Machine learning for multivariate and functional anomaly detection: orderings and data depth

Pavlo Mozharovskyi

LTCI, Telecom Paris, Institut Polytechnique de Paris

Datacraft Seminar

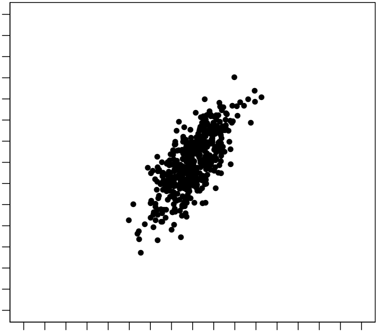
Paris, October 12, 2020

Contents

Contents

A real task

Regard two measurements during a test in a production process:

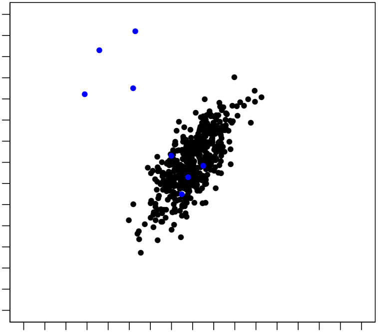


Given **training data**, polluted or not with anomalies:

- 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:

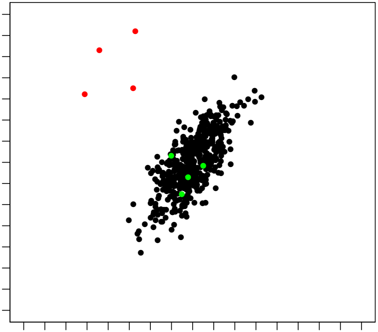
- detect **anomalies** in the given data.

For **new data**, determine:

- Whether new observations are **normal** data or **anomalies**?

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:

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

of observations in the d-dimensional Euclidean space.

Multivariate framework

- A training data set:

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

of observations in the d-dimensional Euclidean space.

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

Multivariate framework

- A training data set:

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

  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:

$$\mathbb{R}^d \to \{-1, +1\} : \boldsymbol{x} \mapsto g(\boldsymbol{x}),$$

  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.*, -1).

Multivariate framework

- A training data set:

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

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:

$$\mathbb{R}^d \to \{-1, +1\} : \boldsymbol{x} \mapsto g(\boldsymbol{x}),$$

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.*, -1).
- It is more useful to provide an ordering on $\mathbb{R}^d$:

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

such that abnormal observations obtain higher anomaly score.

Practical session

Notebooks:

- `anomdet_simulation1.Rmd`,
- `anomdet_hurricanes.Rmd`,
- `anomdet_brainimaging.Rmd`,
- `anomdet_cars.ipynb`,
- `anomdet_airbus.ipynb`.

Data sets:

- `carsanom.csv`: Data set on anomaly detection for cars.
- `airbus_data.csv`: Data set from Airbus.
- `hurdat2-1851-2019-052520.txt`: Historical hurricane data.
- `101_1_dwi_fa.nii`: Anatomical brain volume data.
- `101_1_dwi.voxelcoordsL.txt`: Left brain fiber's bundle.
- `101_1_dwi.voxelcoordsR.txt`: Right brain fiber's bundle.

Supplementary scripts:

- `depth_routines.py`: Routines for data depth calculation.
- `FIF.py`: Implementation of the functional isolation forest.
- `depth_routines.R`: Routines for curves' parametrization.
- `DTI.R`: Routines for input of brain imaging data.

Contents

Contents

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Generalized portrait:

- The method of the **generalized portrait** was introduced by Vapnik & Lerner (1963) and Vapnik & Chervonenkis (1974).

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Generalized portrait:

- The method of the **generalized portrait** was introduced by Vapnik & Lerner (1963) and Vapnik & Chervonenkis (1974).
- Generalized portrait is the vector:

$$\psi = \frac{\boldsymbol{\varphi}}{\min_{\boldsymbol{x} \in \boldsymbol{X}} \langle \boldsymbol{x}, \boldsymbol{\varphi} \rangle} \quad \text{with } \boldsymbol{\varphi} \text{ from} \quad \max_{\|\boldsymbol{\varphi}\|=1} \min_{\boldsymbol{x} \in \boldsymbol{X}} \langle \boldsymbol{x}, \boldsymbol{\varphi} \rangle .$$

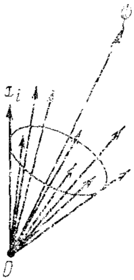


Рис. 24.

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Kernel trick (Boser, Guyon, Vapnik; 1992):

- Let Φ be a feature map: $\mathbb{R}^d \mapsto \mathcal{H}$.

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Kernel trick (Boser, Guyon, Vapnik; 1992):

- Let Φ be a feature map: $\mathbb{R}^d \mapsto \mathcal{H}$.
- Due to the kernel trick, the dot product in the image of φ can be computed by evaluation of a kernel K:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle .$$

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Kernel trick (Boser, Guyon, Vapnik; 1992):

- Let Φ be a feature map: $\mathbb{R}^d \mapsto \mathcal{H}$.
- Due to the kernel trick, the dot product in the image of φ can be computed by evaluation of a kernel K:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle .$$

- Example: Gaussian kernel

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{\gamma \|\boldsymbol{x}_i, \boldsymbol{x}_j\|}$$

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Kernel trick (Boser, Guyon, Vapnik; 1992):

- Let Φ be a feature map: $\mathbb{R}^d \mapsto \mathcal{H}$.
- Due to the kernel trick, the dot product in the image of φ can be computed by evaluation of a kernel K:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle .$$

- Example: Gaussian kernel

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{\gamma \|\boldsymbol{x}_i, \boldsymbol{x}_j\|}$$

Soft margin (Cortes, Vapnik; 1995):

- Allow for a portion of points from $\boldsymbol{X}$ to be beyond the margin, label points far from the origin by "1", those close by "-1".

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Kernel trick (Boser, Guyon, Vapnik; 1992):

- Let Φ be a feature map: $\mathbb{R}^d \mapsto \mathcal{H}$.
- Due to the kernel trick, the dot product in the image of φ can be computed by evaluation of a kernel K:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle .$$

- Example: Gaussian kernel

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{\gamma \|\boldsymbol{x}_i, \boldsymbol{x}_j\|}$$

Soft margin (Cortes, Vapnik; 1995):

- Allow for a portion of points from $\boldsymbol{X}$ to be beyond the margin, label points far from the origin by "1", those close by "-1".
- Controlled by a parameter $\nu \in (0, 1)$ (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999).

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Idea 1: Separate points from the origin.

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Idea 1: Separate points from the origin.

This can be formulated as a quadratic programming problem

$$\min_{\boldsymbol{\psi}\in\mathcal{H},\boldsymbol{\xi}\in\mathbb{R}^n,\rho\in\mathbb{R}} \quad \frac{1}{2}\|\boldsymbol{\psi}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho$$

$$\text{subject to} \quad \langle \boldsymbol{\xi}, \Phi(\boldsymbol{x}_i)\rangle \geq \rho - \xi_i\,,\ \xi_i \geq 0 \text{ for } i = 1, \dots, n\,,$$

with $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$.

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Idea 1: Separate points from the origin.

This can be formulated as a quadratic programming problem

$$\min_{\boldsymbol{\psi}\in\mathcal{H},\boldsymbol{\xi}\in\mathbb{R}^n,\rho\in\mathbb{R}} \quad \frac{1}{2}\|\boldsymbol{\psi}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho$$

$$\text{subject to} \quad \langle \boldsymbol{\xi}, \Phi(\boldsymbol{x}_i)\rangle \geq \rho - \xi_i\,,\ \xi_i \geq 0 \text{ for } i = 1, \ldots, n\,,$$

with $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$.

The solution $(\boldsymbol{\psi}^*, \boldsymbol{\xi}^*, \rho^*)$ yields the following decision function:

$$g_{OCSVM}(\boldsymbol{x}) = \operatorname{sgn}(\langle \boldsymbol{\xi}^*, \Phi(\boldsymbol{x})\rangle - \rho^*)\,.$$

One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Idea 1: Separate points from the origin.

This can be formulated as a quadratic programming problem

$$\min_{\boldsymbol{\psi}\in\mathcal{H},\boldsymbol{\xi}\in\mathbb{R}^n,\rho\in\mathbb{R}} \quad \frac{1}{2}\|\boldsymbol{\psi}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho$$

$$\text{subject to} \quad \langle\boldsymbol{\xi}, \Phi(\boldsymbol{x}_i)\rangle \geq \rho - \xi_i\,,\ \xi_i \geq 0 \text{ for } i = 1,\dots,n\,,$$

with $\boldsymbol{\xi} = (\xi_1,\dots,\xi_n)^\top$.

The solution $(\boldsymbol{\psi}^*, \boldsymbol{\xi}^*, \rho^*)$ yields the following decision function:

$$g_{OCSVM}(\boldsymbol{x}) = \operatorname{sgn}(\langle\boldsymbol{\xi}^*, \Phi(\boldsymbol{x})\rangle - \rho^*)\,.$$

One can reformulate the optimization problem to employ the kernel trick.

One-class support vector machines (Schölkopf *et al.*, 1999)

In dual formulation, using the Lagrangian, one can restate the optimization problem as follows:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu n} \text{ for } i = 1, ..., n, \ \sum_{i=1}^{n}\alpha_i = 1\,,$$

with $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^\top$.

One-class support vector machines (Schölkopf *et al.*, 1999)

In dual formulation, using the Lagrangian, one can restate the optimization problem as follows:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu n} \text{ for } i = 1, \ldots, n, \sum_{i=1}^{n}\alpha_i = 1\,,$$

with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$.

The decision function is then:

$$g_{OCSVM}(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) - \rho\right),$$

where ρ can be recovered from any $\boldsymbol{x}_j$ such that $0 < \alpha_j < \frac{1}{\nu n}$:

$$\rho = \langle \boldsymbol{\psi}, \Phi(\boldsymbol{x}_j)\rangle = \sum_{i=1}^{n}\alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_j)\,.$$

One-class support vector machines (Schölkopf *et al.*, 1999)

Idea 2: Put points into a small ball.

$$\min_{R \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{c} \in \mathcal{H},} \quad R^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i$$

$$\text{subject to} \quad \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}\| \leq R^2 + \xi_i, \ \xi_i \geq 0 \text{ for } i = 1, ..., n\,.$$

One-class support vector machines (Schölkopf *et al.*, 1999)

Idea 2: Put points into a small ball.

$$\min_{R\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^n,\boldsymbol{c}\in\mathcal{H},} \quad R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i$$

$$\text{subject to} \quad \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}\| \leq R^2 + \xi_i,\ \ \xi_i \geq 0 \text{ for } i = 1, ..., n\,.$$

This leads to the dual:

$$\min_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{n}\alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu n},\ \text{ for } i = 1, ..., n,\ \sum_{i=1}^{n}\alpha_i = 1\,.$$

One-class support vector machines (Schölkopf *et al.*, 1999)

Idea 2: Put points into a small ball.

$$\min_{R\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^n,\boldsymbol{c}\in\mathcal{H},} \quad R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i$$

$$\text{subject to} \quad \|\Phi(\boldsymbol{x}_i) - \boldsymbol{c}\| \leq R^2 + \xi_i,\ \ \xi_i \geq 0 \text{ for } i = 1, \ldots, n\,.$$

This leads to the dual:

$$\min_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{i=1}^{n}\alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu n},\ \text{for } i = 1, \ldots, n,\ \sum_{i=1}^{n}\alpha_i = 1\,.$$

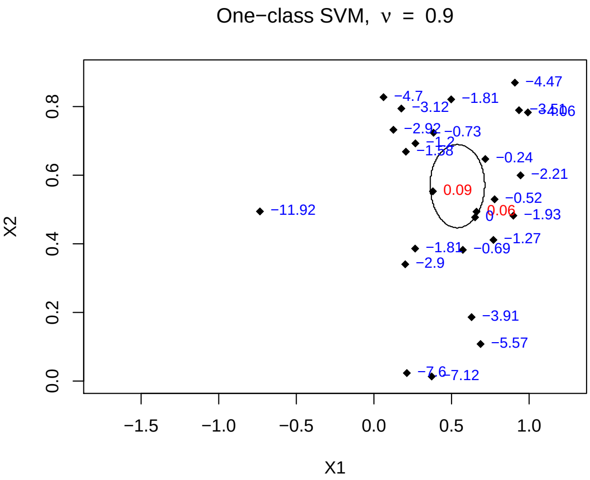
which leads to the decision function:

$$g_{OCSVM}(\boldsymbol{x}) = \left(R^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + 2\sum_{i=1}^{n}\alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) - K(\boldsymbol{x}, \boldsymbol{x})\right),$$

with $R^2 = \sum_{i,j}\alpha_i\alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - 2\sum_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_k) + K(\boldsymbol{x}_k, \boldsymbol{x}_k)$ for any $\boldsymbol{x}_k$ such that $0 < \alpha_k < 1/(\nu n)$.

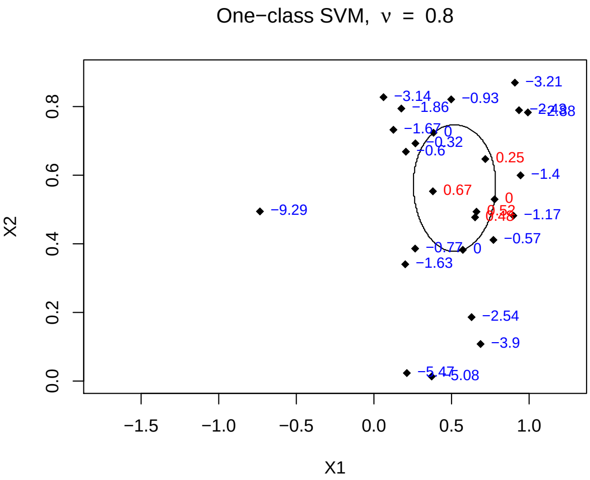
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 1



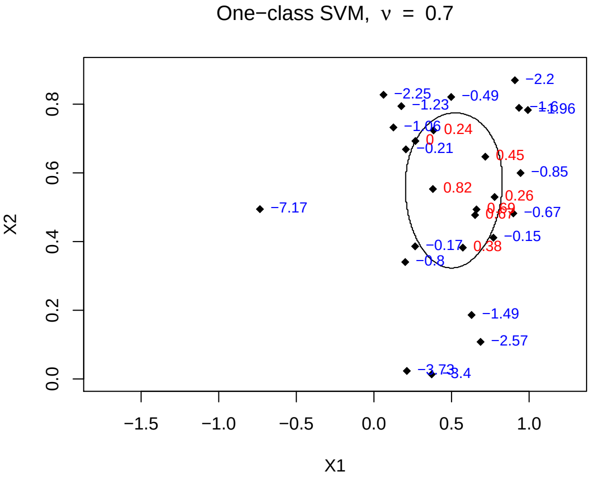
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 1



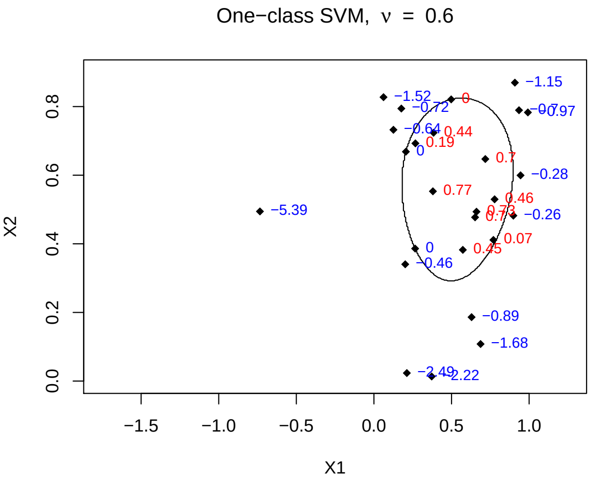
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 1



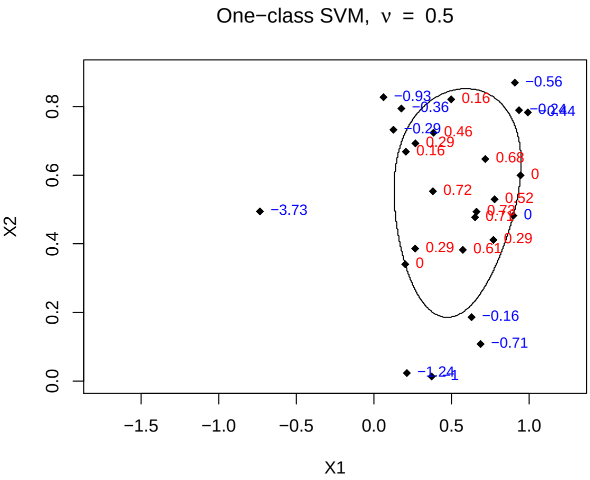
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 1



