On algorithms for computation of the Tukey depth

Rainer Dyckerhoff[a] Xiaohui Liu[b] Karl Mosler[a]
Pavlo Mozharovskyi[c]

[a]Institute of Econometrics and Statistics, University of Cologne
[b]School of Statistics,
Jiangxi University of Finance and Economics;
Research Center of Applied Statistics,
Jiangxi University of Finance and Economics
[c]LTCI, Telecom Paris, Institut Polytechnique de Paris

LTCI Data Science seminar

Paris, June 6, 2019

Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

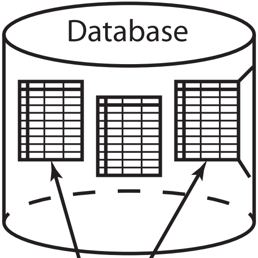
Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

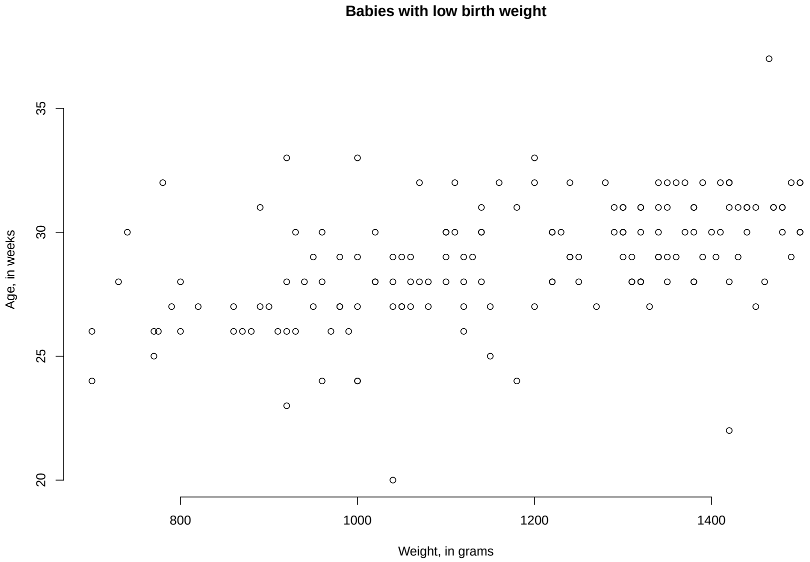
Multivariate data



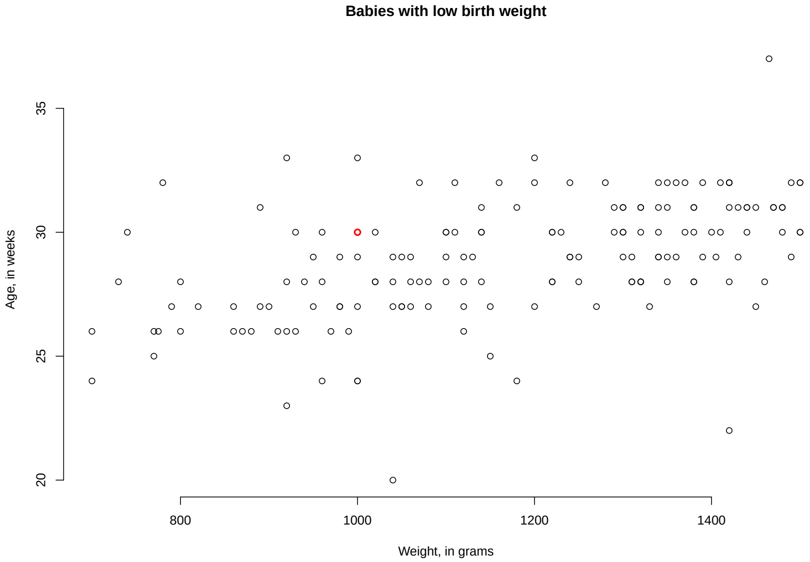
Tables containing data:
- rows = objects
- columns = properties

	Weight	Age
Subject 1	1350	32
Subject 2	1500	32
Subject 162	1320	28
Subject 161	1150	27

Tukey (=halfspace, location) data depth



Tukey (=halfspace, location) data depth



Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

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 $P \in \mathcal{P}$, a data depth is a function

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

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 $\boldsymbol{x}^* \in \operatorname{argmax}_{\boldsymbol{x} \in \mathbb{R}^d} D(\boldsymbol{x}|X)$, 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 $D_\tau(X) = \{\boldsymbol{x} \in \mathbb{R}^d : D(\boldsymbol{x}|X) \geq \tau\}$ are closed for all τ;
- (**quasiconcave in x**): the upper-level sets are convex for all τ.

Tukey (halfspace, location) depth

Tukey (1975) — "Mathematics and the picturing of data"

Tukey depth of $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. a d-variate random vector X distributed as P is defined as the smallest probability mass of a closed halfspace containing $\mathbf{x}$:

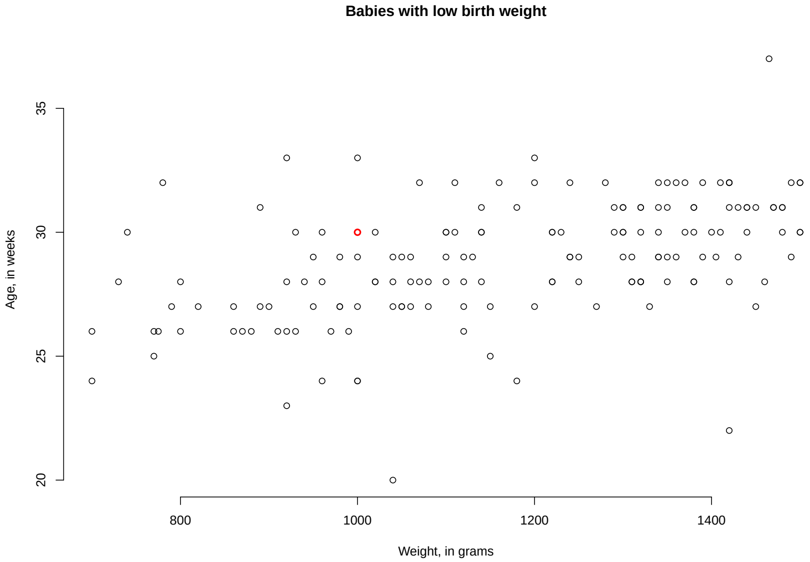
$$D^T(\boldsymbol{x}|X) = \inf\{P(H)\,:\,H \text{ is a closed halfspace, } \boldsymbol{x} \in H\},$$

and w.r.t. a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subset \mathbb{R}^d$:

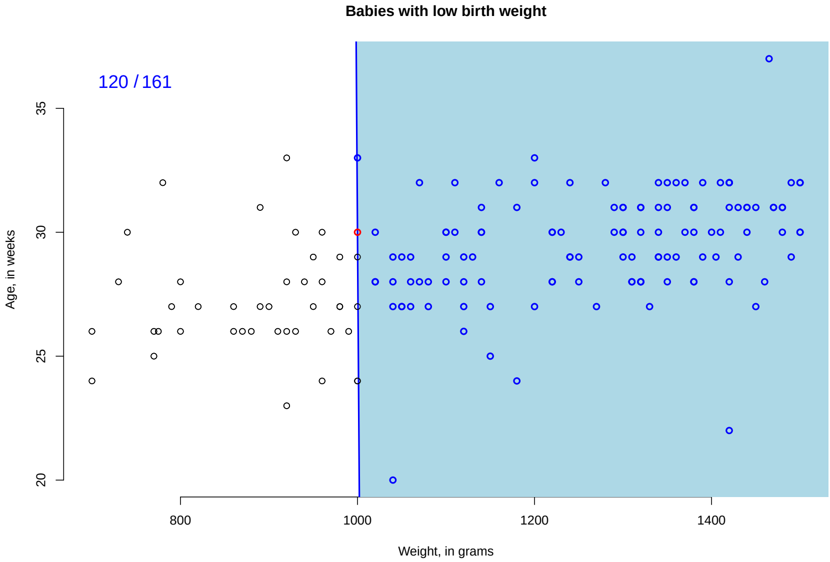
$$D^{T(n)}(\boldsymbol{x}|\boldsymbol{X}) = \frac{1}{n} \min_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \sharp\{i : \boldsymbol{u}'\boldsymbol{x}_i \geq \boldsymbol{u}'\boldsymbol{x}\}.$$

Other depth notions: **Mahalanobis** ('36), **projection** (Stahel, '81; Donoho, '82), **simplicial volume** (Oja, '83), **simplicial** (Liu, '90), **zonoid** (Koshevoy, Mosler, '97), **spatial** (Vardi, Zhang, '00; Serfling, '02) depth.

