Anomaly detection using data depth: multivariate case

Pavlo Mozharovskyi

LTCI, Télécom Paris, Institut Polytechnique de Paris

Journée de la recherche du LTCI

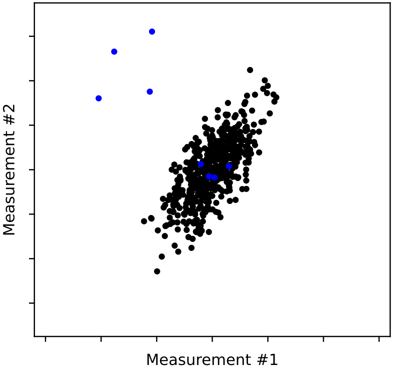
Palaiseau, October 14, 2022

Contents

Contents

A real task

Regard two measurements during a test in a production process:

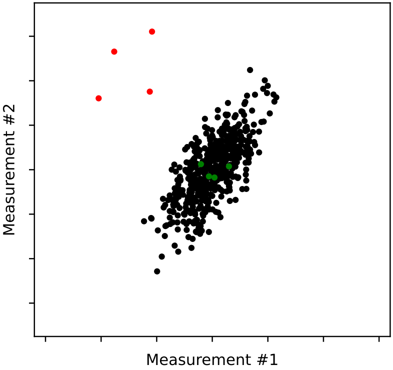


Given **training data**, polluted or not with anomalies:

- detect **anomalies** in the given data.

A real task

Regard two measurements during a test in a production process:



Given **training data**, polluted or not with anomalies:

- detect **anomalies** in the given data.

For **new data**, determine:

- Whether new observations are normal data or anomalies?

Multivariate framework

- A training data set:

$$\boldsymbol{X}_{tr} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subset \mathbb{R}^d$$

  of observations in the d-dimensional Euclidean space.

- Typical example: a table from a data base, with lines being observations (=individuals, items,...).
- Construct a decision function:

$$\mathbb{R}^d \rightarrow \{0, 1\} : \boldsymbol{x} \mapsto g(\boldsymbol{x}) \,,$$

  which attributes to any (possible) $\boldsymbol{x} \in \mathbb{R}^d$ a label whether it is an anomaly (*e.g.*, 1) or a normal observation (*e.g.*, 0).

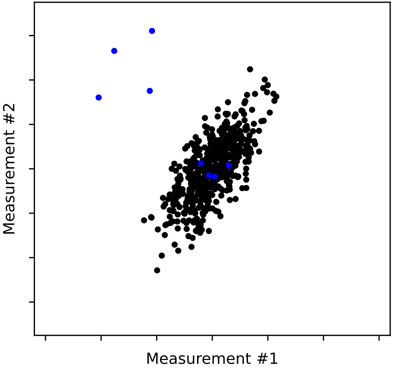
- It is more useful to provide an ordering on $\mathbb{R}^d$:

$$\mathbb{R}^d \rightarrow \mathbb{R} : \boldsymbol{x} \mapsto g(\boldsymbol{x}) \,,$$

  such that abnormal observations obtain differing anomaly score.

Teaser

Same data set with two measurements:

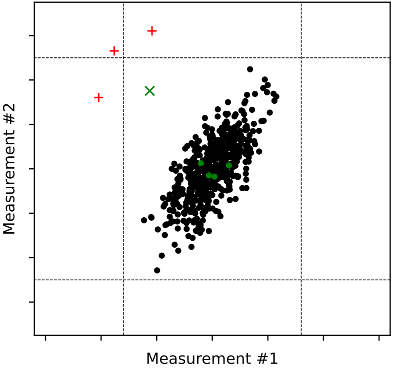


Given **training data**, polluted or not with anomalies:

- detect **anomalies** in the given data.

Teaser

Same data set with two measurements:



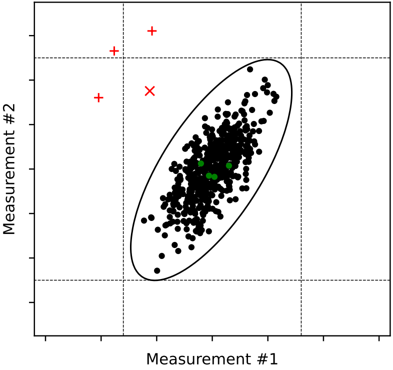
For **new data**, employ:

- Anomaly detection rule using bounding box:

$$g_{\text{box}}(\boldsymbol{x}|\boldsymbol{X}_{tr}) = \begin{cases} \text{anoamly}\,(=1)\,, & \text{if } \boldsymbol{x} \notin \bigcap_{j=1,\ldots,d}(\underline{H}_{j,l_j} \cap \overline{H}_{j,u_j})\,, \\ \text{normal}\,(=0)\,, & \text{otherwise.} \end{cases}$$

Teaser

Same data set with two measurements:



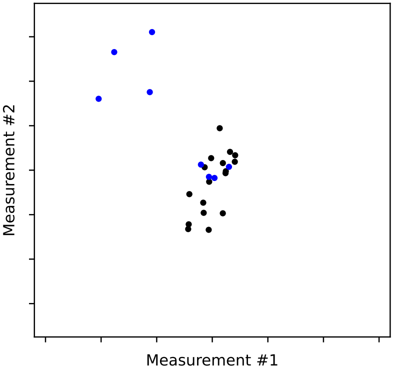
For **new data**, employ:

- Anomaly detection rule using Mahalanobis depth:

$$g_{\text{Mah}}(\boldsymbol{x}|\boldsymbol{X}_{tr}) = \begin{cases} \text{anomaly}\,, & \text{if } D^{\text{Mah}}(\boldsymbol{x}|\boldsymbol{X}_{tr}) < t_{\text{Mah},\boldsymbol{X}_{tr}}\,, \\ \text{normal}\,, & \text{otherwise.} \end{cases}$$

Teaser

Same data set with two measurements, but **less observations**:

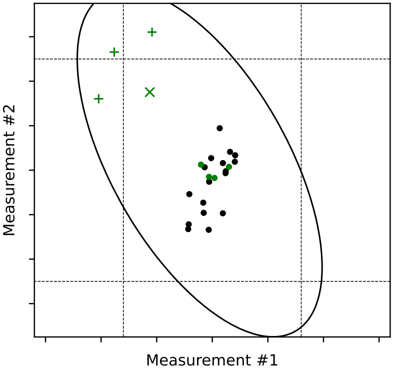


Given **training data**, polluted or not with anomalies:

- detect **anomalies** in the given data.

Teaser

Same data set with two measurements, but **less observations**:



Given **training data**, polluted or not with anomalies:

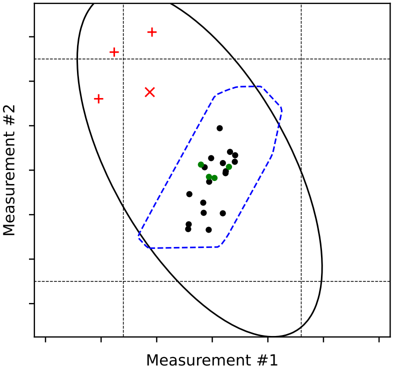
- detect **anomalies** in the given data.

For **new data**, employ:

- Anomaly detection rule using Mahalanobis depth.

Teaser

Same data set with two measurements, but **less observations**:



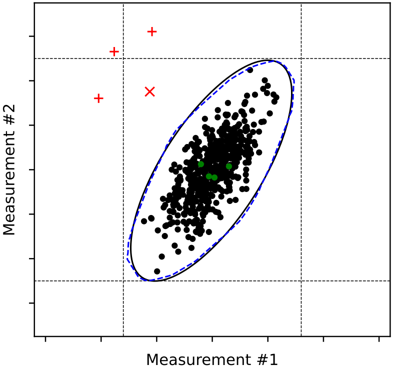
For **new data**, employ:

- Anomaly detection rule using projection depth:

$$g_{\text{prj}}(\boldsymbol{x}|\boldsymbol{X}_{tr}) = \begin{cases} \text{anomaly}, & \text{if } D^{\text{prj}}(\boldsymbol{x}|\boldsymbol{X}_{tr}) < t_{\text{prj},\boldsymbol{X}_{tr}}, \\ \text{normal}, & \text{otherwise.} \end{cases}$$

Teaser

Now **back to big data** set with two measurements:



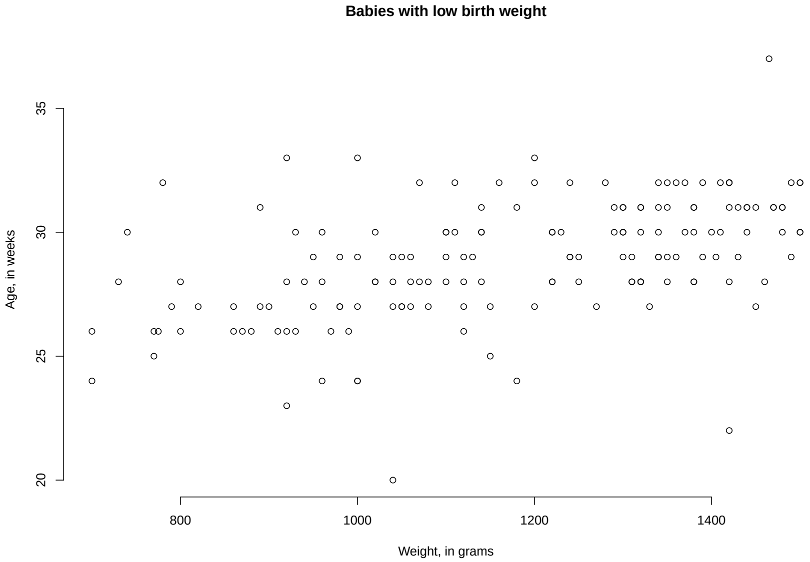
For **new data**, employ:

- Anomaly detection rule using projection depth:

