depth lift (10 points)

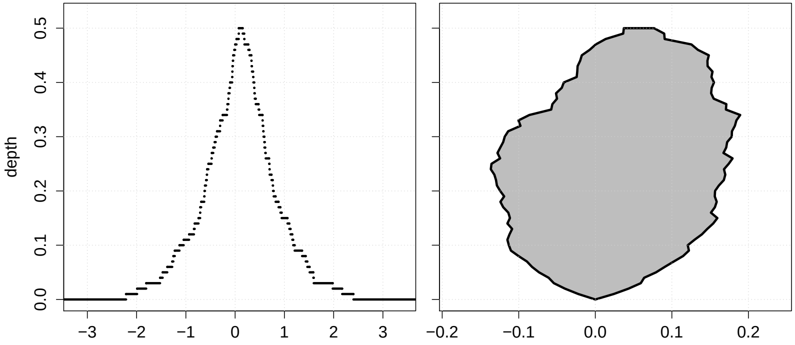


Depth lift

By adding a real dimension to the trimmed regions $D_\alpha(X)$, $\alpha \in [0, \alpha_{max}]$, define the **depth lift**:

$$D(X) = \{(\alpha, \boldsymbol{y}) \in [0, \alpha_{max}] \times \mathbb{R}^d : \boldsymbol{y} = \alpha \boldsymbol{x}, \boldsymbol{x} \in D_\alpha(X), \alpha \in [0, \alpha_{max}]\}$$

Example: Tukey depth lift (100 points)

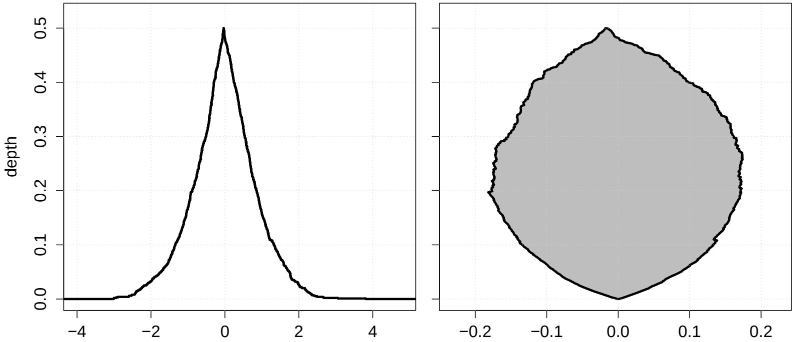


Depth lift

By adding a real dimension to the trimmed regions $D_\alpha(X)$, $\alpha \in [0, \alpha_{max}]$, define the **depth lift**:

$$D(X) = \{(\alpha, \boldsymbol{y}) \in [0, \alpha_{max}] \times \mathbb{R}^d : \boldsymbol{y} = \alpha \boldsymbol{x}, \boldsymbol{x} \in D_\alpha(X), \alpha \in [0, \alpha_{max}]\}$$

Example: Tukey depth lift (1000 points)

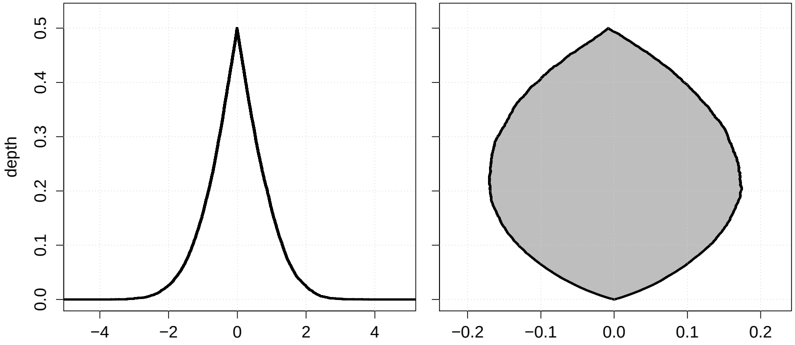


Depth lift

By adding a real dimension to the trimmed regions $D_\alpha(X)$, $\alpha \in [0, \alpha_{max}]$, define the **depth lift**:

$$D(X) = \{(\alpha, \boldsymbol{y}) \in [0, \alpha_{max}] \times \mathbb{R}^d : \boldsymbol{y} = \alpha \boldsymbol{x}, \boldsymbol{x} \in D_\alpha(X), \alpha \in [0, \alpha_{max}]\}$$