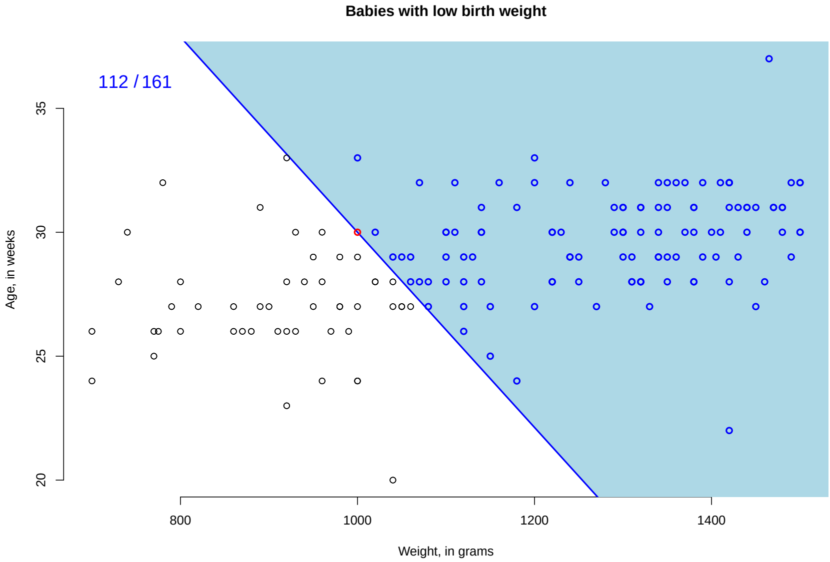
Tukey (=halfspace, location) data depth



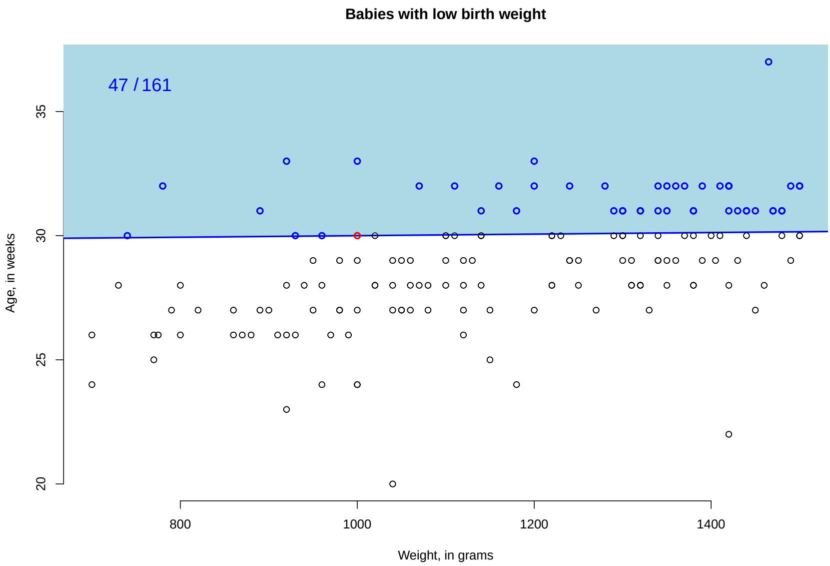
Tukey (=halfspace, location) data depth



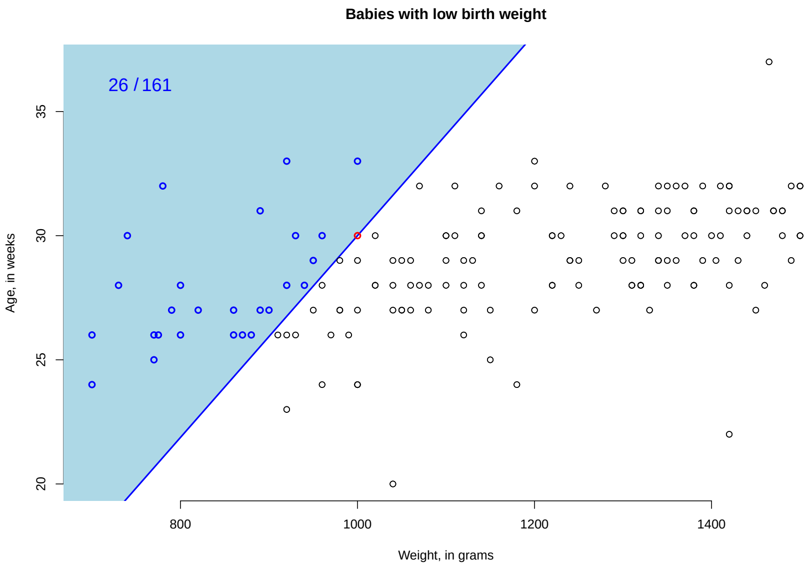
Tukey (=halfspace, location) data depth



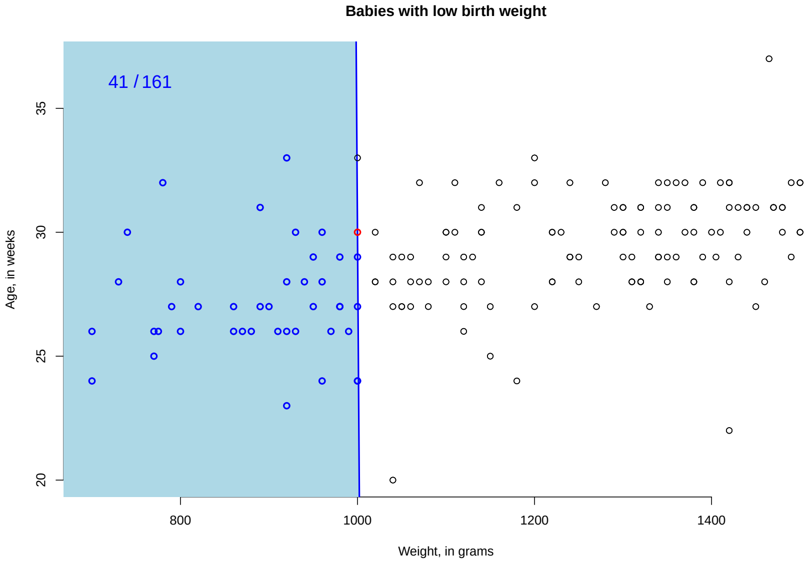
Tukey (=halfspace, location) data depth



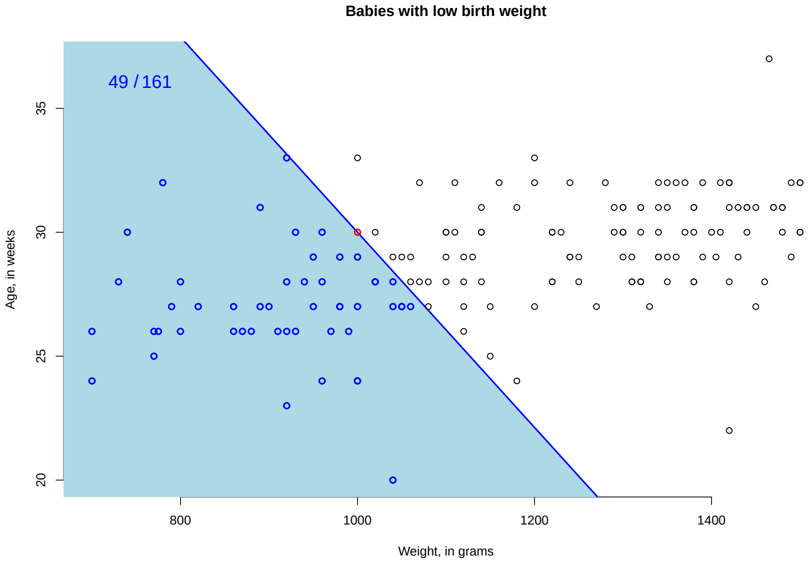
Tukey (=halfspace, location) data depth



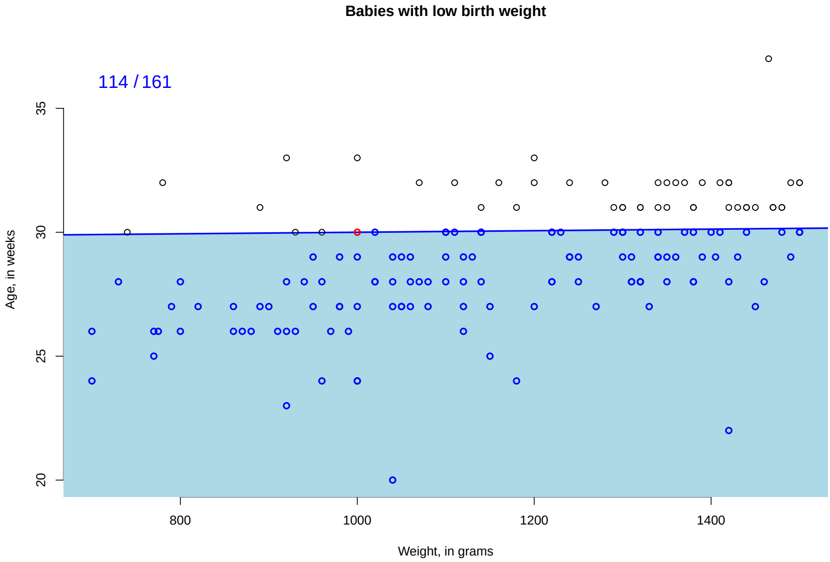
Tukey (=halfspace, location) data depth



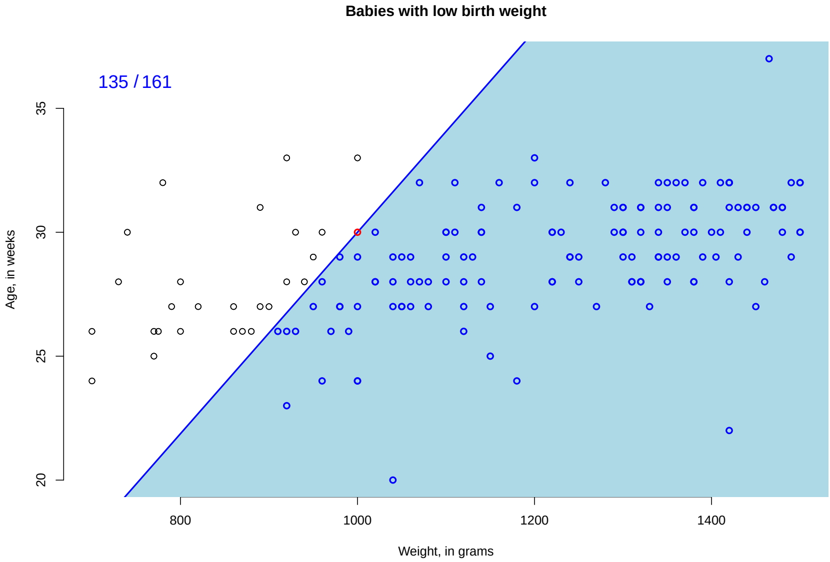
Tukey (=halfspace, location) data depth



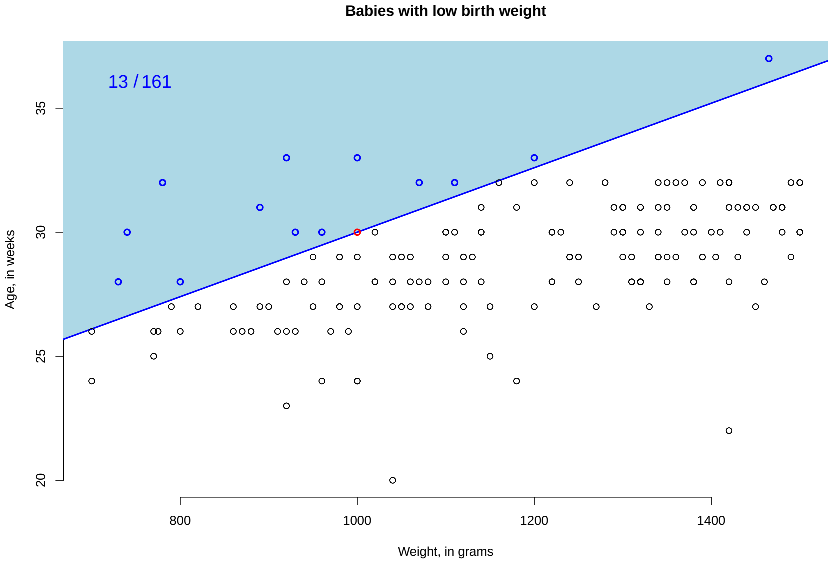
Tukey (=halfspace, location) data depth



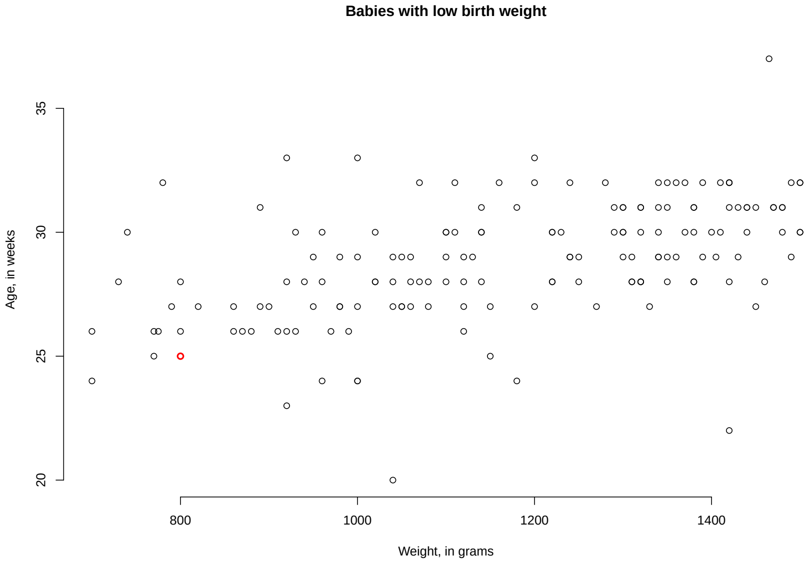
Tukey (=halfspace, location) data depth



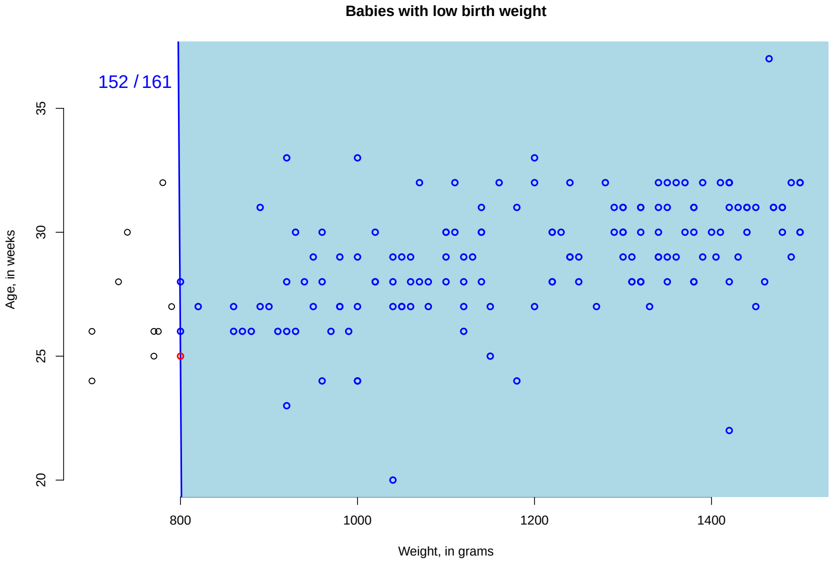
Tukey (=halfspace, location) data depth



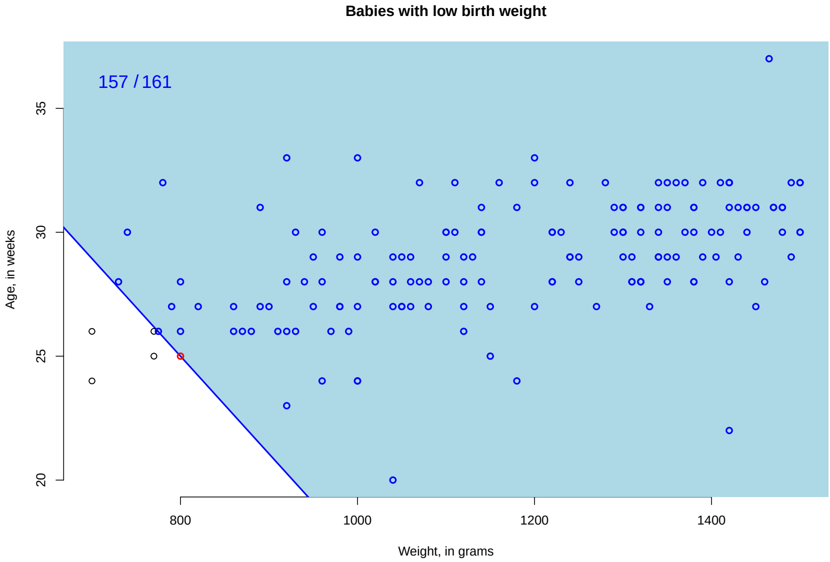
Tukey (=halfspace, location) data depth



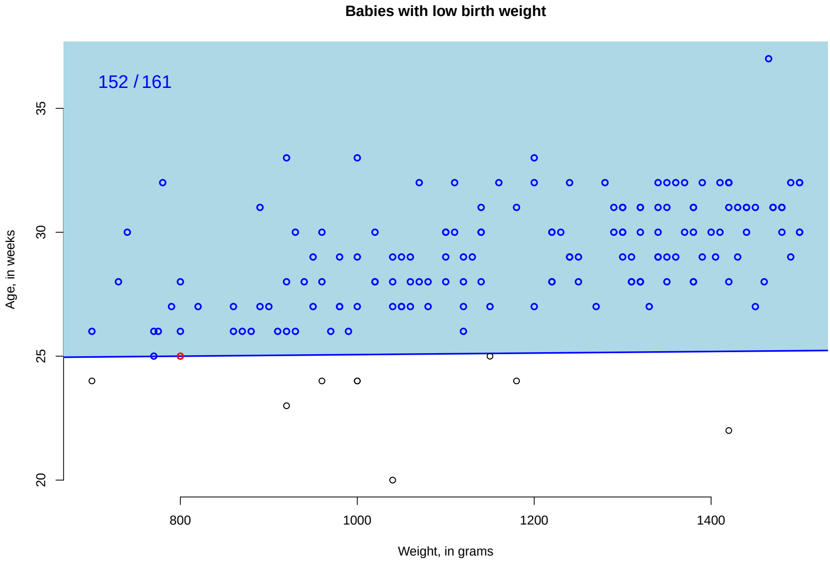
Tukey (=halfspace, location) data depth



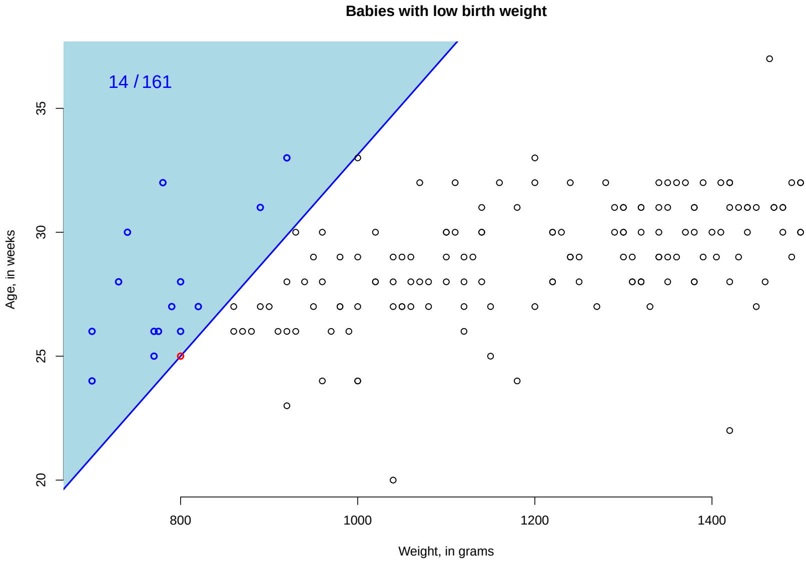
Tukey (=halfspace, location) data depth



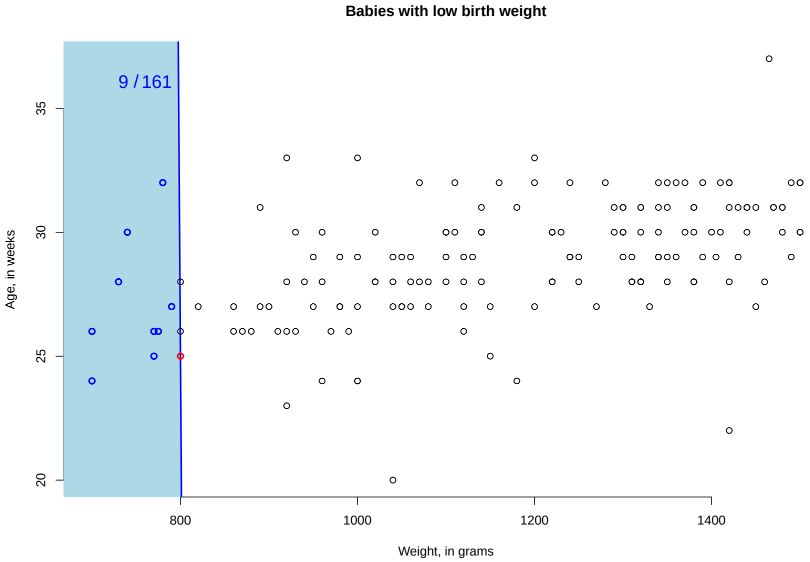
Tukey (=halfspace, location) data depth



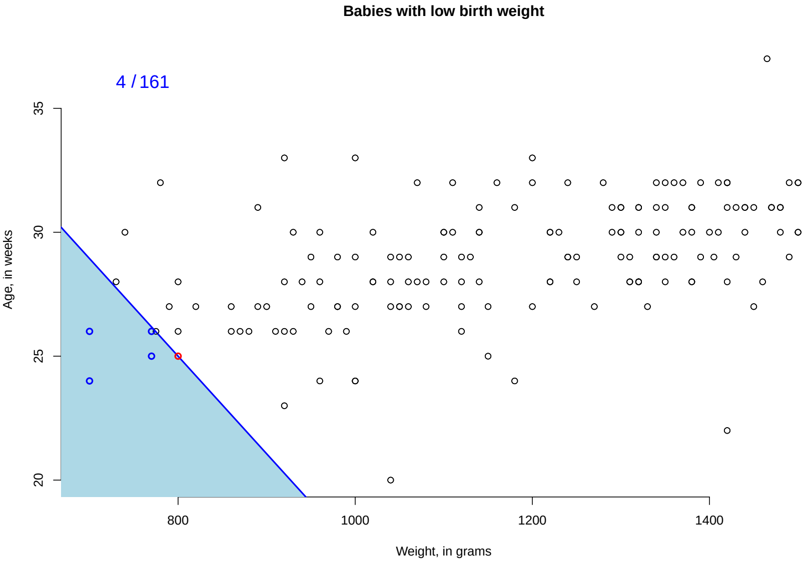
Tukey (=halfspace, location) data depth



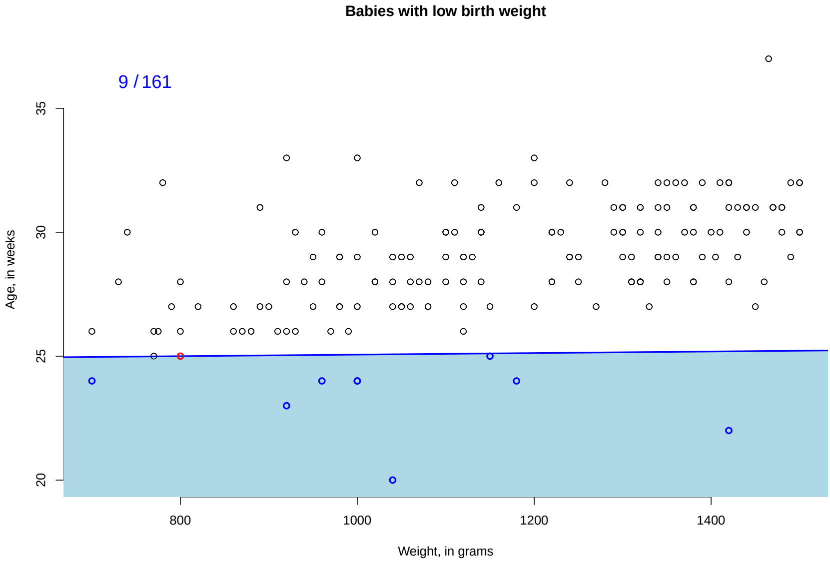
Tukey (=halfspace, location) data depth



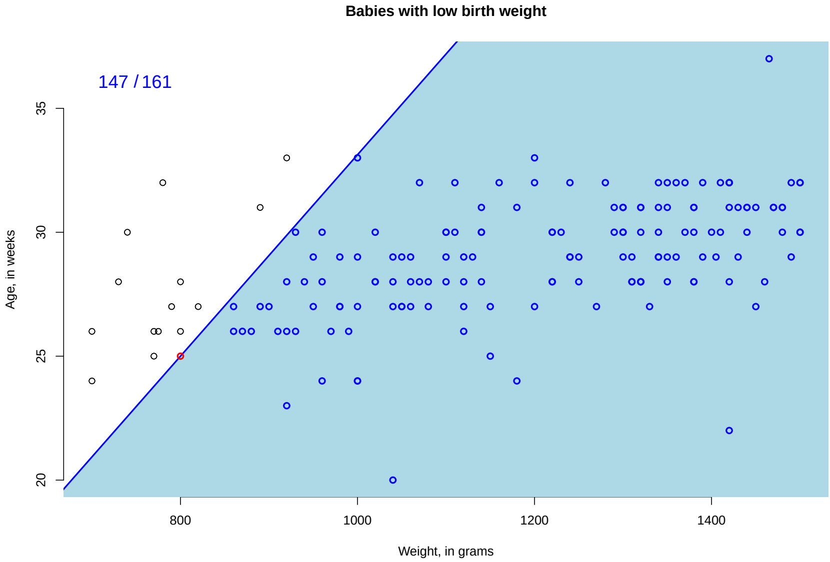
Tukey (=halfspace, location) data depth



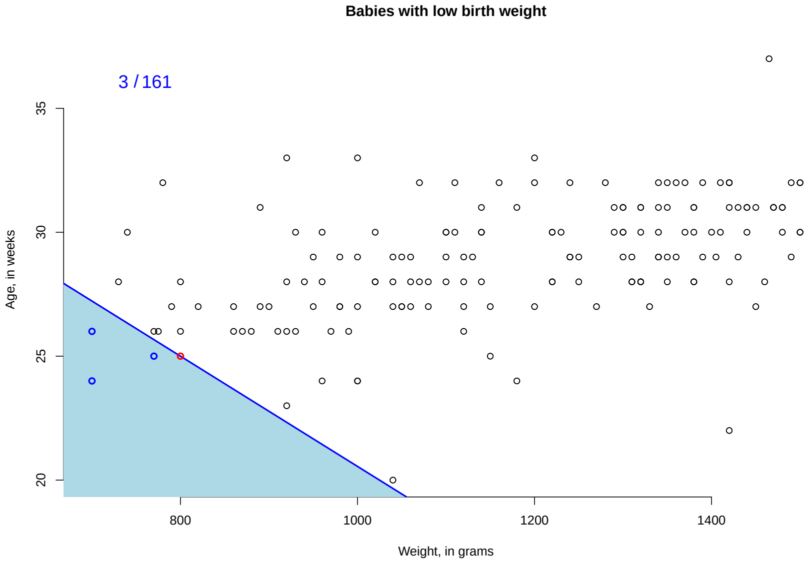
Tukey (=halfspace, location) data depth



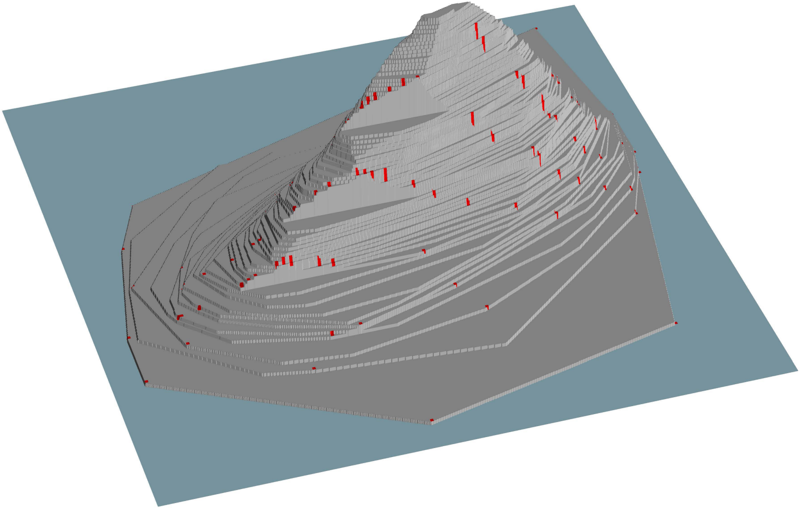
Tukey (=halfspace, location) data depth



Tukey (=halfspace, location) data depth



Tukey (=halfspace, location) data depth



Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

Tukey-trimmed regions

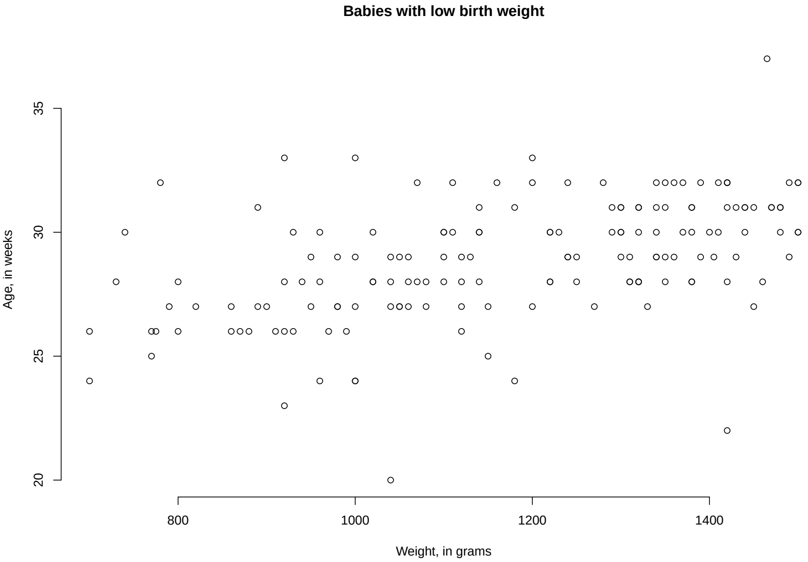
Tukey depth defines a family of (depth-)trimmed (central) regions $D_\tau^T(X)$, the upper-level sets of the depth function:

$$D_\tau^T(X) = \big\{\boldsymbol{x} \in \mathbb{R}^d \,:\, D^T(\boldsymbol{x}|X) \geq \tau\big\}.$$

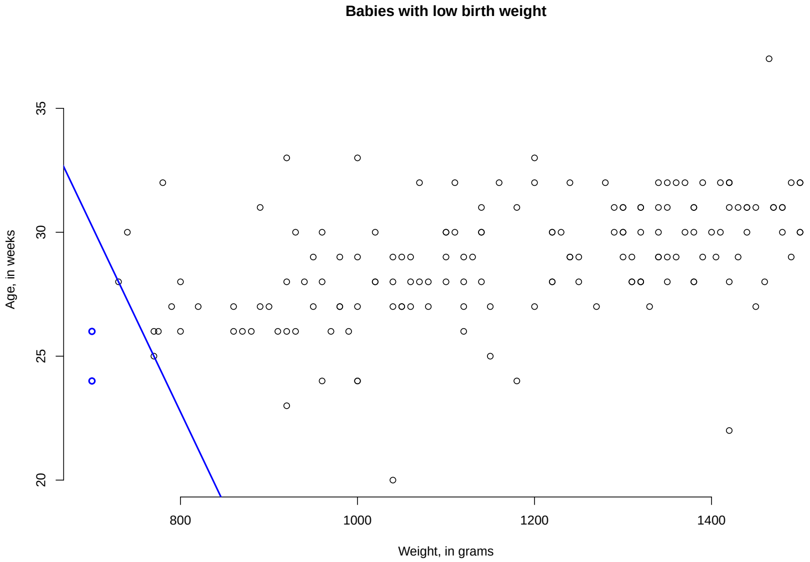
Properties:

Depth:	**Regions:**
▶ Affine invariant;	Affine equivariant;
▶ Vanishing at infinity;	Bounded;
▶ Monotone w.r.t. deepest point;	Nested;
▶ Upper-semicontinuous;	Closed;
▶ Quasiconcave.	Convex.

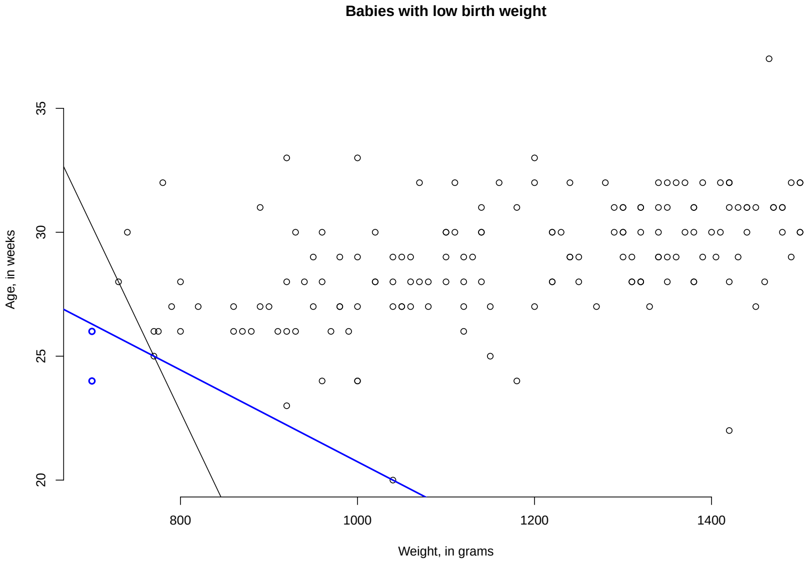
Tukey (=halfspace, location) depth-trimmed regions



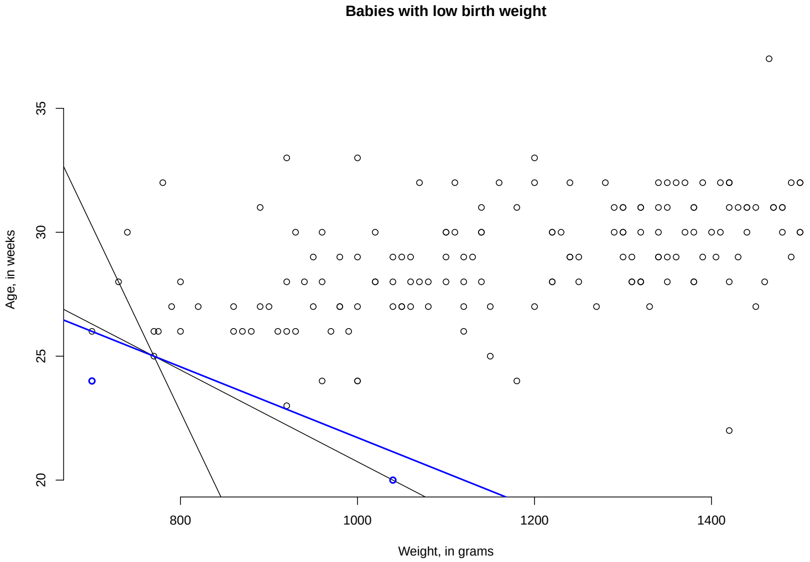
Tukey (=halfspace, location) depth-trimmed regions



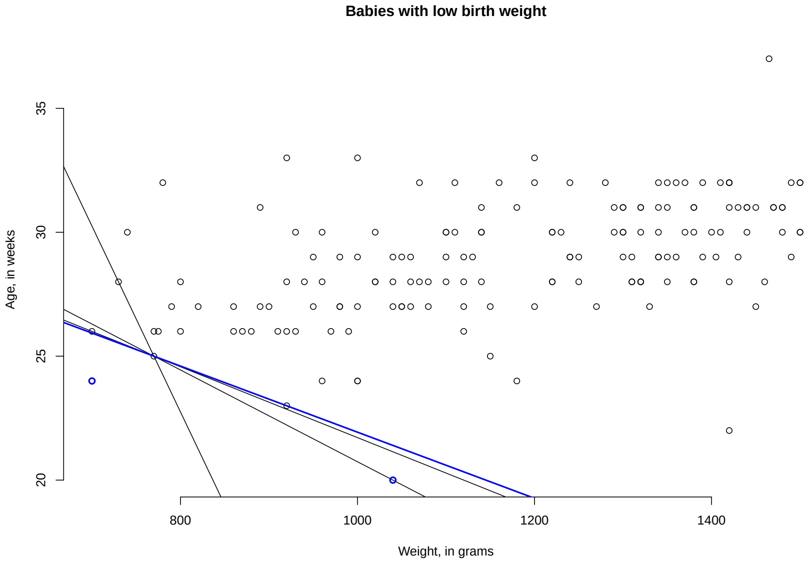
Tukey (=halfspace, location) depth-trimmed regions



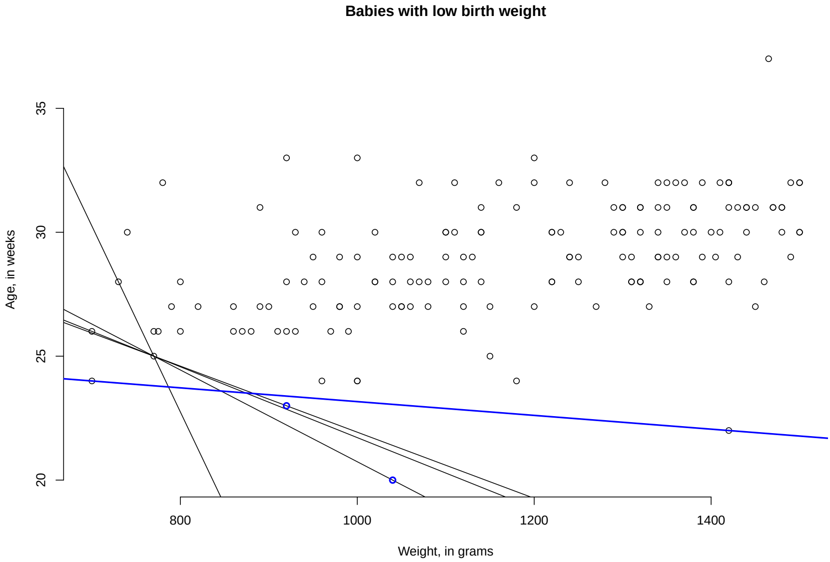
Tukey (=halfspace, location) depth-trimmed regions