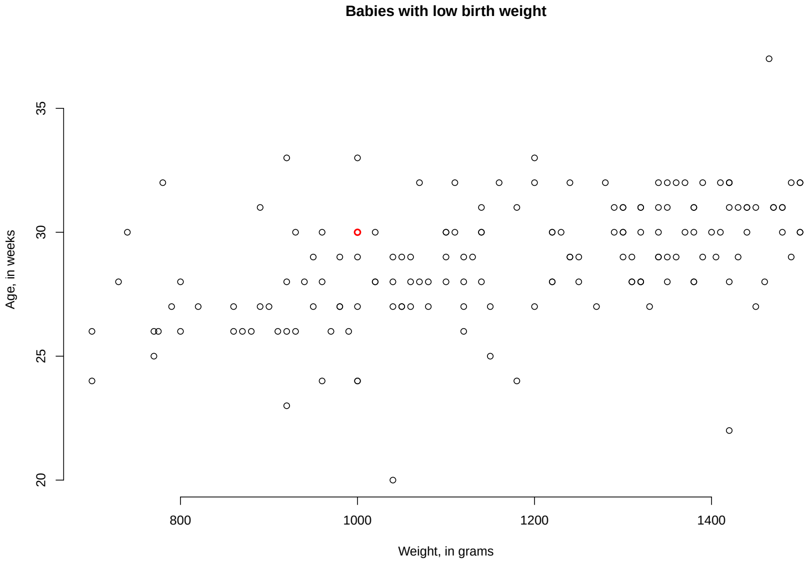
$$g_{\text{prj}}(\boldsymbol{x}|\boldsymbol{X}_{tr}) = \begin{cases} \text{anomaly}, & \text{if } D^{\text{prj}}(\boldsymbol{x}|\boldsymbol{X}_{tr}) < t_{\text{prj},\boldsymbol{X}_{tr}}, \\ \text{normal}, & \text{otherwise.} \end{cases}$$

Contents

Data depth


Babies with low birth weight
Age, in weeks
20
25
30
35
800
1000
1200
1400
Weight, in grams

Data depth



Babies with low birth weight
35
30
25
20
Age, in weeks
800
1000
1200
1400
Weight, in grams

Statistical data depth

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

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

that is:

D1 **translation invariant:** $D(\boldsymbol{x} + b|X + b) = D(\boldsymbol{x}|X)$ for any $b \in \mathbb{R}^p$;

D2 **linear invariant:** $D(A\boldsymbol{x}|AX) = D(\boldsymbol{x}|X)$ for any $p \times p$ non-singular matrix A;

D3 **vanishing at infinity:** $\lim_{||\boldsymbol{x}|| \to \infty} D(\boldsymbol{x}|X) = 0$;

D4 **monotone on rays:** for any $\boldsymbol{x}^* \in \operatorname{argmax}_{\boldsymbol{x} \in \mathbb{R}^p} D(\boldsymbol{x}|X)$, any $\boldsymbol{x} \in \mathbb{R}^p$, and any $0 \leq \alpha \leq 1$ it holds: $D(\boldsymbol{x}|X) \leq D(\boldsymbol{x}^* + \alpha(\boldsymbol{x} - \boldsymbol{x}^*)|X)$;

D5 **upper semicontinuous in $\boldsymbol{x}$:** the upper-level sets $D_\alpha(X) = \{\boldsymbol{x} \in \mathbb{R}^p : D(\boldsymbol{x}|X) \geq \alpha\}$ are closed for all α.

Halfspace (=Tukey, location) depth

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

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

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

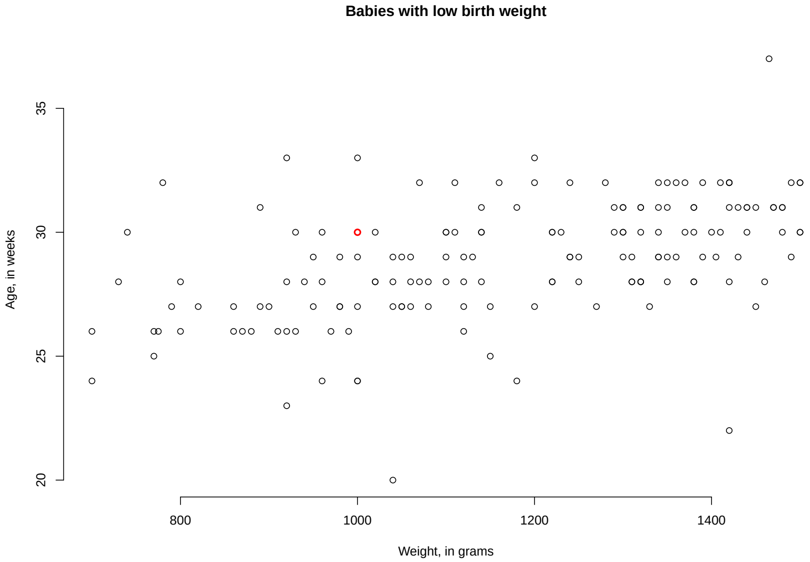
and w.r.t. a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subset \mathbb{R}^p$:

$$D^{h(n)}(\boldsymbol{x}|\boldsymbol{X}) = \frac{1}{n} \min_{\boldsymbol{u} \in \mathbb{S}^{p-1}} \sharp\{i : \boldsymbol{u}'\boldsymbol{x}_i \geq \boldsymbol{u}'\boldsymbol{x}\}.$$

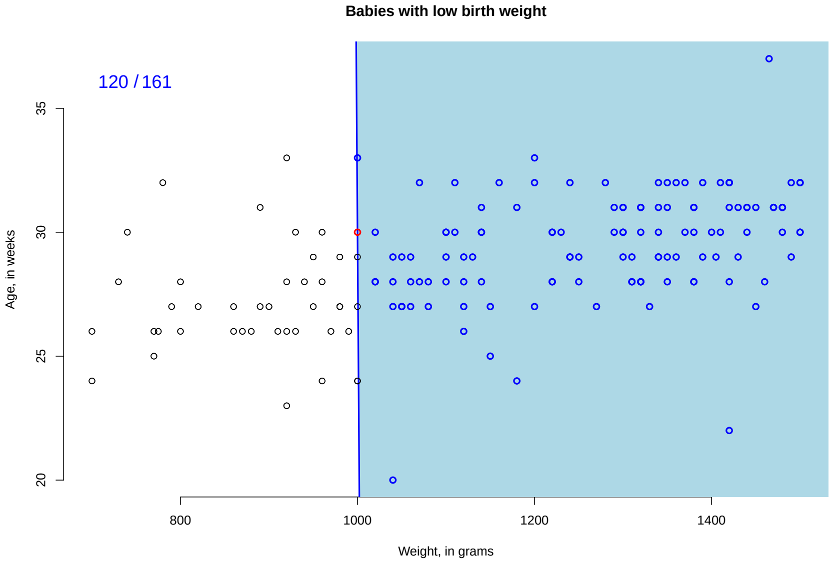
Halfspace depth

- satisfies all the above postulates,
- is purely non-parametric and robust,
- has direct connection to quantiles and many applications.