Example: Tukey depth lift (10000 points)



Contents

The idea

- A **compromise** between **global** (regression) and **local** (kNN) imputation.
- Unsureness in normality **away from the center** appeals to **tending to the unconditional mean** rather than the conditional one.
- Cautious to **introduce new variance** into the data.

Solution: min. **covariance determinant**. Why **not?**

- take into account **data geometry**,
- get rid of **moment assumptions**.

Given a data set $\boldsymbol{X} = (\boldsymbol{X}_{miss}, \boldsymbol{X}_{obs})$, impute with $\boldsymbol{Y}$:

$$\boldsymbol{Y} \in \underset{\boldsymbol{Y}_{obs}=\boldsymbol{X}_{obs}, \boldsymbol{Y}_{miss} \in \mathbb{R}^{\#\boldsymbol{X}_{miss}}}{\operatorname{argmin}} \operatorname{mes}\left(D^{(n)}(\boldsymbol{Y})\right),$$

where "mes(A)" denotes the Lebesgue measure of A.

Properties

Finite-sample extreme-value theorem

Let $\boldsymbol{X} = (\boldsymbol{X}_{miss}, \boldsymbol{X}_{obs})$ be a data set with missing values. Then, for any $0 < M < \infty$,

$$f(\boldsymbol{Y}_{miss}) = \text{mes}\Big(D^{(n)}(\boldsymbol{Y}|\boldsymbol{Y}_{obs} = \boldsymbol{X}_{obs})\Big)$$

attains its minimum on $[-M, M]^{\#\boldsymbol{X}_{miss}}$ for **Tukey**, **zonoid**, and **Mahalanobis** (if $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ is invertable for all $\boldsymbol{X}_{miss} \in [-M, M]^{\#\boldsymbol{X}_{miss}}$) depths.

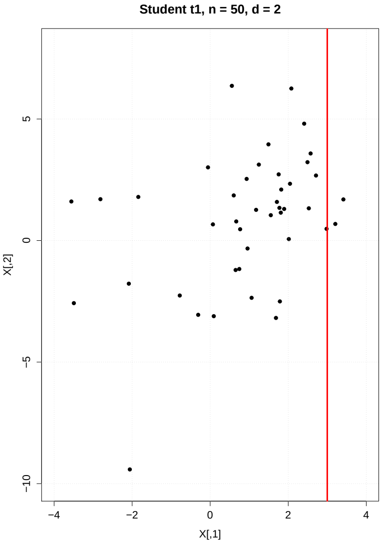
Infinite-observed extreme-value theorem

Futher, let $\boldsymbol{X}_{obs}^{(n)}$ being sampled from absolutely continuous $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, F)$. Then, for any $0 < M < \infty$, for **Tukey**, **zonoid** (finite 1st moment), and **Mahalanobis** (finite 2nd moment) depths the above holds, and for M large enough in the minimum for all $\boldsymbol{y}$ with missingness

$$\lim_{n\to\infty} \boldsymbol{y}_{miss} \xrightarrow{a.s.} \boldsymbol{\mu}_{miss} + \boldsymbol{\Sigma}_{miss,obs}\boldsymbol{\Sigma}_{obs,obs}^{-1}(\boldsymbol{y}_{obs} - \boldsymbol{\mu}_{obs})\,.$$