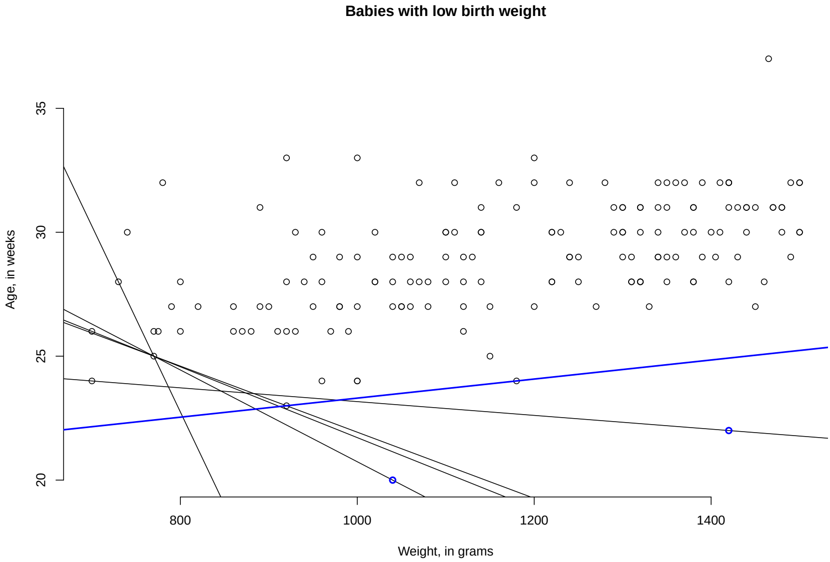
Tukey (=halfspace, location) depth-trimmed regions



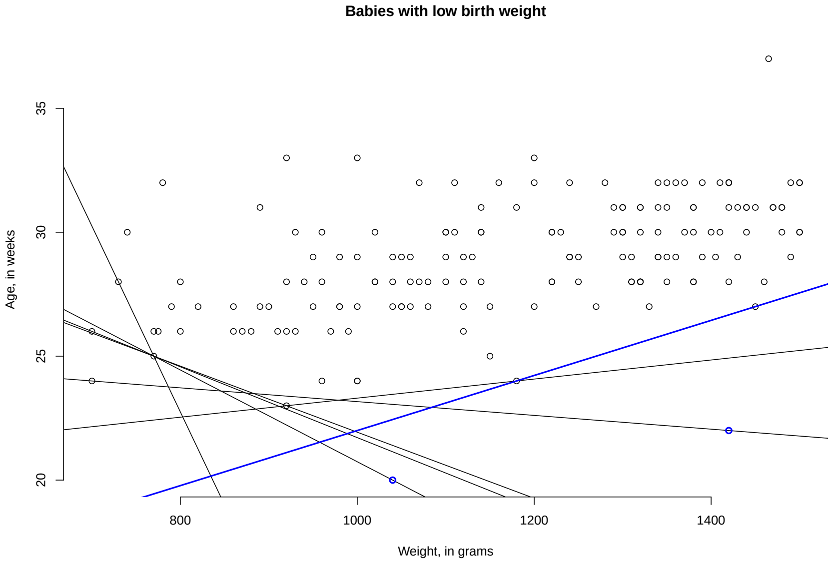
Tukey (=halfspace, location) depth-trimmed regions



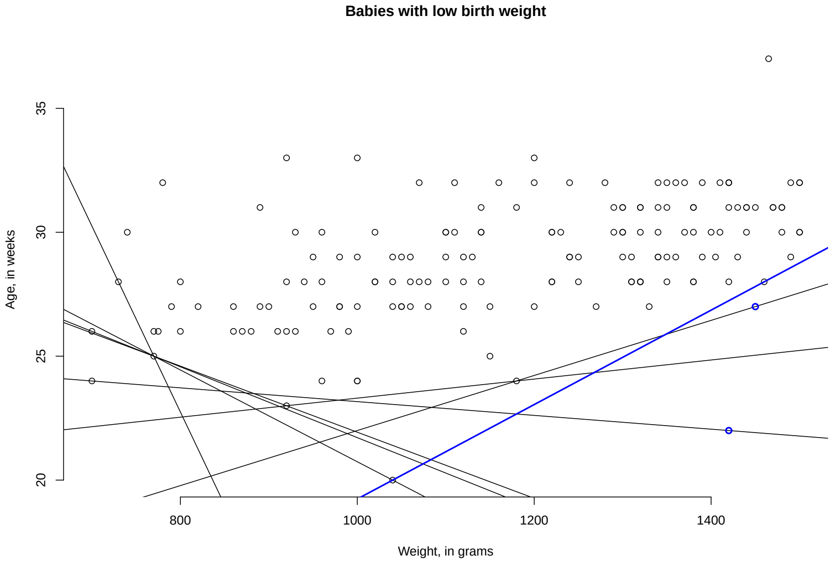
Tukey (=halfspace, location) depth-trimmed regions



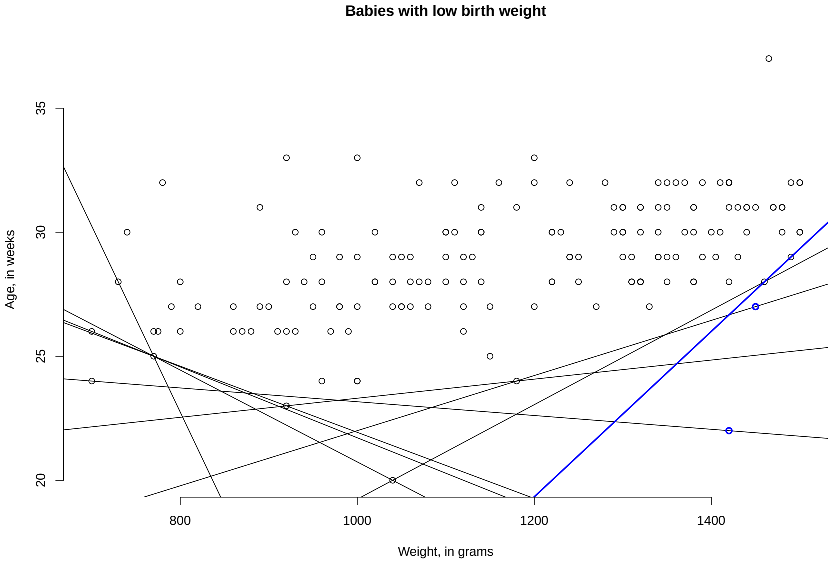
Tukey (=halfspace, location) depth-trimmed regions



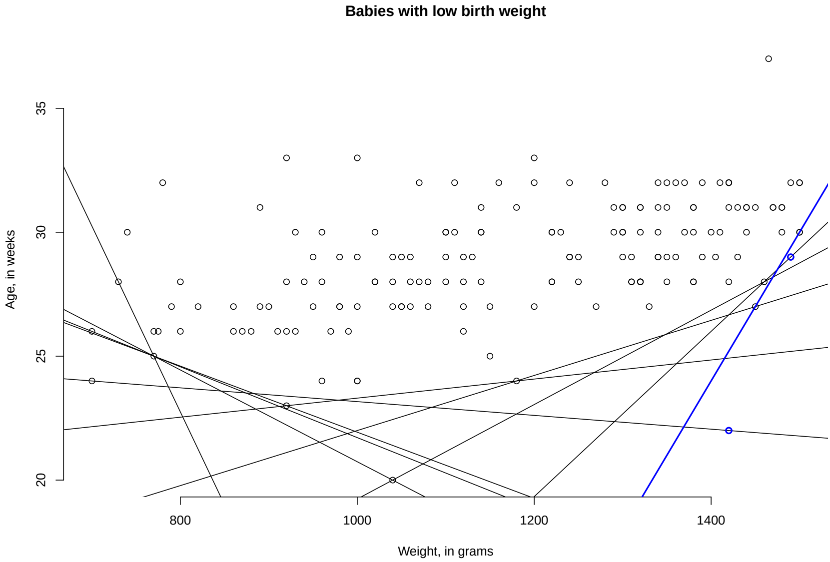
Tukey (=halfspace, location) depth-trimmed regions



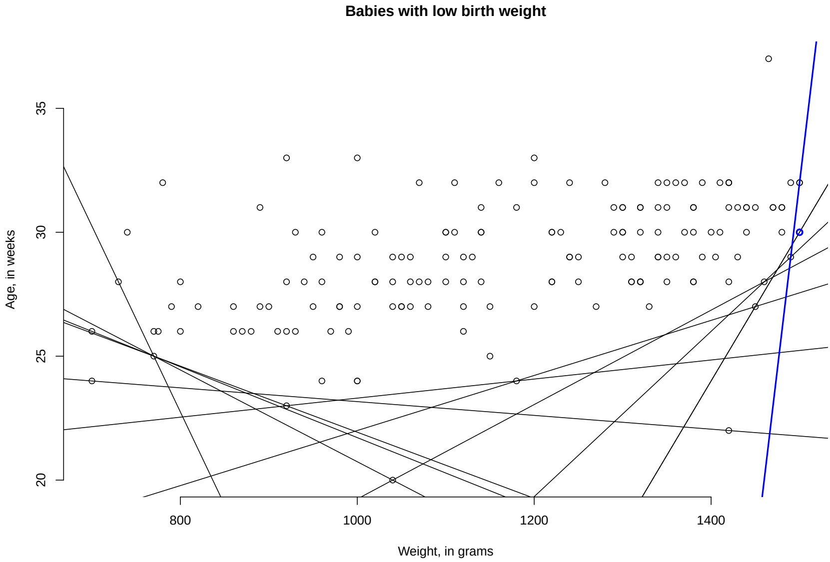
Tukey (=halfspace, location) depth-trimmed regions



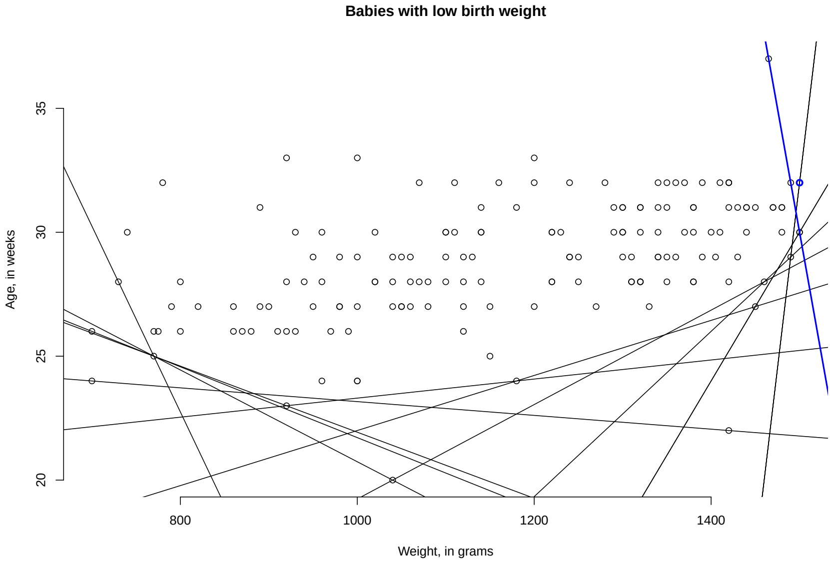
Tukey (=halfspace, location) depth-trimmed regions



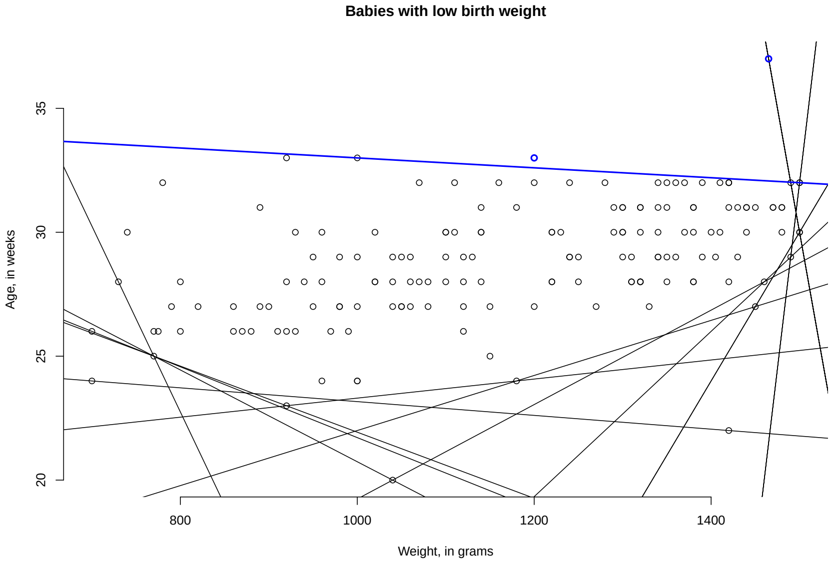
Tukey (=halfspace, location) depth-trimmed regions



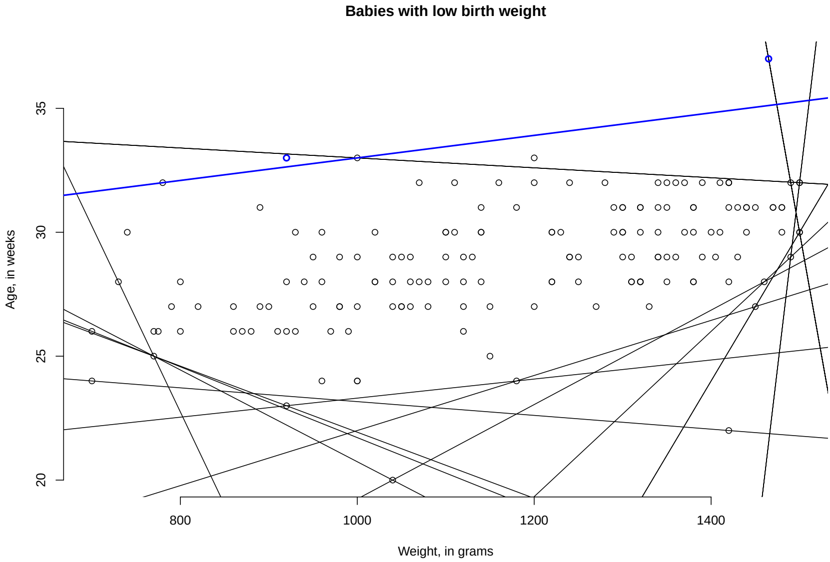
Tukey (=halfspace, location) depth-trimmed regions



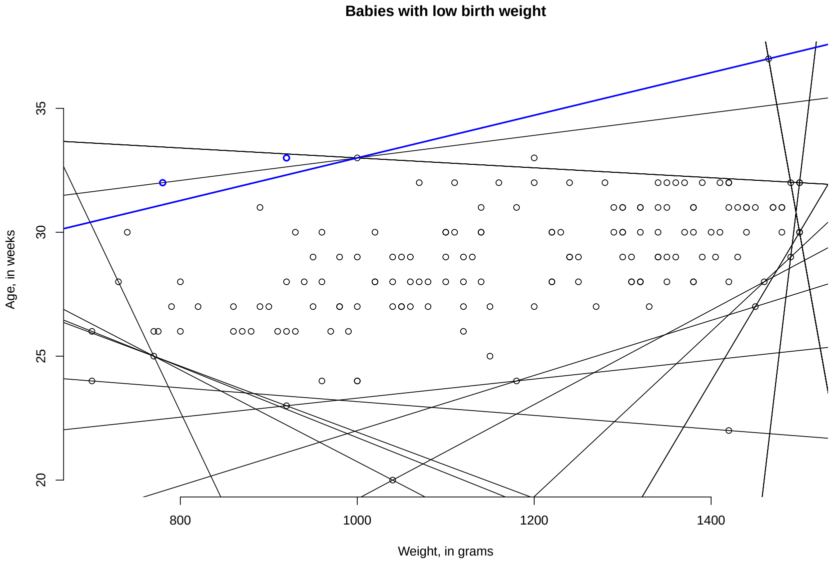
Tukey (=halfspace, location) depth-trimmed regions



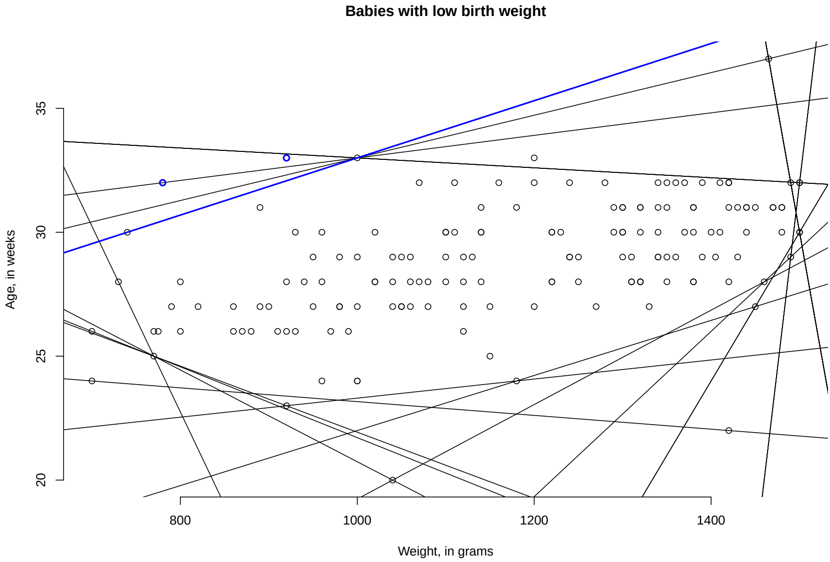
Tukey (=halfspace, location) depth-trimmed regions



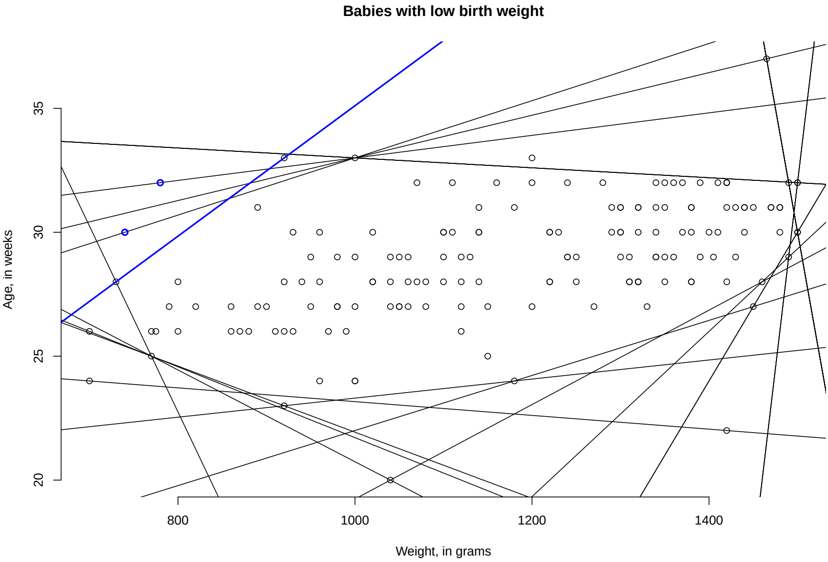
Tukey (=halfspace, location) depth-trimmed regions



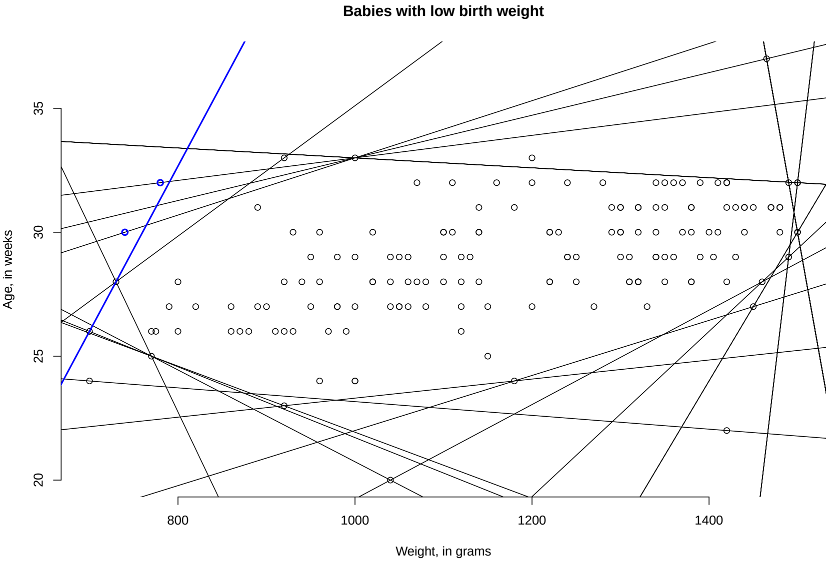
Tukey (=halfspace, location) depth-trimmed regions



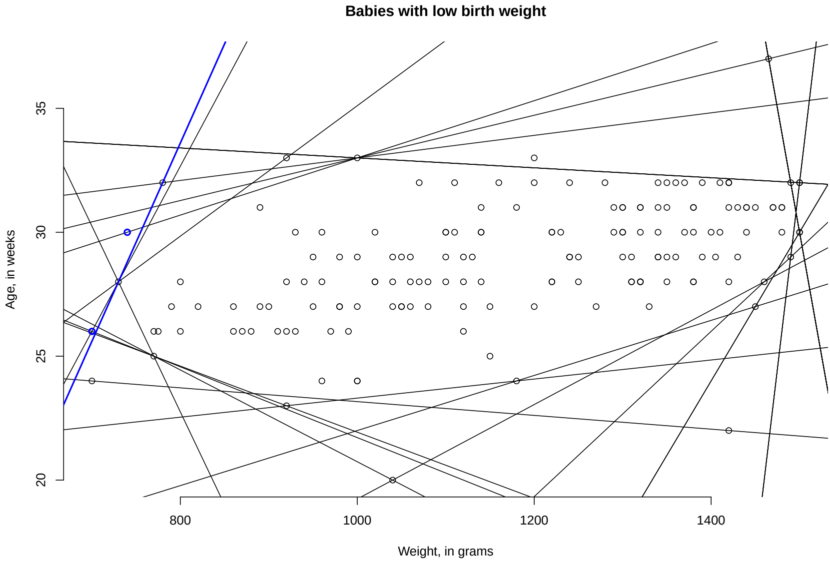
Tukey (=halfspace, location) depth-trimmed regions



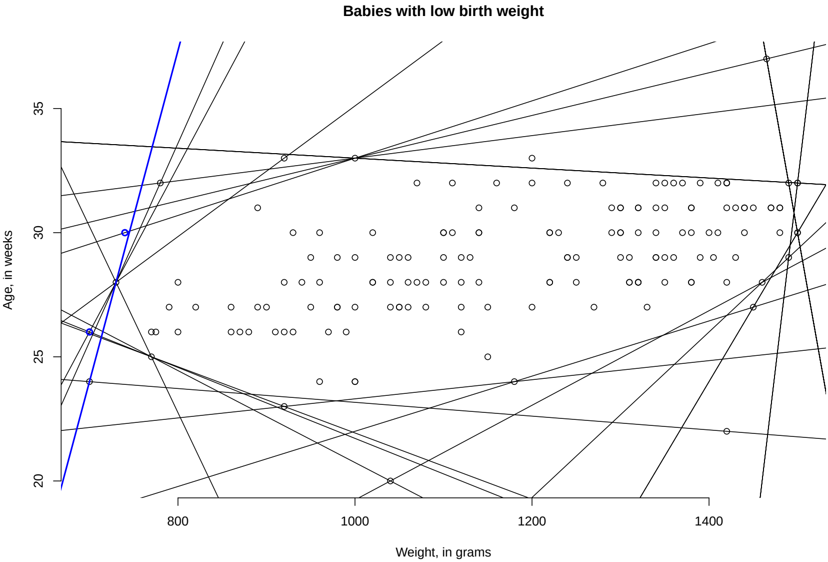
Tukey (=halfspace, location) depth-trimmed regions