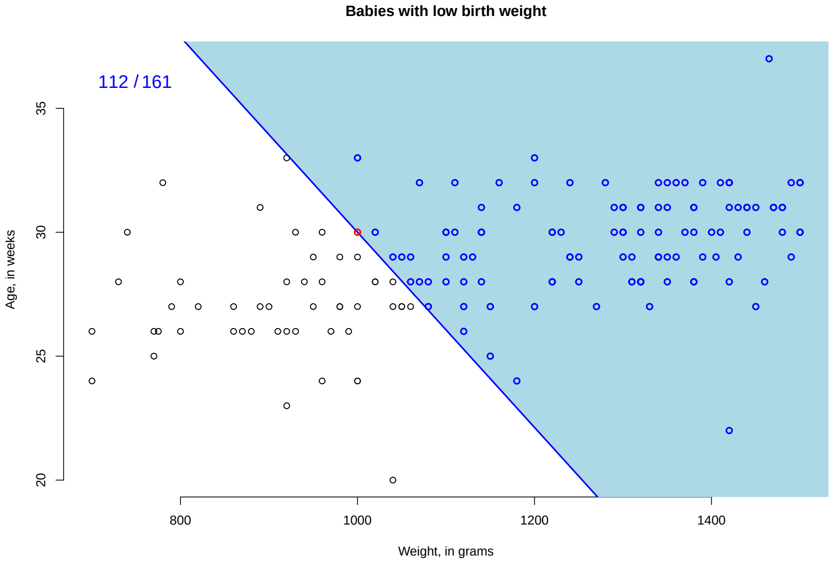
Halfspace (=Tukey, location) data depth



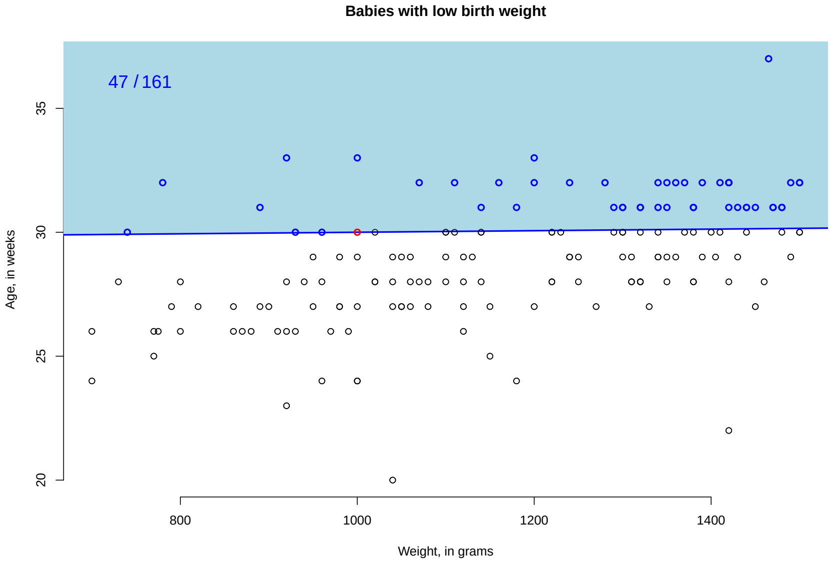
Halfspace (=Tukey, location) data depth



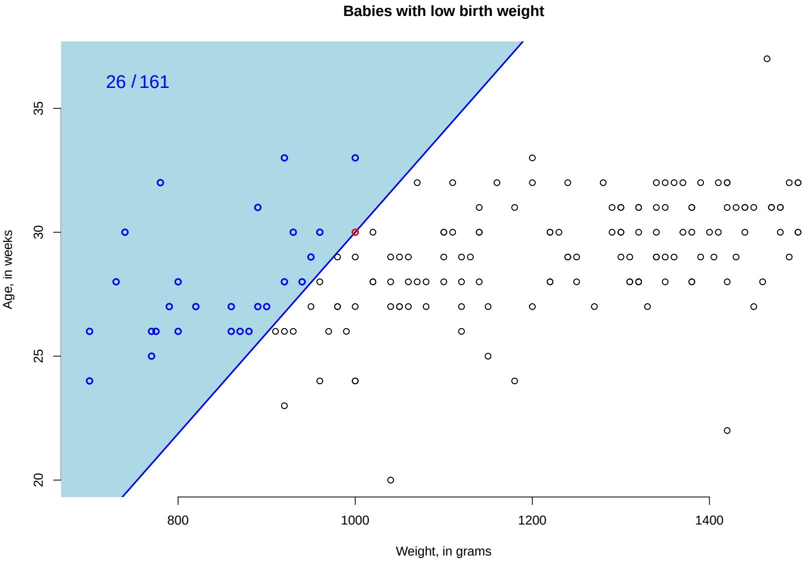
Halfspace (=Tukey, location) data depth



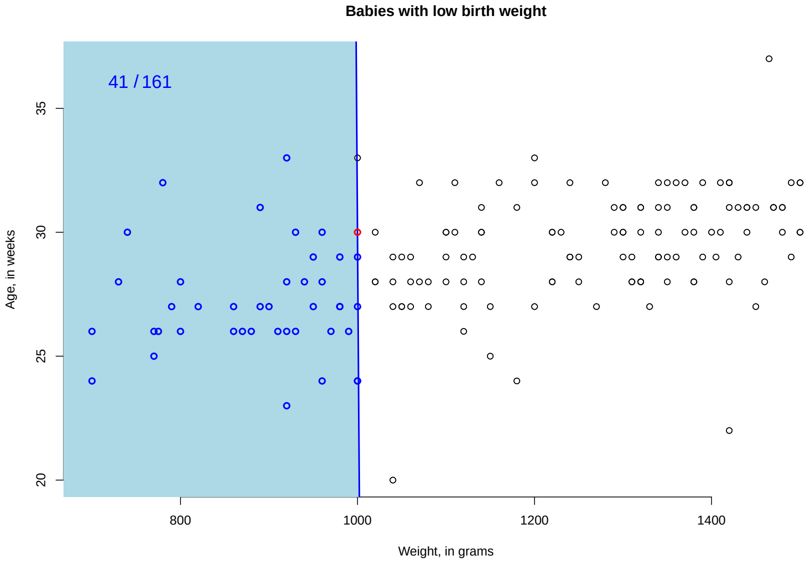
Halfspace (=Tukey, location) data depth



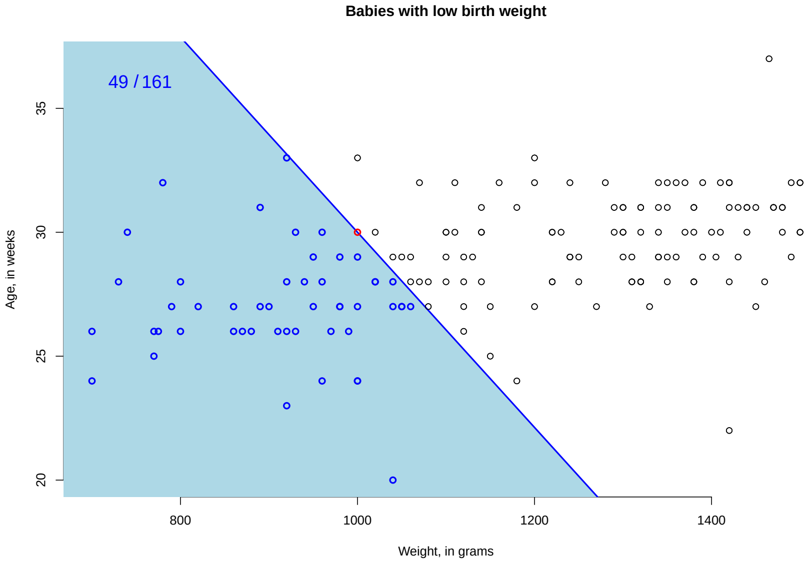
Halfspace (=Tukey, location) data depth



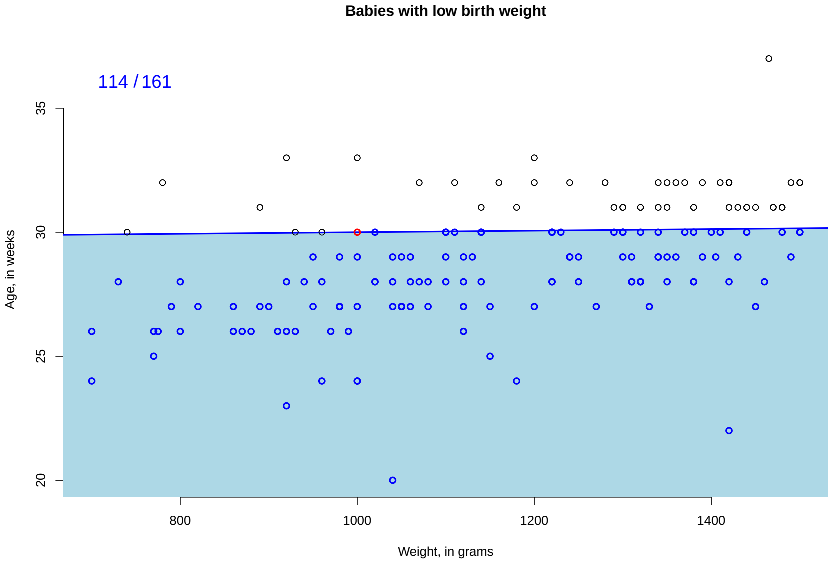
Halfspace (=Tukey, location) data depth



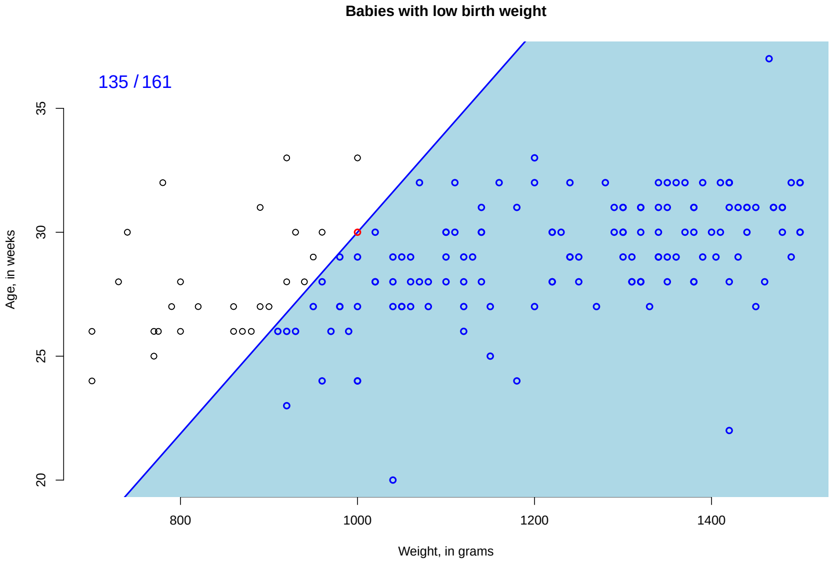
Halfspace (=Tukey, location) data depth



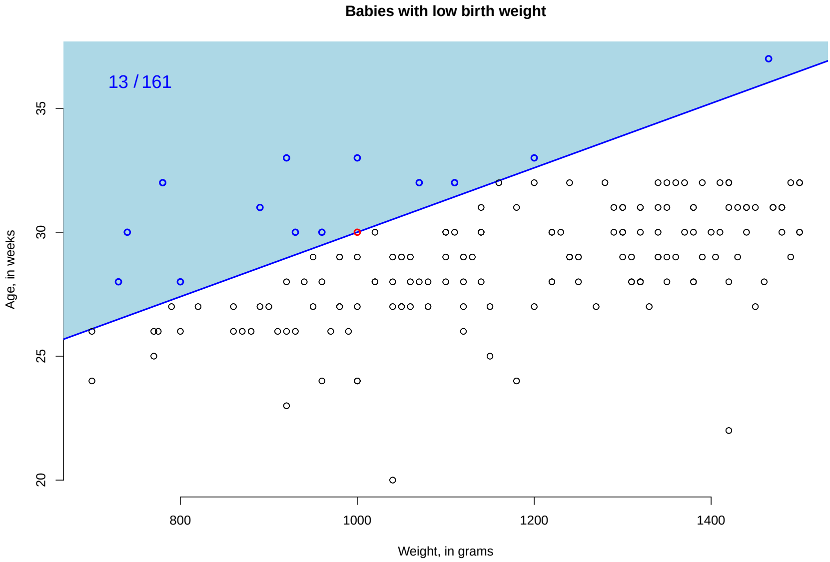
Halfspace (=Tukey, location) data depth