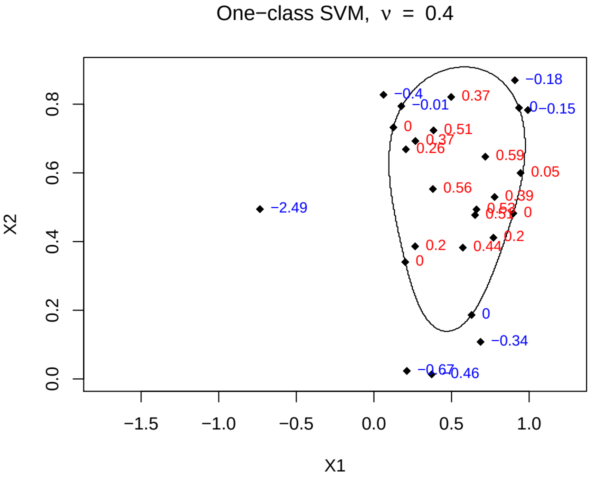
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 1



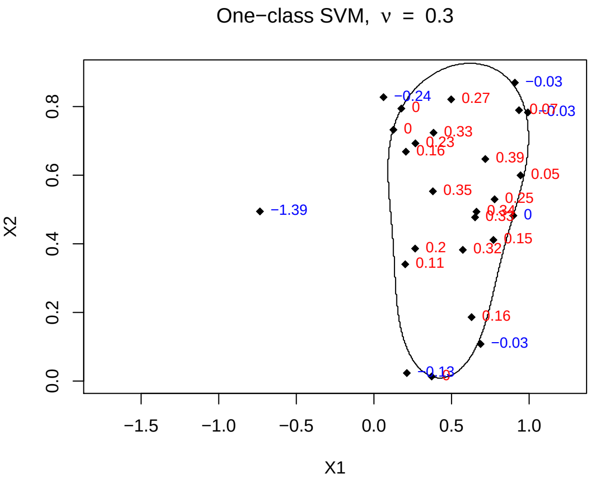
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 1



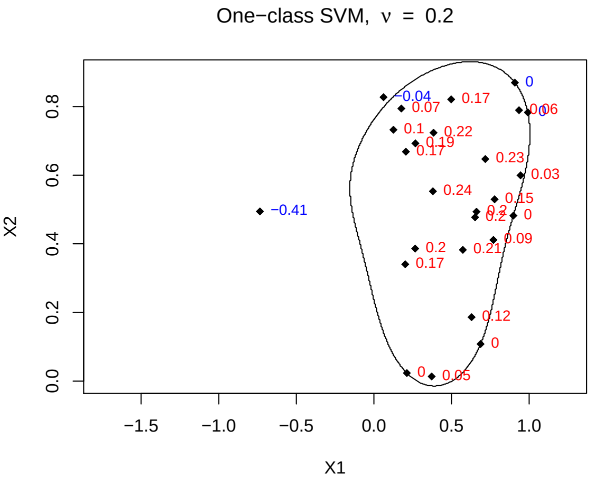
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 1



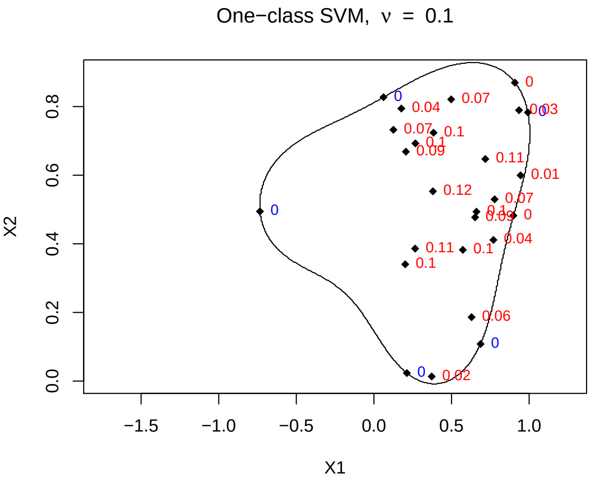
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 1



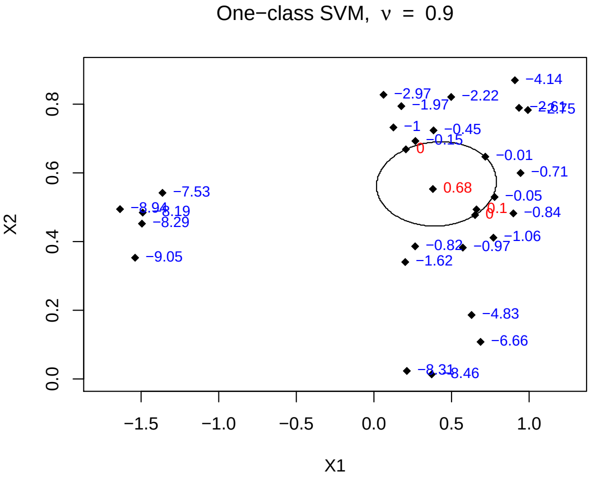
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 1



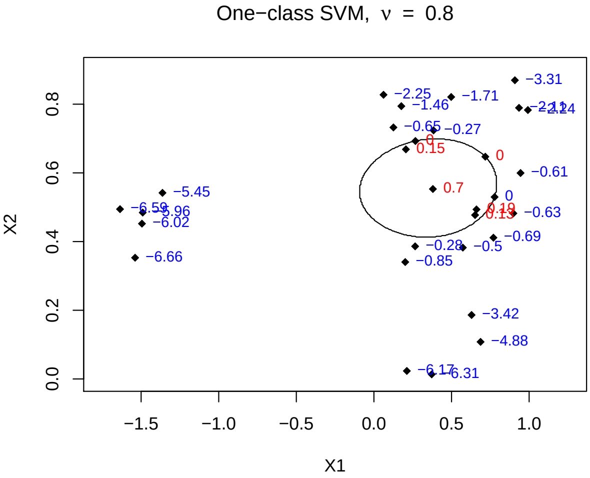
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 2



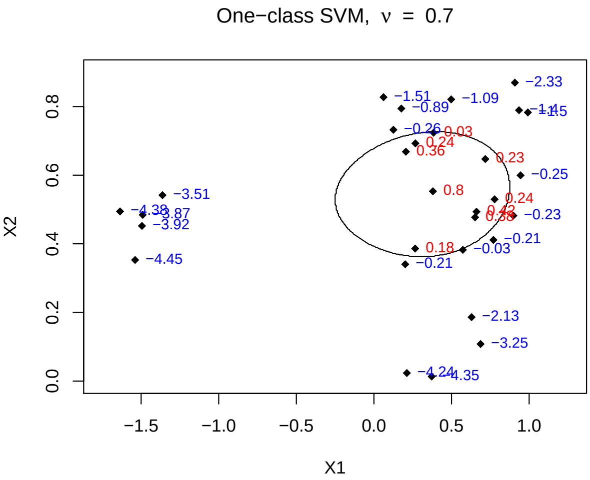
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 2



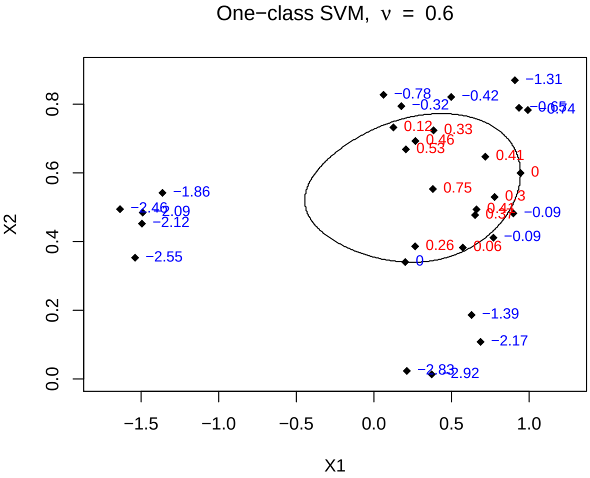
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 2



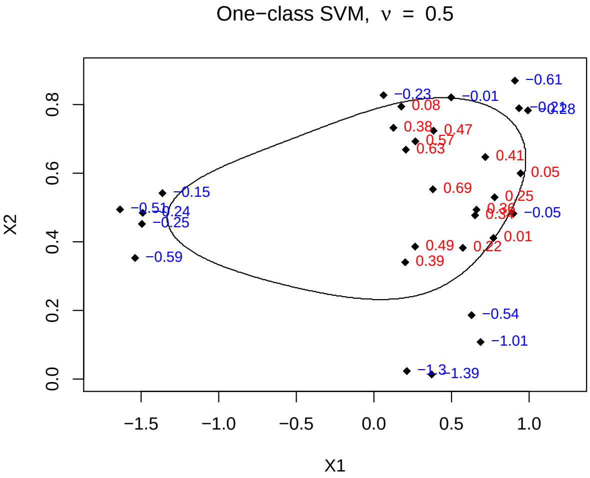
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 2



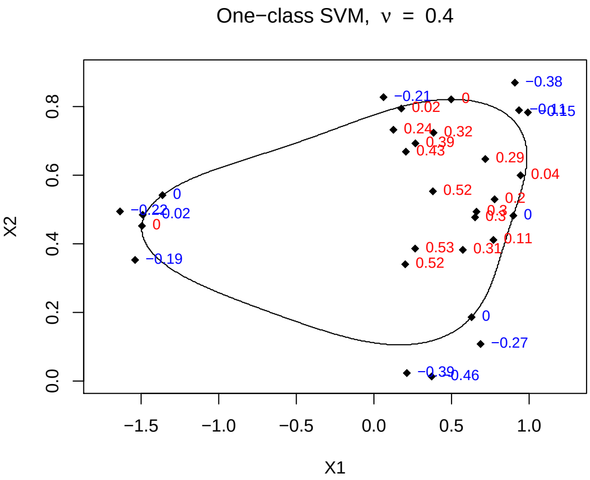
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 2



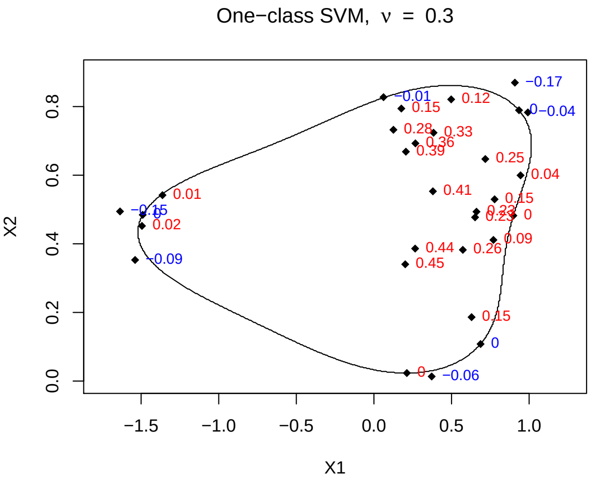
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 2



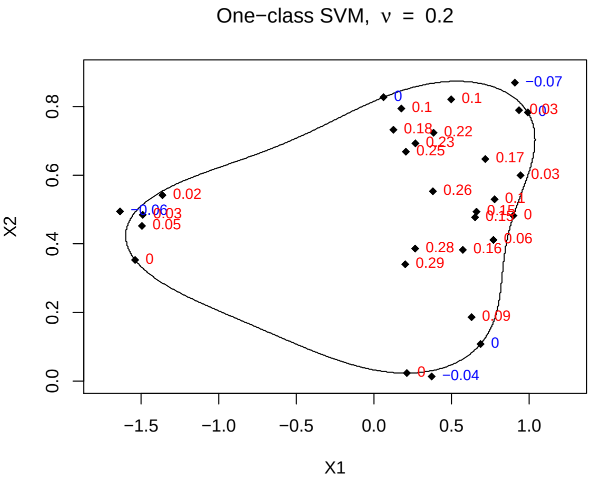
One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 2



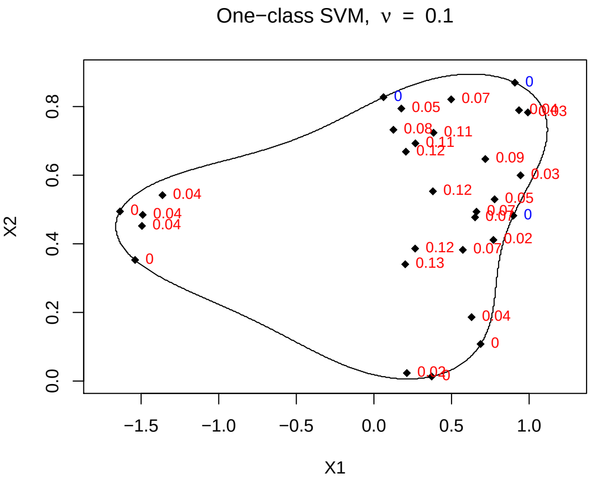
One-class support vector machines
(Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 2



One-class support vector machines (Schölkopf, Platt, Shawe-Taylor, Smola, Williamson; 1999)

Illustration: Case 2



Contents

Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

k-**distance** of a point $\boldsymbol{x}$:

For any integer $k > 0$, the k-distance of point $\boldsymbol{x}$, denoted as $k\text{-}dist(\boldsymbol{x})$, is defined as the distance $d(\boldsymbol{x}, \boldsymbol{o})$ between $\boldsymbol{x}$ and a point $\boldsymbol{o} \in \boldsymbol{X}$ such that:

- for at least k points $\boldsymbol{o}' \in \boldsymbol{X} \setminus \{\boldsymbol{x}\}$ it holds that $d(\boldsymbol{x}, \boldsymbol{o}') \leq d(\boldsymbol{x}, \boldsymbol{o})$, and
- for at most $k - 1$ points $\boldsymbol{o}' \in \boldsymbol{X} \setminus \{\boldsymbol{x}\}$ it holds that $d(\boldsymbol{x}, \boldsymbol{o}') < d(\boldsymbol{x}, \boldsymbol{o})$.

Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

k-**distance** of a point $\boldsymbol{x}$:

For any integer $k > 0$, the k-distance of point $\boldsymbol{x}$, denoted as $k\text{-}dist(\boldsymbol{x})$, is defined as the distance $d(\boldsymbol{x}, \boldsymbol{o})$ between $\boldsymbol{x}$ and a point $\boldsymbol{o} \in \boldsymbol{X}$ such that:

- for at least k points $\boldsymbol{o}' \in \boldsymbol{X} \setminus \{\boldsymbol{x}\}$ it holds that $d(\boldsymbol{x}, \boldsymbol{o}') \leq d(\boldsymbol{x}, \boldsymbol{o})$, and
- for at most $k - 1$ points $\boldsymbol{o}' \in \boldsymbol{X} \setminus \{\boldsymbol{x}\}$ it holds that $d(\boldsymbol{x}, \boldsymbol{o}') < d(\boldsymbol{x}, \boldsymbol{o})$.

(=Distance from $\boldsymbol{x}$ to its kth neighbor.)

Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

k-**distance** of a point $\boldsymbol{x}$:
For any integer $k > 0$, the k-distance of point $\boldsymbol{x}$, denoted as $k\text{-}dist(\boldsymbol{x})$, is defined as the distance $d(\boldsymbol{x}, \boldsymbol{o})$ between $\boldsymbol{x}$ and a point $\boldsymbol{o} \in \boldsymbol{X}$ such that:

- for at least k points $\boldsymbol{o}' \in \boldsymbol{X} \setminus \{\boldsymbol{x}\}$ it holds that $d(\boldsymbol{x}, \boldsymbol{o}') \leq d(\boldsymbol{x}, \boldsymbol{o})$, and
- for at most $k - 1$ points $\boldsymbol{o}' \in \boldsymbol{X} \setminus \{\boldsymbol{x}\}$ it holds that $d(\boldsymbol{x}, \boldsymbol{o}') < d(\boldsymbol{x}, \boldsymbol{o})$.

(=Distance from $\boldsymbol{x}$ to its kth neighbor.)