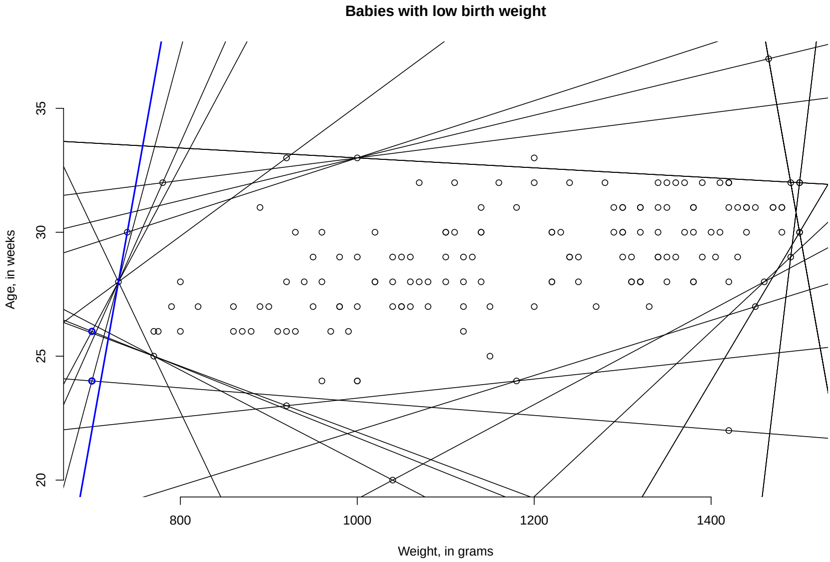
Tukey (=halfspace, location) depth-trimmed regions



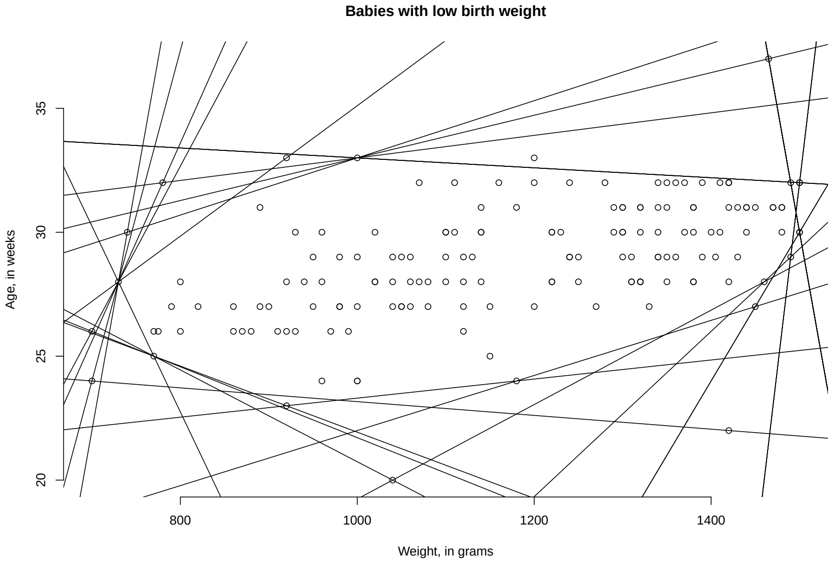
Tukey (=halfspace, location) depth-trimmed regions



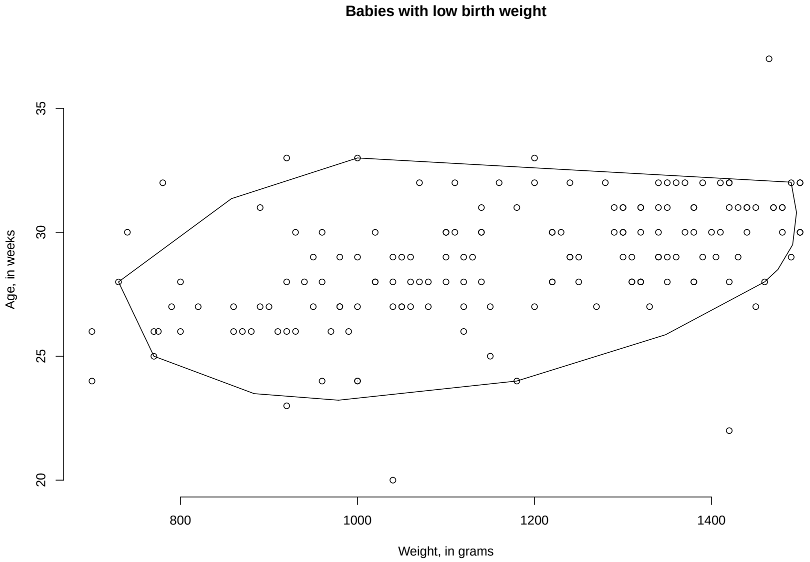
Tukey (=halfspace, location) depth-trimmed regions



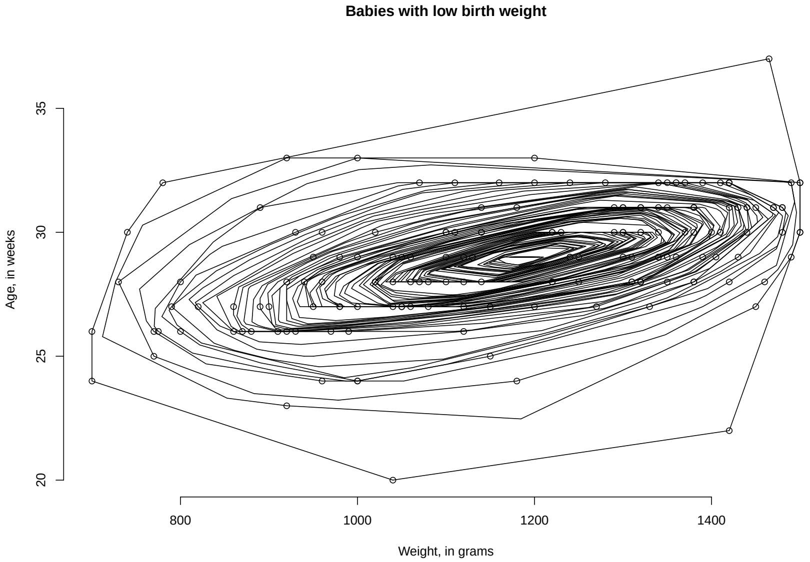
Tukey (=halfspace, location) depth-trimmed regions



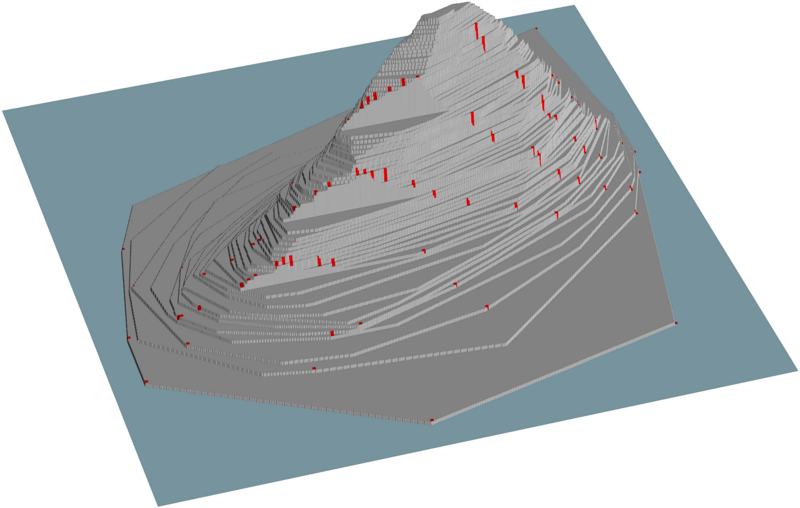
Tukey (=halfspace, location) depth-trimmed regions



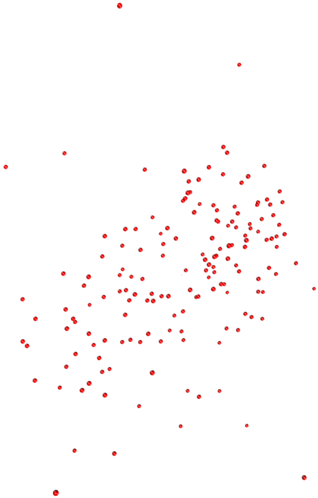
Tukey (=halfspace, location) depth-trimmed regions



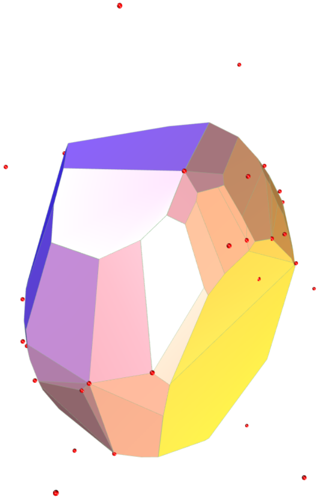
Tukey (=halfspace, location) data depth



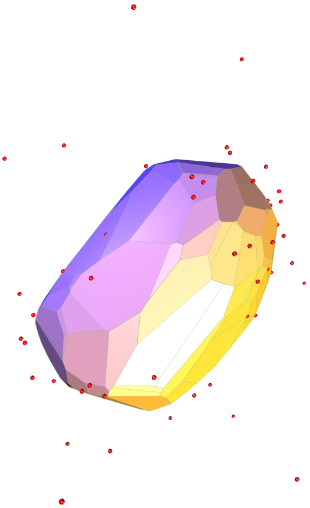
Tukey (=halfspace, location) depth region



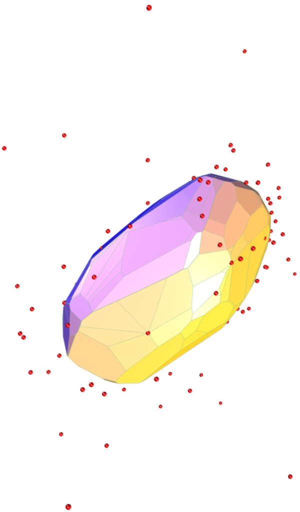
Tukey (=halfspace, location) depth region: $\tau = 2/161$



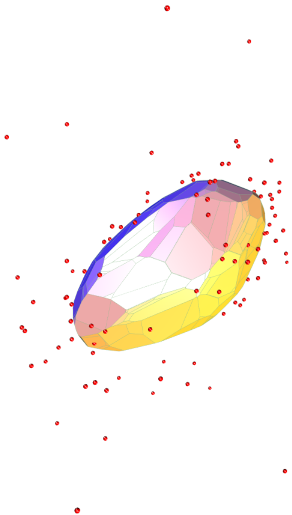
Tukey (=halfspace, location) depth region: $\tau = 5/161$



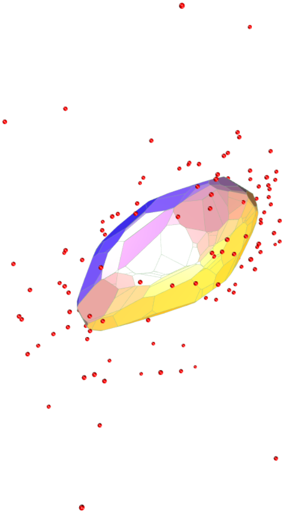
Tukey (=halfspace, location) depth region: $\tau = 9/161$



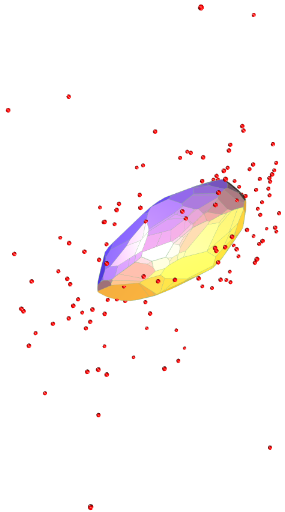
Tukey (=halfspace, location) depth region: $\tau = 13/161$



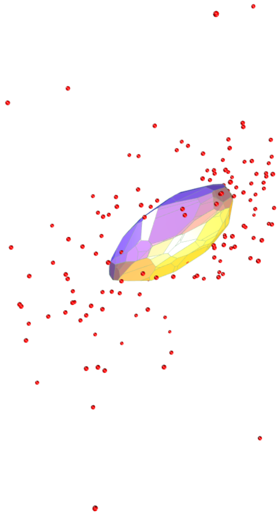
Tukey (=halfspace, location) depth region: $\tau = 17/161$



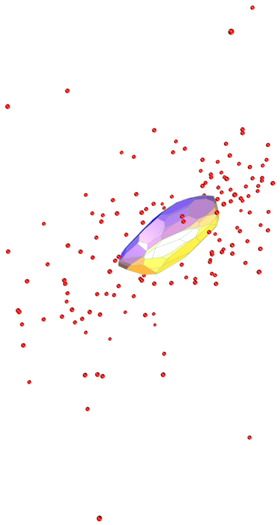
Tukey (=halfspace, location) depth region: τ =25/161



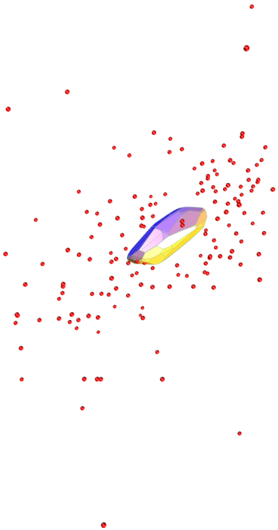
Tukey (=halfspace, location) depth region: $\tau = 33/161$



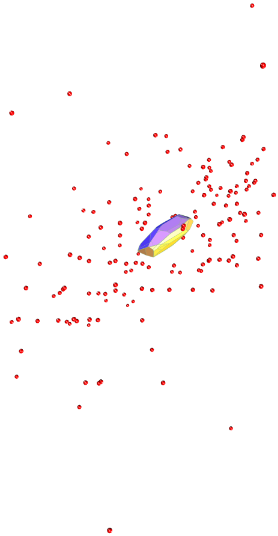
Tukey (=halfspace, location) depth region: $\tau = 41/161$



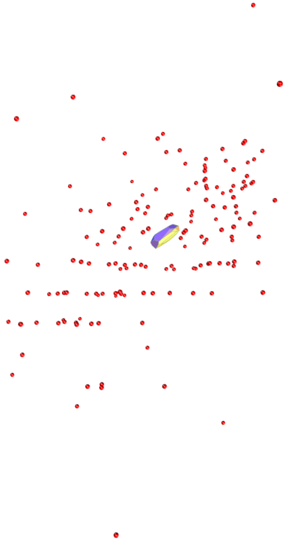
Tukey (=halfspace, location) depth region: $\tau = 49/161$



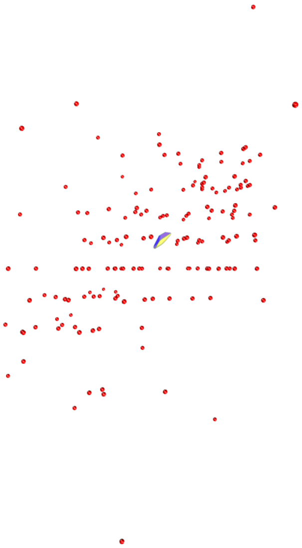
Tukey (=halfspace, location) depth region: $\tau = 57/161$



Tukey (=halfspace, location) depth region: $\tau = 65/161$



Tukey (=halfspace, location) depth region: τ =68/161



Computation of the Tukey depth: literature

- Rousseeuw, P.J. and Ruts, I. (1996).
  "Algorithm AS 307: Bivariate location depth."
  Journal of the Royal Statistical Society, Series C, 45, 516-526.
- Ruts, I. and Rousseeuw, P.J. (1996).
  "Computing depth contours of bivariate point clouds."
  Computational Statistics and Data Analysis, 23, 153-168.
- Rousseeuw, P.J. and Struyf, A. (1998).
  "Computing location depth and regression depth in higher dimensions."
  Statistics and Computing, 8, 193-203.
- Hallin, M., Paindaveine, D., and Šiman, M. (2010).
  "Multivariate quantiles and multiple-output regression quantiles:
  From L_1 optimization to halfspace depth."
  The Annals of Statistics, 38, 635-669.
- Paindaveine, D. and Šiman, M. (2011).
  "On directional multiple-output quantile regression."
  Journal of Multivariate Analysis, 102, 193–212.
- Liu, X. and Zuo, Y. (2014).
  "Computing halfspace depth and regression depth."
  Communications in Statistics - Simulation and Computation, 43, 969-985.

Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

Applications of data depth:

- **Multivariate data analysis** (Liu, Parelius, Singh '99);
- **Statistical quality control** (Liu, Singh '93);
- **Cluster analysis and classification** (Mosler, Hoberg '06; Li, Cuesta-Albertos, Liu '12; M., Mosler, Lange '15);
- **Tests for multivariate location, scale, symmetry** (Liu '92; Dyckerhoff '02; Dyckerhoff, Ley, Paindaveine '15);
- **Outlier detection** (Hubert, Rousseeuw, Segaert '15);
- **Multivariate risk measurement** (Cascos, Mochalov '07);
- **Robust linear programming** (Bazovkin, Mosler '15);
- **Missing data imputation** (M., Josse, Husson '18);
- etc.

R-package **ddalpha** (Pokotylo, M., Dyckerhoff, Nagy):
calculates a number of depths; performs depth-based classification of multivariate and functional data; contains 50 multivariate and 5 functional data sets.

Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

Theoretical background

- **Reminder:** Tukey depth of $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. a data set $\boldsymbol{X}$ is:

$$D^{T(n)}(\boldsymbol{x} \,|\, \boldsymbol{X}) = \frac{1}{n} \min_{\boldsymbol{p} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}} \sharp\{i : \boldsymbol{p}'\boldsymbol{x}_i \geq \boldsymbol{p}'\boldsymbol{x}\}\,.$$

- Due to the affine-invariance property we can rewrite:

$$D^{T(n)}(\boldsymbol{z} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = D^{T(n)}(\boldsymbol{0} \,|\, \boldsymbol{x}_1 - \boldsymbol{z}, \ldots, \boldsymbol{x}_n - \boldsymbol{z})\,.$$

- We call a vector $\boldsymbol{p} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$ *optimal for the data set* $\boldsymbol{X}$ if

$$D^{T(n)}(\boldsymbol{0} \,|\, \boldsymbol{X}) = \frac{1}{n} \#\{i \,|\, \boldsymbol{p}'\boldsymbol{x}_i \geq 0\}\,.$$

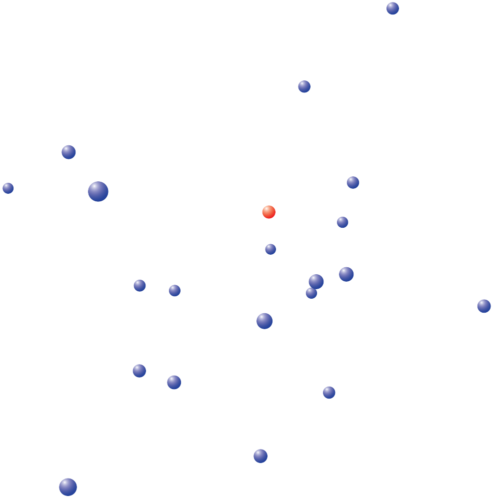
- For $\boldsymbol{p} \neq \boldsymbol{0}$ we define:

$$I_{\boldsymbol{p}}^{+} = \{i \,|\, \boldsymbol{p}'\boldsymbol{x}_i > 0\}\,, \quad I_{\boldsymbol{p}}^{0} = \{i \,|\, \boldsymbol{p}'\boldsymbol{x}_i = 0\}\,, \quad I_{\boldsymbol{p}}^{-} = \{i \,|\, \boldsymbol{p}'\boldsymbol{x}_i < 0\}\,.$$

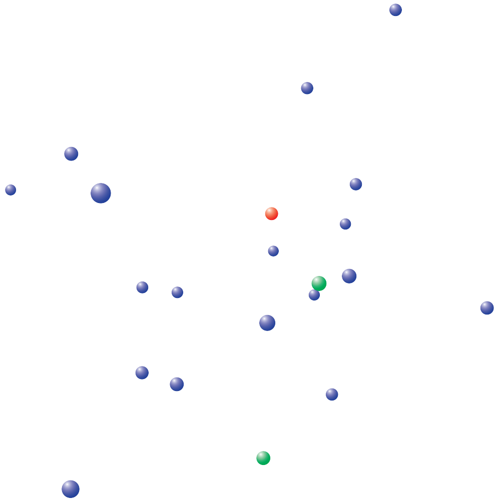
Proposition

If $\boldsymbol{p} \neq 0$ is optimal for $\boldsymbol{X}$, then $I_{\boldsymbol{p}}^{0} = \emptyset$, i.e., no data points lie on the boundary of the closed halfspace defined by $\boldsymbol{p}$.

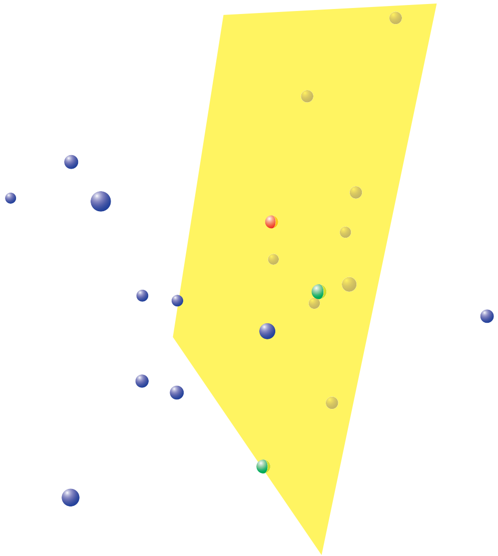
Theoretical background: Illustration



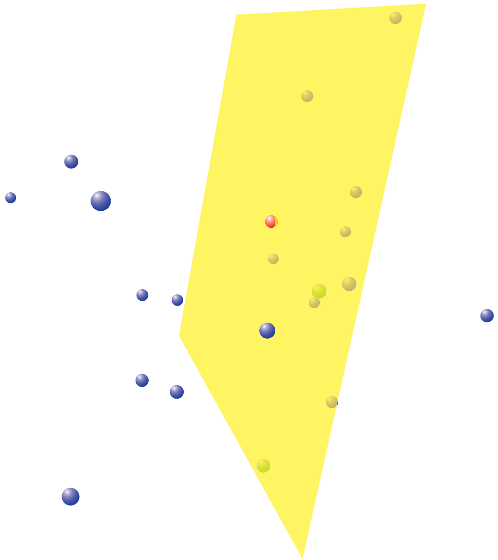
Theoretical background: Illustration



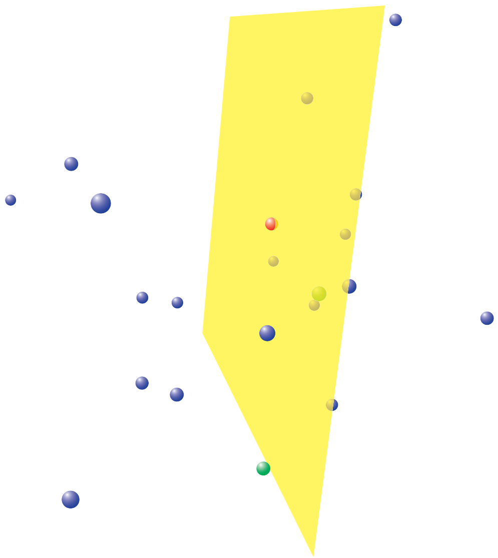
Theoretical background: Illustration



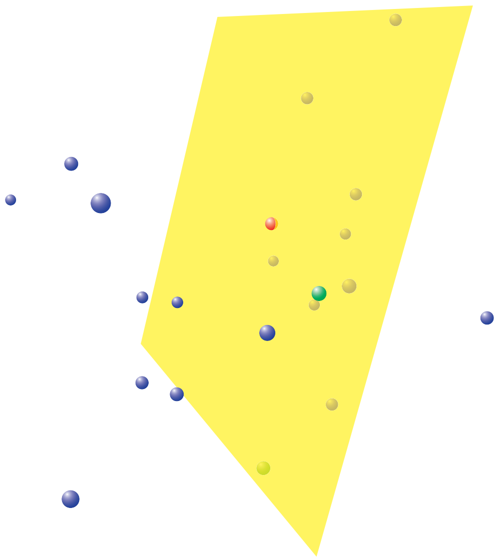
Theoretical background: Illustration



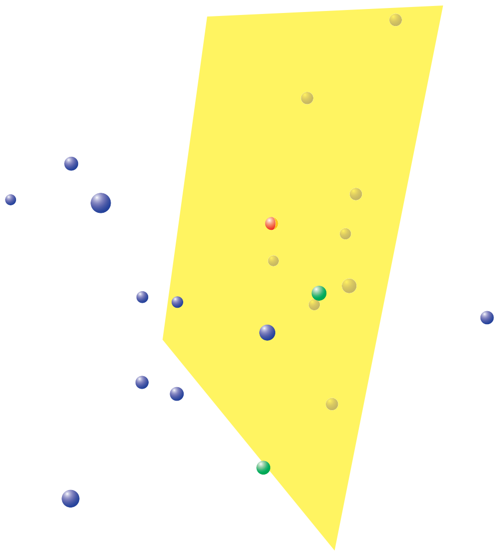
Theoretical background: Illustration



Theoretical background: Illustration



Theoretical background: Illustration



Main result

- Denote by $\mathcal{L}_k$ the set of all subsets I of order k of $\{1, \ldots, n\}$ such that the points $(\boldsymbol{x}_i)_{i \in I}$ are linearly independent.
- For any $I \in \mathcal{L}_k$ with $I = \{i_1, \ldots, i_k\}$, let $\boldsymbol{P}_I = [\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_k}]$.
- For any $I \in \mathcal{L}_k$ with $I = \{i_1, \ldots, i_k\}$, let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{d-k}$ be a basis of the orthogonal complement of $\mathrm{span}(\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_k})$ and $\boldsymbol{A}_I$ the matrix whose columns are the $\boldsymbol{a}_i$.
- Denote by I^c the complement of a set I.
- For a set $I = \{i_1, \ldots, i_k\}$ of indices, we denote by $I^* = \{j \mid \boldsymbol{x}_j \in \mathrm{span}(\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_k}), j = 1, \ldots, n\}$ the set of all indices j such that $\boldsymbol{x}_j$ is contained in the linear hull of $\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_k}$.