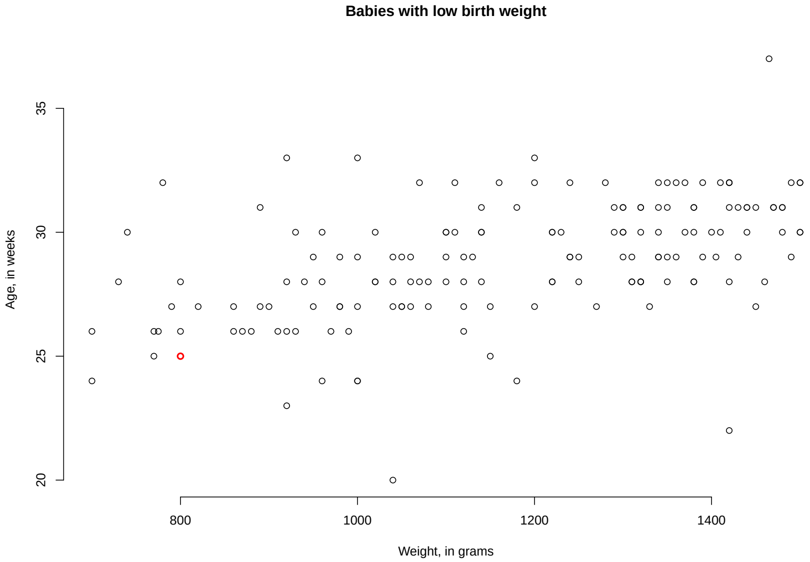
Halfspace (=Tukey, location) data depth



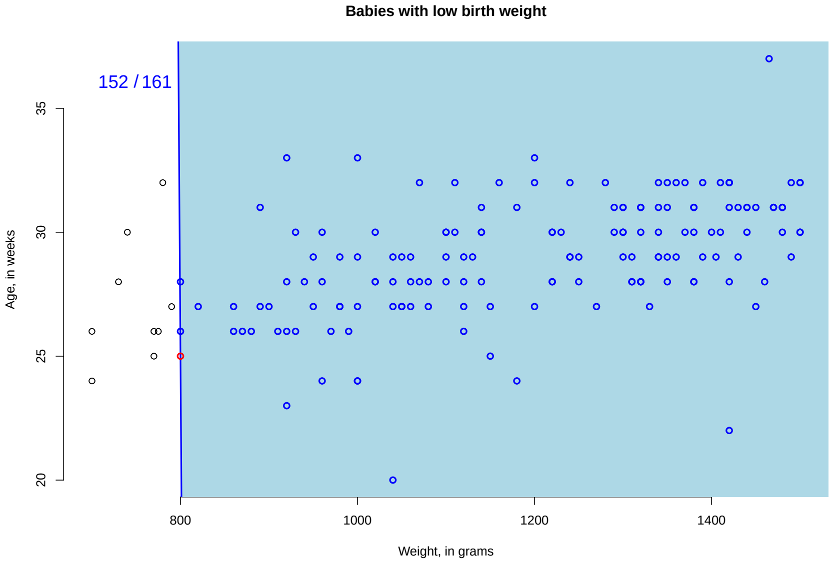
Halfspace (=Tukey, location) data depth



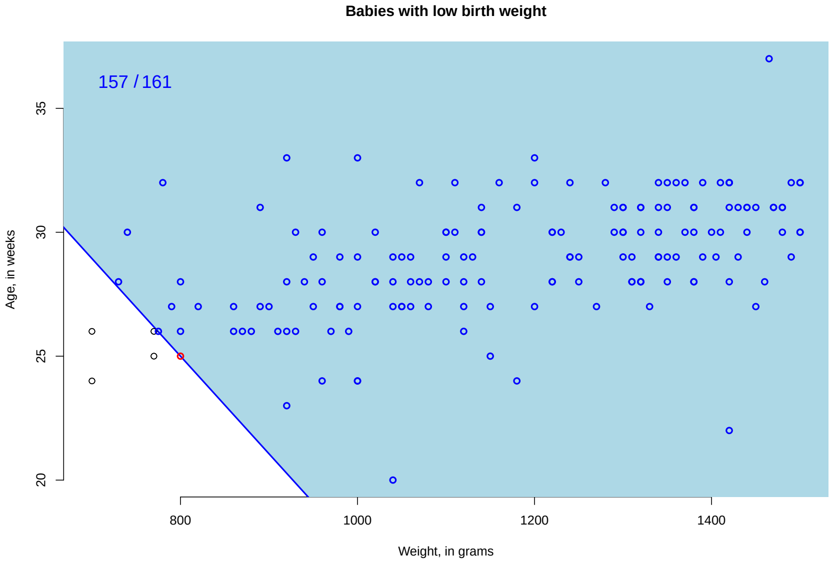
Halfspace (=Tukey, location) data depth



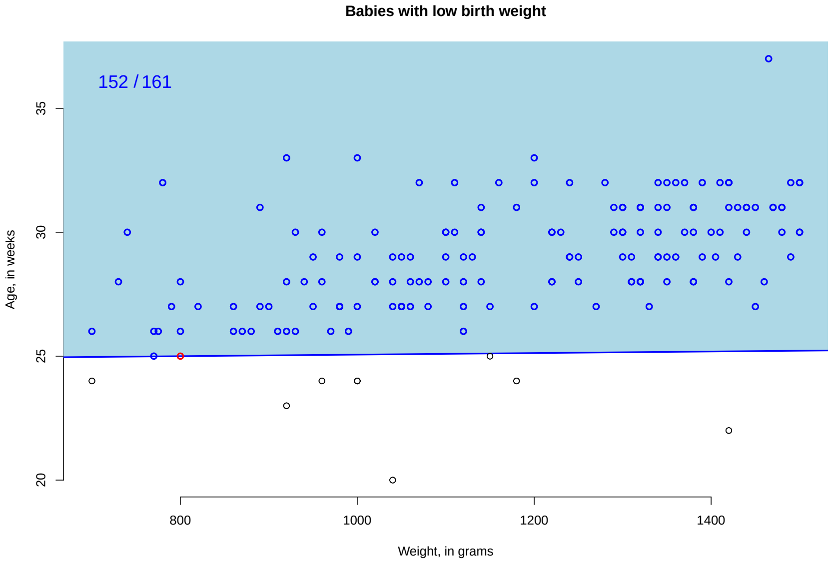
Halfspace (=Tukey, location) data depth



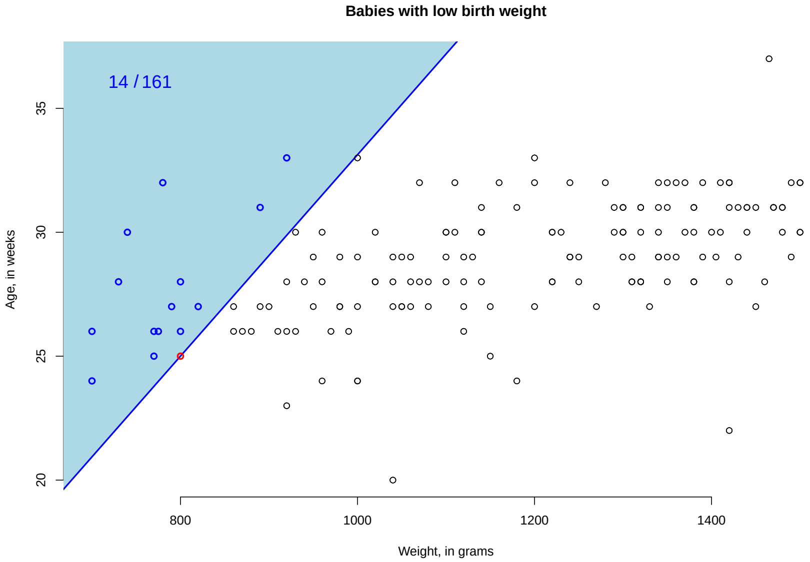
Halfspace (=Tukey, location) data depth



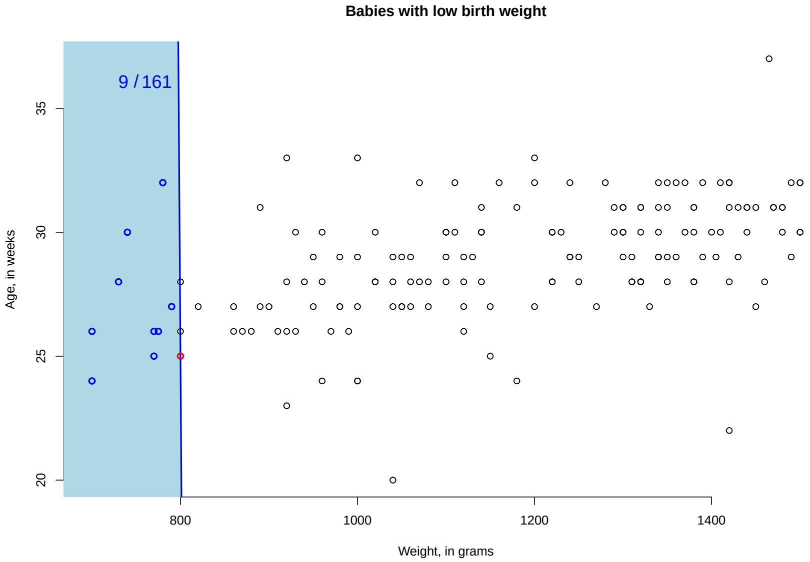
Halfspace (=Tukey, location) data depth



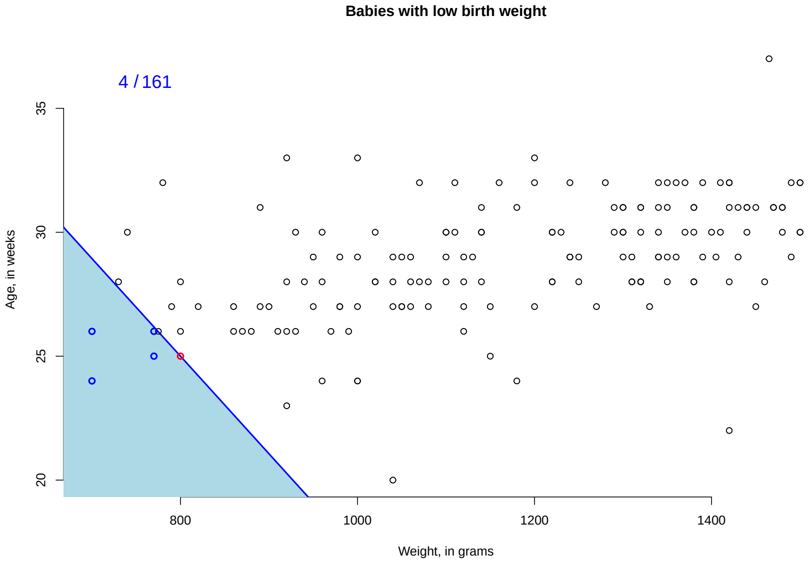
Halfspace (=Tukey, location) data depth



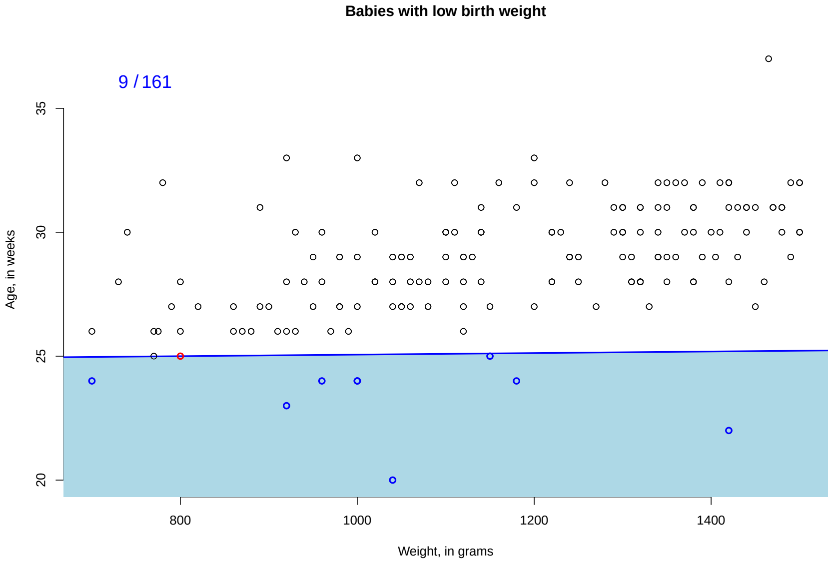
Halfspace (=Tukey, location) data depth