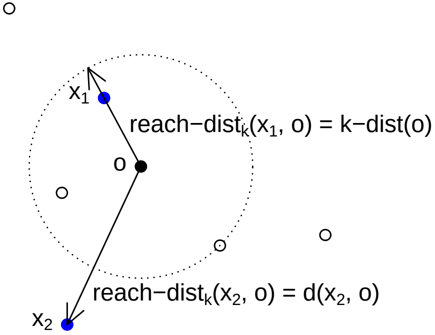
k-**neighborhood** of a point $\boldsymbol{x}$:
Given the $k\text{-}dist(\boldsymbol{x})$, the k-neighborhood of $\boldsymbol{x}$, denoted $N_k(\boldsymbol{x})$, contains every point whose distance from $\boldsymbol{x}$ is not greater than the $k\text{-}dist(\boldsymbol{x})$, *i.e.*:

$$N_k(\boldsymbol{x}) = \left\{ \boldsymbol{q} \in \boldsymbol{X} \setminus \{\boldsymbol{x}\} \,|\, d(\boldsymbol{x}, \boldsymbol{q}) \leq k\text{-}dist(\boldsymbol{x}) \right\}.$$

Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Reachability distance of order k of point $\boldsymbol{x}$ w.r.t. point $\boldsymbol{o}$:
For $k \in \mathbb{N}$, the reachability distance of order k of point $\boldsymbol{x}$ with respect to point $\boldsymbol{o}$ is defined as:

$$\textit{reach-dist}_k(\boldsymbol{x}, \boldsymbol{o}) = \max\{\textit{k-dist}(\boldsymbol{o}), d(\boldsymbol{x}, \boldsymbol{o})\}\,.$$



Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

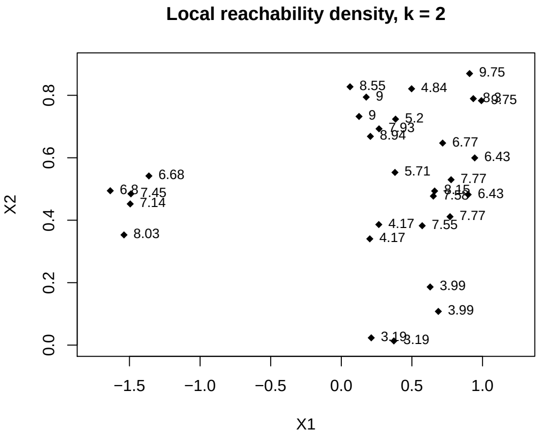
$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \textit{reach-dist}_k(\boldsymbol{x}, \boldsymbol{o})}.$$

Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} reach\text{-}dist_k(\boldsymbol{x}, \boldsymbol{o})}.$$

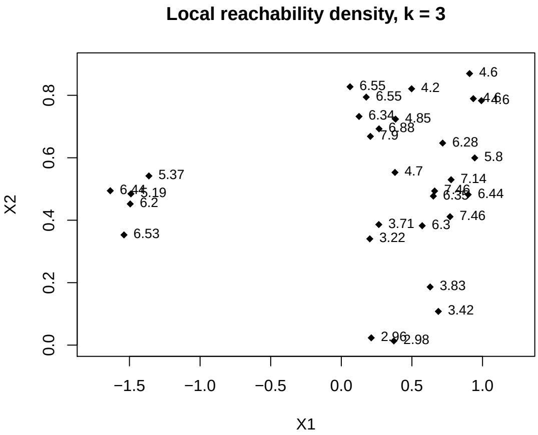


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} reach\text{-}dist_k(\boldsymbol{x}, \boldsymbol{o})} .$$

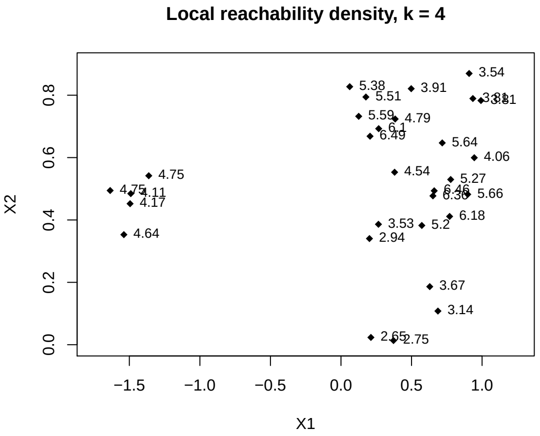


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \textit{reach-dist}_k(\boldsymbol{x}, \boldsymbol{o})} .$$

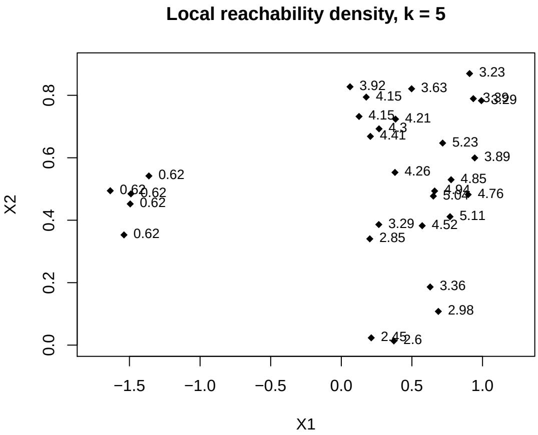


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} reach\text{-}dist_k(\boldsymbol{x}, \boldsymbol{o})}.$$



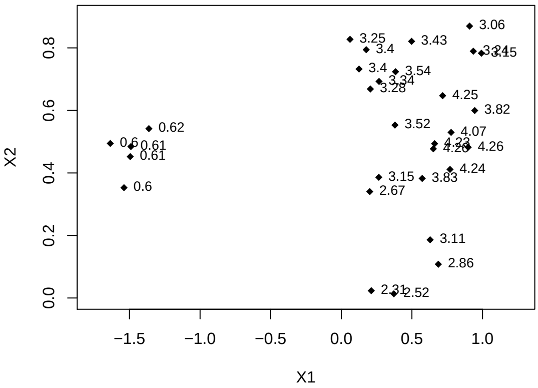
Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o}\in N_k(\boldsymbol{x})} reach\text{-}dist_k(\boldsymbol{x},\boldsymbol{o})}.$$

Local reachability density, k = 6

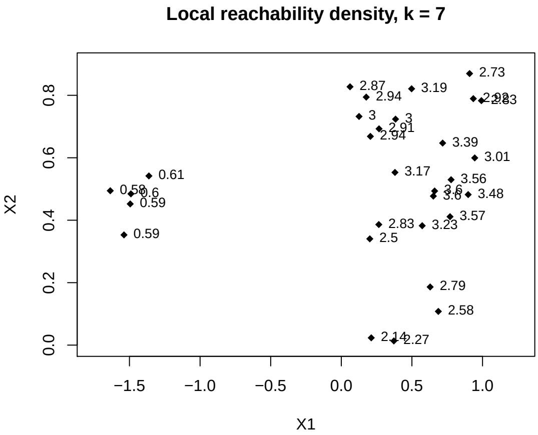


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} reach\text{-}dist_k(\boldsymbol{x}, \boldsymbol{o})}.$$

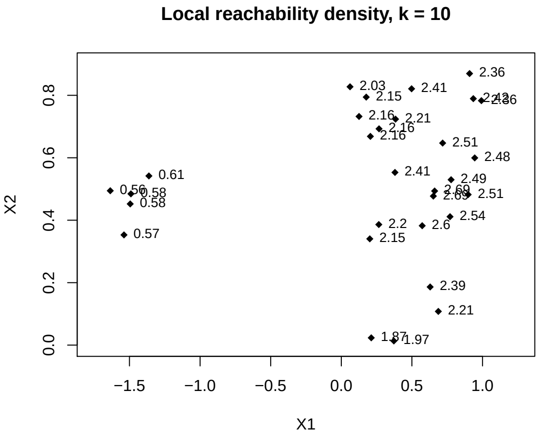


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} reach\text{-}dist_k(\boldsymbol{x}, \boldsymbol{o})} .$$

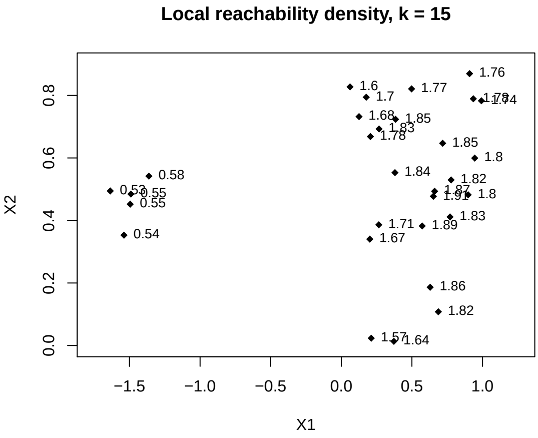


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} reach\text{-}dist_k(\boldsymbol{x}, \boldsymbol{o})}.$$

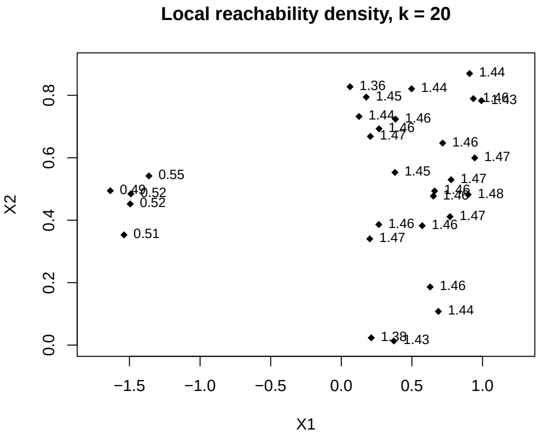


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} reach\text{-}dist_k(\boldsymbol{x}, \boldsymbol{o})} .$$

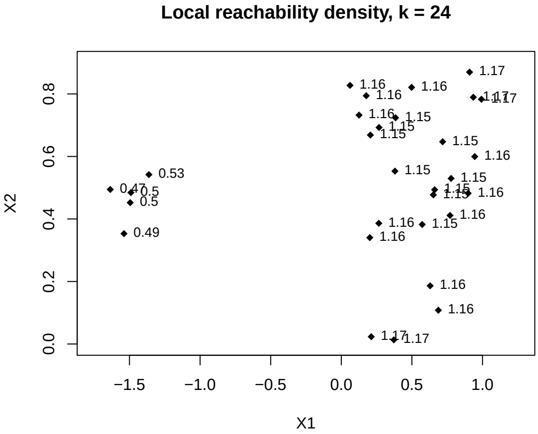


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} reach\text{-}dist_k(\boldsymbol{x}, \boldsymbol{o})} \,.$$

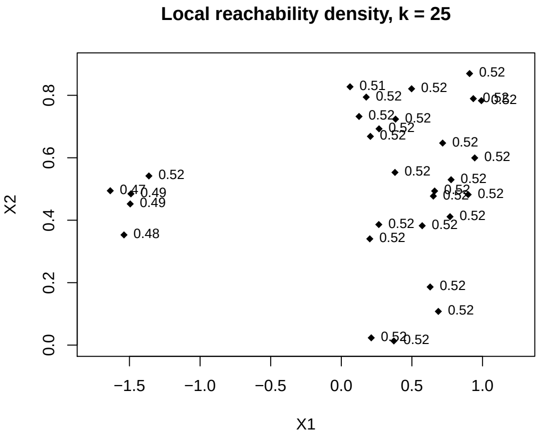


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} reach\text{-}dist_k(\boldsymbol{x}, \boldsymbol{o})}.$$

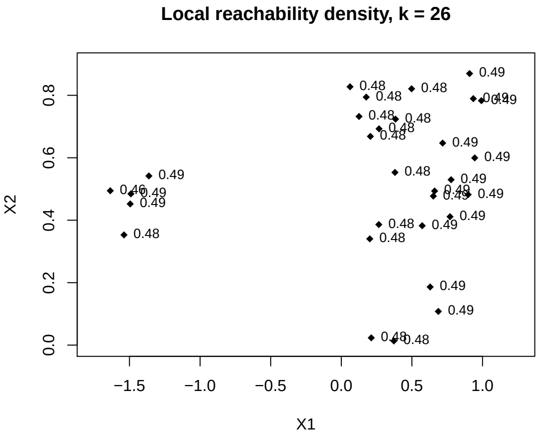


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \textit{reach-dist}_k(\boldsymbol{x}, \boldsymbol{o})} .$$

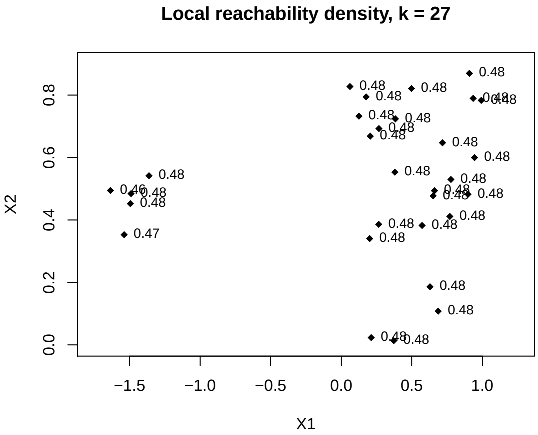


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local reachability density of a point $\boldsymbol{x}$:

The local reachability density of $\boldsymbol{x}$ is defined as:

$$lrd_k(\boldsymbol{x}) = \frac{|N_k(\boldsymbol{x})|}{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \textit{reach-dist}_k(\boldsymbol{x}, \boldsymbol{o})} .$$



Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

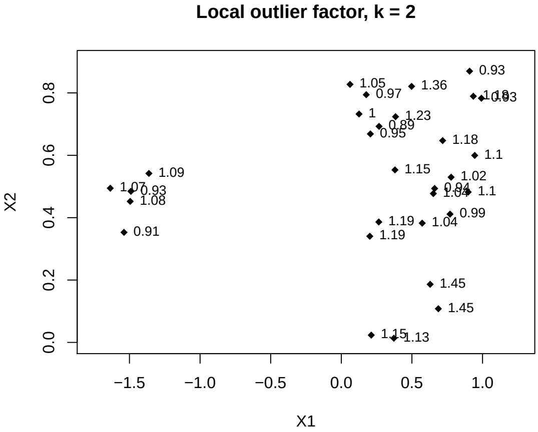
$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} \, .$$

Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$

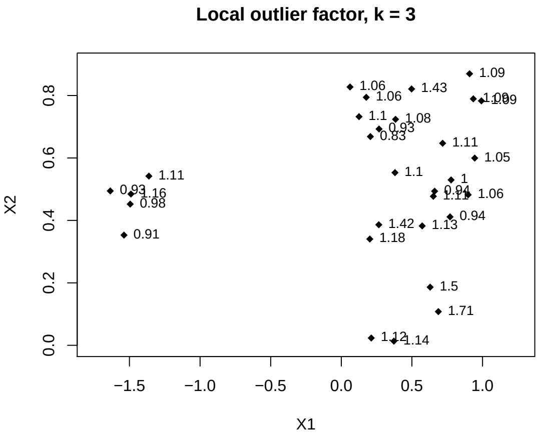


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$

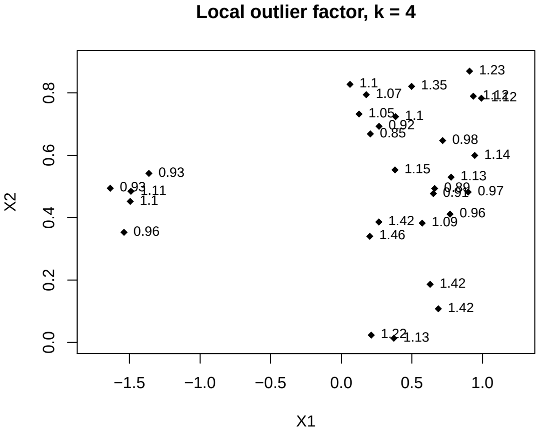


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$

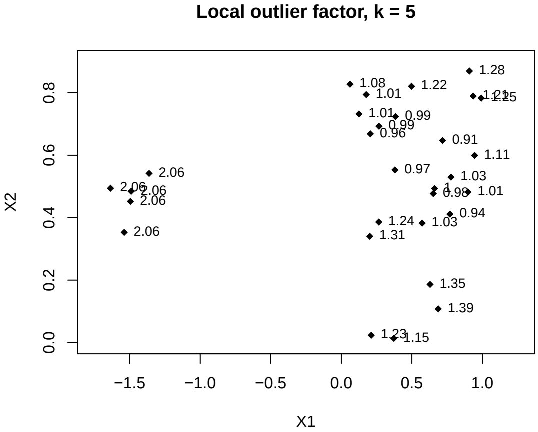


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$

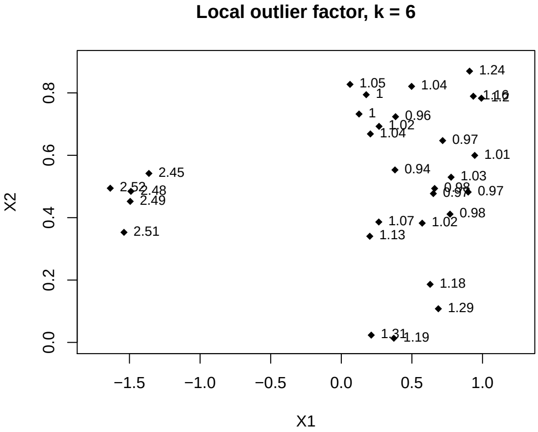


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$

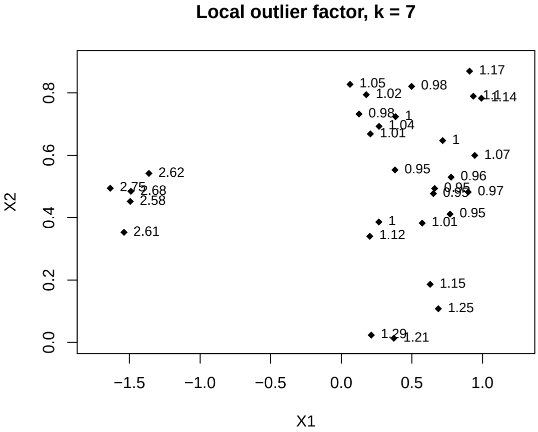


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$

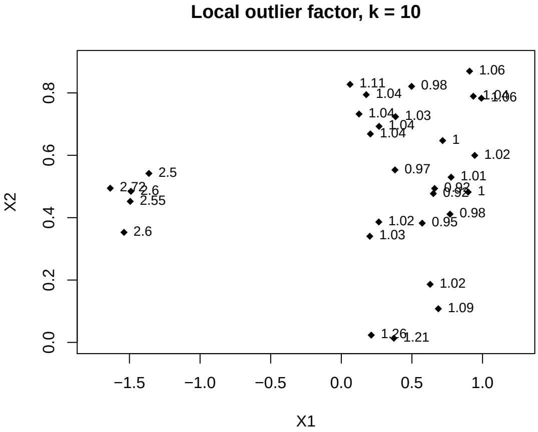


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|}.$$

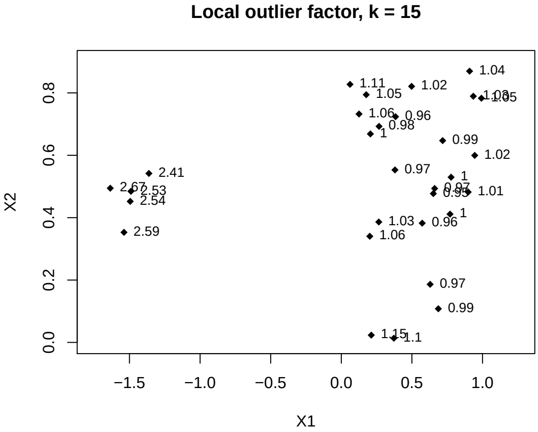


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|}.$$

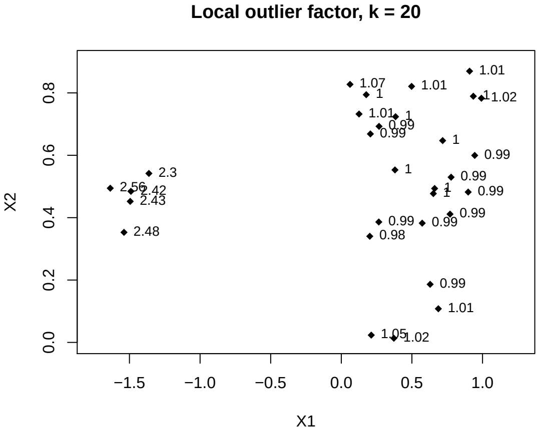


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$

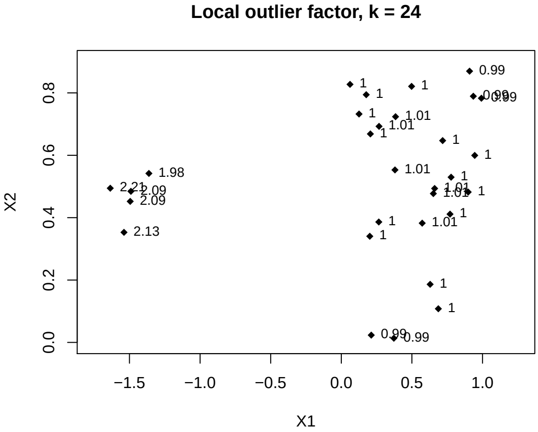


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$

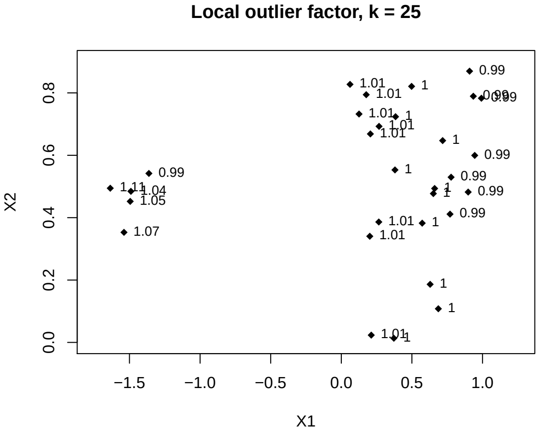


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$

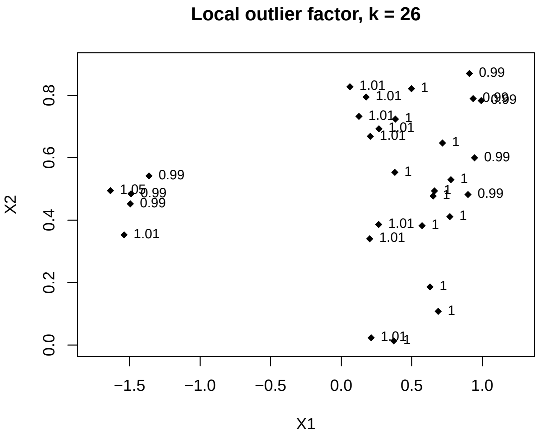


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$

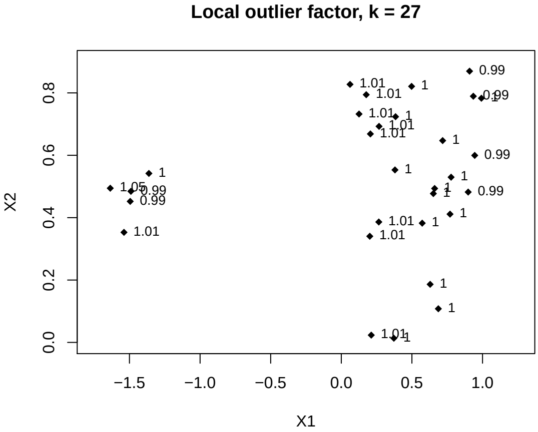


Local outlier factor (Breunig, Kriegel, Ng, Sander; 2000)

Local outlier factor of a point $\boldsymbol{x}$:

The local outlier factor of $\boldsymbol{x}$ is defined as:

$$LOF_k(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{o} \in N_k(\boldsymbol{x})} \frac{lrd_k(\boldsymbol{o})}{lrd_k(\boldsymbol{x})}}{|N_k(\boldsymbol{x})|} .$$



Contents

Isolation forest (Liu, Ting, Zhou; 2008)

- **Isolation forest** (Liu, Ting, Zhou; 2008) is an anomaly detection method inherited from the famous random forest algorithm (Breiman, 2001).
- Since no supervised feedback is given, isolation forest is based on purely random (uniform) variable-based partitioning.

Isolation forest (Liu, Ting, Zhou; 2008)

- **Isolation forest** (Liu, Ting, Zhou; 2008) is an anomaly detection method inherited from the famous random forest algorithm (Breiman, 2001).
- Since no supervised feedback is given, isolation forest is based on purely random (uniform) variable-based partitioning.
- **Main idea:** Outlying observations are isolated faster.

Isolation forest (Liu, Ting, Zhou; 2008)

- **Isolation forest** (Liu, Ting, Zhou; 2008) is an anomaly detection method inherited from the famous random forest algorithm (Breiman, 2001).
- Since no supervised feedback is given, isolation forest is based on purely random (uniform) variable-based partitioning.
- **Main idea:** Outlying observations are isolated faster.
- Tree-kind partitioning is done until "full isolation": outlying observations will have smaller depth (on an average) in the isolation tree.
- A monotone transform is usually applied to the aggregated estimate.
- To reduce both masking effect and computation cost, small-size sub-sampling is used instead of bootstrap.

Isolation forest (Liu, Ting, Zhou; 2008)

- Each isolation tree is grown **recursively** using the described below node-construction procedure

Isolation forest (Liu, Ting, Zhou; 2008)

- Each isolation tree is grown **recursively** using the described below node-construction procedure

Non-terminal **node** (j, k), **subspace** $\mathcal{C}_{j,k}$, **training subset** $\mathcal{S}_{j,k}$:

1. Choose a split variable l uniformly from $\{1, ..., d\}$.

Isolation forest (Liu, Ting, Zhou; 2008)

- Each isolation tree is grown **recursively** using the described below node-construction procedure

Non-terminal **node** (j, k), **subspace** $\mathcal{C}_{j,k}$, **training subset** $\mathcal{S}_{j,k}$:

1. Choose a split variable l uniformly from $\{1, ..., d\}$.
2. Choose randomly and uniformly a split value κ in the interval

$$\left[\min_{\boldsymbol{x} \in \mathcal{S}_{j,k}} \langle \boldsymbol{x}, \boldsymbol{e}_l \rangle, \max_{\boldsymbol{x} \in \mathcal{S}_{j,k}} \langle \boldsymbol{x}, \boldsymbol{e}_j \rangle\right].$$

Isolation forest (Liu, Ting, Zhou; 2008)

- Each isolation tree is grown **recursively** using the described below node-construction procedure

Non-terminal **node** (j, k), **subspace** $\mathcal{C}_{j,k}$, **training subset** $\mathcal{S}_{j,k}$:

1. Choose a split variable l uniformly from $\{1, ..., d\}$.
2. Choose randomly and uniformly a split value κ in the interval

$$\left[\min_{\boldsymbol{x}\in\mathcal{S}_{j,k}} \langle \boldsymbol{x}, \boldsymbol{e}_l \rangle, \max_{\boldsymbol{x}\in\mathcal{S}_{j,k}} \langle \boldsymbol{x}, \boldsymbol{e}_j \rangle\right] .$$

3. Form the children subsets

$$\begin{aligned} \mathcal{C}_{j+1,2k} &= \mathcal{C}_{j,k} \cap \{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{e}_l \rangle \leq \kappa\}, \\ \mathcal{C}_{j+1,2k+1} &= \mathcal{C}_{j,k} \cap \{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{e}_l \rangle > \kappa\}. \end{aligned}$$

as well as the children training datasets

$$\mathcal{S}_{j+1,2k} = \mathcal{S}_{j,k} \cap \mathcal{C}_{j+1,2k} \text{ and } \mathcal{S}_{j+1,2k+1} = \mathcal{S}_{j,k} \cap \mathcal{C}_{j+1,2k+1}.$$

Isolation forest (Liu, Ting, Zhou; 2008)

- Each isolation tree is grown **recursively** using the described below node-construction procedure

Non-terminal **node** (j, k), **subspace** $\mathcal{C}_{j,k}$, **training subset** $\mathcal{S}_{j,k}$:

1. Choose a split variable l uniformly from $\{1, ..., d\}$.
2. Choose randomly and uniformly a split value κ in the interval

$$\left[\min_{\boldsymbol{x} \in \mathcal{S}_{j,k}} \langle \boldsymbol{x}, \boldsymbol{e}_l \rangle, \max_{\boldsymbol{x} \in \mathcal{S}_{j,k}} \langle \boldsymbol{x}, \boldsymbol{e}_j \rangle\right].$$

3. Form the children subsets

$$\begin{aligned} \mathcal{C}_{j+1,2k} &= \mathcal{C}_{j,k} \cap \{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{e}_l \rangle \leq \kappa\}, \\ \mathcal{C}_{j+1,2k+1} &= \mathcal{C}_{j,k} \cap \{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{e}_l \rangle > \kappa\}. \end{aligned}$$

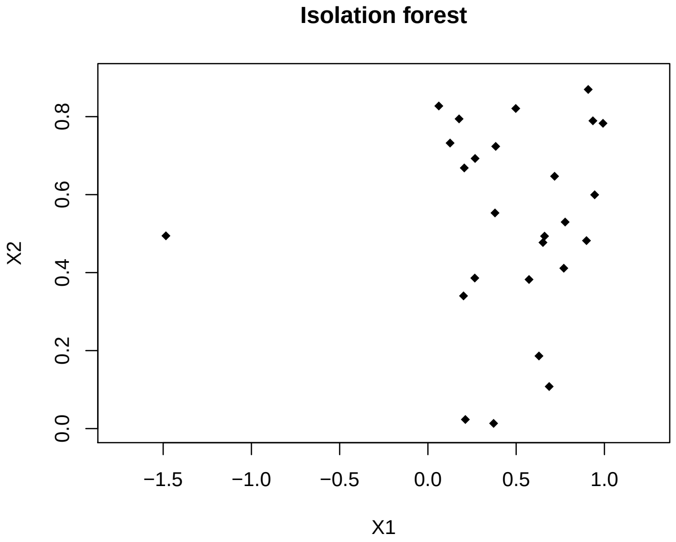
as well as the children training datasets

$$\mathcal{S}_{j+1,2k} = \mathcal{S}_{j,k} \cap \mathcal{C}_{j+1,2k} \text{ and } \mathcal{S}_{j+1,2k+1} = \mathcal{S}_{j,k} \cap \mathcal{C}_{j+1,2k+1}.$$

Stop when only one observation is in each node: isolation.

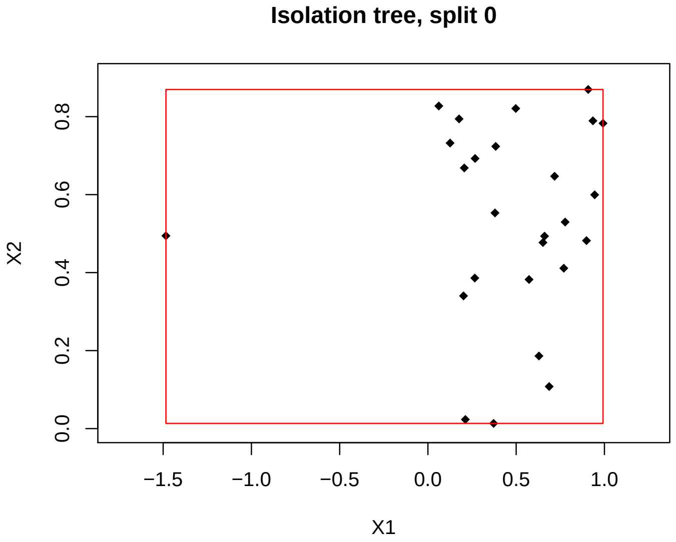
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



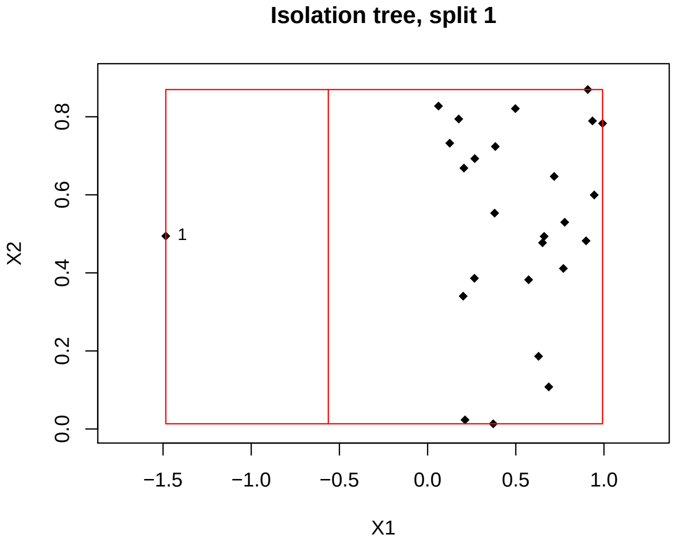
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



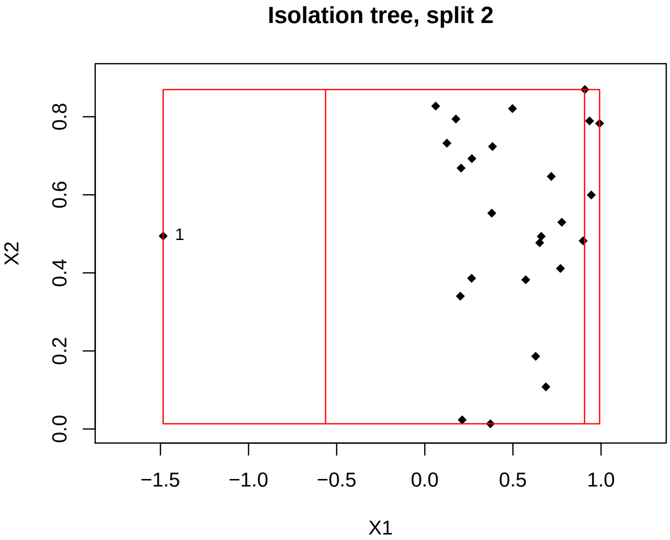
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



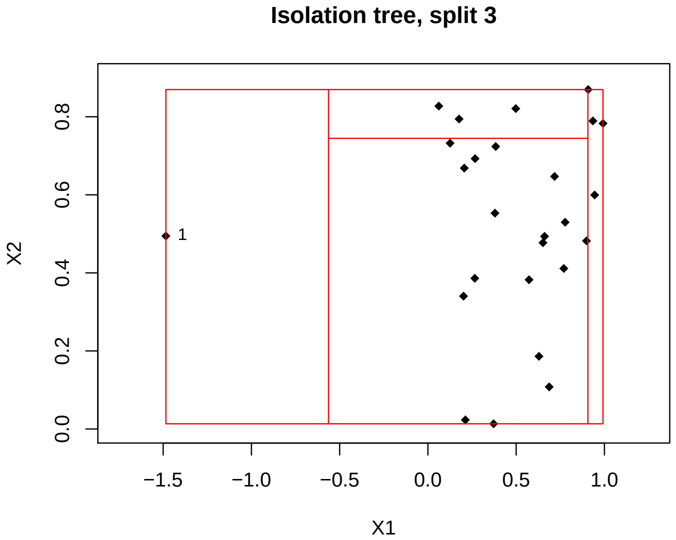
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