Theorem

For each k such that $1 \leq k < d$ it holds that

$$n \cdot D^{T(n)}(\boldsymbol{0} \mid \boldsymbol{X}) = \min_{I \in \mathcal{L}_k} \left[n \cdot D^{T(n)}(\boldsymbol{0} \mid \boldsymbol{A}_I' \boldsymbol{X}_{(I^*)^c}) + n \cdot D^{T(n)}(\boldsymbol{0} \mid \boldsymbol{P}_I' \boldsymbol{X}_{I^*}) \right].$$

Contents

Data depth
Tukey depth: definition
Tukey trimmed regions
Applications

Computation of the Tukey depth
Theoretical background
Algorithm and simulations

Computation of Tukey trimmed regions
Existing approaches
The proposed algorithm
Tukey median

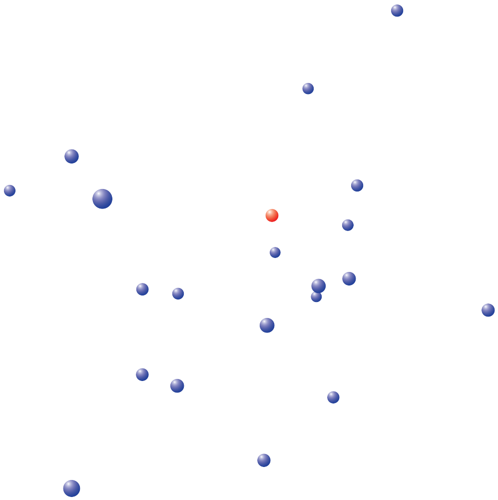
Outlook for approximations

Conclusions

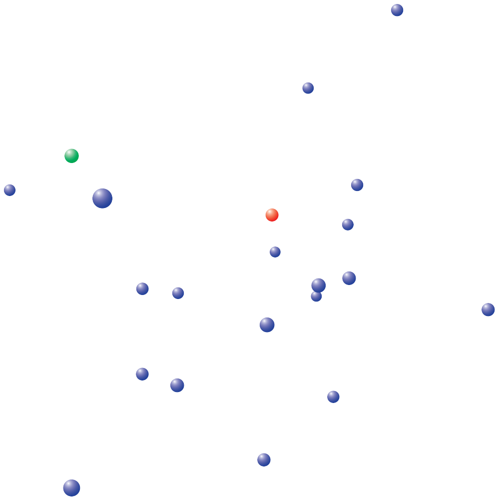
Example: combinatorial algorithm with $k = d - 2$

function $\text{nHD\_Comb2}(d, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ ▷ Halfspace depth of 0
    if $d = 1$ **then return** $\text{nHD1}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$
    if $d = 2$ **then return** $\text{nHD2}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$
    $n_{min} \leftarrow n$
    for each subset $I \subset \{1, \ldots, n\}$ of order $d - 2$ **do**
        if $(\boldsymbol{x}_i)_{i \in I}$ linearly independent **then**
            Compute basis $\boldsymbol{a}_1, \boldsymbol{a}_2$ of orthogonal complement of $(\boldsymbol{x}_i)_{i \in I}$
            $\boldsymbol{A}_I \leftarrow \text{Matrix}[\boldsymbol{a}_1, \boldsymbol{a}_2]$
            $\boldsymbol{P}_I \leftarrow \text{Matrix}[\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_{d-2}}]$
            for all $\boldsymbol{x}_j$ **do**
                if $\boldsymbol{A}_I' \boldsymbol{x}_j \neq \boldsymbol{0}$ **then** $\boldsymbol{y}_j \leftarrow \boldsymbol{A}_I' \boldsymbol{x}_j$
                else $\boldsymbol{z}_j \leftarrow \boldsymbol{P}_I' \boldsymbol{x}_j$
            $l \leftarrow \#\{j : \boldsymbol{A}_I' \boldsymbol{x}_j \neq \boldsymbol{0}\}$
            $n_{new} \leftarrow \text{nHD2}(\boldsymbol{y}_{j_1}, \ldots, \boldsymbol{y}_{j_l})$
            if $n - l > d - 2$ **then**
                $n_{new} \leftarrow n_{new} + \text{nHD\_Comb2}(d - 2, \boldsymbol{z}_{j_1}, \ldots, \boldsymbol{z}_{j_{n-l}})$
            if $n_{new} < n_{min}$ **then** $n_{min} \leftarrow n_{new}$
    return n_{min}

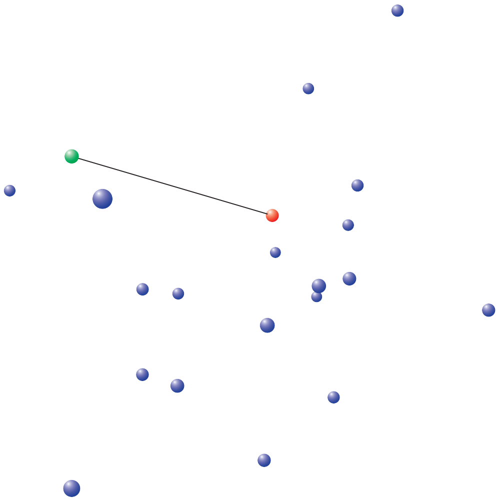
Combinatorial algorithm: Illustration



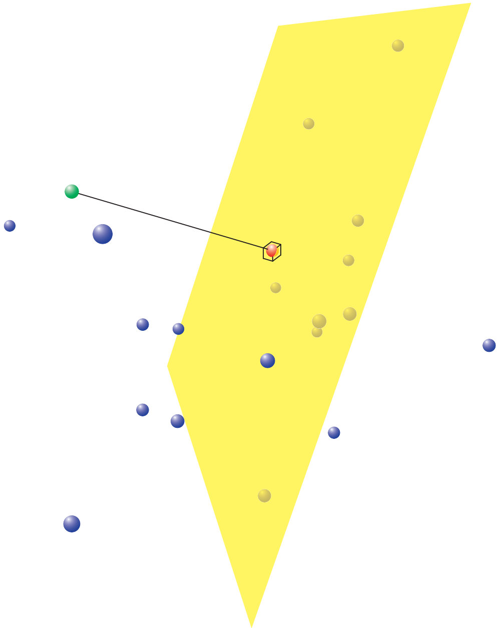
Combinatorial algorithm: Illustration



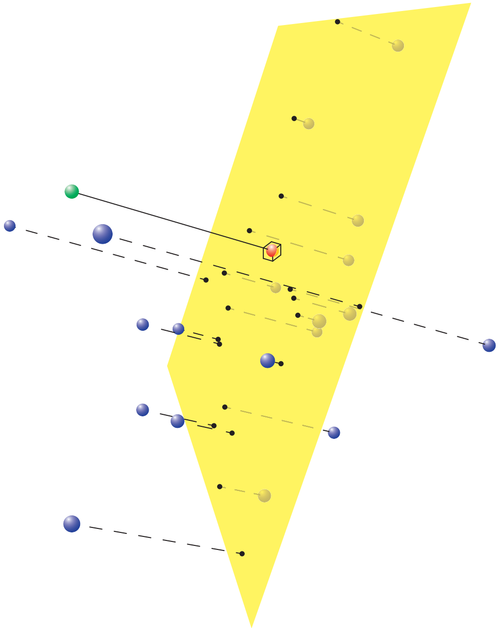
Combinatorial algorithm: Illustration



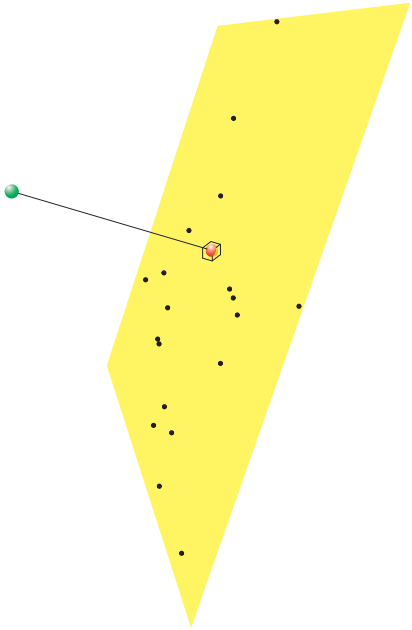
Combinatorial algorithm: Illustration



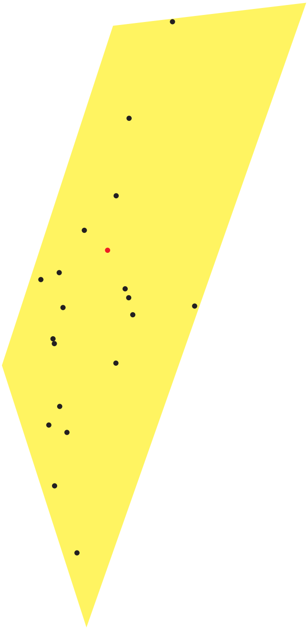
Combinatorial algorithm: Illustration



Combinatorial algorithm: Illustration



Combinatorial algorithm: Illustration



Time of depth calculation (sec.): $k = 1$, $d - 2$, $d - 1$

	$n = 40$	80	160	320	640	1280	2560	5120	10240	20480	40960	81920
$d = 3$	0.000	0.000	0.000	0.011	0.047	0.184	0.780	3.19	13.2	54.5	225	933
	0.002	0.002	0.003	0.016	0.063	0.250	1.03	4.22	17.4	72.2	293	1210
	0.000	0.003	0.014	0.117	0.936	7.60	61.3	519	—	—	—	—
4	0.006	0.048	0.402	3.36	28.2	235	1960	—	—	—	—	—
	0.005	0.038	0.302	2.50	20.4	166	1360	—	—	—	—	—
	0.005	0.055	0.784	12.3	203	3290	—	—	—	—	—	—
5	0.205	3.62	62.5	1070	—	—	—	—	—	—	—	—
	0.055	0.952	16.1	269	—	—	—	—	—	—	—	—
	0.047	1.24	35.7	1110	—	—	—	—	—	—	—	—
6	7.32	275	—	—	—	—	—	—	—	—	—	—
	0.506	18.4	633	—	—	—	—	—	—	—	—	—
	0.392	21.6	1250	—	—	—	—	—	—	—	—	—
7	257	—	—	—	—	—	—	—	—	—	—	—
	3.60	278	—	—	—	—	—	—	—	—	—	—
	2.66	305	—	—	—	—	—	—	—	—	—	—
8	—	—	—	—	—	—	—	—	—	—	—	—
	21.4	3550	—	—	—	—	—	—	—	—	—	—
	14.3	3470	—	—	—	—	—	—	—	—	—	—
9	—	—	—	—	—	—	—	—	—	—	—	—
	107	—	—	—	—	—	—	—	—	—	—	—
	69.8	—	—	—	—	—	—	—	—	—	—	—
10	—	—	—	—	—	—	—	—	—	—	—	—
	439	—	—	—	—	—	—	—	—	—	—	—
	289	—	—	—	—	—	—	—	—	—	—	—

Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

Directional quantiles

Kong, Mizera (2008)

For a random vector $X \in \mathbb{R}^d$ define the $\boldsymbol{u}$-directional τ-th quantile:

$$Q(\tau, \boldsymbol{u}, X) = \inf\{x \,:\, P(\boldsymbol{u}'X \leq x) \geq \tau\}.$$

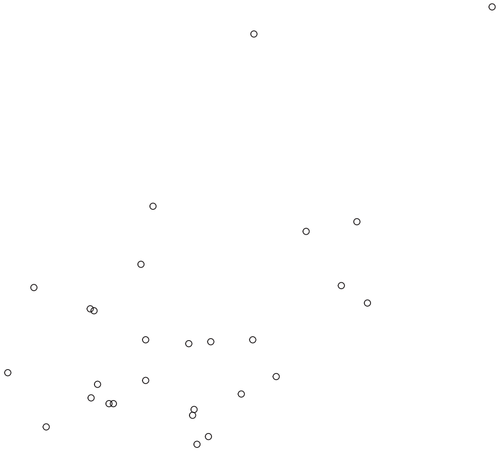
Directional quantile envelope:

$$R_\tau(X) = \bigcap_{\boldsymbol{u} \in \mathbb{S}^{d-1}} H\big(\boldsymbol{u}, Q(\tau, \boldsymbol{u}, X)\big),$$

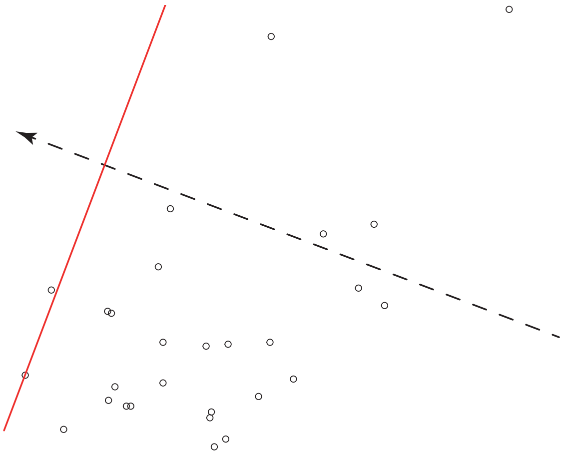
where $H(\boldsymbol{u}, q) = \{x : \boldsymbol{u}'x \geq q\}$ is the *supporting halfspace* determined by $\boldsymbol{u} \in \mathbb{S}^{d-1}$ and $q \in \mathbb{R}$.
Then for every $p \in (0; \frac{1}{2}]$ directional quantile envelopes coincide with the Tukey regions:

$$R_\tau(X) = D_\tau^T(X) \text{ for every } \tau \in (0; \frac{1}{2}].$$

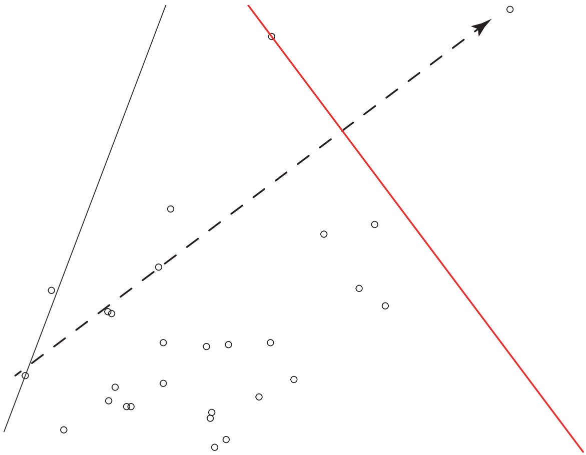
Directional quantiles (illustration)



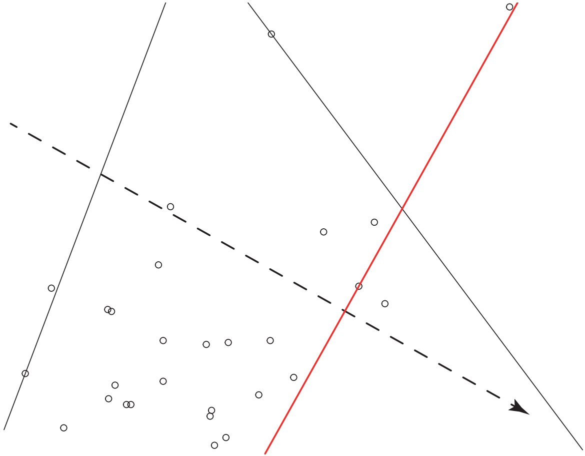
Directional quantiles (illustration)



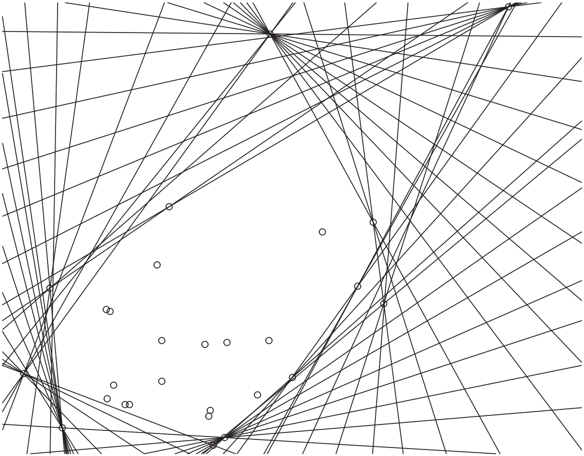
Directional quantiles (illustration)



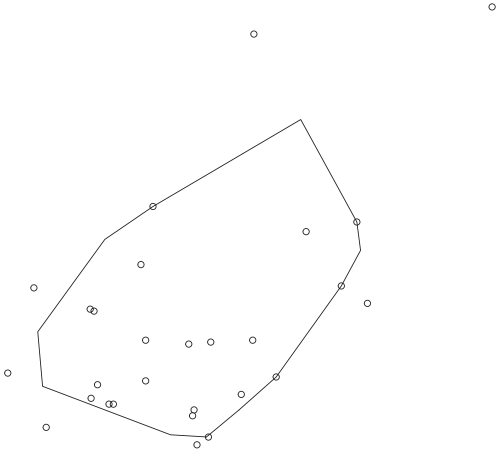
Directional quantiles (illustration)



Directional quantiles (illustration)



Directional quantiles (illustration)



Multiple-output regression quantiles

Hallin, Paindaveine, Šiman (2010); Paindaveine, Šiman (2011)

Regress $Y \in \mathbb{R}^m$ on $X = (1, W')' \in \mathbb{R}^p$.

For a data set $(\boldsymbol{X}, \boldsymbol{Y}) = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathbb{R}^p \times \mathbb{R}^m; i = 1, ..., n\}$, for $\tau \in (0, 1)$, and for $\boldsymbol{u} \in S^{m-1}$ a $(\tau \boldsymbol{u})$-quantile positive halfspace is any

$$H_{\tau\boldsymbol{u}}^{(n)+} := \{(\boldsymbol{w}', \boldsymbol{y}')' \in \mathbb{R}^{p-1} \times \mathbb{R}^m : \hat{\boldsymbol{b}}'_{\tau\boldsymbol{u}} \boldsymbol{y} - \hat{\boldsymbol{a}}'_{\tau\boldsymbol{u}}(1, \boldsymbol{w}')' \geq 0\} \text{ with}$$

$$(\hat{\boldsymbol{a}}'_{\mathrm{HP\check{S}};\tau\boldsymbol{u}}, \hat{\boldsymbol{b}}'_{\mathrm{HP\check{S}};\tau\boldsymbol{u}})' = \operatorname{argmin} \sum_{i=1}^{n} \rho_\tau(\boldsymbol{b}'\boldsymbol{y}_i - \boldsymbol{a}'\boldsymbol{x}_i) \text{ subject to } \boldsymbol{u}'\boldsymbol{b} = 1,$$

or

$$(\hat{\boldsymbol{a}}'_{\mathrm{proj};\tau\boldsymbol{u}}, \hat{\boldsymbol{b}}'_{\mathrm{proj};\tau\boldsymbol{u}})' = \operatorname{argmin} \sum_{i=1}^{n} \rho_\tau(\boldsymbol{b}'\boldsymbol{y}_i - \boldsymbol{a}'\boldsymbol{x}_i) \text{ subject to } \boldsymbol{u} = \boldsymbol{b},$$

where $\rho_\tau(x) = x\big(\tau - I(x < 0)\big)$ is the τ-quantile check function.

Multiple-output regression quantiles (location)

Hallin, Paindaveine, Šiman (2010); Paindaveine, Šiman (2011)

Regress $X \in \mathbb{R}^d$ on $1 \in \mathbb{R}$.