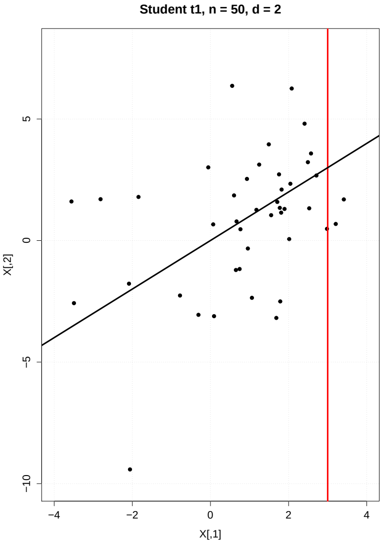
Asymptotics

-1.84206859	1.79279953
0.55599704	6.36432907
1.06121143	-2.35623912
1.75508455	2.72156758
2.01510259	0.05933206
1.71022037	1.58864924
-2.05965874	-9.41484259
3.41324550	1.69117380
1.49315060	3.95687823
2.49331460	3.22377039
2.57342204	3.58135948
....	....
0.07041123	0.66430372
0.74614531	-1.17108834
-3.49366370	-2.57489754
1.78620809	-2.50428772
2.04467078	2.33675844
2.98223544	0.48124126
1.17635526	1.26232904
0.93401823	2.53479658
1.82078645	2.09722487
2.71138114	2.67442090
-10.71645912	-12.16363642
3.00000000	NA



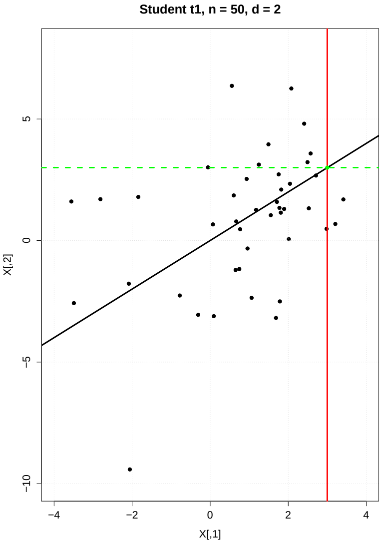
Asymptotics

-1.84206859	1.79279953
0.55599704	6.36432907
1.06121143	-2.35623912
1.75508455	2.72156758
2.01510259	0.05933206
1.71022037	1.58864924
-2.05965874	-9.41484259
3.41324550	1.69117380
1.49315060	3.95687823
2.49331460	3.22377039
2.57342204	3.58135948
....	....
0.07041123	0.66430372
0.74614531	-1.17108834
-3.49366370	-2.57489754
1.78620809	-2.50428772
2.04467078	2.33675844
2.98223544	0.48124126
1.17635526	1.26232904
0.93401823	2.53479658
1.82078645	2.09722487
2.71138114	2.67442090
-10.71645912	-12.16363642
3.00000000	NA