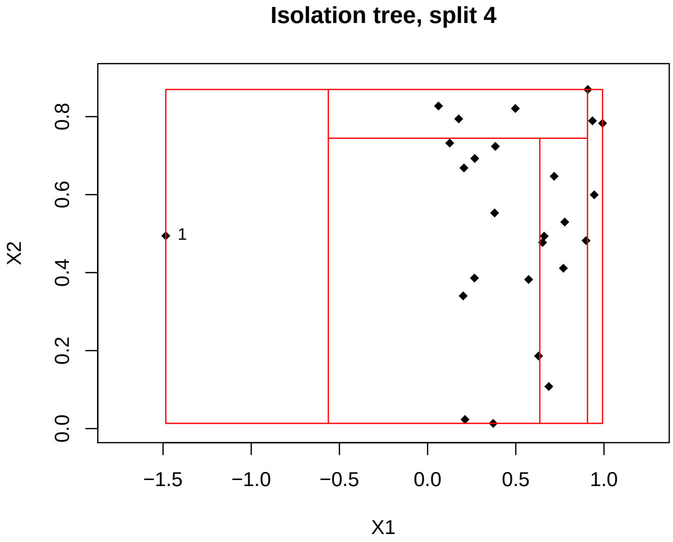
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



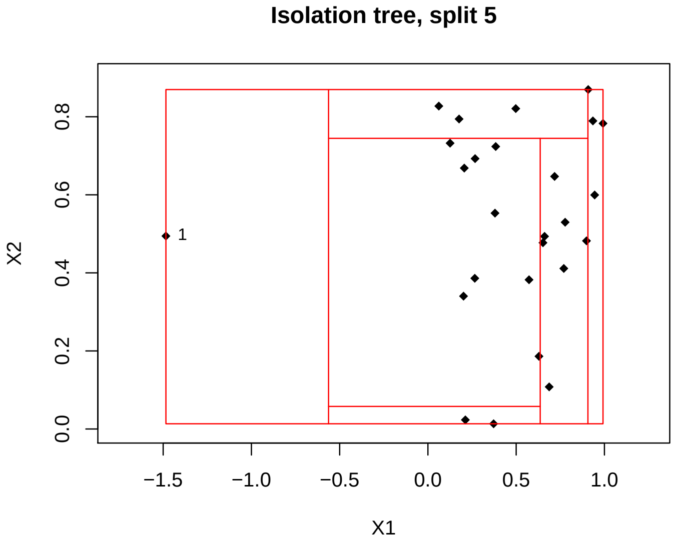
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



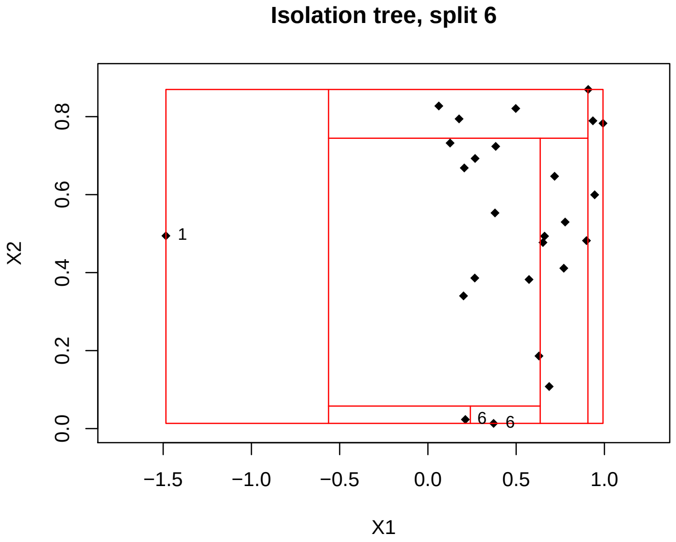
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



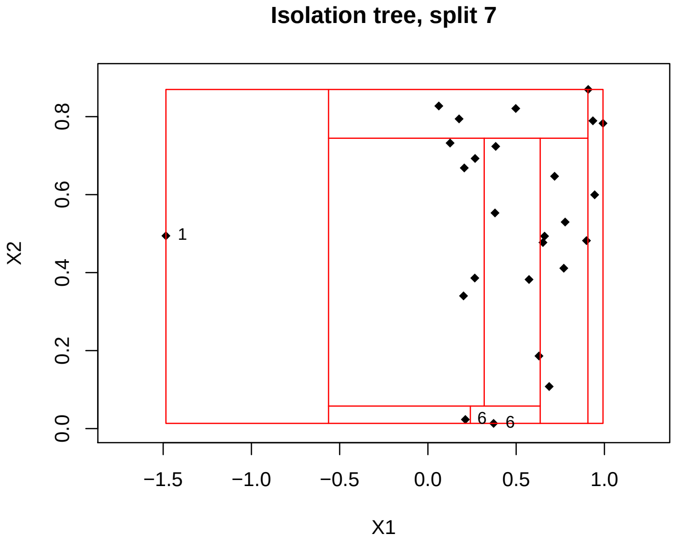
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



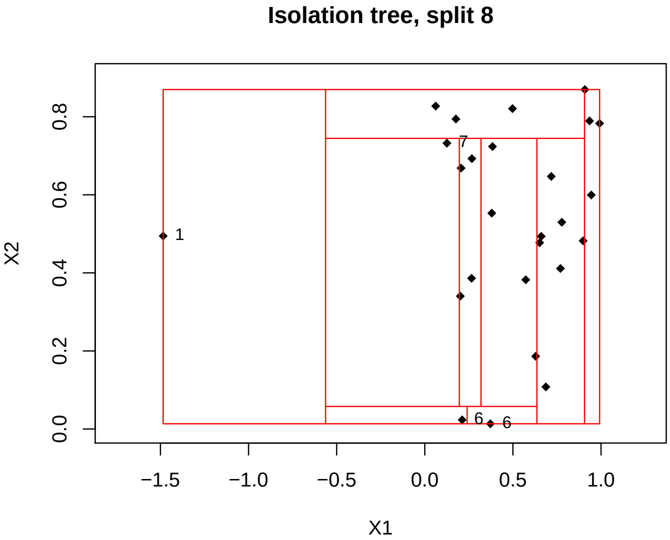
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



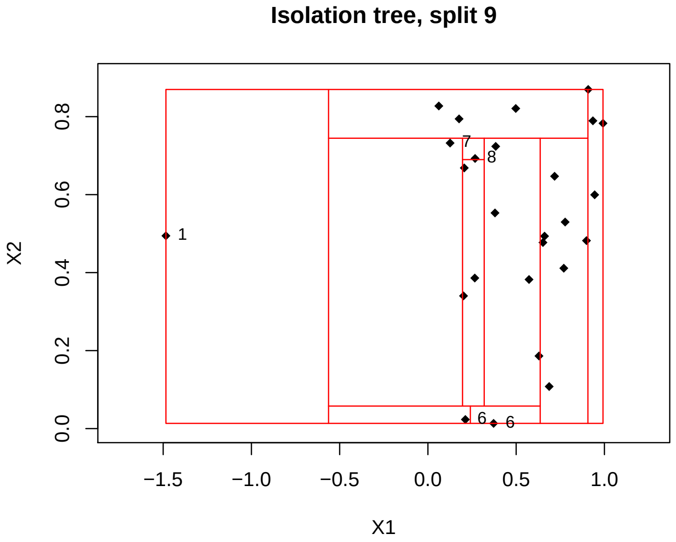
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



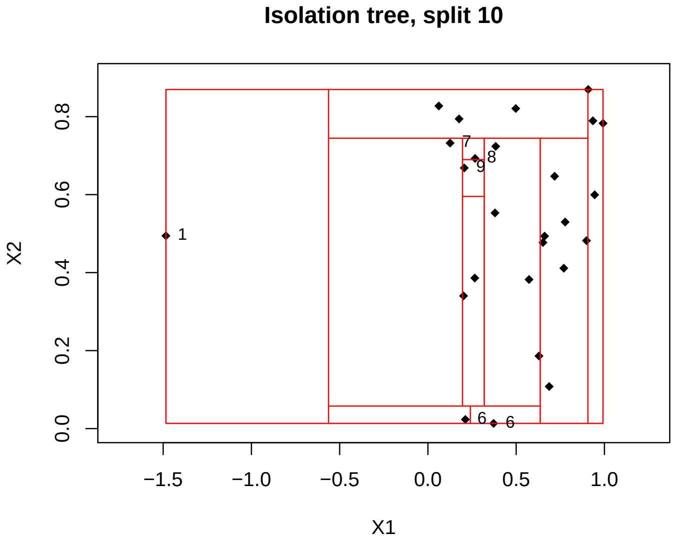
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



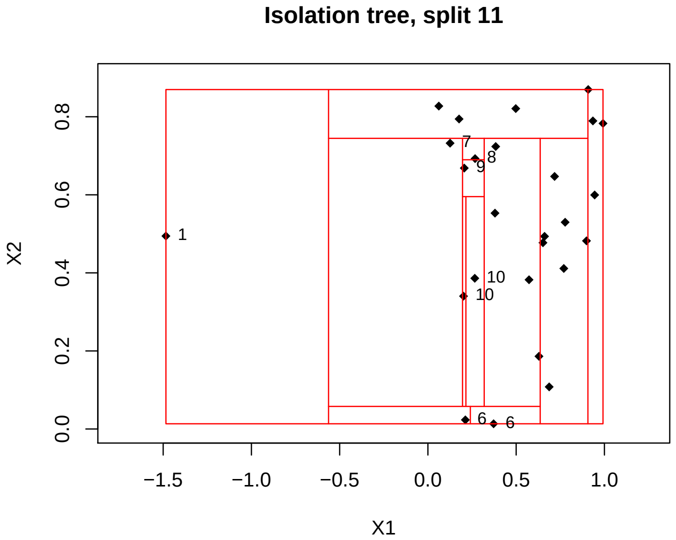
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



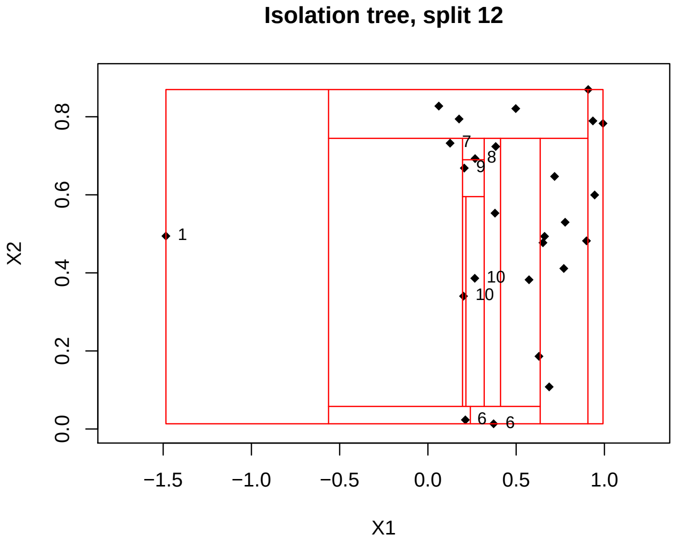
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



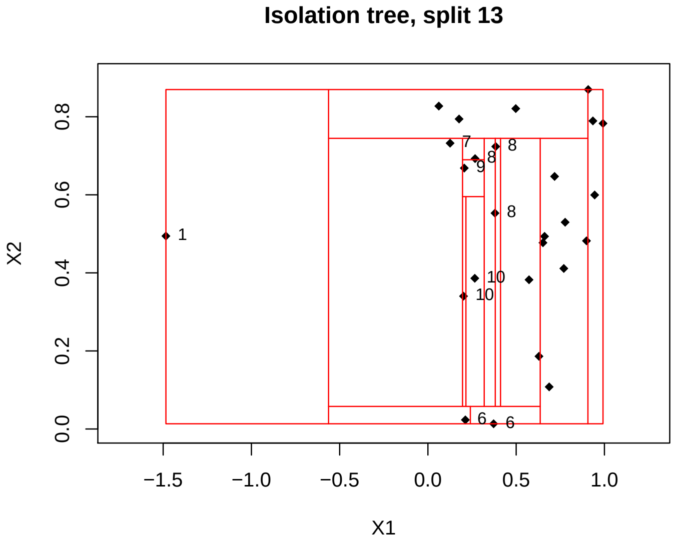
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



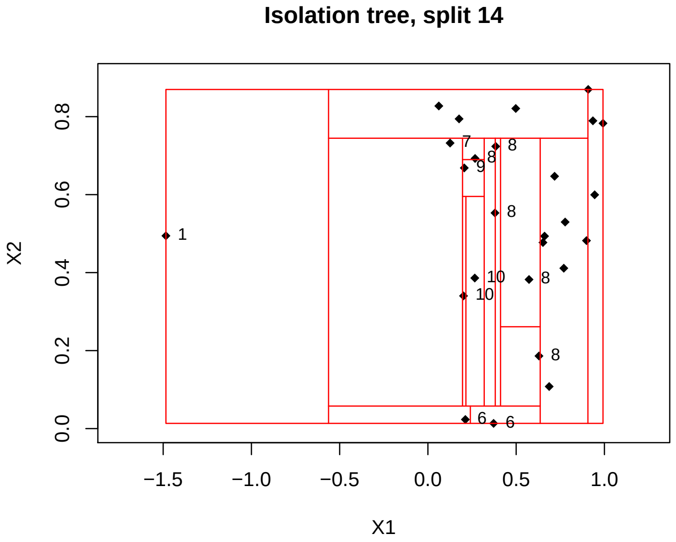
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



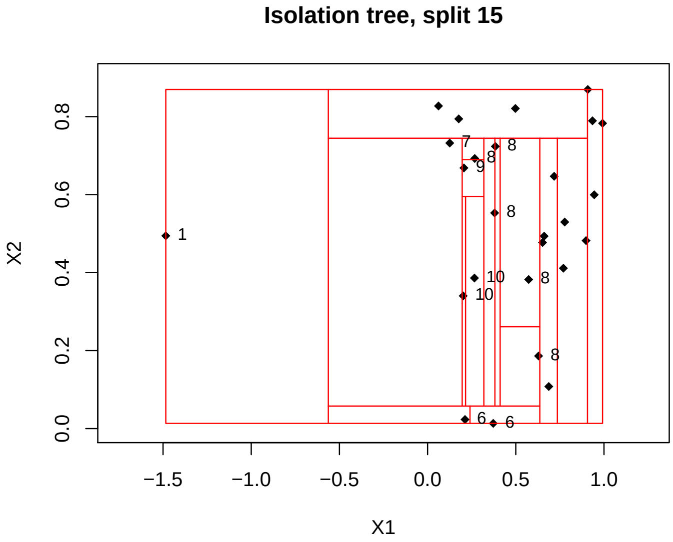
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



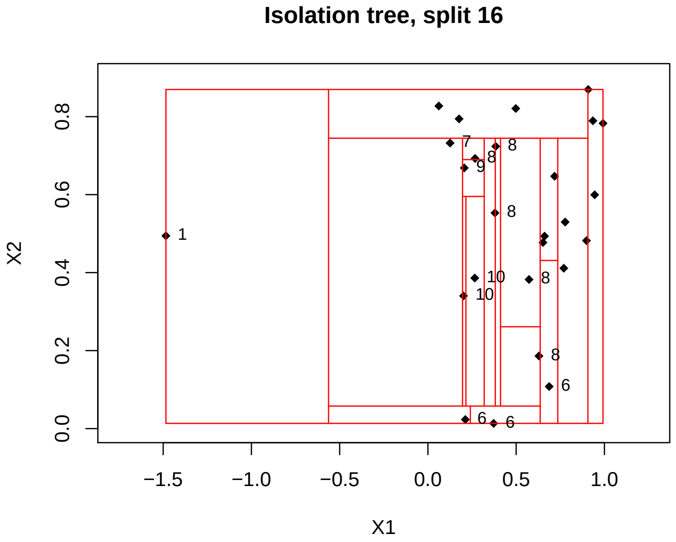
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



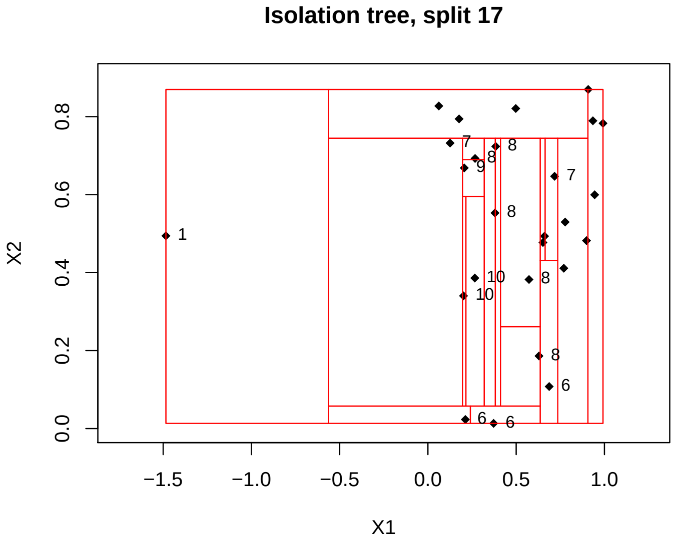
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



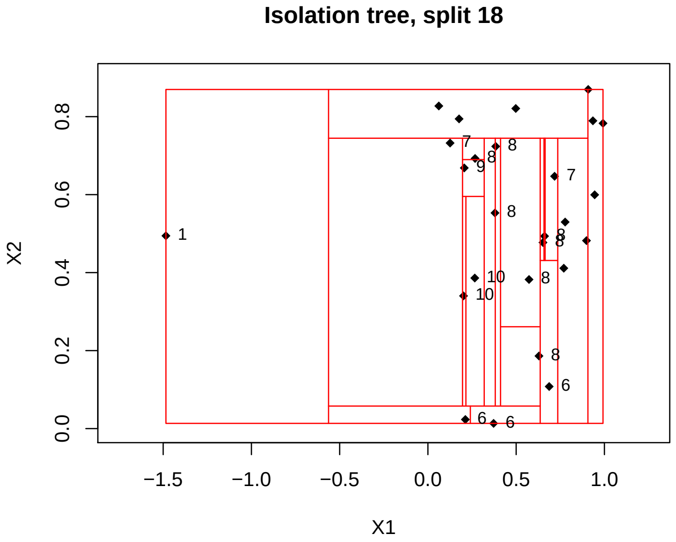
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