For a data set $\boldsymbol{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d; i = 1, ..., n\}$, for $\tau \in (0, 1)$, and for $\boldsymbol{u} \in \mathbb{S}^{d-1}$ a $(\tau \boldsymbol{u})$-quantile positive halfspace is any

$$H_{\tau \boldsymbol{u}}^{(n)+} := \{\boldsymbol{x}' \in \mathbb{R}^d \ : \ \hat{\boldsymbol{b}}_{\tau \boldsymbol{u}}' \boldsymbol{x} - \hat{a}_{\tau \boldsymbol{u}} \geq 0\} \text{ with}$$

$$(\hat{a}_{\text{HPŠ};\tau \boldsymbol{u}}, \hat{\boldsymbol{b}}_{\text{HPŠ};\tau \boldsymbol{u}}')' = \operatorname{argmin} \sum_{i=1}^{n} \rho_\tau(\boldsymbol{b}' \boldsymbol{x}_i - a) \ \text{ subject to } \boldsymbol{u}' \boldsymbol{b} = 1,$$

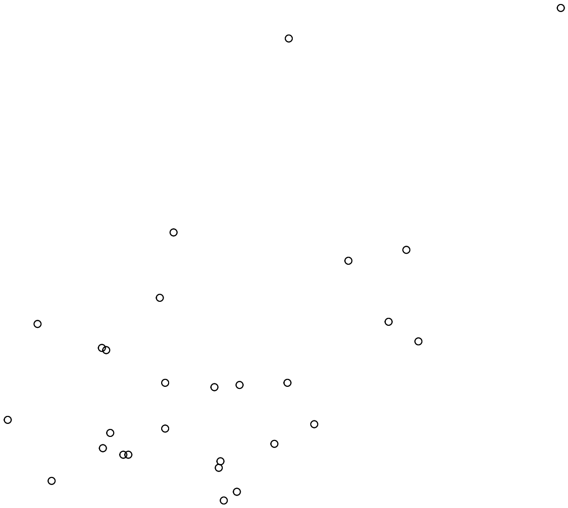
or

$$(\hat{a}_{\text{proj};\tau \boldsymbol{u}}, \hat{\boldsymbol{b}}_{\text{proj};\tau \boldsymbol{u}}')' = \operatorname{argmin} \sum_{i=1}^{n} \rho_\tau(\boldsymbol{b}' \boldsymbol{x}_i - a) \ \text{ subject to } \boldsymbol{u} = \boldsymbol{b},$$

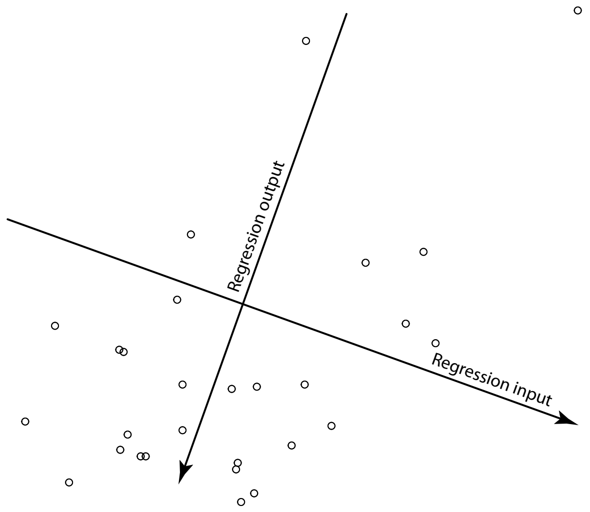
where $\rho_\tau(x) = x\big(\tau - I(x < 0)\big)$ is the τ-quantile check function. Then for some $\mathcal{L}_{\boldsymbol{X},\tau}$ with $\#\mathcal{L}_{\boldsymbol{X},\tau} < \infty$ it holds:

$$D_\tau^{T(n)}(\boldsymbol{X}) = R_\tau^{(n)}(\boldsymbol{X}) = \bigcap_{\boldsymbol{u} \in \mathbb{S}^{d-1}} \{H_{\tau \boldsymbol{u}}^{(n)+}\} = \bigcap_{\boldsymbol{u} \in \mathcal{L}_{\boldsymbol{X},\tau}} \{H_{\tau \boldsymbol{u}}^{(n)+}\}.$$

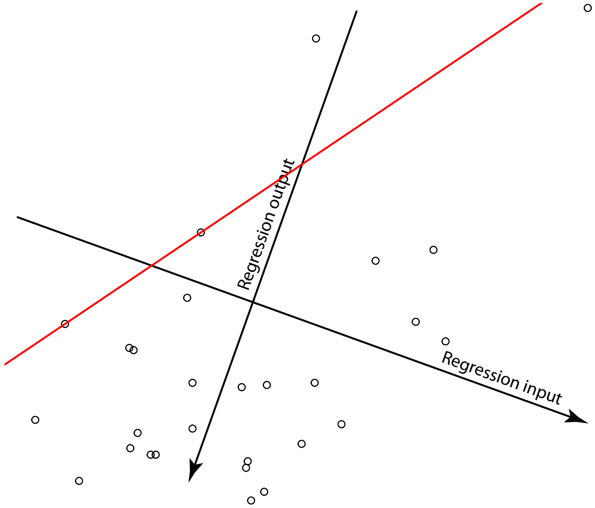
Multiple-output regression quantiles (illustration)



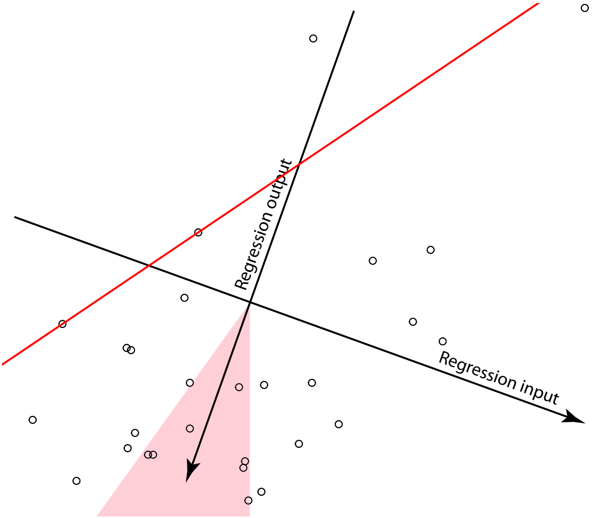
Multiple-output regression quantiles (illustration)



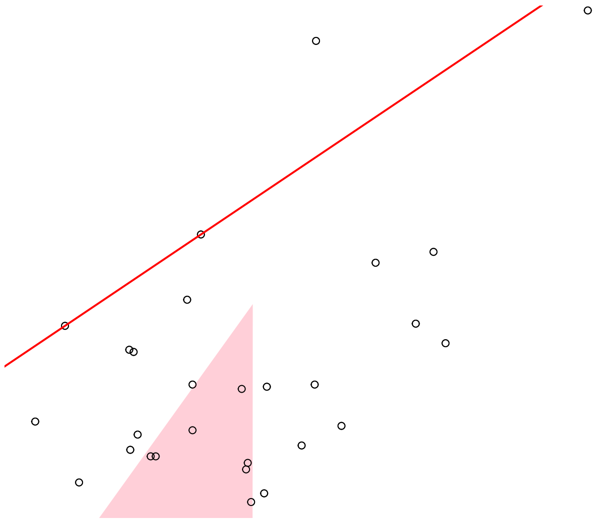
Multiple-output regression quantiles (illustration)



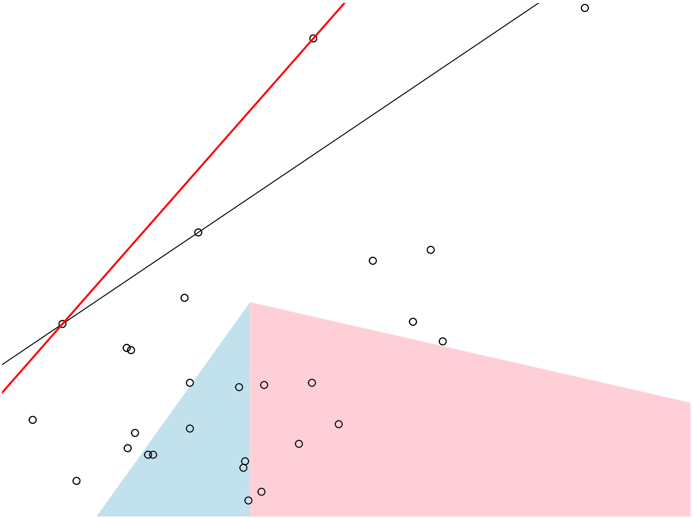
Multiple-output regression quantiles (illustration)



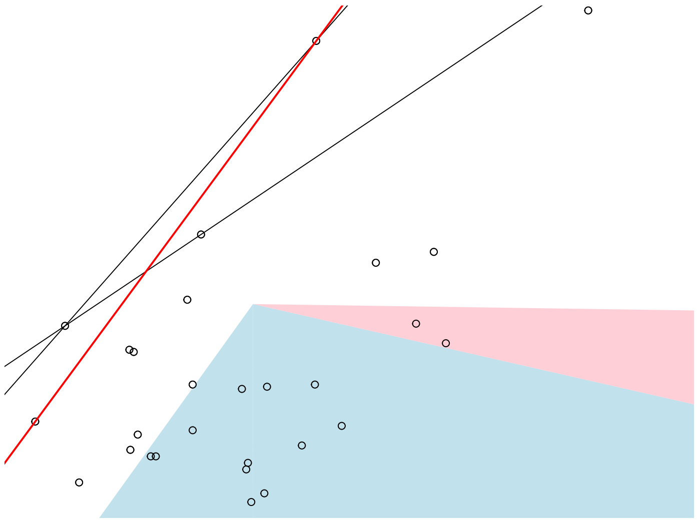
Multiple-output regression quantiles (illustration)



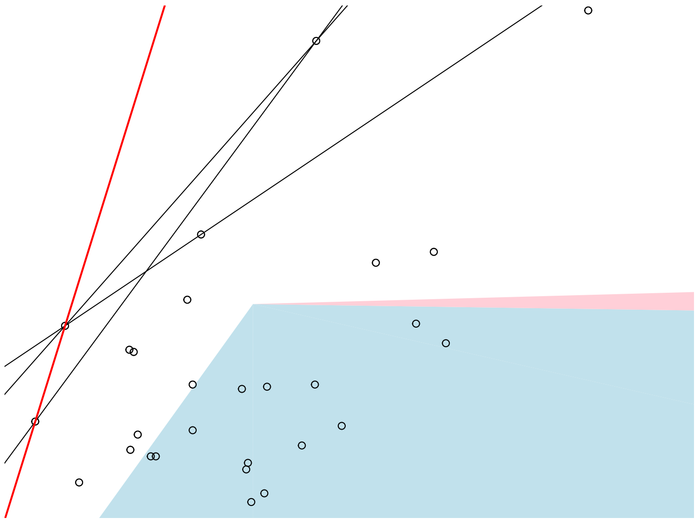
Multiple-output regression quantiles (illustration)



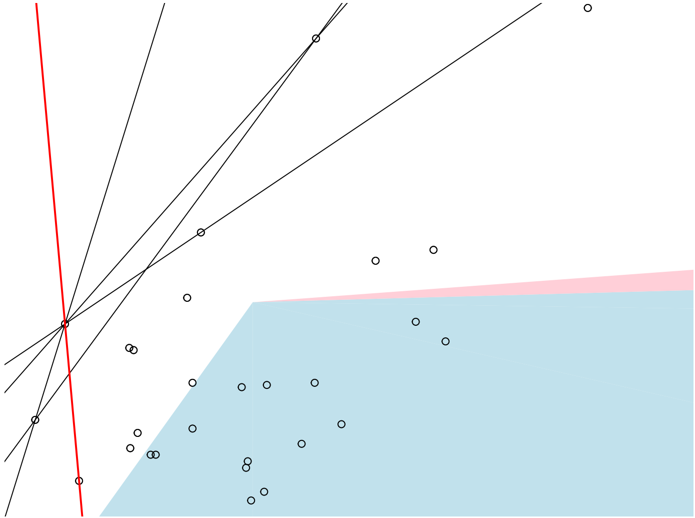
Multiple-output regression quantiles (illustration)



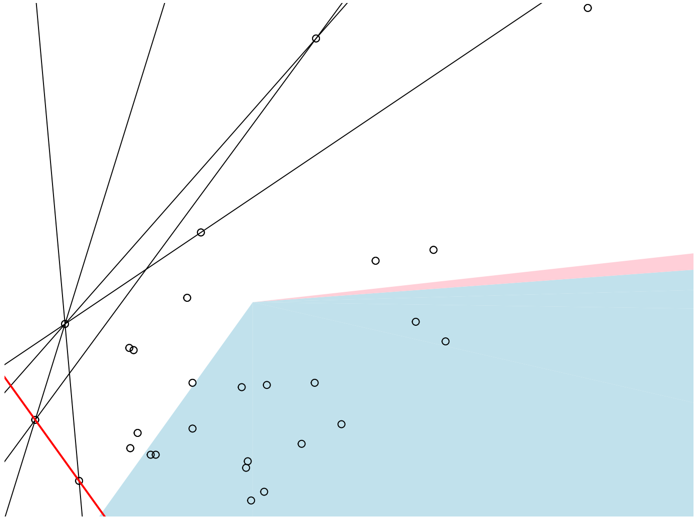
Multiple-output regression quantiles (illustration)



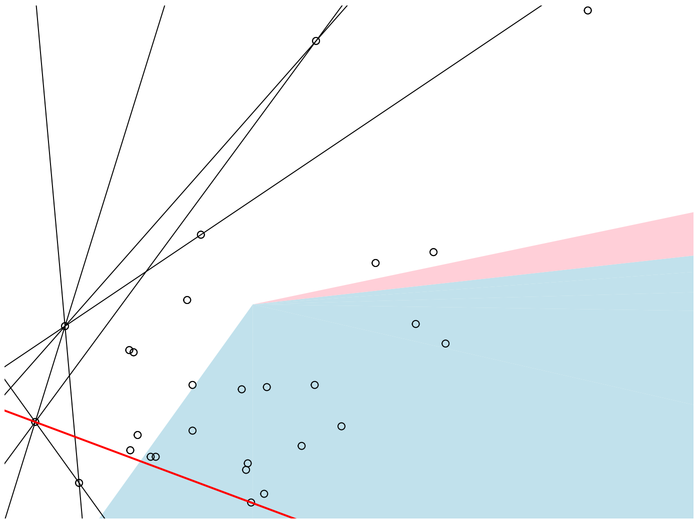
Multiple-output regression quantiles (illustration)



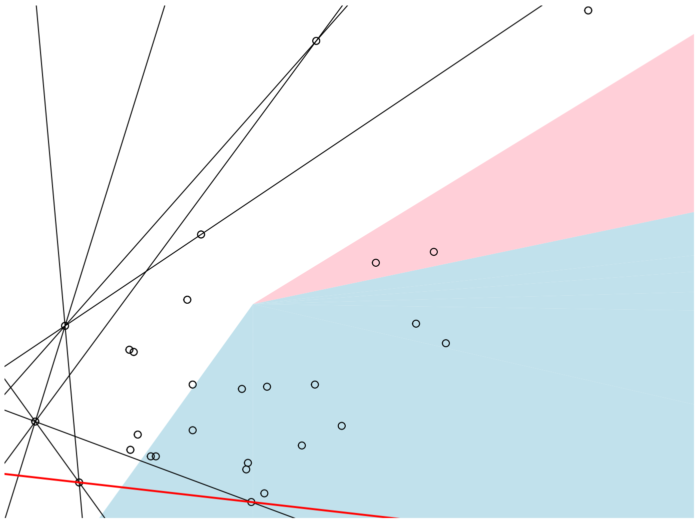
Multiple-output regression quantiles (illustration)



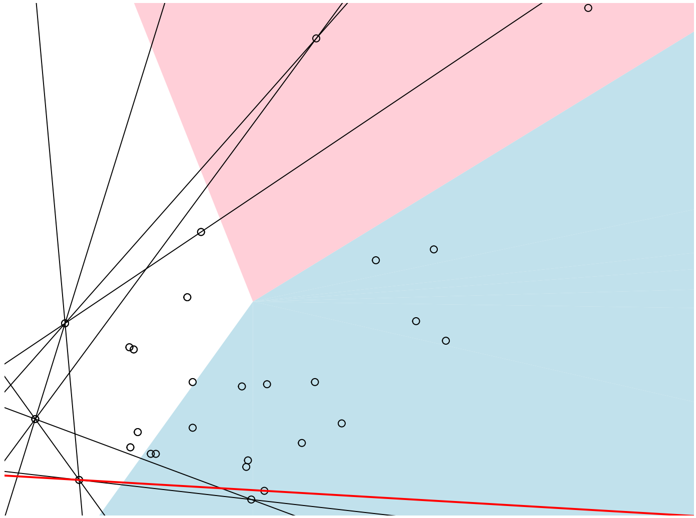
Multiple-output regression quantiles (illustration)



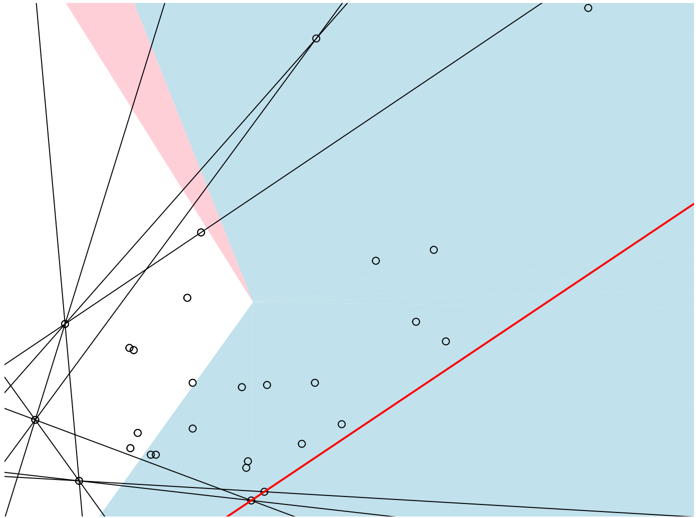
Multiple-output regression quantiles (illustration)



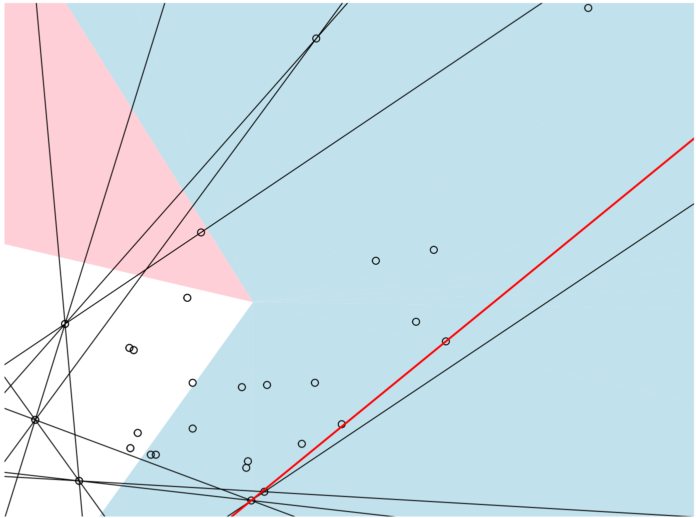
Multiple-output regression quantiles (illustration)



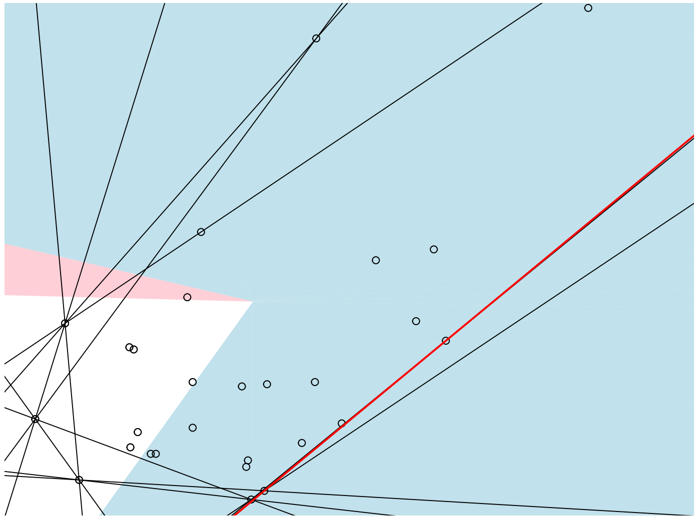
Multiple-output regression quantiles (illustration)



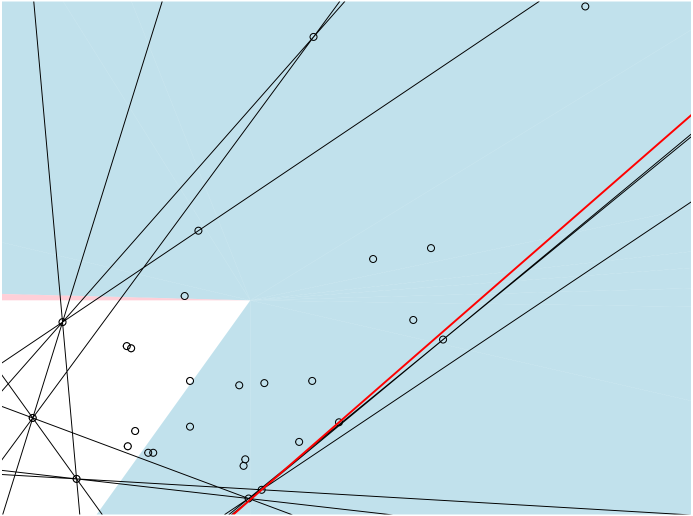
Multiple-output regression quantiles (illustration)



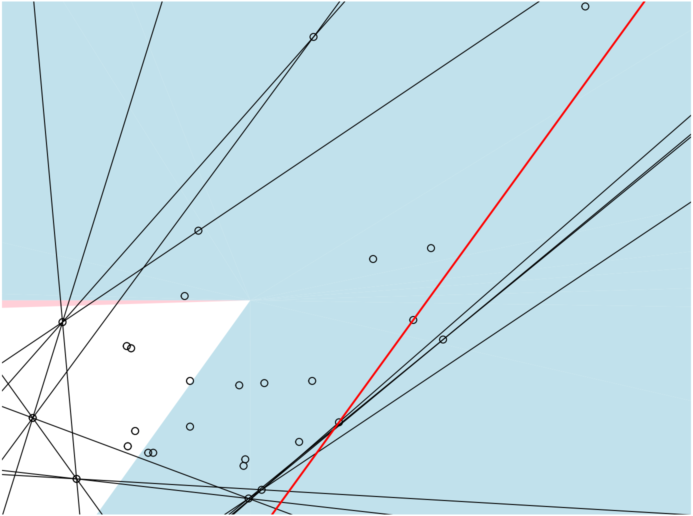
Multiple-output regression quantiles (illustration)



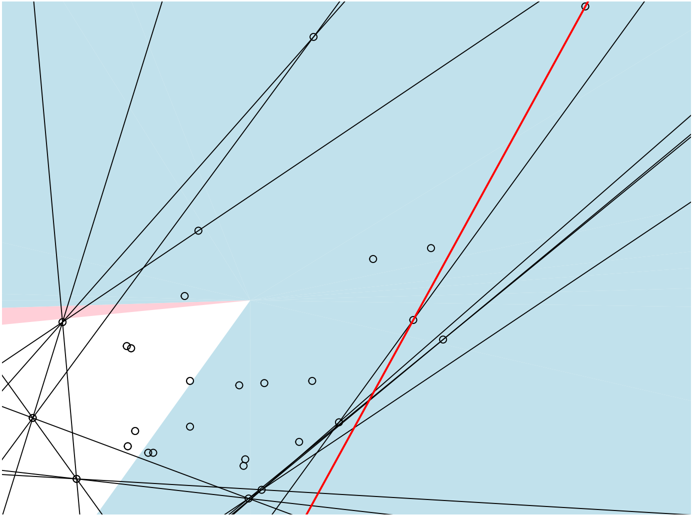
Multiple-output regression quantiles (illustration)



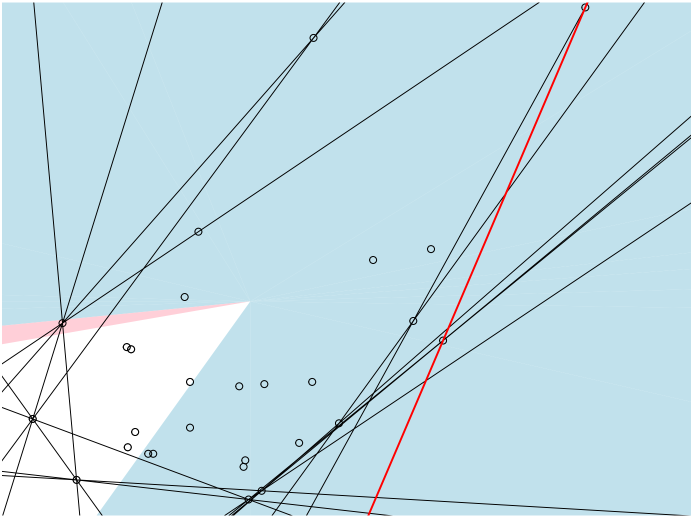
Multiple-output regression quantiles (illustration)