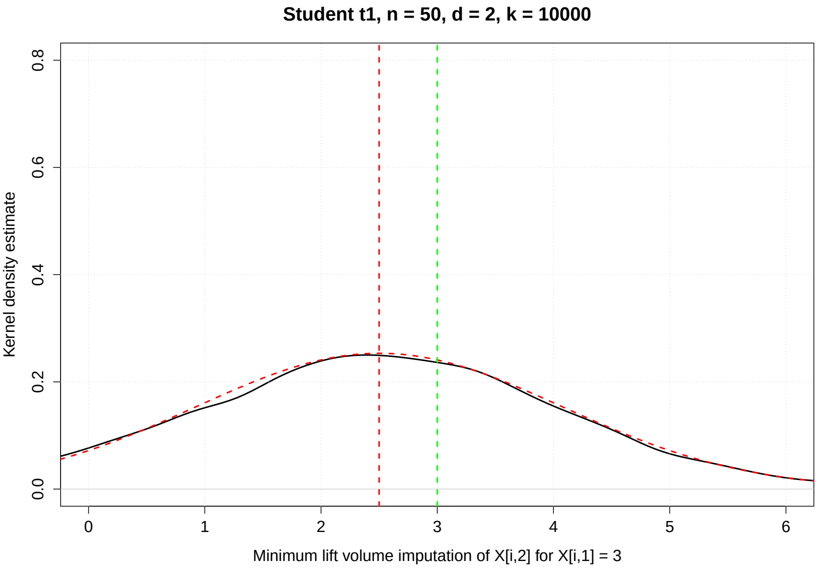


Asymptotics

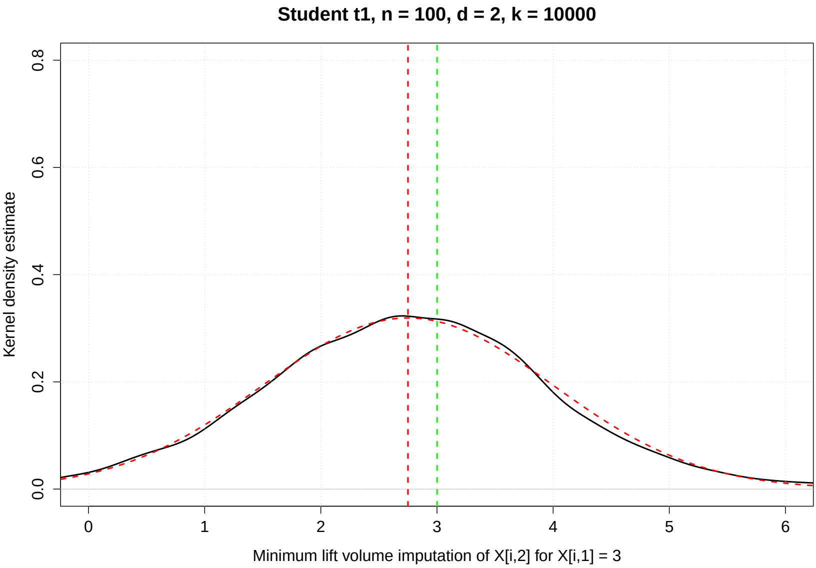
-1.84206859	1.79279953
0.55599704	6.36432907
1.06121143	-2.35623912
1.75508455	2.72156758
2.01510259	0.05933206
1.71022037	1.58864924
-2.05965874	-9.41484259
3.41324550	1.69117380
1.49315060	3.95687823
2.49331460	3.22377039
2.57342204	3.58135948
....	....
0.07041123	0.66430372
0.74614531	-1.17108834
-3.49366370	-2.57489754
1.78620809	-2.50428772
2.04467078	2.33675844
2.98223544	0.48124126
1.17635526	1.26232904
0.93401823	2.53479658
1.82078645	2.09722487
2.71138114	2.67442090
-10.71645912	-12.16363642
3.00000000	3.00000000



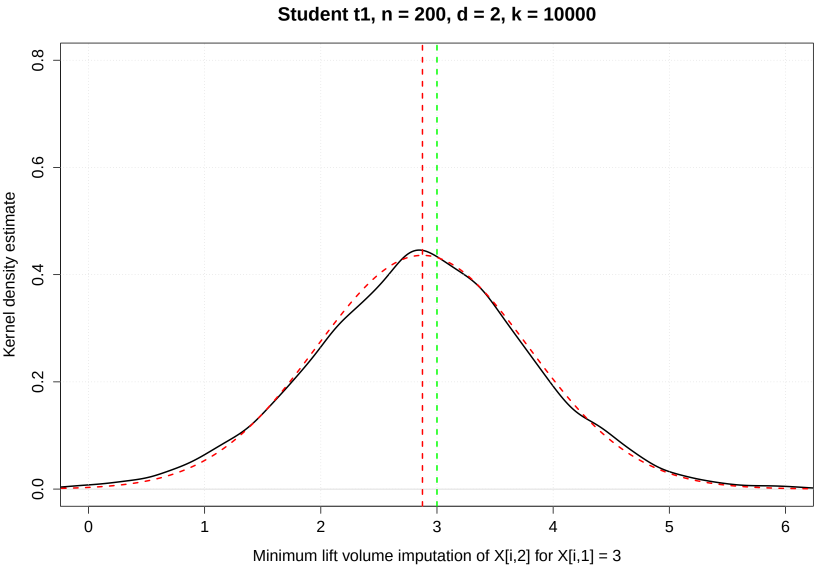
Asymptotics (Tukey depth)