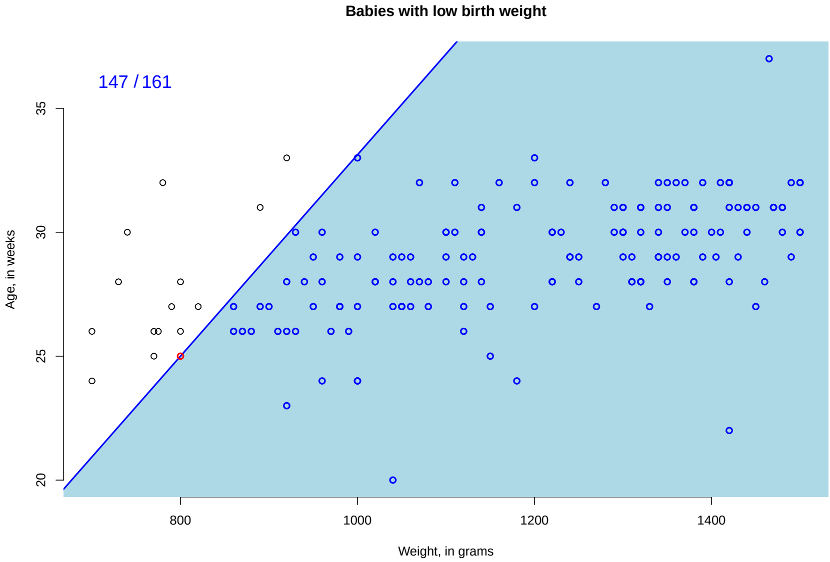
Halfspace (=Tukey, location) data depth



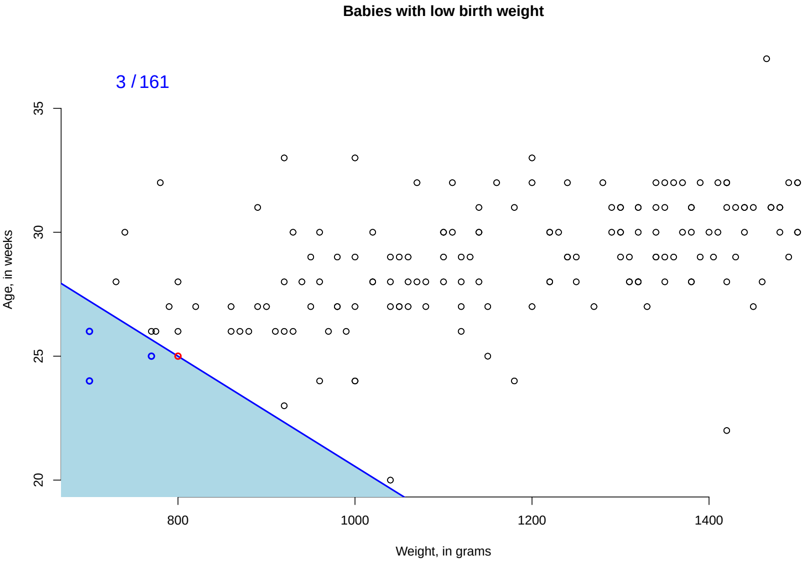
Halfspace (=Tukey, location) data depth



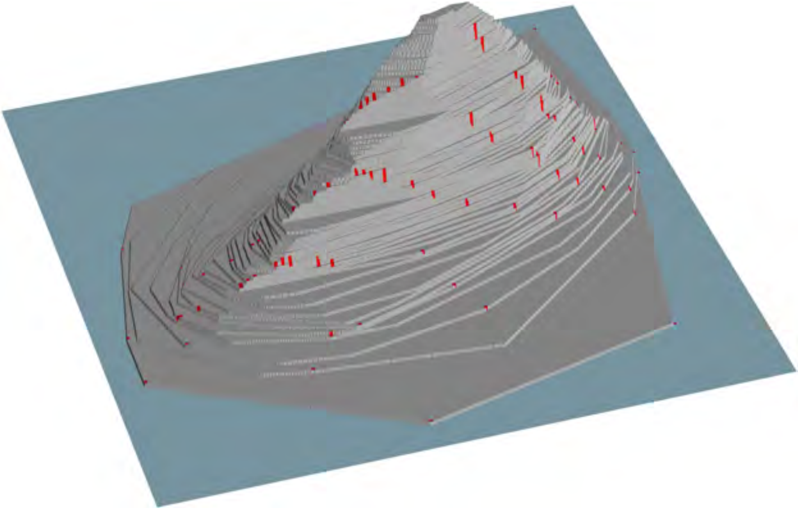
Halfspace (=Tukey, location) data depth



Halfspace (=Tukey, location) data depth



Halfspace (=Tukey, location) data depth



Halfspace-trimmed regions

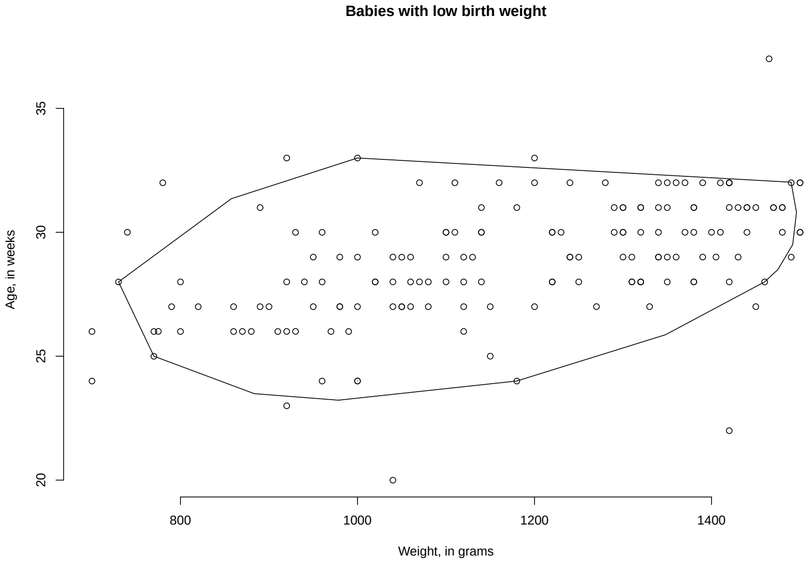
Halfspace depth defines a family of (depth-)trimmed (central) regions $D^h_\tau(X)$, the upper-level sets of the depth function:

$$D^h_\tau(X) = \{\boldsymbol{x} \in \mathbb{R}^p \,:\, D^h(\boldsymbol{x}|X) \geq \tau\}.$$

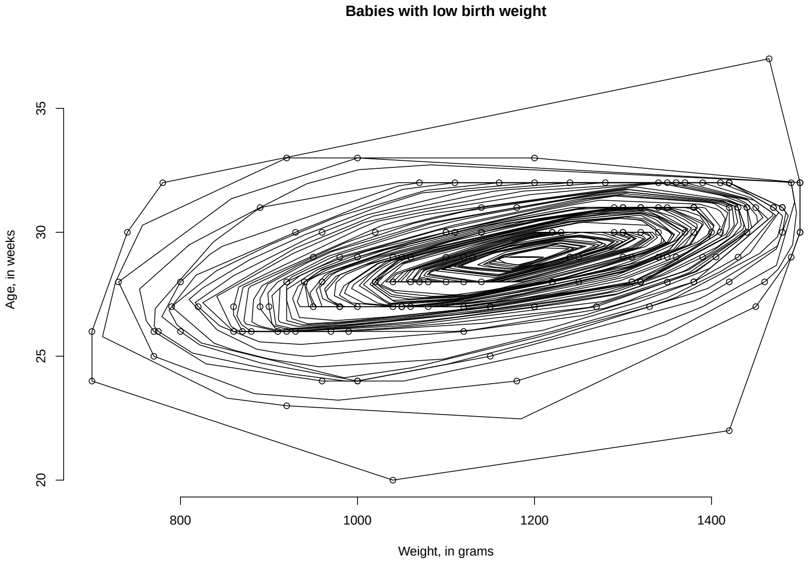
Properties:

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

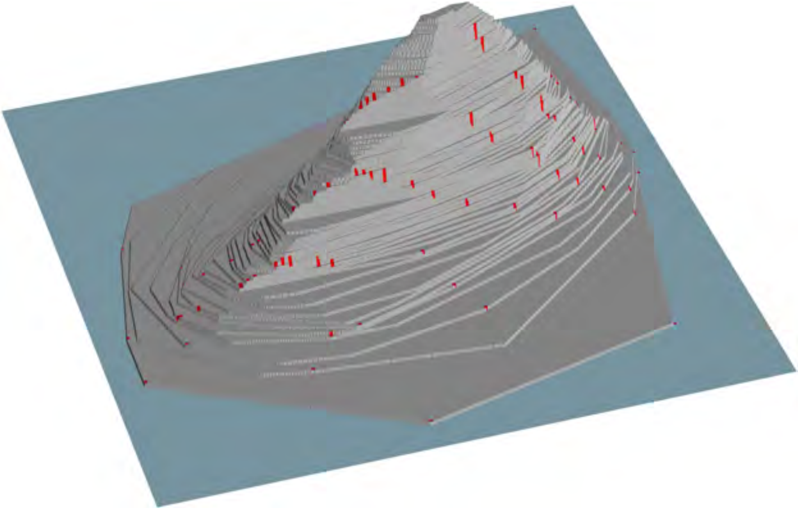
Halfspace (=Tukey, location) depth-trimmed regions



Halfspace (=Tukey, location) depth-trimmed regions



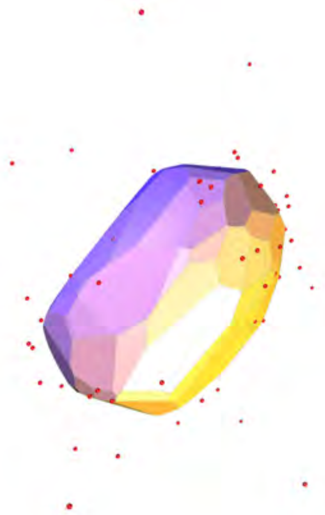
Halfspace (=Tukey, location) data depth



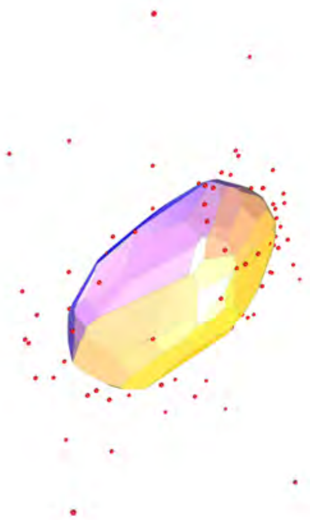
Halfspace (=Tukey, location) depth region

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

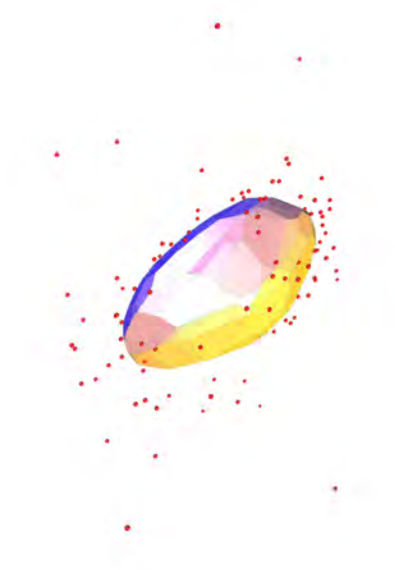
Halfspace (=Tukey, location) depth region: $\tau = 5/161$



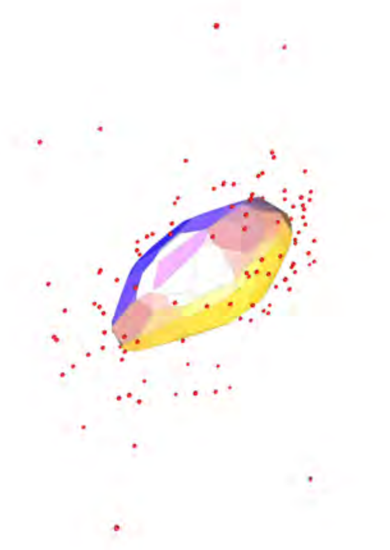
Halfspace (=Tukey, location) depth region: $\tau = 9/161$



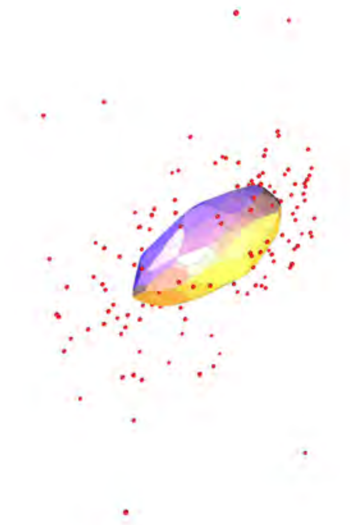
Halfspace (=Tukey, location) depth region: $\tau = 13/161$



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



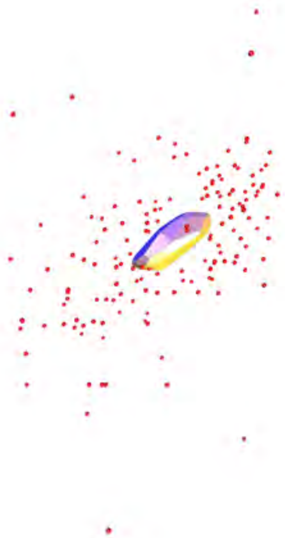
Halfspace (=Tukey, location) depth region: $\tau = 25/161$



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

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

Halfspace (=Tukey, location) depth region: τ =49/161



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

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

Halfspace (=Tukey, location) depth region: $\tau = 68/161$

Further depth notions

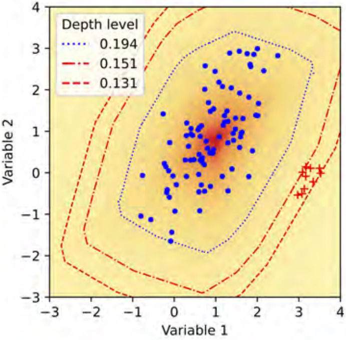
- **Mahalanobis depth** (Mahalanobis, 1936)
- **Convex hull peeling depth** (Barnett, 1976; Eddy, 1981)
- **Projection depth** (Stahel, 1981; Donoho, 1982)
- **Simplicial volume depth** (Oja, 1983)
- **Simplicial depth** (Liu, 1990)
- **Majority depth** (Singh, 1991)
- **Zonoid depth** (Koshevoy and Mosler, 1997)
- $\mathbb{L}_p$**-depth** (Zuo and Serfling, 2000)
- **Spatial depth** (Serfling, 2002)
- **Expected convex hull depth** (Cascos, 2007)
- **Geometrical depth** (Dyckerhoff and Mosler, 2011)
- **Lens depth** (Liu and Modarres, 2011)

Contents

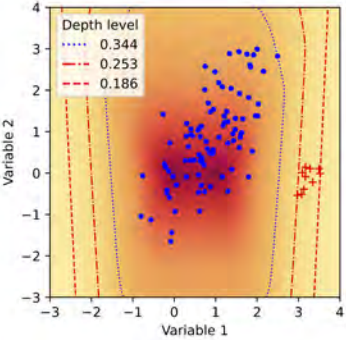
Projection *vs.* halfspace depth

- Normal data (90 obs.): $\mathcal{N}\Big((1,1)^\top, \left(\begin{smallmatrix}1 & 1\\ 1 & 2\end{smallmatrix}\right)\Big)$.
- Anomalies (10 obs.): $\mathcal{N}\Big((3.181, -0.222)^\top, \left(\begin{smallmatrix}1 & 1\\ 1 & 2\end{smallmatrix}\right)/36\Big)$.

Projection depth

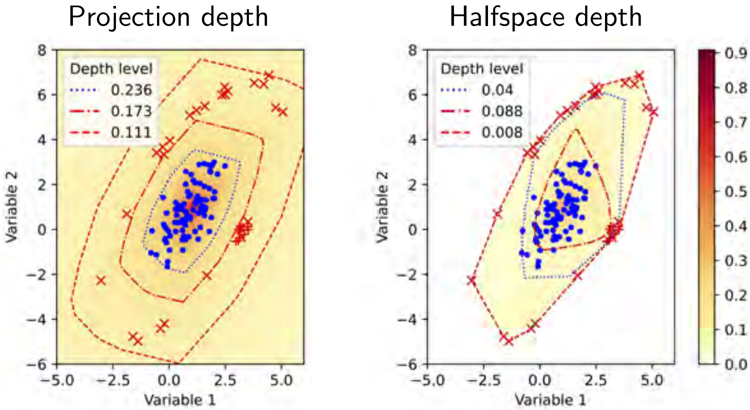


Simplicial volume depth



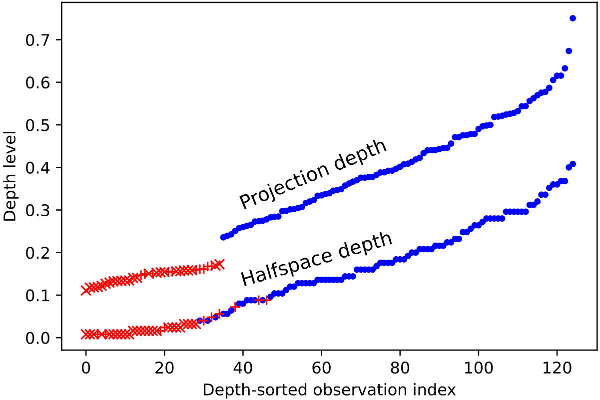
Projection *vs.* halfspace depth

- Normal data (90 obs.): $\mathcal{N}\Big((1,1)^\top, \left(\begin{smallmatrix}1 & 1\\ 1 & 2\end{smallmatrix}\right)\Big)$.
- Anomalies (10 obs.): $\mathcal{N}\Big((3.181, -0.222)^\top, \left(\begin{smallmatrix}1 & 1\\ 1 & 2\end{smallmatrix}\right)/36\Big)$.
- Anomalies (25 obs.): masking anomalies.