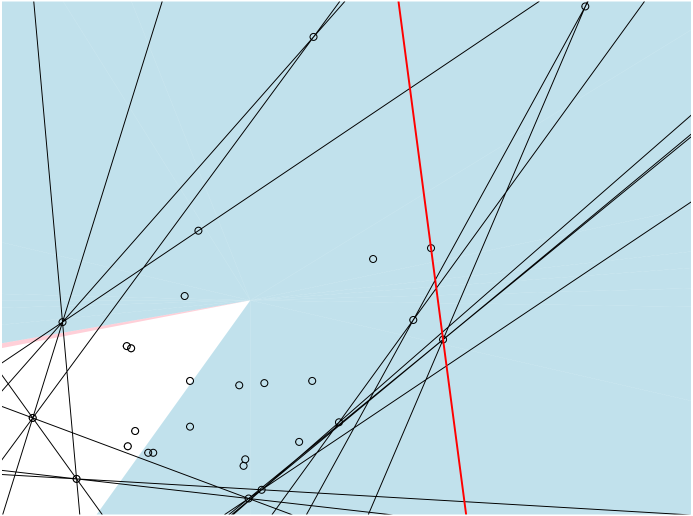
Multiple-output regression quantiles (illustration)



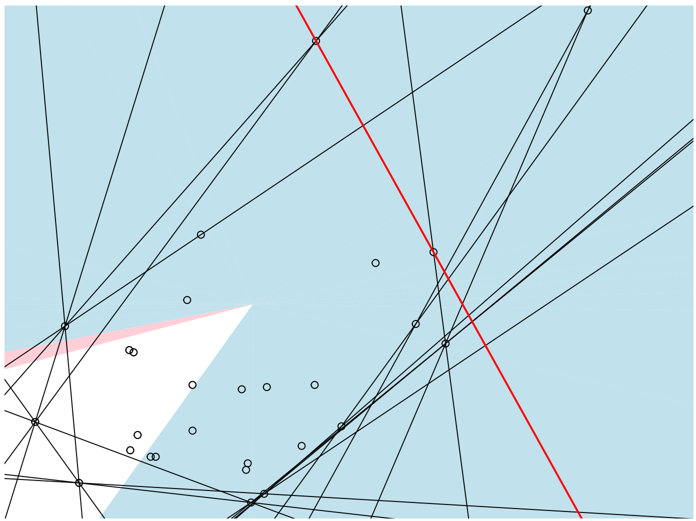
Multiple-output regression quantiles (illustration)



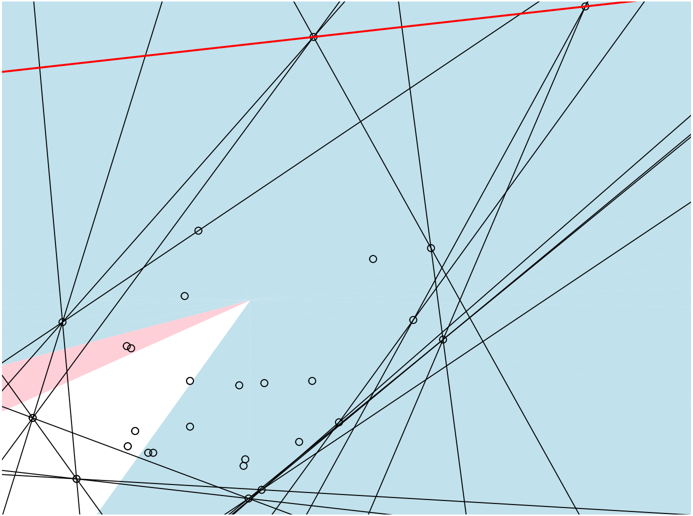
Multiple-output regression quantiles (illustration)



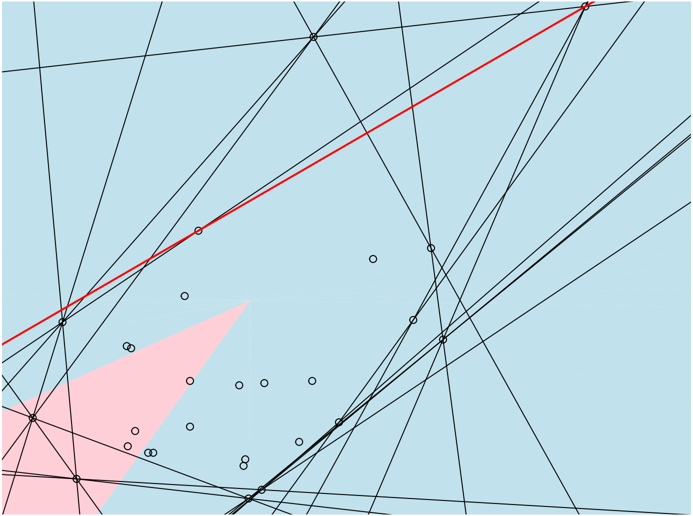
Multiple-output regression quantiles (illustration)



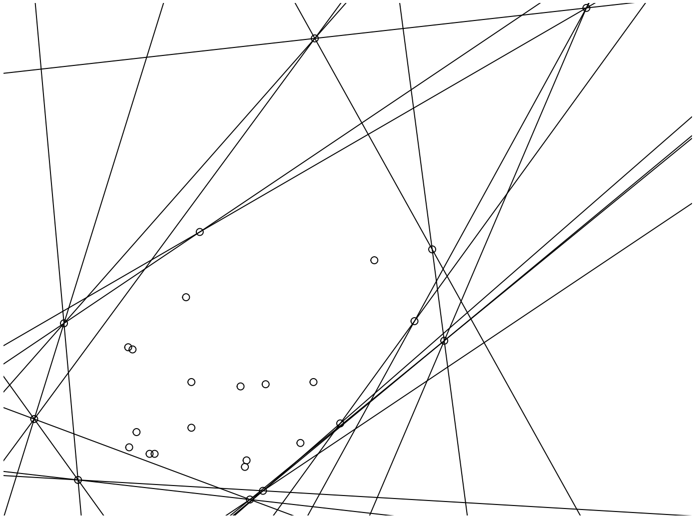
Multiple-output regression quantiles (illustration)



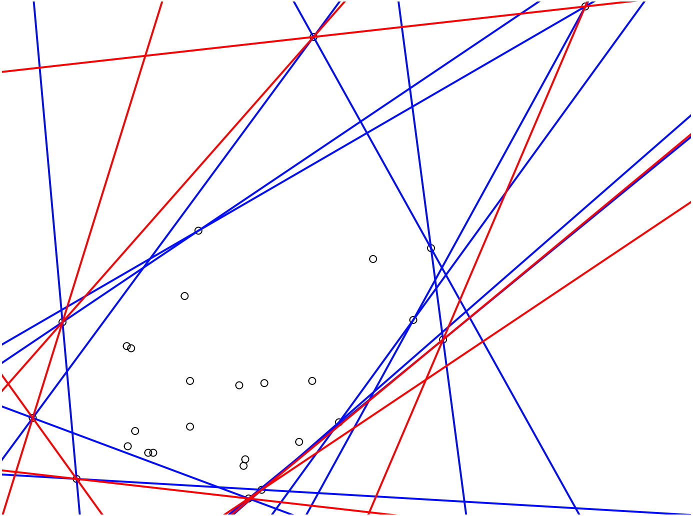
Multiple-output regression quantiles (illustration)



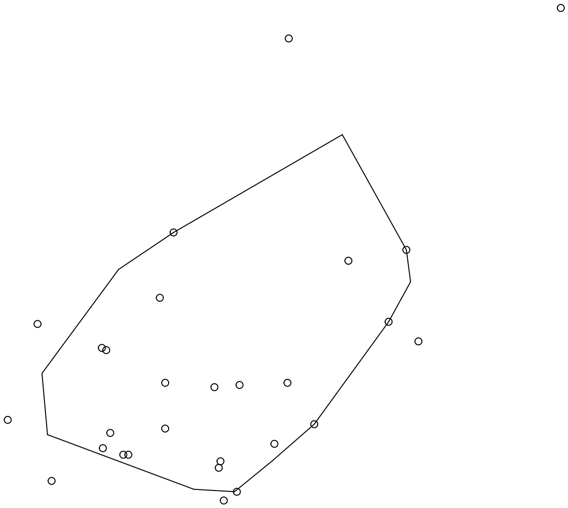
Multiple-output regression quantiles (illustration)



Multiple-output regression quantiles (illustration)



Multiple-output regression quantiles (illustration)



Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

Univariate projection quantiles (algorithm)

Input: $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \in \mathbb{R}^d$, $2 < d < n < \infty$.

Step 1: Set matrix $\mathcal{A} = (\texttt{false}_n)^{d-1}$, tree $\mathcal{G}_\tau = \emptyset$, queue $\mathcal{Q} = \emptyset$.

Step 2: Generate a sufficient initial set of $(d-1)$-tuples (*e.g.* ridges of the data convex hull), push them into $\mathcal{Q}$ and set them `true` in $\mathcal{A}$.

Step 3: Pop a $(d-1)$-tuple, say $[i_1, \cdots, i_{d-1}]$, from $\mathcal{Q}$.

Step 4: Find all subscripts $i_d \in \{1, \cdots, n\} \setminus \{i_1, \cdots, i_{d-1}\}$ s. t. $\{\boldsymbol{x}_{i_1}, \cdots, \boldsymbol{x}_{i_{d-1}}, \boldsymbol{x}_{i_d}\}$ determine a τ-th halfspace, store these in $\mathcal{T}$.

Step 5: For each j(determining halfspace $\mathbf{g}_j$)$\in \mathcal{T}$ do:

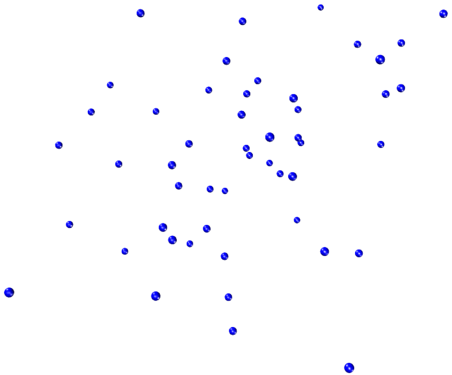
- **5.1:** If $\mathbf{g}_j \notin \mathcal{G}_\tau$, add $\mathbf{g}_j$ to $\mathcal{G}_\tau$, else go to the next element in $\mathcal{T}$.
- **5.2:** For each ridge of $\mathbf{g}_j$, except $[i_1, \cdots, i_{d-1}]$, say $[j_1, ..., j_{d-1}]$, do:
  - **5.2.1:** If $\mathcal{A}_{[j_1,...,j_{d-1}]} = \texttt{false}$, set $\mathcal{A}_{[j_1,...,j_{d-1}]} = \texttt{true}$, push $[j_1, ..., j_{d-1}]$ into $\mathcal{Q}$.

Step 6: If $\mathcal{Q}$ is not empty, go to **Step 3**, else stop.

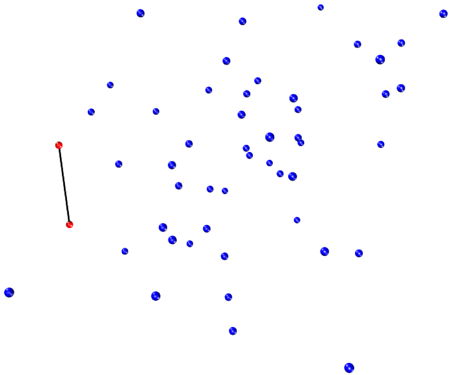
Step 7: Eliminate redundant halfspaces, compute region's facets.

Output: $\mathcal{G}_\tau$, τ-region's facets.

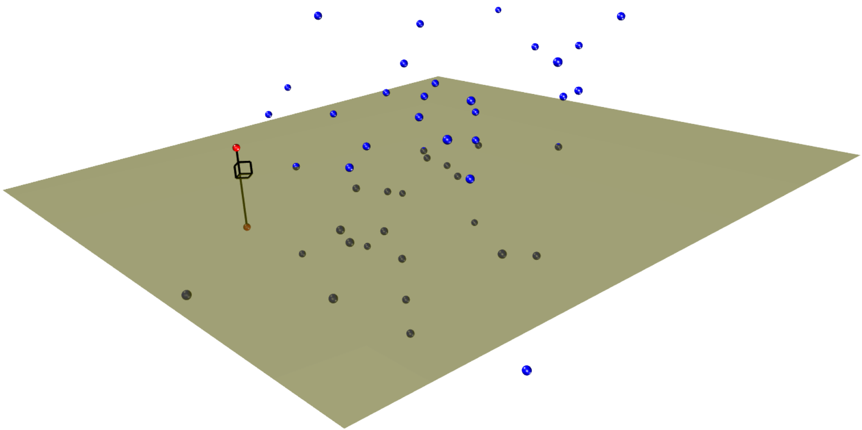
Univariate projection quantiles (illustration)



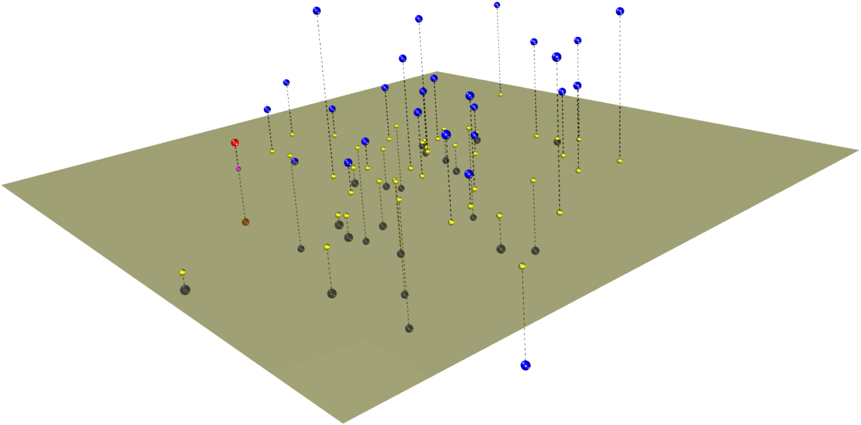
Univariate projection quantiles (illustration)



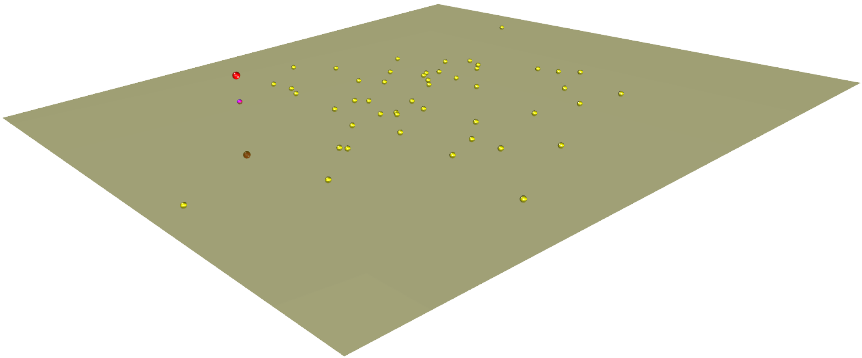
Univariate projection quantiles (illustration)



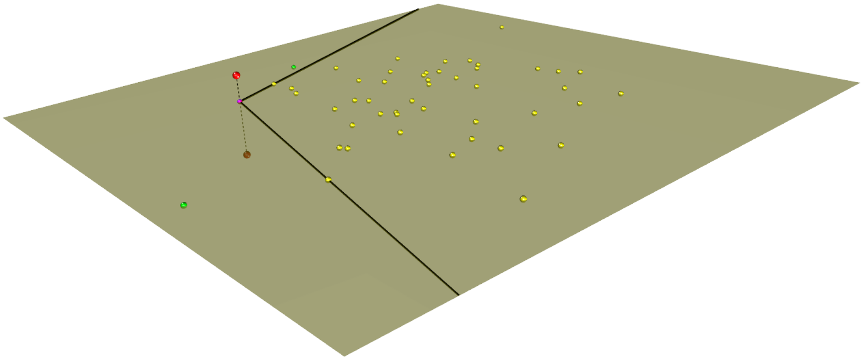
Univariate projection quantiles (illustration)



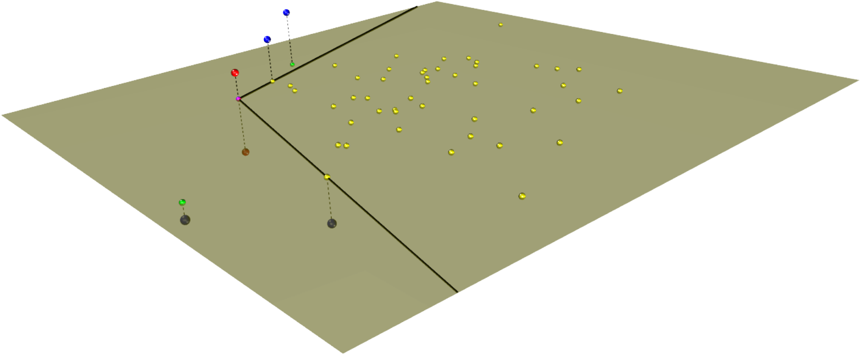
Univariate projection quantiles (illustration)



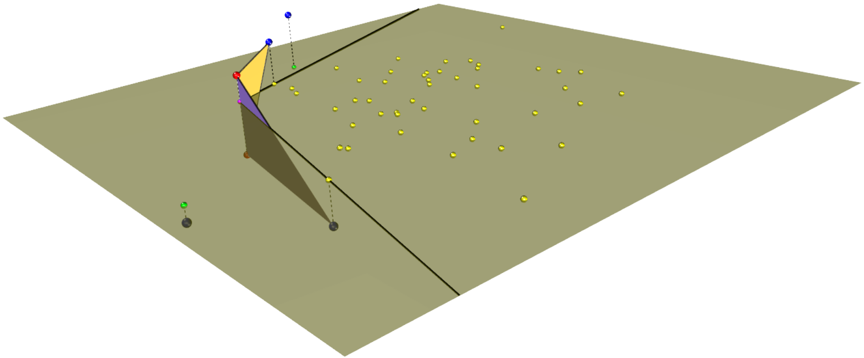
Univariate projection quantiles (illustration)



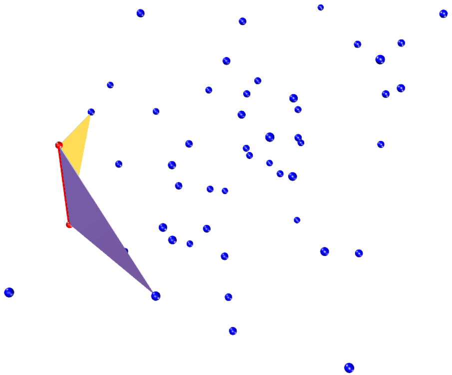
Univariate projection quantiles (illustration)



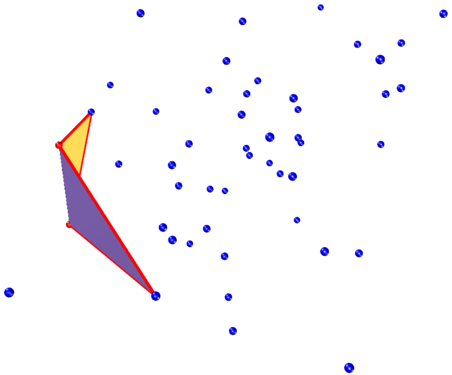
Univariate projection quantiles (illustration)



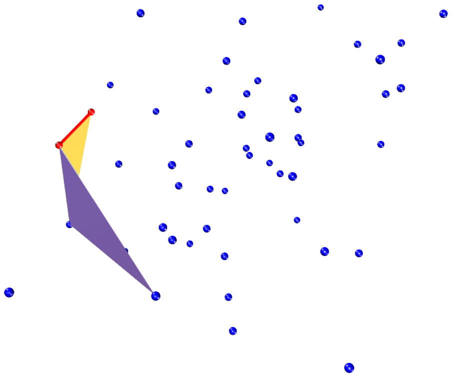
Univariate projection quantiles (illustration)



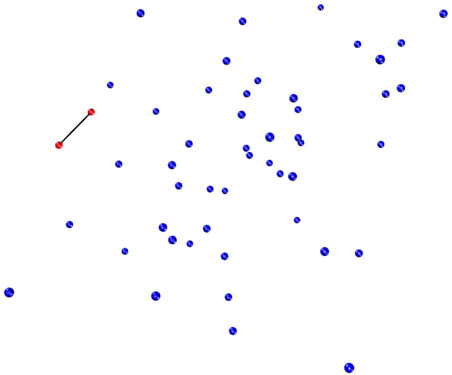
Univariate projection quantiles (illustration)



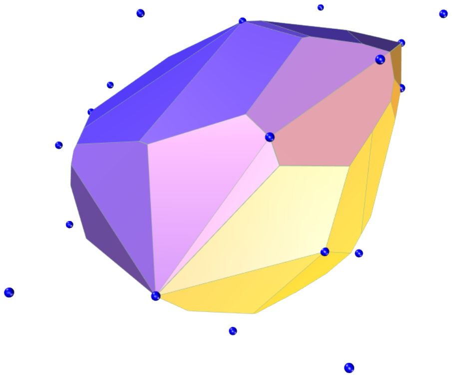
Univariate projection quantiles (illustration)



Univariate projection quantiles (illustration)



Univariate projection quantiles (illustration)



Complexity

Notation:

- n — number of points;
- d — dimension;
- $\sum_{\boldsymbol{u}}$ — number of relevant directions.

Algorithmic complexity in terms of time:

- Directional quantiles: $O(n \sum_{\boldsymbol{u}})$
- Multiple-output regression quantiles: $O_i + O(n \sum_{\boldsymbol{u}})$

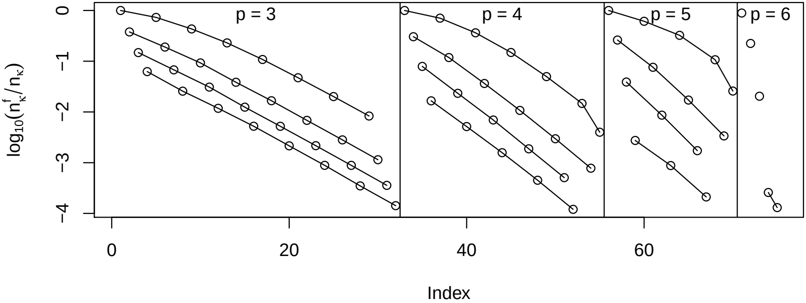
  $\sum_{\boldsymbol{u}}$ is never larger than $O(n^d)$ $\Longrightarrow$ $O(n^{d+1})$

  $\sum_{\boldsymbol{u}}$ is on an average $O(n^{d-1})$ $\Longrightarrow$ av. $O(n^d)$
- Univariate projection quantiles: $O_i + O(n \log n \sum_{\boldsymbol{u}})$

  $\sum_{\boldsymbol{u}}$ is never larger than $\binom{n}{d-1}$ $\Longrightarrow$ $O(n^d \log n)$

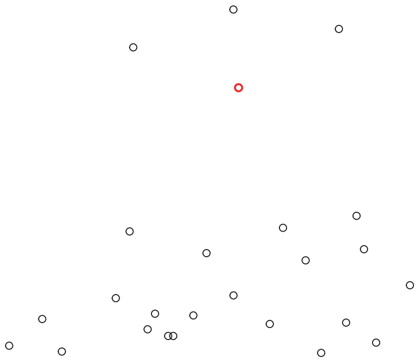
Relevant hyperplanes/facets

Logarithmized (average) ratio of the number of facets of the Tukey region to the number of relevant hyperplanes:

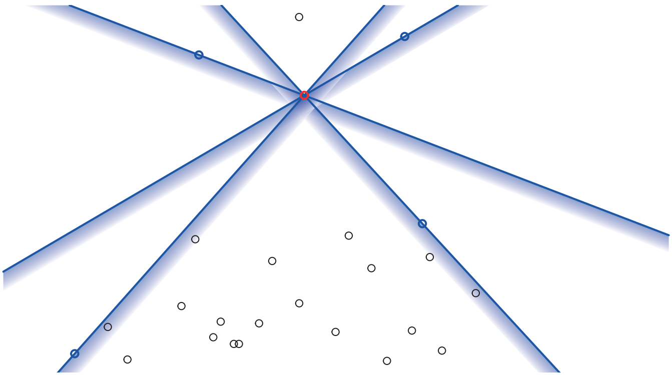


- Cases correspond to dimensions: $p = 3, 4, 5, 6$.
- Lines correspond to depths: $\tau = 0.025, 0.1, 0.2, 0.3$.
- Points correspond to sample sizes:
  $n = 40, 80, 160, 320, 640, 1280, 2560, 5120$.