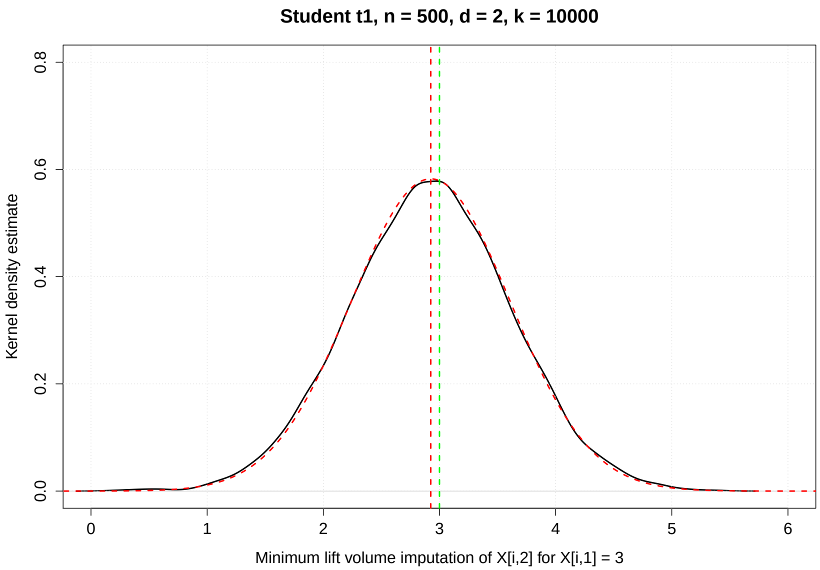
Asymptotics (Tukey depth)



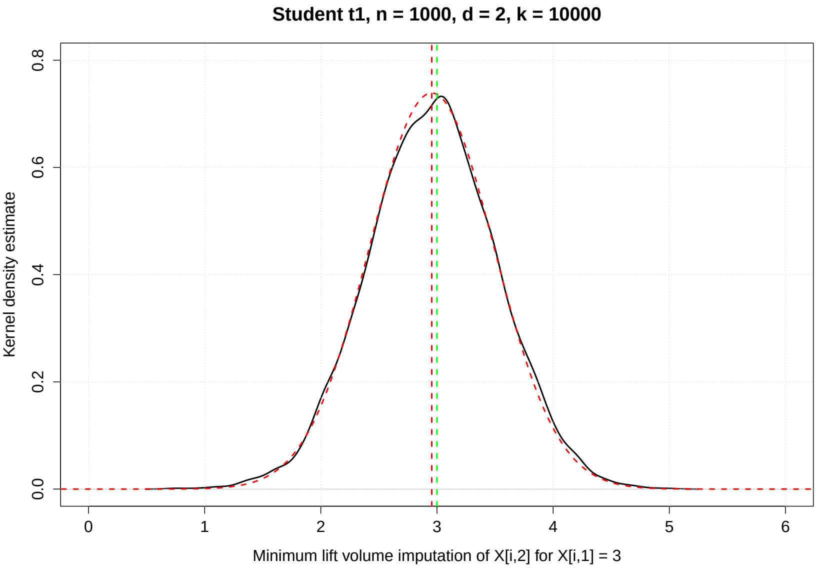
Asymptotics (Tukey depth)



Asymptotics (Tukey depth)



Asymptotics (Tukey depth)



Special case: Mahalanobis depth

Impute $\boldsymbol{X} = (\boldsymbol{X}_{miss}, \boldsymbol{X}_{obs})$ in $\mathbb{R}^d$ with $\boldsymbol{Y}$ so that for each $\boldsymbol{y}$ having missing values holds:

$$\boldsymbol{y}_{miss} = \underset{\boldsymbol{y}_{obs} = \boldsymbol{x}_{obs}}{\operatorname{argmin}} \; \operatorname{mes}\bigl(D^M(\boldsymbol{Y})\bigr) .$$