Projection *vs.* halfspace depth

- Normal data (90 obs.): $\mathcal{N}\Big((1,1)^\top,\ \left(\begin{smallmatrix}1 & 1\\ 1 & 2\end{smallmatrix}\right)\Big)$.
- Anomalies (10 obs.): $\mathcal{N}\Big((3.181,-0.222)^\top,\ \left(\begin{smallmatrix}1 & 1\\ 1 & 2\end{smallmatrix}\right)/36\Big)$.
- Anomalies (25 obs.): masking anomalies.



Contents

Illustration of properties

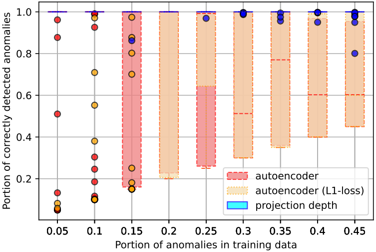
Properties of data depth:

- **Robustness**, on comparison with:
  - Auto-encoder.
- **Extrapolation abilities**, on comparison with:
  - Local outlier factor (LOF).
  - One-class support vector machine (OC-SVM).
  - Isolation forest (IF).
- **Explainability** of anomalies.

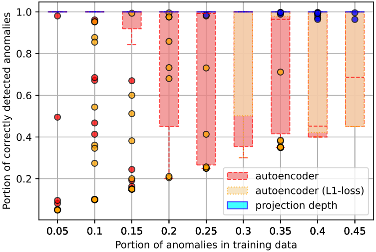
Autoencoder *vs.* depth

- Normal data: $\mathcal{N}(\boldsymbol{i}_d, \boldsymbol{I}_{d\times d})$.
- Anomalies: ellicpical Cauchy distribution.

$d = 10$



$d = 20$

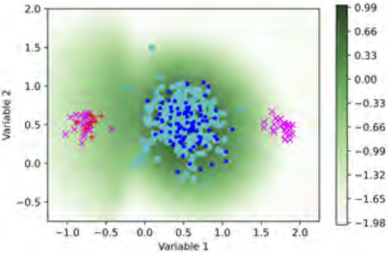


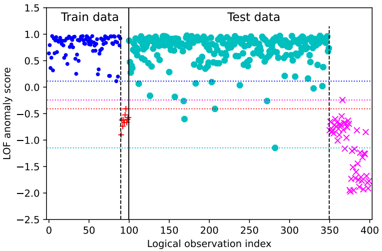
Quality measure: Portion of anomalies if we detect all of them.
Autoencoders:

- For $d = 10$: neuronal layers 10–5–2–5–10.
- For $d = 20$: neuronal layers 20–10–5–10–20.

Local outlier factor

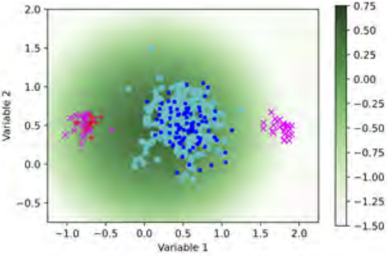
- **Training data**: polluted with anomalies.
- **Test data**: same + new anomalies.

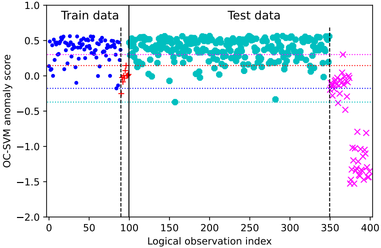




One-class support vector machine

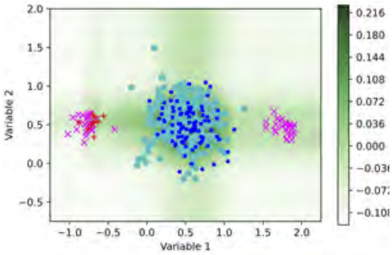
- **Training data**: polluted with anomalies.
- **Test data**: same + new anomalies.

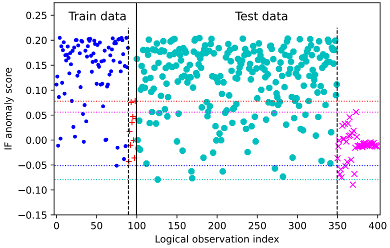




Isolation forest

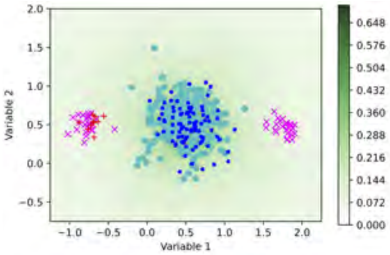
- **Training data**: polluted with anomalies.
- **Test data**: same + new anomalies.

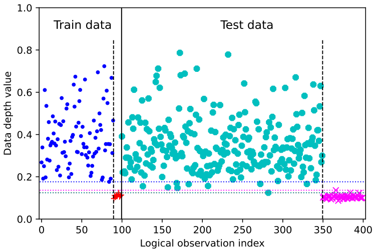




Data depth (projection depth notion)

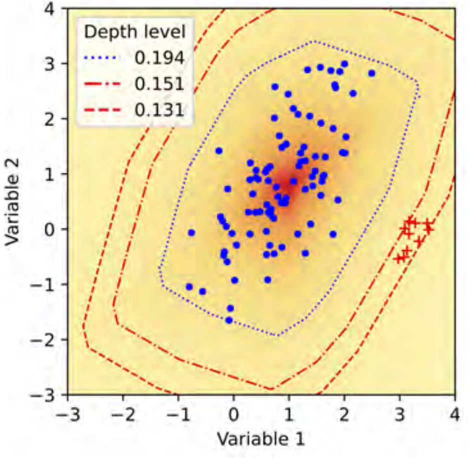
- **Training data**: polluted with anomalies.
- **Test data**: same + new anomalies.





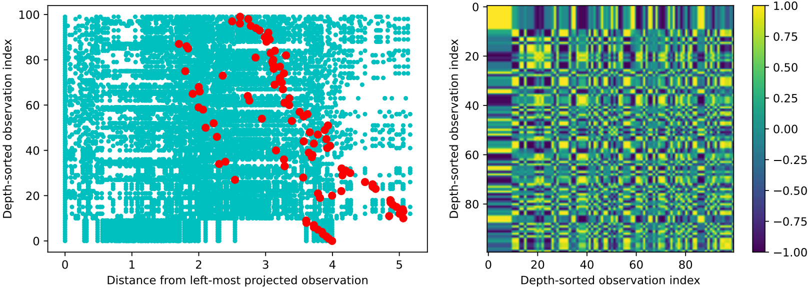
Explainability

- Let us take the previous example.



Explainability

- **Optimizing direction**: variables contribution, *e.g.*, $(0.863, -0.505)^\top$.
- **Directions' plot**: compare abnormalities.
- **Angles' heatmap**: Allows to detect clustered anomalies.



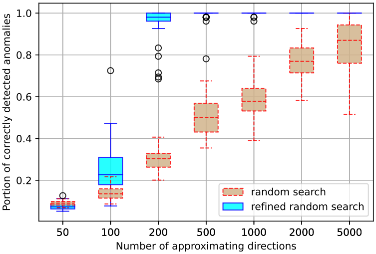
Contents

Numerical approximation: number of directions

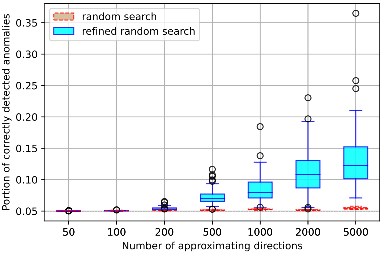
Employing **approximating algorithms** for data depth: Dyckerhoff, Mozharovskyi, Nagy (2021).

- Normal data (950 obs.): $\mathcal{N}\left(\mathbf{0}_d, \text{Toeplitz}_{d\times d}\right)$.
- Anomalies (50 obs.): $\mathcal{N}\left(\mathbf{0}_d + 1.25\cdot\lambda\cdot\min\text{PC}, \boldsymbol{I}_{d\times d}\right)$.

$d = 10$



$d = 50$

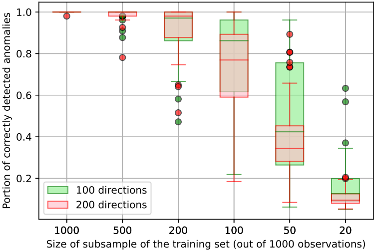


Statistical approximation: sub-sampling

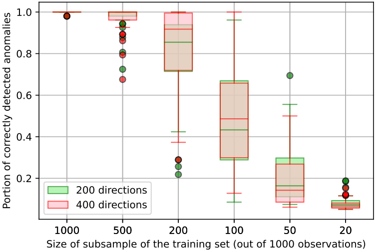
Employing **approximating algorithms** for data depth: Dyckerhoff, Mozharovskyi, Nagy (2021).

- Normal data (950 obs.): $\mathcal{N}\left(\mathbf{0}_d, \text{Toeplitz}_{d\times d}\right)$.
- Anomalies (50 obs.): $\mathcal{N}\left(\mathbf{0}_d + 1.25 \cdot \lambda \cdot \min \text{PC}, \boldsymbol{I}_{d\times d}\right)$.