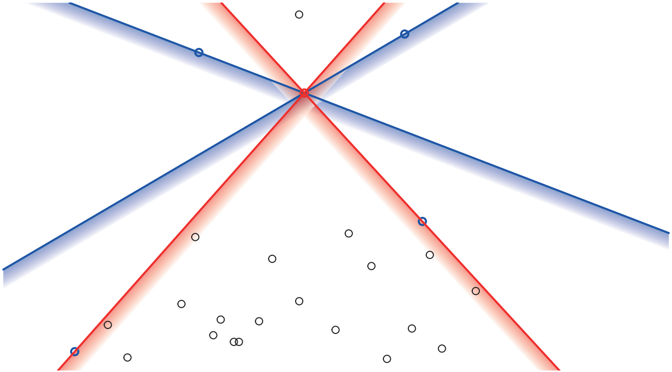
Upper bound on the number of facets of $D_\tau^{T(n)}$



Upper bound on the number of facets of $D_\tau^{T(n)}$



Upper bound on the number of facets of $D_\tau^{T(n)}$



Proposition

For a given data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ in $\mathbb{R}^d$, the number of the (non-redundant) facets of the Tukey region $D_\tau^{T(n)}(\boldsymbol{X})$ is bounded from above by $2\binom{n}{d-1}$.

Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

Tukey median (algorithm)

Input: $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \in \mathbb{R}^d$, $2 < d < n < \infty$.

Step 1: Initialize bounds on τ^*:

1.1: Compute coordinate-wise median $\boldsymbol{x}_0$.
1.2: Compute $d_0 = D^T(\boldsymbol{x}_0|\boldsymbol{X})$.
1.3: Set $\tau_{low} = \max\left\{\frac{1}{n}\lceil\frac{n}{p+1}\rceil, d_0\right\}$, $\tau_{up} = \frac{1}{n}\lfloor\frac{n-p+2}{2}\rfloor + \frac{1}{n}^{**}$.

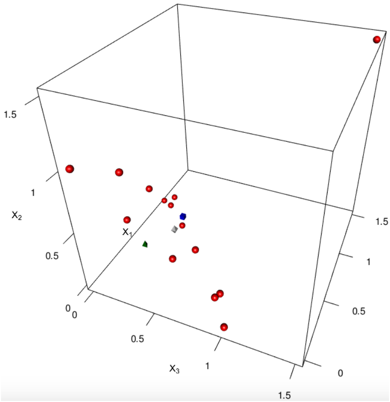
Step 2: Update bounds:

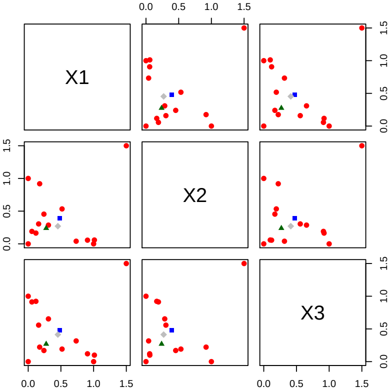
2.1: Let $\overline{\tau} = \frac{1}{n}\lfloor\frac{n(\tau_{low}+\tau_{up})}{2}\rfloor$, and compute the region $\mathcal{D}(\overline{\tau}) = D^T_{\overline{\tau}}(\boldsymbol{X})$.
2.2: If $\mathcal{D}(\overline{\tau})$ does not exist then set $\tau_{up} = \overline{\tau}$,
otherwise: calculate the barycenter **c** of $\mathcal{D}(\overline{\tau})$ and set $\tau_{low} = D^T(\boldsymbol{c}|\boldsymbol{X})$.
2.3: If $\tau_{low} < \tau_{up} - \frac{1}{n}$, then repeat **Step 2**, else stop.

Output: The Tukey median $D^T_{\tau_{low}}(\boldsymbol{X})$.

**Liu, Luo, Zuo (2017). Some results on the computing of Tukey's halfspace median. Statistical Papers. `https://doi.org/10.1007/s00362-017-0941-5`.

Tukey median (illustration)





	location estimators	x_1	x_2	x_3	depth
(cube)	mean	0.480	0.393	0.476	$\frac{1}{14}$
(tetrahedron)	coordinate-wise median	0.275	0.239	0.269	0
(rhombus)	Tukey median	0.454	0.27	0.413	$\frac{4}{14}$

Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

The projection property

A depth D satisfies the **projection property** (Dyckerhoff, '04) if

$$D(\boldsymbol{z}|X) = \inf_{\boldsymbol{p}\in\mathbb{S}^{d-1}} D(\boldsymbol{p}^T\boldsymbol{z}|\boldsymbol{p}^T X)\,.$$

Proposition (Dyckerhoff, '04)

- *Let D be a depth satisfying* **projection property***,*
- $\mathbb{S}^{d-1} \to [0,\infty)$, $\boldsymbol{p} \mapsto D(\boldsymbol{p}^T\boldsymbol{z}|\boldsymbol{p}^T X)$ **upper-semicontinuous***,*
- $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots$ *a sequence of independent identically uniformly on $\mathbb{S}^{d-1}$ distributed random vectors.*

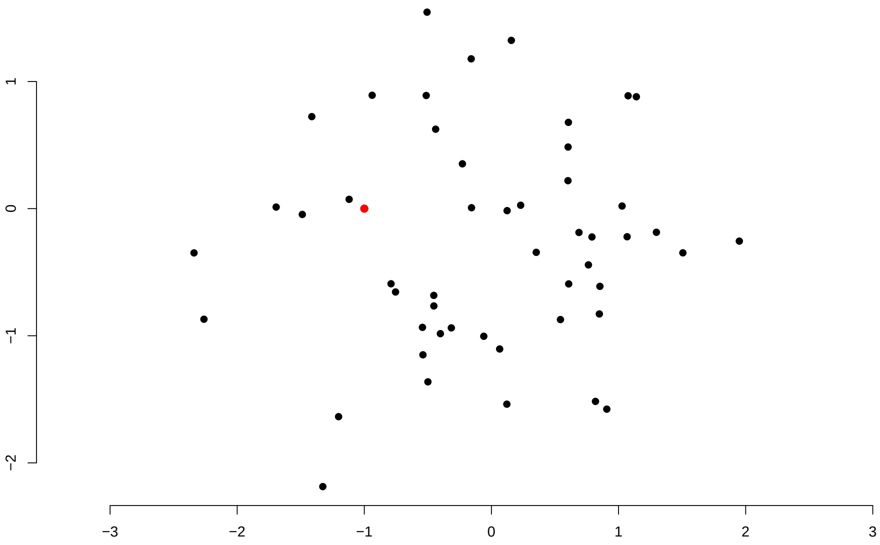
Then with probability one

$$\min_{1\le i\le m} D(\boldsymbol{p}_i^T\boldsymbol{z}|\boldsymbol{p}_i^T X) \xrightarrow{m\to\infty} D(\boldsymbol{z}|X)\,.$$

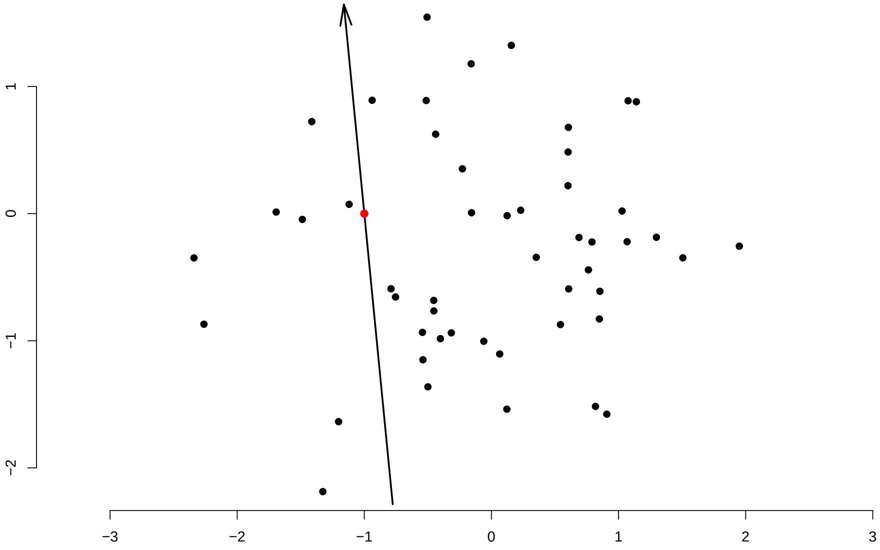
Application to the Tukey depth:

$$\begin{aligned}\underbrace{D^T(\boldsymbol{z}|X)}_{d\text{-variate depth}} &= \inf_{\boldsymbol{p}\in\mathbb{S}^{d-1}} \overbrace{D^T(\boldsymbol{p}^T\boldsymbol{z}|\boldsymbol{p}^T X)}^{\text{univariate depth}} \\ &= \inf_{\boldsymbol{p}\in\mathbb{S}^{d-1}} \min\{\, F_{\boldsymbol{p}^T X}(\boldsymbol{p}^T\boldsymbol{z})\,,\, 1 - F_{\boldsymbol{p}^T X}(\boldsymbol{p}^T\boldsymbol{z}^-)\,\}\,.\end{aligned}$$

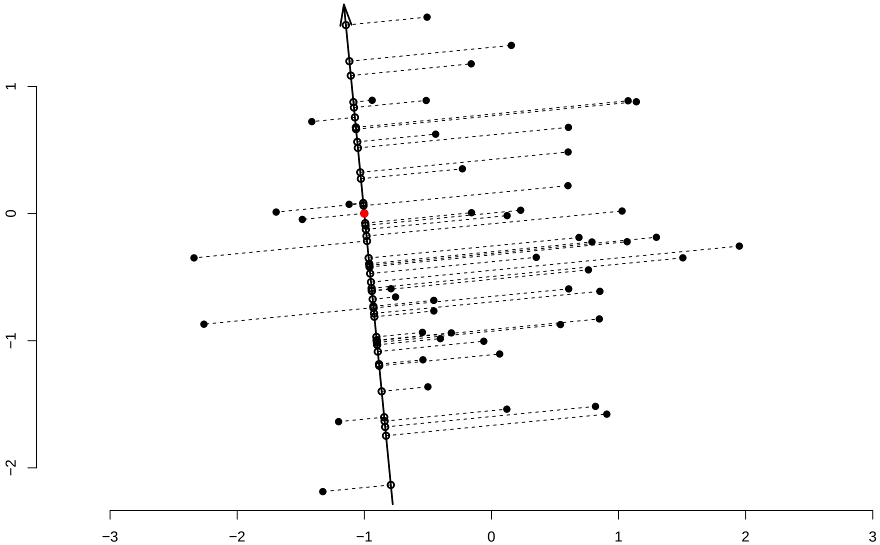
Halfspace depth: approximate computation



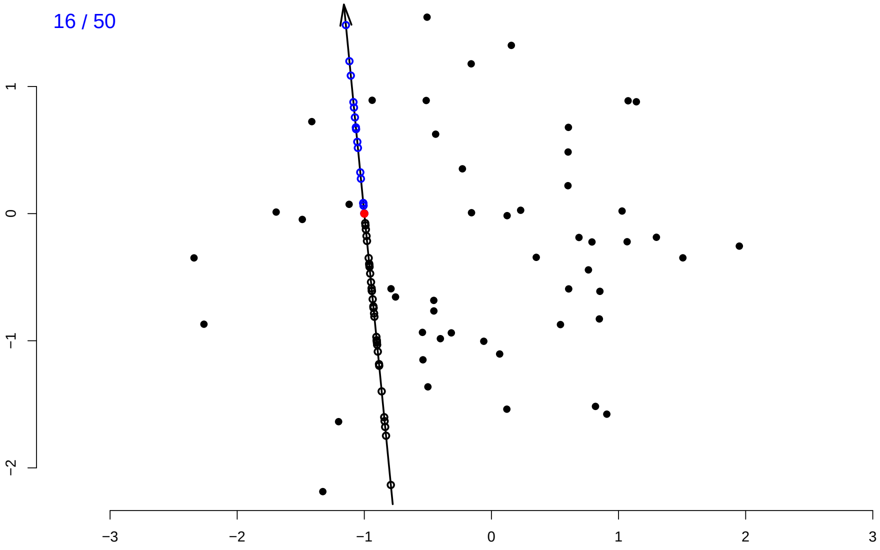
Halfspace depth: approximate computation



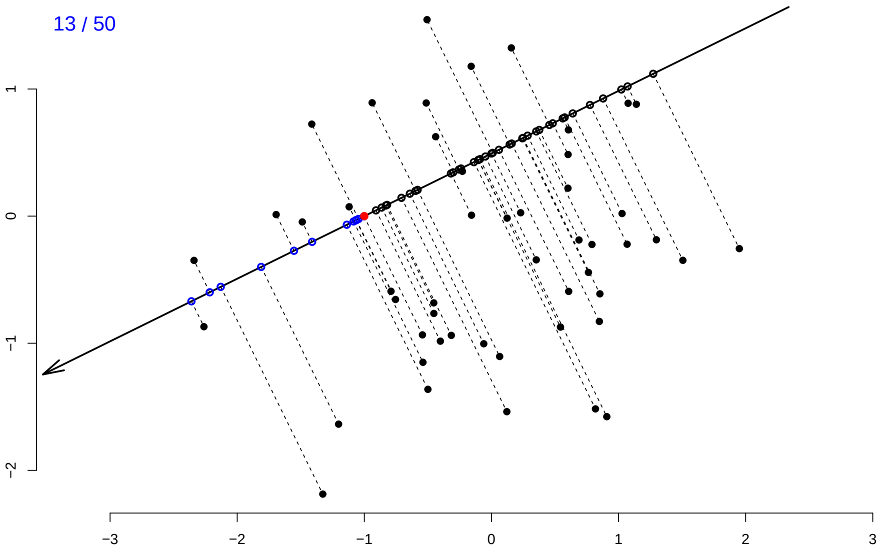
Halfspace depth: approximate computation



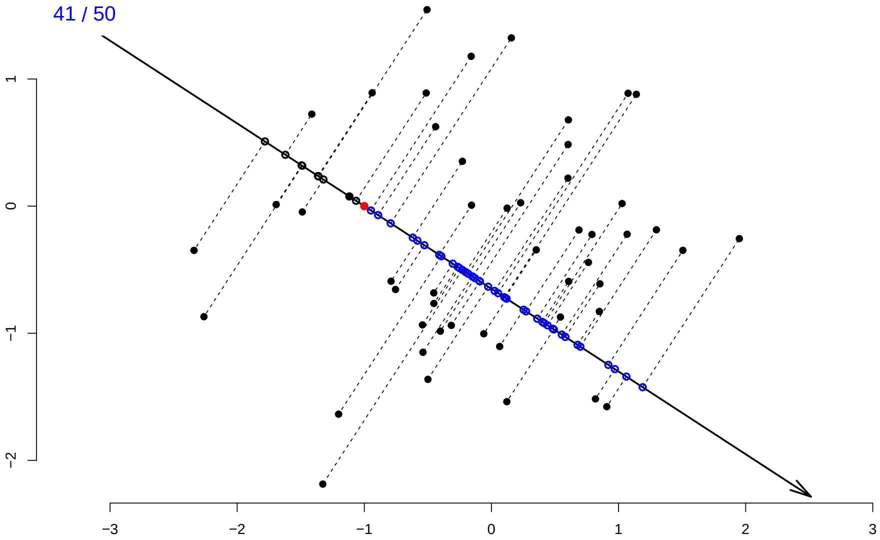
Halfspace depth: approximate computation



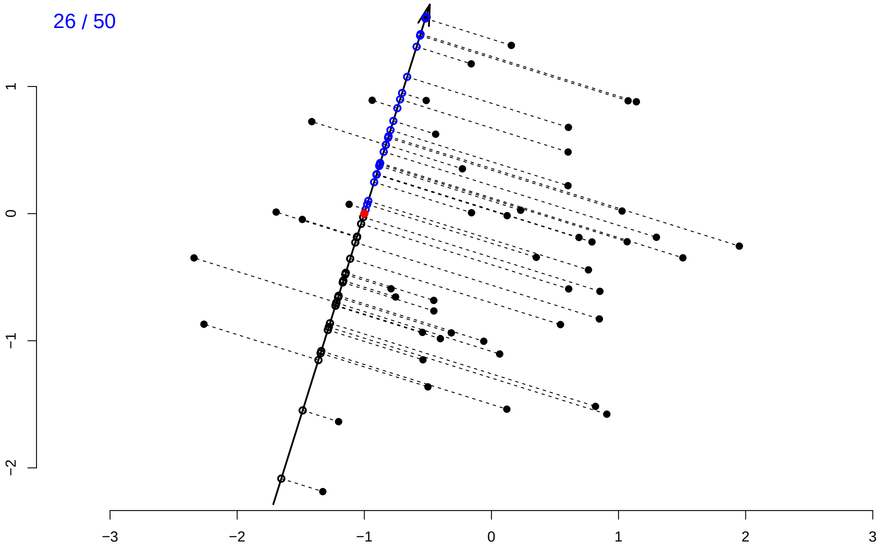
Halfspace depth: approximate computation



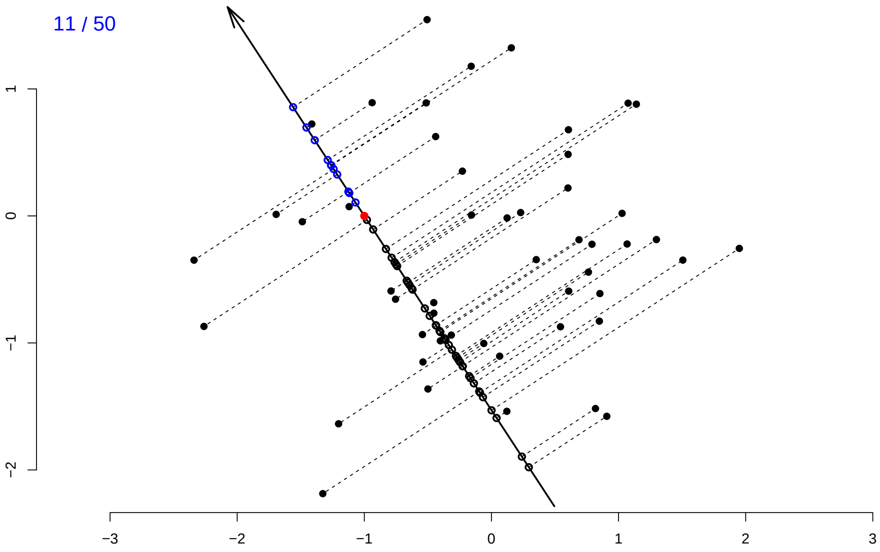
Halfspace depth: approximate computation



Halfspace depth: approximate computation



Halfspace depth: approximate computation



Computation of projection depths: Remarks

- A number of depths (**Mahalanobis**, **halfspace**, **zonoid**, **ect.**) satisfy the assumptions of the preceding proposition.
- The projection property enables the approximate computation of depths **even in high dimensions**.
- Due to affine invariance one can restrict $\boldsymbol{p}$ to a **hemisphere**.
- Since a large number of **univariate depths** have to be computed there is **need in efficient algorithms** for them.
- Many univariate depths can be exactly computed with time **complexity** $O(n)$.
- Computing the approximate depth leads to **optimization** of a (non-differentiable) **function on the sphere** $\mathbb{S}^{d-1}$. So what does the function

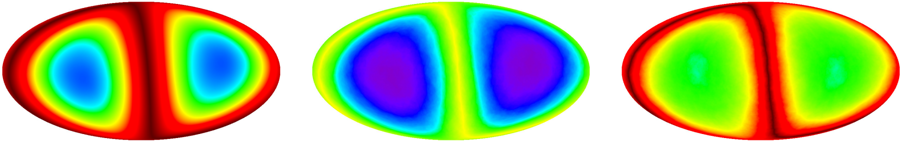
$$\mathbb{S}^{d-1} \to \mathbb{R},\, \boldsymbol{p} \mapsto D(\boldsymbol{p}^T \boldsymbol{z} | \boldsymbol{p}^T X)$$

look like? For example, does it have many local minima?

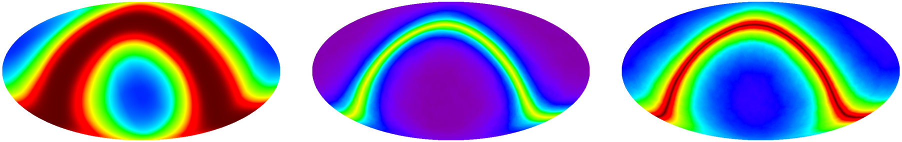
Illustration: The map $\boldsymbol{p} \mapsto D(\boldsymbol{p}^T \boldsymbol{z} | \boldsymbol{p}^T X)$ in $\mathbb{R}^3$

Zonoid depth	Tukey depth	Projection depth

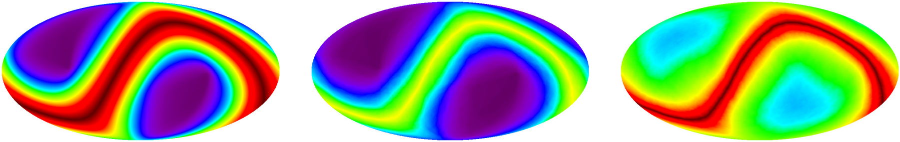
Trivariate normal distribution, $n = 1000$



Trivariate Cauchy distribution, $n = 1000$



Trivariate uniform distribution, $n = 1000$



Contents

Data depth
- Tukey depth: definition
- Tukey trimmed regions
- Applications

Computation of the Tukey depth
- Theoretical background
- Algorithm and simulations

Computation of Tukey trimmed regions
- Existing approaches
- The proposed algorithm
- Tukey median

Outlook for approximations

Conclusions

Some conclusions

- Developed a **family of** exact **algorithms for** computing **Tukey depth** in the Euclidean space of any dimension, which are **implemented in** `R`-package **ddalpha** accessible on CRAN.
- Exact **algorithm for** computing **Tukey trimmed regions**, **implemented in** `R`-package **TukeyRegion**.
- *Ongoing work:* development of **approximate algorithms for** computation of bf depths that satisfy the projection property.

Open challenges:

- Still **high complexity** and computation time scales **exponentially with dimension**.
- **Precision challenge**: too many potentially relevant hyperplanes, precision-constant problems.
- For approximation: adaptation of the algorithm(s) to **optimization over the surface of a hypersphere**.

Thank you for your attention! Questions?

- Rousseeuw, P.J. and Ruts, I. (1996). Algorithm AS 307: Bivariate location depth. *Journal of the Royal Statistical Society, Series C*, 45, 516-526.
- Ruts, I. and Rousseeuw, P.J. (1996). Computing depth contours of bivariate point clouds. *Computational Statistics and Data Analysis*, 23, 153-168.
- Rousseeuw, P.J. and Struyf, A. (1998). Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 8, 193-203.
- Hallin, M., Paindaveine, D., and Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: From L_1 optimization to halfspace depth. *The Annals of Statistics*, 38, 635-669.
- Paindaveine, D. and Šiman, M. (2011). On directional multiple-output quantile regression. *Journal of Multivariate Analysis*, 102, 193–212.
- Liu, X. and Zuo, Y. (2014). Computing halfspace depth and regression depth. *Communications in Statistics - Simulation and Computation*, 43, 969-985.
- Liu, X. (2017). Fast implementation of the Tukey depth. *Computational Statistics*, 32, 1395-1410.