Then for each $\boldsymbol{y}$ having missing values holds as well:

- $\boldsymbol{y}$ is imputed with the **conditional mean**:

  $$\boldsymbol{y}_{miss} = \boldsymbol{\mu}(\boldsymbol{Y})_{miss} + \boldsymbol{\Sigma}(\boldsymbol{Y})_{miss,obs} \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y})_{obs,obs} (\boldsymbol{y}_{obs} - \boldsymbol{\mu}(\boldsymbol{Y})_{obs}) ,$$

- $\boldsymbol{y}$ is imputed with single-output **regression**,
- $\boldsymbol{y}$ is imputed with **regularized PCA** by Josse & Husson (2012) with any $0 < \sigma^2 \leq \lambda_d$:

  $$\boldsymbol{y}_{miss(i)} = \sum_{j=1}^{d} \sqrt{\frac{\lambda_j - \sigma^2}{\lambda_j}} \boldsymbol{u}_j \boldsymbol{V}_{miss(i),j}$$

  with $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}$ being the centered SVD,
- $\boldsymbol{y}$ is imputed with **maximum depth**:

  $$\boldsymbol{y}_{miss} = \underset{\boldsymbol{y}_{obs} = \boldsymbol{x}_{obs}}{\operatorname{argmax}} \; D^M(\boldsymbol{y} | \boldsymbol{Y}) .$$

Computation

Direct minimization requires evaluation of empirical version of $\text{mes}\big(D^{(n)}(\boldsymbol{X})\big)$ for n points in $\mathbb{R}^d$, namely

- for Mahalanobis depth:

$$\text{mes}\big(D^{M(n)}(\boldsymbol{X})\big) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \prod_{i=1}^{\lfloor \frac{d}{2} \rfloor} \frac{2i + 2(d \bmod 2)}{i + \frac{1}{2} + d \bmod 2} \times$$
$$\times \big(1 - (1 - \frac{\pi}{8})\mathbb{I}(d \bmod 2 \neq 0)\big)\sqrt{\|\boldsymbol{\Sigma}_{\boldsymbol{X}}\|},$$

- for zonoid depth:

$$\text{mes}\big(D^{z(n)}(\boldsymbol{X})\big) = \frac{1}{n^{d+1}} \sum_{\{i_0,\dots,i_d\} \subset \{1,\dots,n\}} \big\|\big((1, \boldsymbol{x}'_{i_0})', \dots, (1, \boldsymbol{x}'_{i_d})'\big)\big\|,$$

- for Tukey depth:

$$\text{mes}\big(D^{T(n)}(\boldsymbol{X})\big) = \sum_{i=1}^{n\alpha_{max}} \frac{i^{d+1} - (i-1)^{d+1}}{(d+1)n^{d+1}} \text{mes}\big(D^{T(n)}_{\frac{i}{n}}(\boldsymbol{X})\big).$$

Computation

Denote

$$\alpha^* = \inf_{\alpha \in (0;1)} \left\{ \alpha \,|\, int\, D_\alpha(\boldsymbol{X}) \cap \{ \boldsymbol{z} \,|\, \boldsymbol{z} \in \mathbb{R}^d \,,\, \boldsymbol{z}_{obs} = \boldsymbol{x}_{obs} \} = \varnothing \right\},$$

and impute $\boldsymbol{x}$ having missing values with

$$\boldsymbol{y} = ave\Big(\operatorname*{arg\,min}_{\boldsymbol{u} \in \mathbb{R}^d \,,\, \boldsymbol{u}_{obs} = \boldsymbol{x}_{obs}} \{ \|\boldsymbol{u} - \boldsymbol{v}\| \,|\, \boldsymbol{v} \in \mathbb{R}^d \,,\, \boldsymbol{v} \in D_{\alpha^*}(\boldsymbol{X}), \| \} \Big).$$

For continuous depth this is equivalent to:

$$\boldsymbol{y} = \operatorname*{arg\,max}_{\boldsymbol{z} \in \mathbb{R}^d \,,\, \boldsymbol{z}_{obs} = \boldsymbol{x}_{obs}} D(\boldsymbol{z}|\boldsymbol{X}).$$

After an arbitrary **initialization** continue imputation for each point **iteratively** untill **convergence**.