$d = 10$



$d = 50$



Contents

Thank you for your attention! Questions?

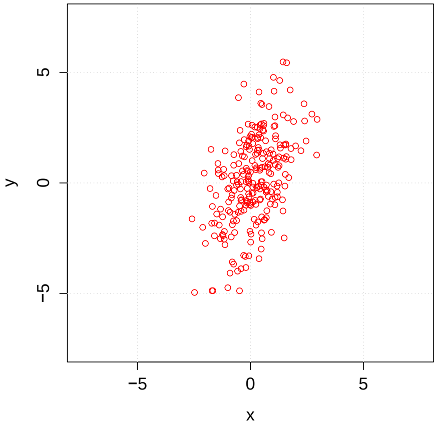
- **Data depth** has undergone substantial theoretical development during recent 30 years and possesses attractive properties, *e.g.*, robustness, affine invariance, *etc.*
- Recently, efficient algorithms (both exact and approximate) have been developed for computation of numerous depths.
- Data **depth** can be used as a powerful tool for anomaly detection.
- When applying data depth for anomaly detection, several aspects should be taken into account, considered in this presentation.
- **Disclaimer**: The presented examples were designed to illustrate advantages of depth-based anomaly detection, their generalization can be limited.

Computational taxonomy

	Exponential time	**Polynomial time**
Affine-invariant	*convex hull peeling depth* *majority depth* expected convex hull depth geometrical depth *halfspace depth* *projection depth* *simplicial depth*	zonoid depth Mahalanobis depth
Not affine-invariant	*simplicial volume depth*	$\mathbb{L}_2$ *spatial depth* *lens depth*

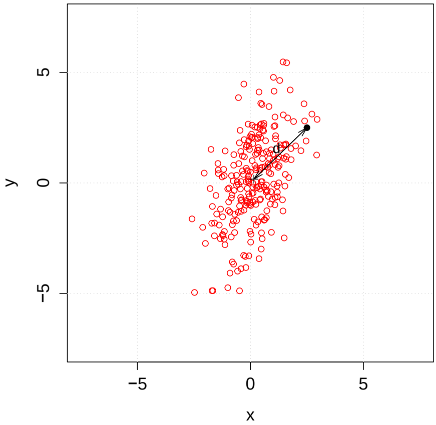
∗ : *Italics* indicate **robust** depth notions.

Mahalanobis depth (Mahalanobis, 1936)



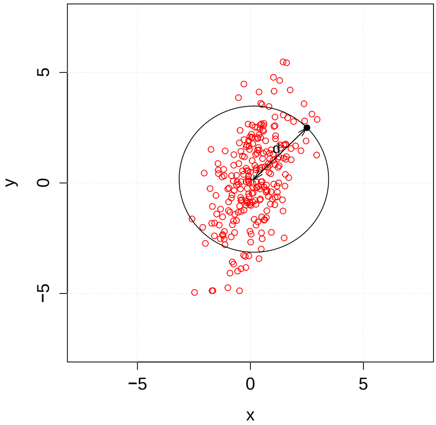
- $X \sim \mathrm{N}(\mu_X, \Sigma_X)$

Mahalanobis depth (Mahalanobis, 1936)



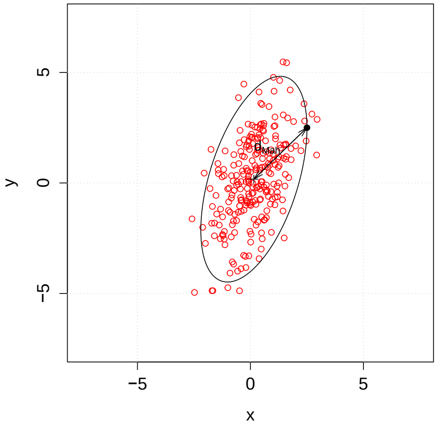
- $X \sim \mathrm{N}(\mu_X, \Sigma_X)$
- $d(\boldsymbol{x}|X) = \|\boldsymbol{x} - \mu_X\|$

Mahalanobis depth (Mahalanobis, 1936)



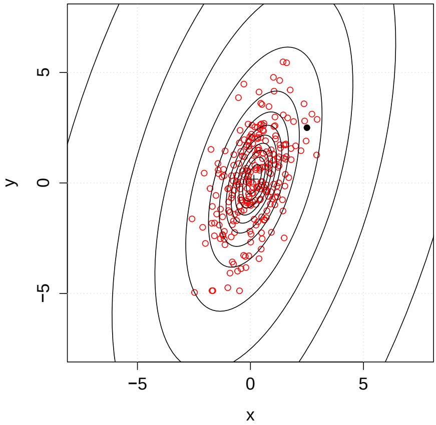
- $X \sim \mathrm{N}(\mu_X, \Sigma_X)$
- $d(\boldsymbol{x}|X) = \|\boldsymbol{x} - \mu_X\|$

Mahalanobis depth (Mahalanobis, 1936)



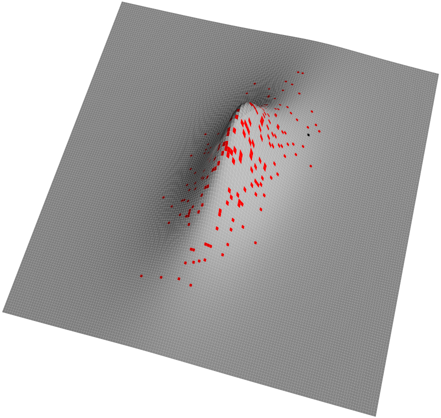
- $X \sim \mathrm{N}(\mu_X, \Sigma_X)$
- $d(\boldsymbol{x}|X) = \|\boldsymbol{x} - \mu_X\|$
- $d^2_{Mah}(\boldsymbol{x}|X) = (\boldsymbol{x} - \mu_X)^\top \Sigma_X^{-1} (\boldsymbol{x} - \mu_X)$

Mahalanobis depth (Mahalanobis, 1936)



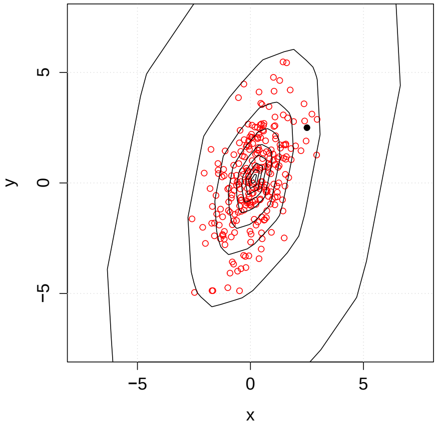
- $X \sim \text{N}(\mu_X, \Sigma_X)$
- $d(\boldsymbol{x}|X) = \|\boldsymbol{x} - \mu_X\|$
- $d^2_{Mah}(\boldsymbol{x}|X) = (\boldsymbol{x} - \mu_X)^\top \Sigma_X^{-1} (\boldsymbol{x} - \mu_X)$
- $D^{\text{Mah}}(\boldsymbol{x}|X) = \frac{1}{1 + d^2_{Mah}(\boldsymbol{x}|X)}$

Mahalanobis depth (Mahalanobis, 1936)



- $X \sim \mathrm{N}(\mu_X, \Sigma_X)$
- $d(\boldsymbol{x}|X) = \|\boldsymbol{x} - \mu_X\|$
- $d^2_{Mah}(\boldsymbol{x}|X) = (\boldsymbol{x} - \mu_X)^\top \Sigma_X^{-1} (\boldsymbol{x} - \mu_X)$
- $D^{\mathrm{Mah}}(\boldsymbol{x}|X) = \frac{1}{1 + d^2_{Mah}(\boldsymbol{x}|X)}$

Projection depth (Zuo, Serfling, 2000)



- A measure of **outlyingness** of $\boldsymbol{x}$ w.r.t. X:

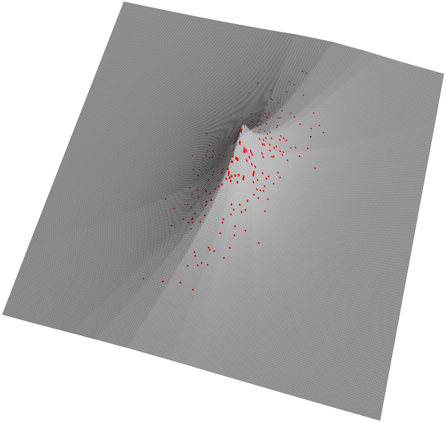
  $$O_{Prj}(\boldsymbol{x}|X) = \sup_{\mathbf{u}\in S^{d-1}} \frac{\left|\mathbf{u}^\top \boldsymbol{x} - m_X(\mathbf{u}'\boldsymbol{x})\right|}{\sigma_X(\mathbf{u}^\top \boldsymbol{x})},$$

  m_X and σ_X are **univariate** location and scatter measures.

- m_X =median and σ_X =MAD (median absolute deviation).

- $D^{\text{prj}}(\boldsymbol{x}|X) = \frac{1}{1+O_{Prj}(\boldsymbol{x}|X)}$.

Projection depth (Zuo, Serfling, 2000)



- A measure of **outlyingness** of $\boldsymbol{x}$ w.r.t. X:

  $$O_{Prj}(\boldsymbol{x}|X) = \sup_{\mathbf{u}\in S^{d-1}} \frac{\left|\mathbf{u}^\top \boldsymbol{x} - m_X(\mathbf{u}'\boldsymbol{x})\right|}{\sigma_X(\mathbf{u}^\top \boldsymbol{x})},$$

  m_X and σ_X are **univariate** location and scatter measures.

- m_X =median and σ_X =MAD (median absolute deviation).

- $D^{\text{prj}}(\boldsymbol{x}|X) = \frac{1}{1+O_{Prj}(\boldsymbol{x}|X)}$.