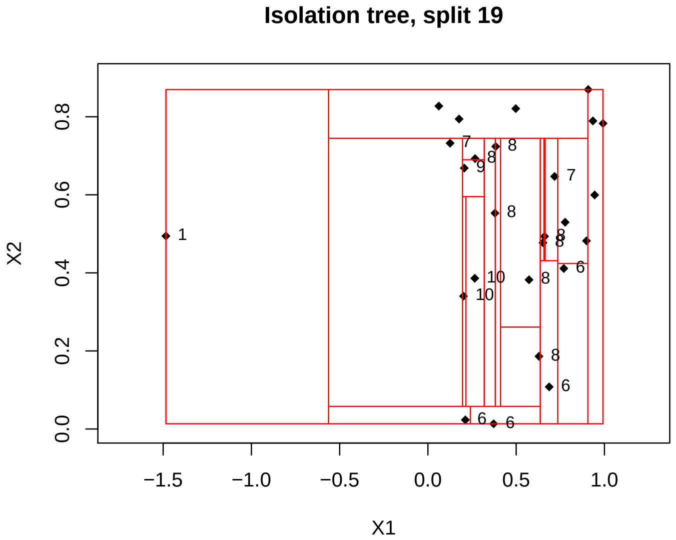
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



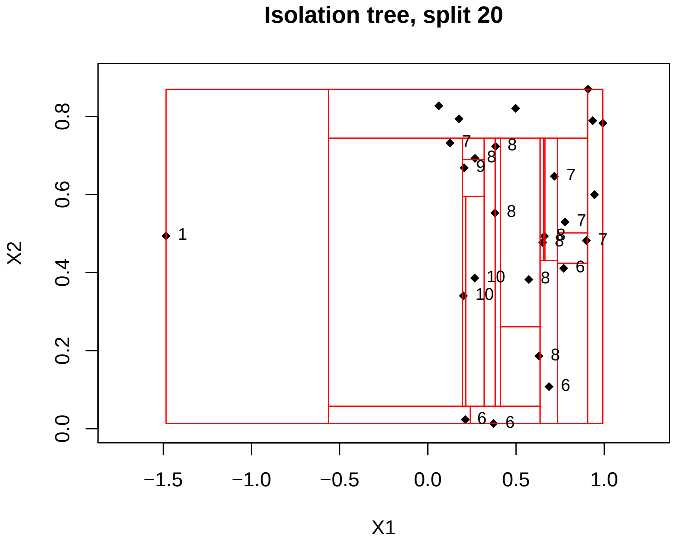
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



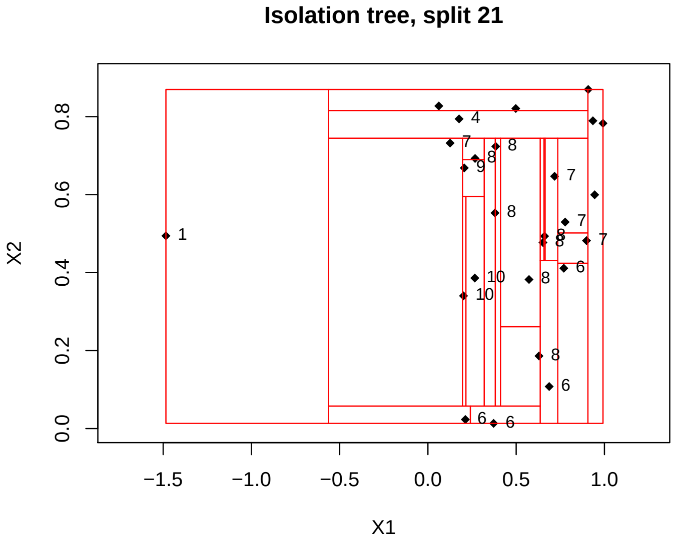
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



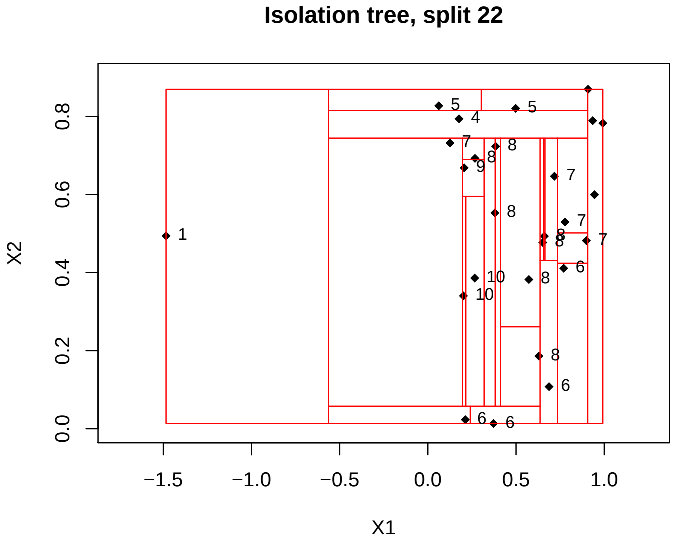
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



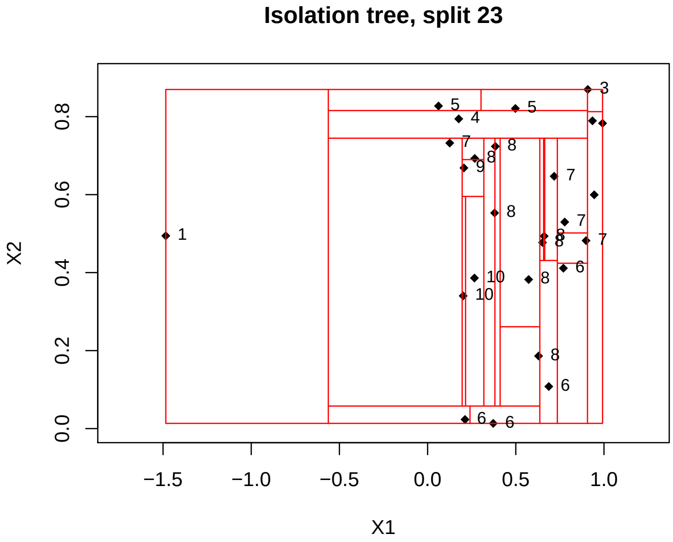
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



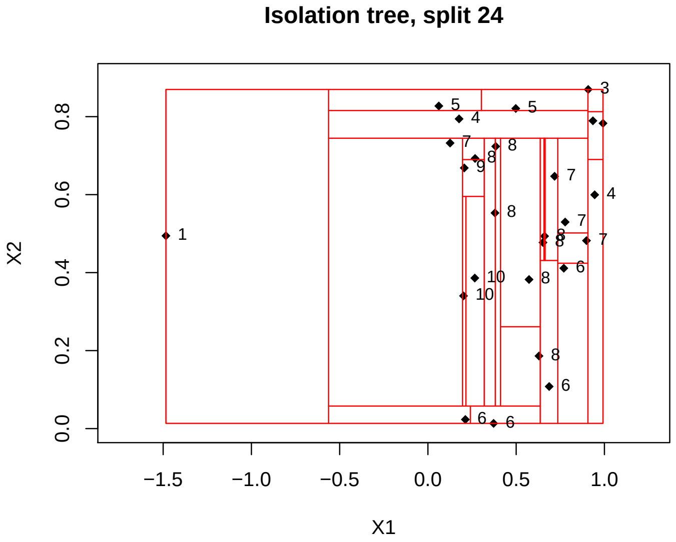
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



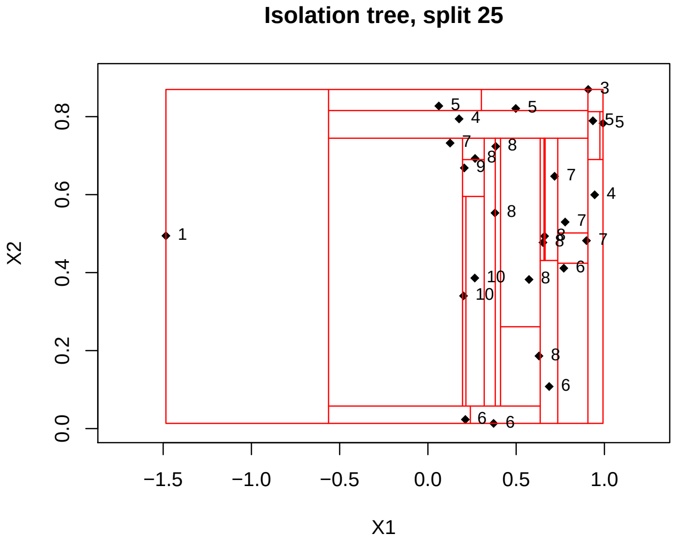
Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



Isolation forest (Liu, Ting, Zhou; 2008)

Illustration: Isolation tree



Isolation forest (Liu, Ting, Zhou; 2008)

Anomaly score calculation for observation $\boldsymbol{x}$:

1. For each isolation tree $i \in \{1, ..., T\}$, locate $\boldsymbol{x}$ in a terminal node and calculate the depth of this node $h_i(\boldsymbol{x})$.
2. Attribute the anomaly score:

$$s(\boldsymbol{x}) = 2^{-\frac{\frac{1}{n}\sum_{i=1}^{T} h_i(\boldsymbol{x})}{c(n)}},$$

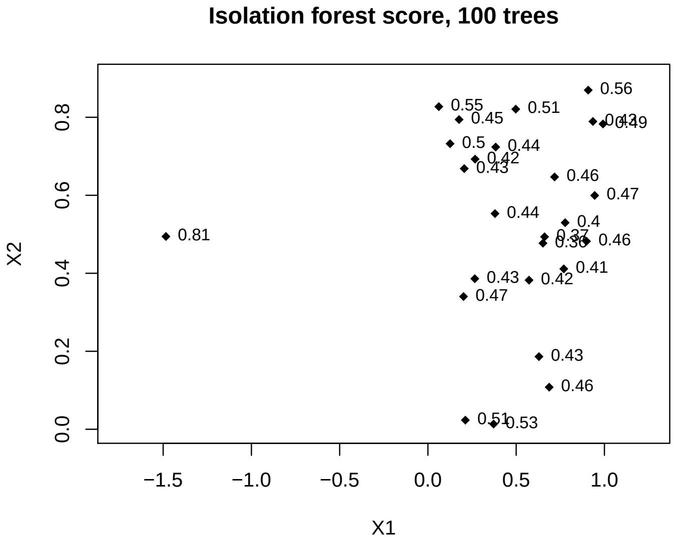
with $c(n) = 2H(n-1) - \frac{2(n-1)}{n}$ where $H(k)$ is the harmonic number and can be estimated by $\ln(k) + 0.5772156649$.

Score behavior:

- when $\frac{1}{n}\sum_{i=1}^{T} h_i(\boldsymbol{x}) \to c(n)$, $s(\boldsymbol{x}) \to 0.5$,
- when $\frac{1}{n}\sum_{i=1}^{T} h_i(\boldsymbol{x}) \to 0$, $s(\boldsymbol{x}) \to 1$,
- when $\frac{1}{n}\sum_{i=1}^{T} h_i(\boldsymbol{x}) \to n-1$, $s(\boldsymbol{x}) \to 0$.

Isolation forest (Liu, Ting, Zhou; 2008)

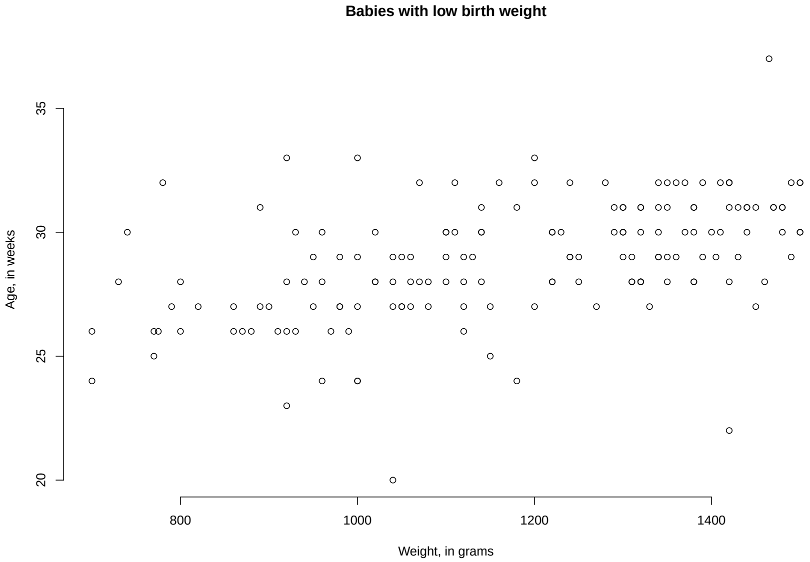
Illustration: Anomaly score



Contents

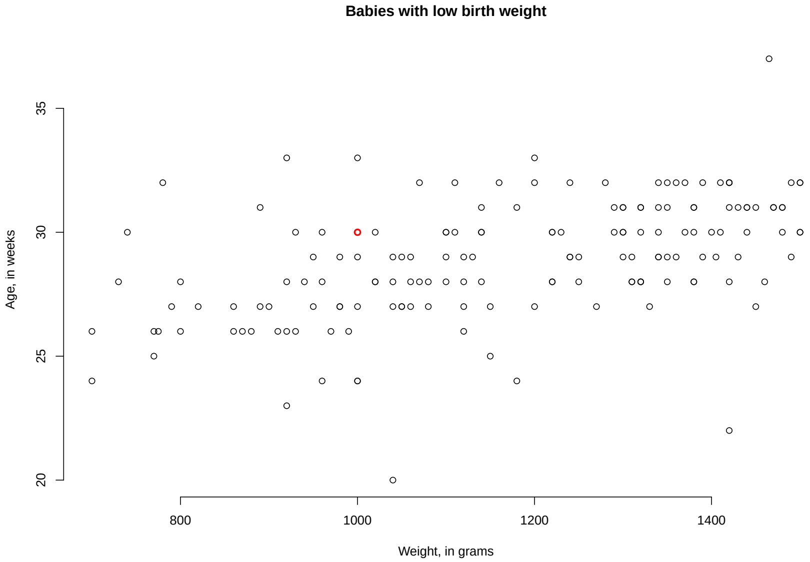
Contents

Data depth



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

Data depth


Babies with low birth weight
Age, in weeks
20
25
30
35
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

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

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 $b \in \mathbb{R}^p$;

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;

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

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

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 $\boldsymbol{x}^* \in \mathrm{argmax}_{\boldsymbol{x} \in \mathbb{R}^p} D(\boldsymbol{x}|X)$, 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.

Statistical data depth

Some remarks:

- **D4** implies star-shaped upper-level sets of D.

Statistical data depth

Some remarks:

- **D4** implies star-shaped upper-level sets of D.

  One can strengthen to:

  - **D4con**: $D(\cdot|X)$ is a **quasiconcave** function, *i.e.* the upper-level sets $D_\alpha(X)$ are convex for all α.

Statistical data depth

Some remarks:

- **D4** implies star-shaped upper-level sets of D.

  One can strengthen to:

  - **D4con**: $D(\cdot|X)$ is a **quasiconcave** function, *i.e.* the upper-level sets $D_\alpha(X)$ are convex for all α.

- **D1** and **D2** define **affine invariante depth**.

Statistical data depth

Some remarks:

- **D4** implies star-shaped upper-level sets of D.

  One can strengthen to:

  - **D4con**: $D(\cdot|X)$ is a **quasiconcave** function, *i.e.* the upper-level sets $D_\alpha(X)$ are convex for all α.

- **D1** and **D2** define **affine invariante depth**.

  One can also weaken to:

  - **D2iso**: $D(A\boldsymbol{x}|AX) = D(\boldsymbol{x}|X)$ for every isometric linear A to define **orthogonal invariant depth**;

Statistical data depth

Some remarks:

- **D4** implies star-shaped upper-level sets of D.

  One can strengthen to:

  - **D4con**: $D(\cdot|X)$ is a **quasiconcave** function, *i.e.* the upper-level sets $D_\alpha(X)$ are convex for all α.

- **D1** and **D2** define **affine invariante depth**.

  One can also weaken to:

  - **D2iso**: $D(A\boldsymbol{x}|AX) = D(\boldsymbol{x}|X)$ for every isometric linear A to define **orthogonal invariant depth**;
  - **D2sca**: $D(\lambda\boldsymbol{x}|\lambda X) = D(\boldsymbol{x}|X)$ for any $\lambda > 0$ to define **scale invariant depth**.

Statistical data depth

Some remarks:

- **D4** implies star-shaped upper-level sets of D.