Contents

Student t-distribution (MCAR)

$$n = 50,\ \mu = (1,1,1)',\ \Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 4 & 8 \end{pmatrix},\ \text{miss} = 25\%.$$

MSE	Normal	$t\,10$	$t\,5$	$t\,3$	$t\,2$	Cauchy
D.Tuk	2.773 (0.8294)	3.392 (1.129)	4.401 (1.742)	6.461 (3.570)	11.49 (8.006)	135.5 (164.7)
D.Zon	2.690 (0.8251)	3.360 (1.161)	4.451 (1.823)	6.844 (4.176)	13.08 (9.493)	179.8 (215.1)
D.Mah	2.719 (0.8593)	3.355 (1.158)	4.451 (1.905)	6.743 (3.872)	13.15 (9.369)	175.1 (207.1)
EM	2.532 (0.7893)	3.124 (1.046)	4.042 (1.647)	6.169 (3.390)	11.55 (8.114)	162.8 (193.1)
RF	3.143 (0.9778)	3.967 (1.402)	5.096 (2.007)	7.592 (4.170)	12.95 (9.193)	153.8 (184.8)
kNN	3.261 (1.0415)	3.919 (1.462)	5.104 (2.292)	7.266 (3.657)	13.10 (9.024)	156.6 (189.2)
rPCA2	2.719 (0.8576)	3.355 (1.158)	4.444 (1.905)	6.752 (3.894)	13.16 (9.404)	175.8 (211.8)
rPCA1	2.712 (0.8222)	3.299 (1.143)	4.310 (1.713)	6.433 (3.619)	12.24 (8.787)	159.4 (192.4)

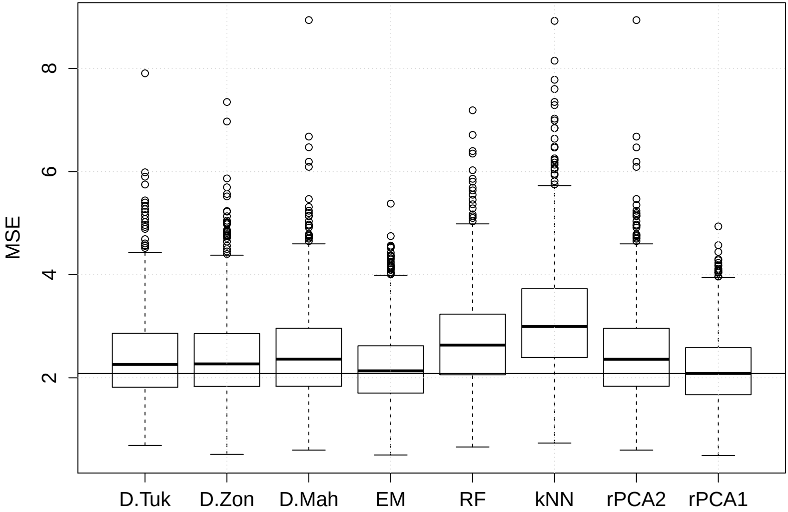
Student t-distribution + Cauchy outliers (MCAR)

$$n = 50,\ \mu = (1,1,1)',\ \Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 4 & 8 \end{pmatrix},\ \text{miss} = 25\%,\ \text{outl} = 15\%.$$

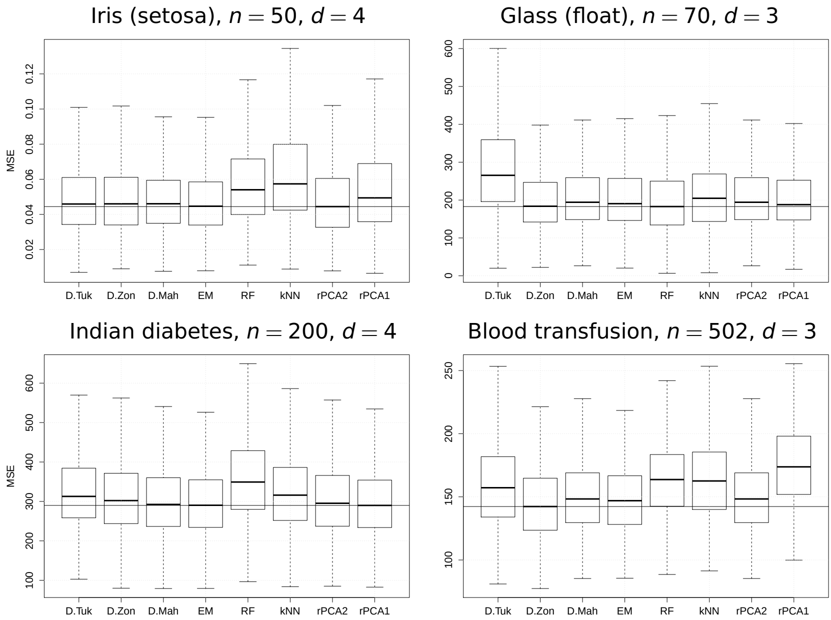
MSE	Normal	$t\,10$	$t\,5$	$t\,3$	$t\,2$	Cauchy
D.Tuk	3.025 (1.066)	3.678 (1.374)	4.676 (1.997)	6.555 (3.563)	12.03 (8.692)	125.5 (154.4)
D.Zon	3.480 (1.463)	4.420 (1.895)	5.371 (2.576)	8.115 (4.784)	14.78 (11.61)	162.0 (198.5)
D.Mah	3.787 (1.848)	4.519 (2.136)	5.614 (2.784)	8.150 (5.013)	14.86 (11.53)	169.8 (210.5)
EM	3.516 (1.673)	4.288 (1.950)	5.261 (2.557)	7.554 (4.622)	13.82 (10.67)	153.4 (189.8)
RF	3.520 (1.204)	4.251 (1.602)	5.324 (2.276)	7.930 (4.398)	13.98 (10.30)	139.9 (172.3)
kNN	3.535 (1.227)	4.294 (1.647)	5.313 (2.285)	7.690 (4.299)	13.55 (9.275)	143.5 (174.6)
rPCA2	3.793 (1.846)	4.532 (2.139)	5.609 (2.796)	8.151 (5.013)	14.86 (11.55)	168.8 (210.3)
rPCA1	3.657 (1.774)	4.424 (2.201)	5.520 (2.736)	8.339 (5.283)	14.51 (11.36)	157.9 (197.0)

Normal distribution (MAR)

$$n = 50,\ \mu = (1,1,1)',\ \Sigma = \begin{pmatrix} 1 & 1.75 & 2 \\ 1.75 & 4 & 4 \\ 2 & 4 & 8 \end{pmatrix},\ \text{miss} = 24\%.$$



Real data (MCAR, 10%)



Contents

Conclusions and outlook

Suggested is a **framework** for **single imputation** based on **data depth**, which incorporates features of both **global** and **local** approaches, and imputes **close to data** geometry, preserves functionality under **missing at random** (MAR) mechanism, is **robust** both in sense of distribution and outliers.

Future work:

- Detailed exploration of **theoretical properties**, especially for the family of **elliptically symmetric** distributions.
- Extension to **higher-dimensional** data, **larger samples**, further **experimental study**.
- Extension to **multiple imputation**.
- Implementation in an **R-package**.

Questions? Suggestions? Ideas?

Thank you for your attention!