  One can strengthen to:

  - **D4con**: $D(\cdot|X)$ is a **quasiconcave** function, *i.e.* the upper-level sets $D_\alpha(X)$ are convex for all α.

- **D1** and **D2** define **affine invariante depth**.

  One can also weaken to:

  - **D2iso**: $D(A\boldsymbol{x}|AX) = D(\boldsymbol{x}|X)$ for every isometric linear A to define **orthogonal invariant depth**;
  - **D2sca**: $D(\lambda\boldsymbol{x}|\lambda X) = D(\boldsymbol{x}|X)$ for any $\lambda > 0$ to define **scale invariant depth**.

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), **lens** (Liu, Modarres, '11), ... depth.

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 '20);
- 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.

Tukey (=halfspace, location) depth

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

Tukey 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**:

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

Tukey (=halfspace, location) depth

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

Tukey 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**:

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

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

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

Tukey (=halfspace, location) depth

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

Tukey 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 $\mathbf{x}$:

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

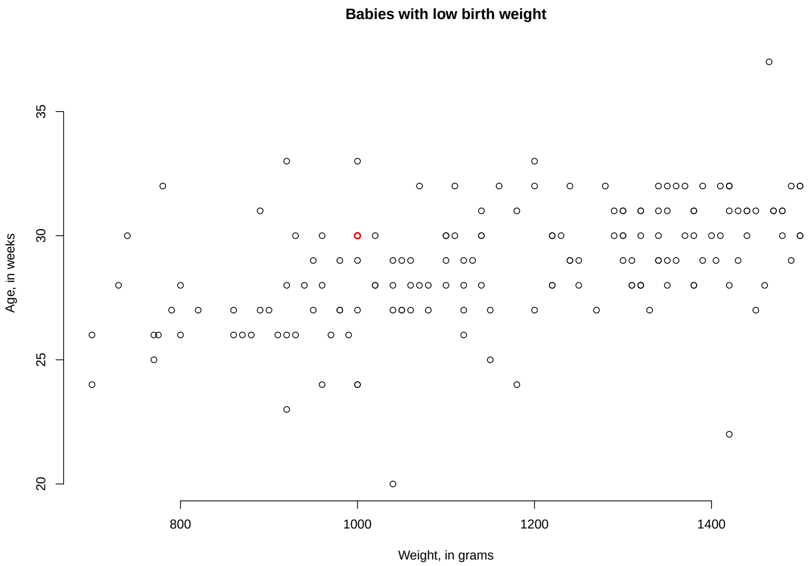
and w.r.t. a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subset \mathbb{R}^p$:

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

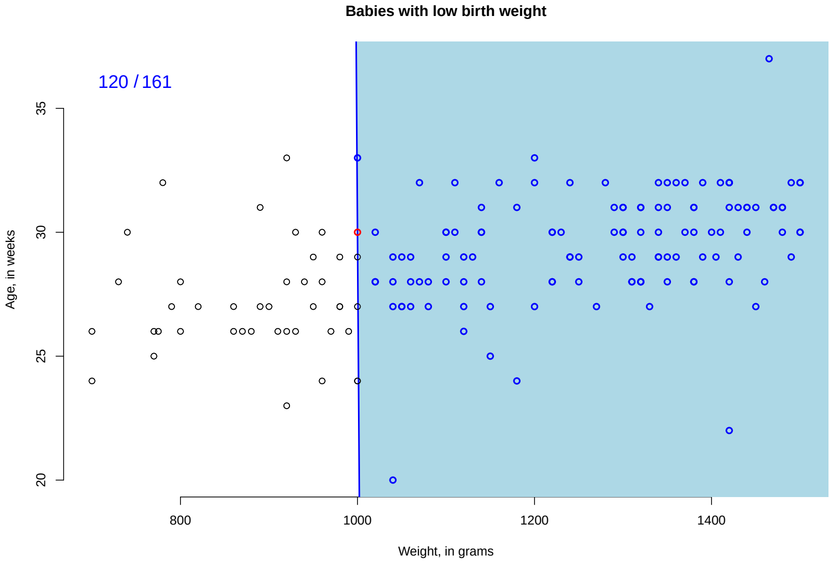
Tukey depth

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

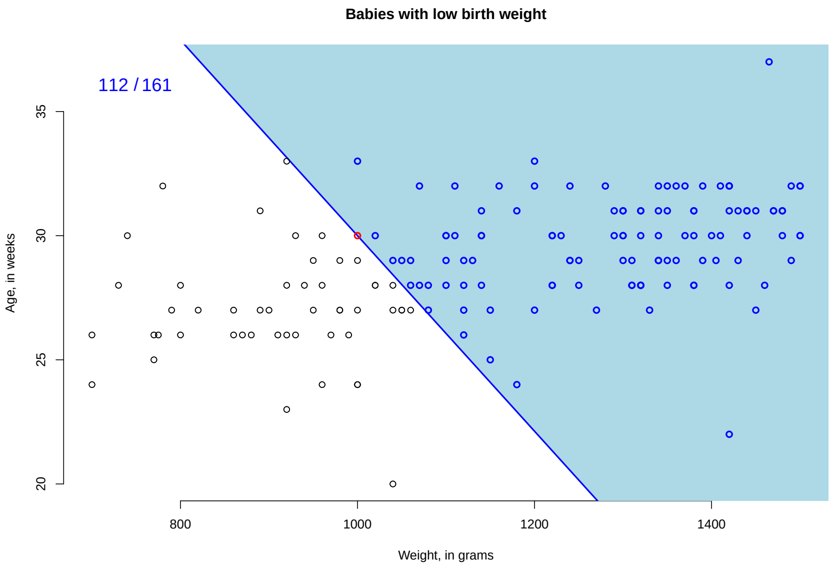
Tukey (=halfspace, location) data depth



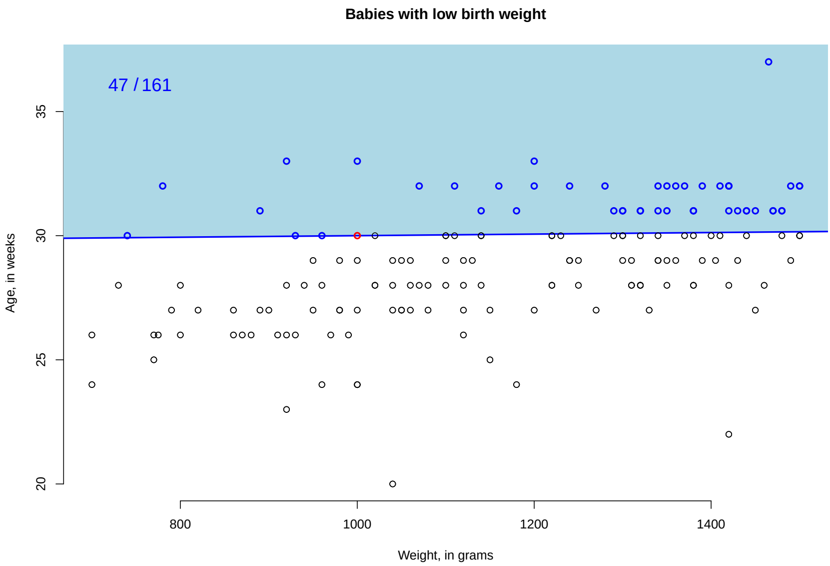
Tukey (=halfspace, location) data depth



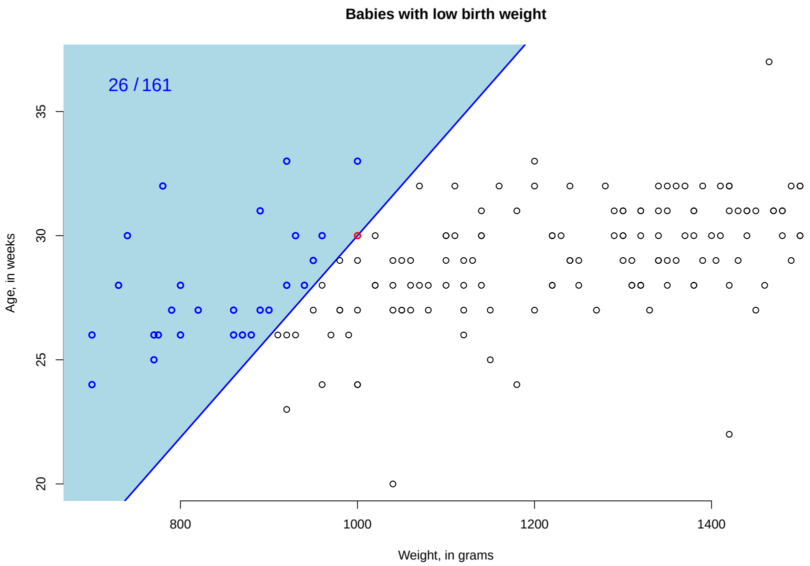
Tukey (=halfspace, location) data depth



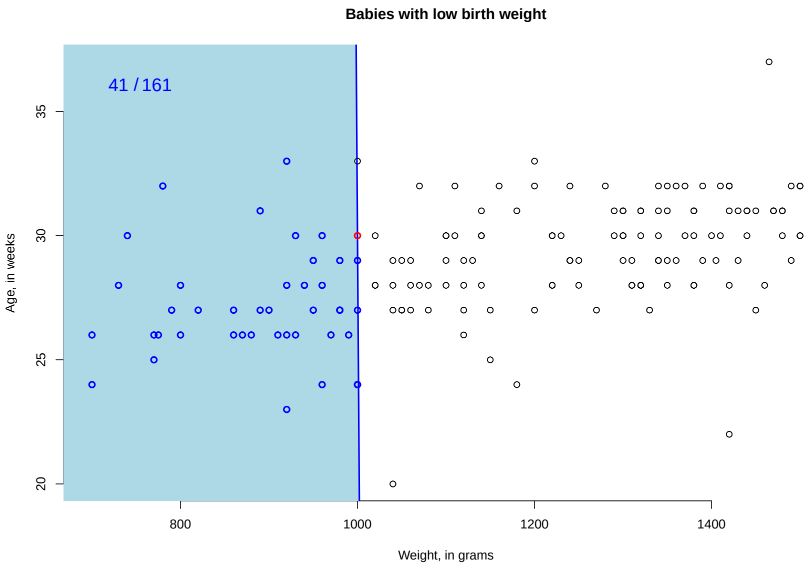
Tukey (=halfspace, location) data depth



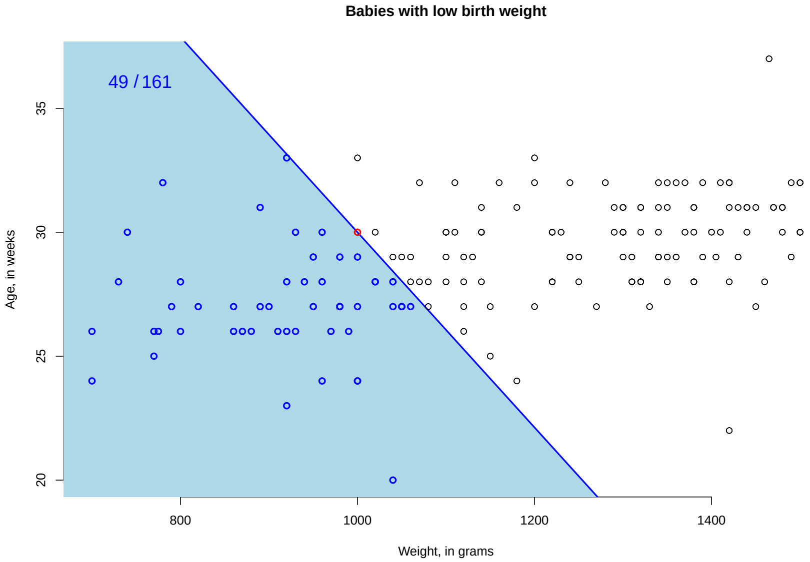
Tukey (=halfspace, location) data depth



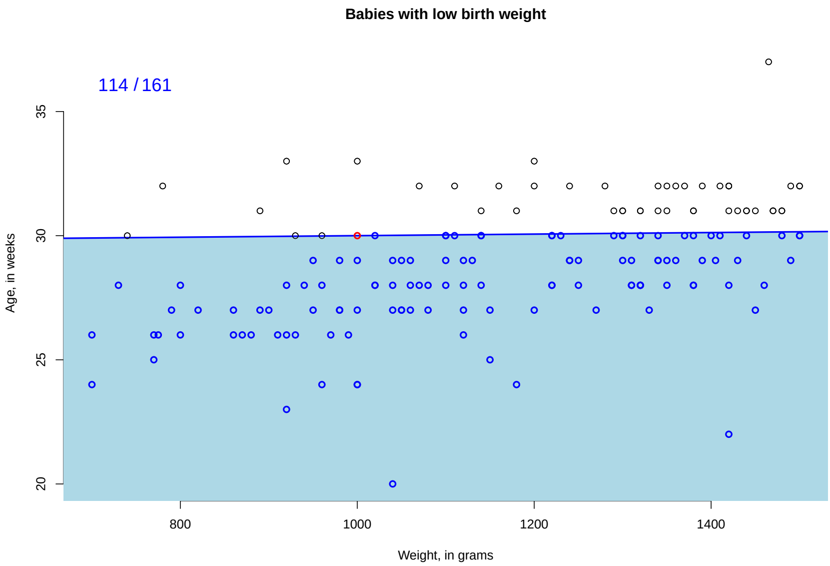
Tukey (=halfspace, location) data depth



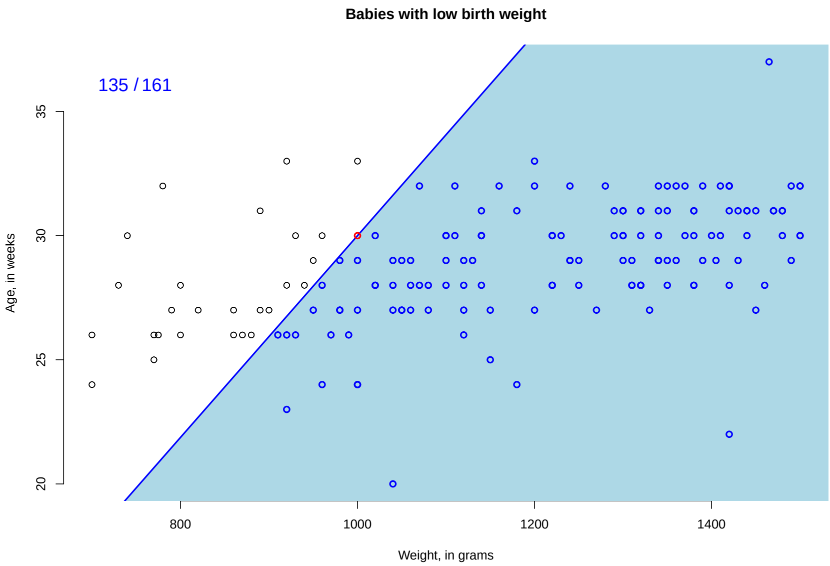
Tukey (=halfspace, location) data depth



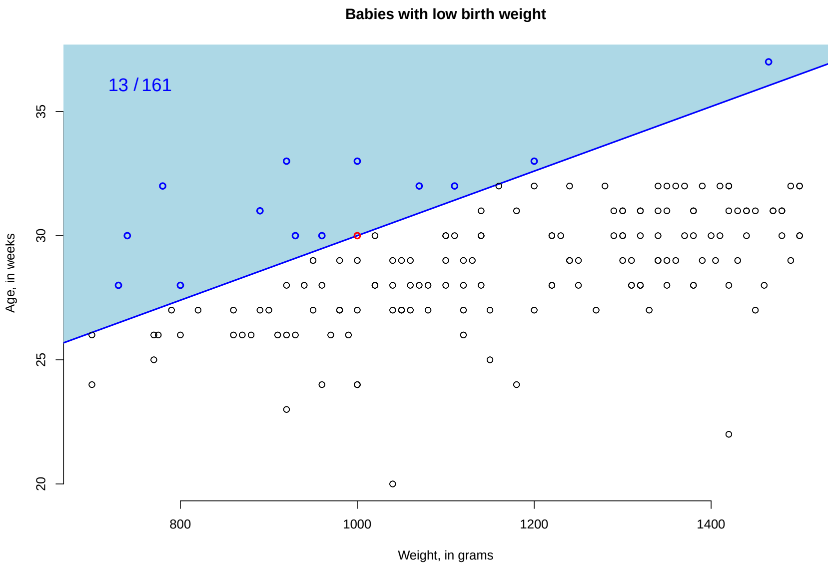
Tukey (=halfspace, location) data depth



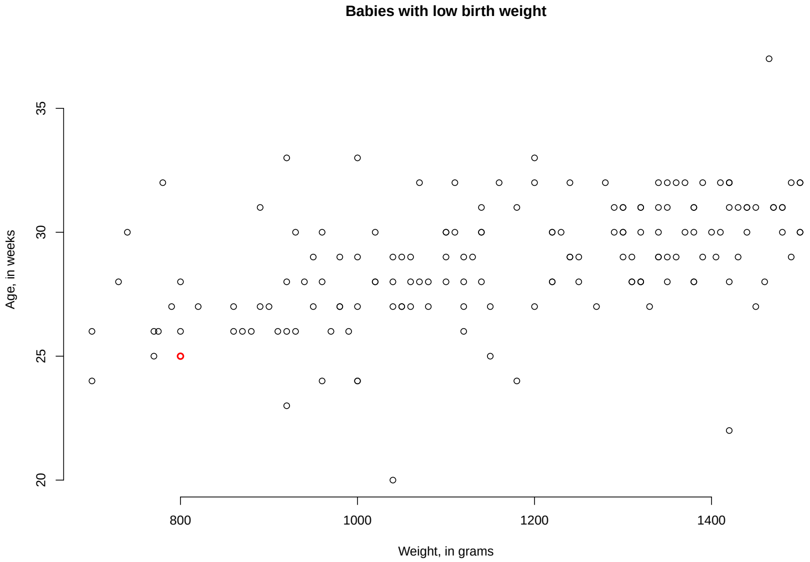
Tukey (=halfspace, location) data depth



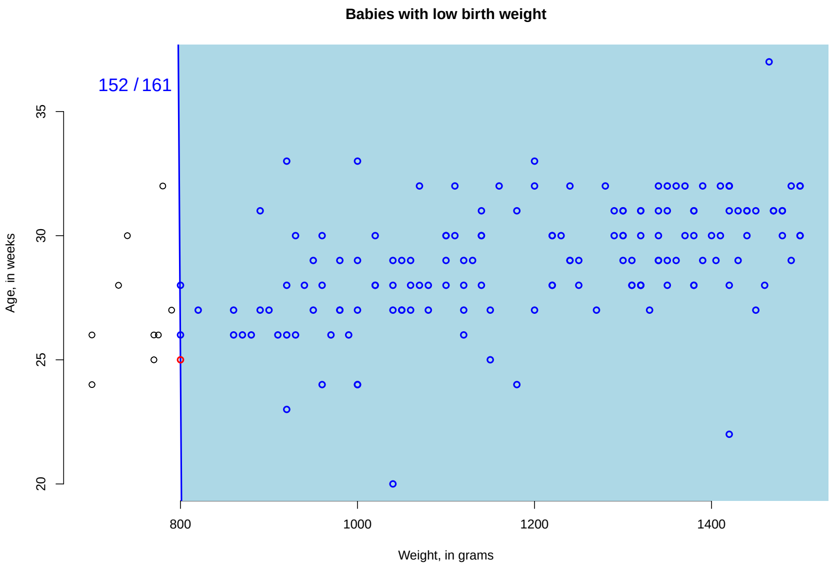
Tukey (=halfspace, location) data depth



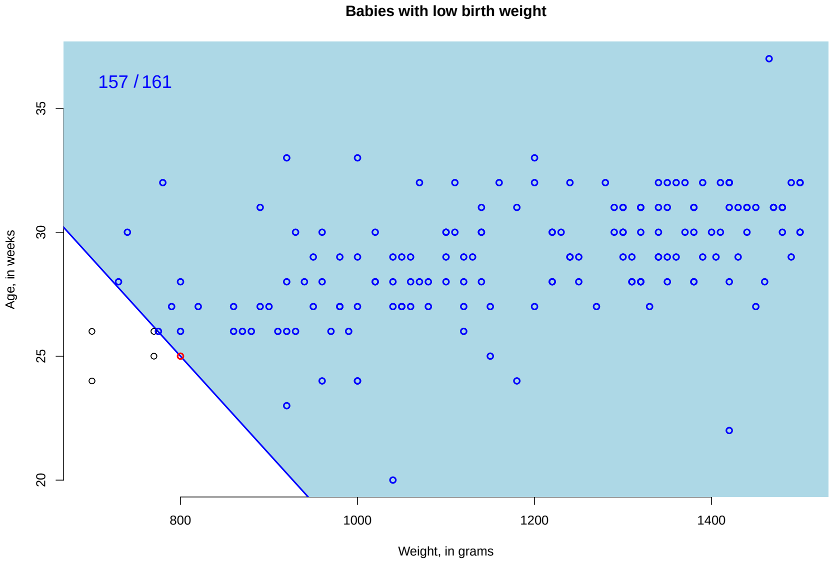
Tukey (=halfspace, location) data depth



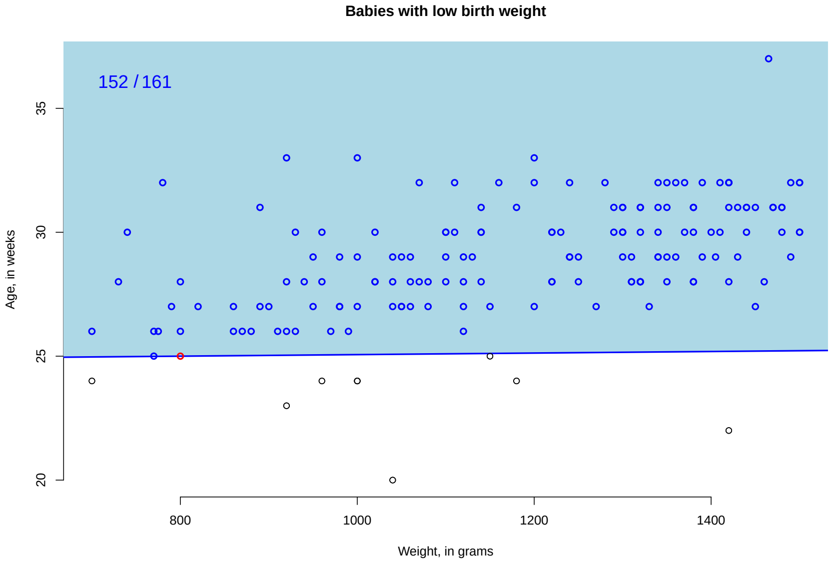
Tukey (=halfspace, location) data depth



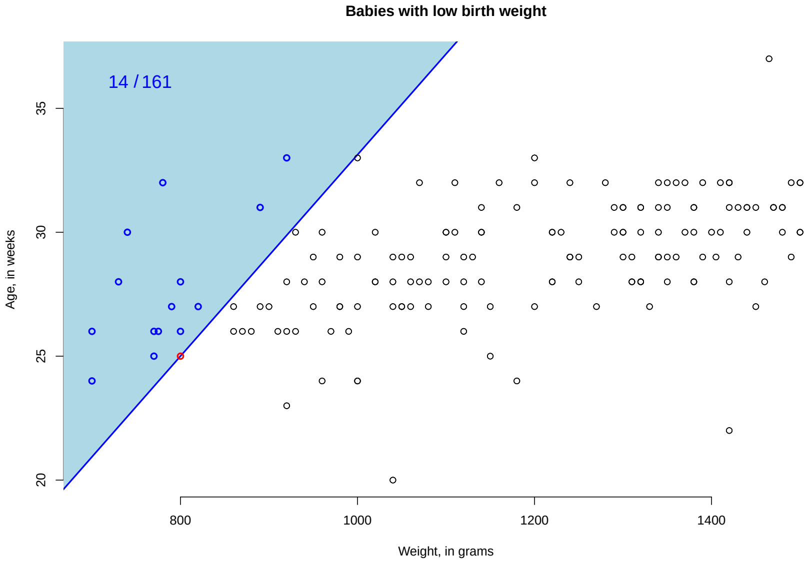
Tukey (=halfspace, location) data depth



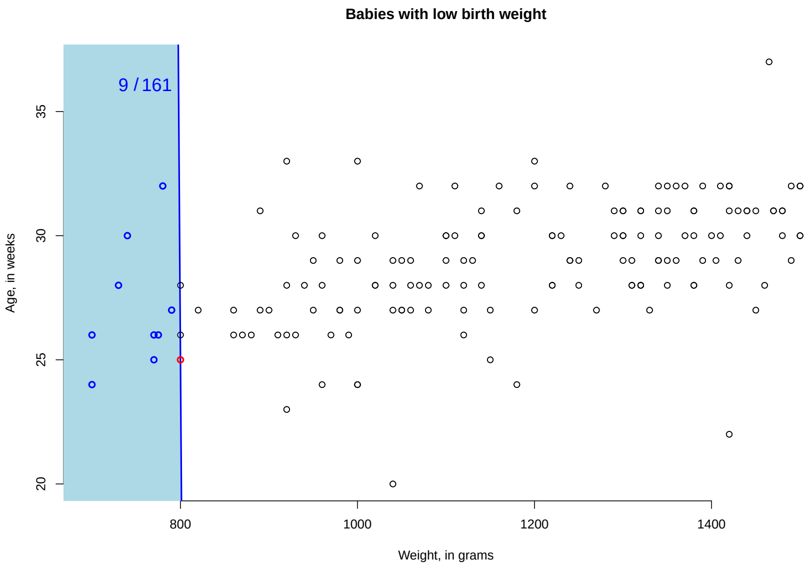
Tukey (=halfspace, location) data depth



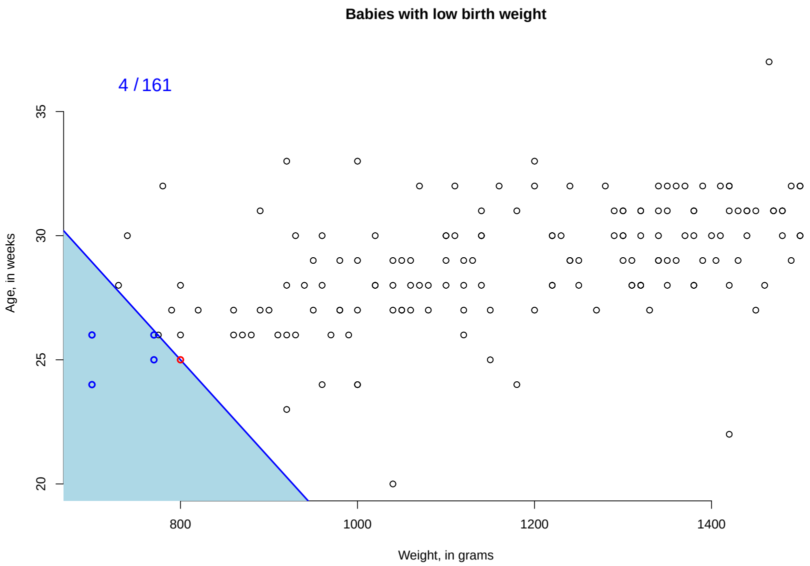
Tukey (=halfspace, location) data depth



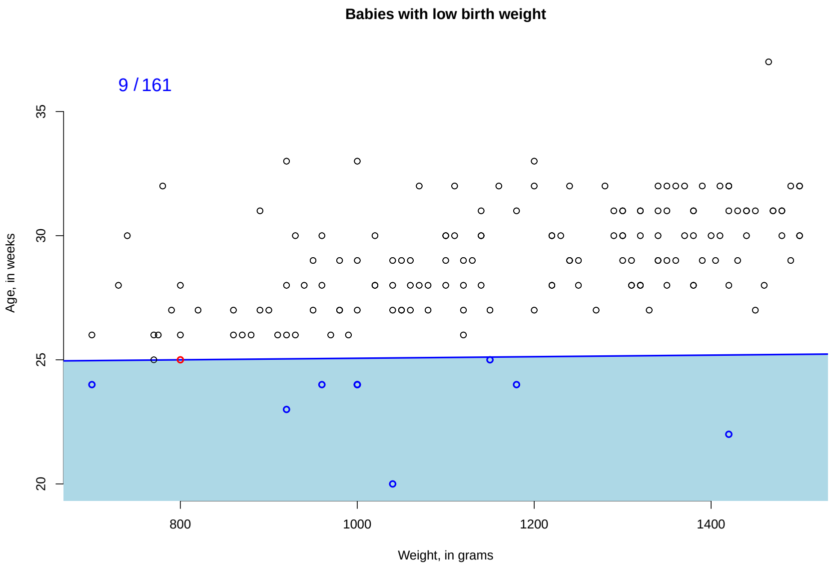
Tukey (=halfspace, location) data depth



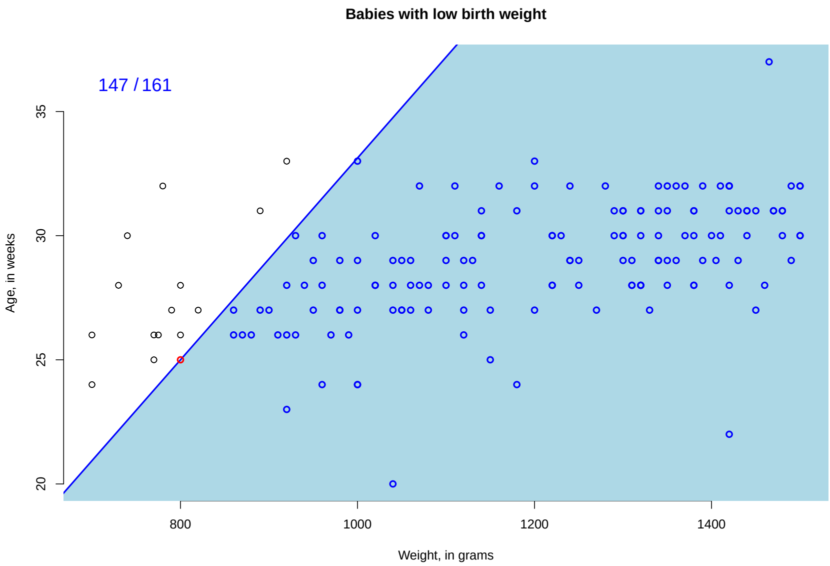
Tukey (=halfspace, location) data depth



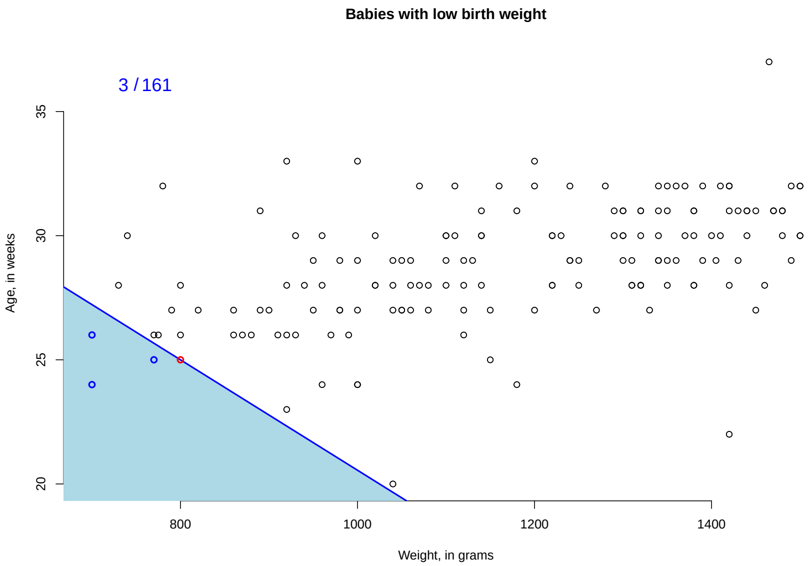
Tukey (=halfspace, location) data depth



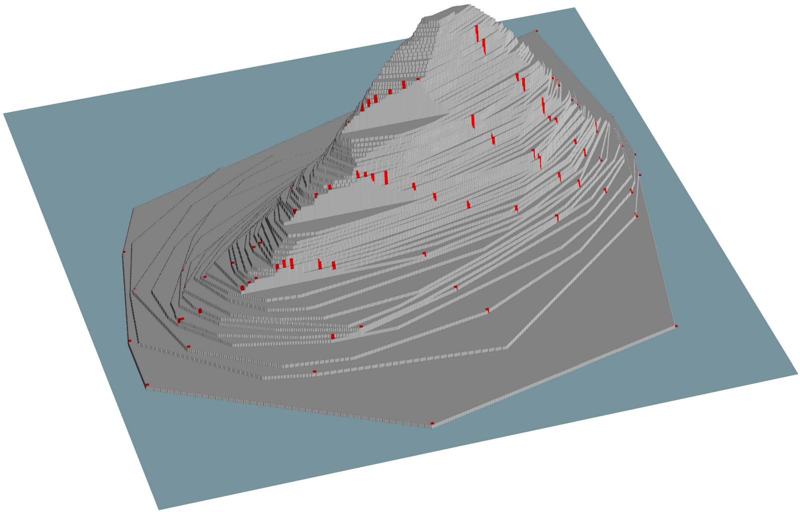
Tukey (=halfspace, location) data depth



Tukey (=halfspace, location) data depth



Tukey (=halfspace, location) data depth



Contents

Central regions

- For given distribution P and $\alpha \in [0, 1]$, the level sets $D_\alpha(P)$ form a family of **depth-trimmed** of **central regions**.

Central regions

- For given distribution P and $\alpha \in [0, 1]$, the level sets $D_\alpha(P)$ form a family of **depth-trimmed** of **central regions**.
- The innermost region arises at some depth $\alpha_{\mathsf{max}} \leq 1$, which depends on the depth notion D and distribution P. Then $D_\alpha(X)$ is the set of **deepest points**.

Central regions

- For given distribution P and $\alpha \in [0, 1]$, the level sets $D_\alpha(P)$ form a family of **depth-trimmed** of **central regions**.
- The innermost region arises at some depth $\alpha_{\max} \leq 1$, which depends on the depth notion D and distribution P. Then $D_\alpha(X)$ is the set of **deepest points**.
- Central regions describe distribution w.r.t. location, dispersion, and shape.

Central regions

- For given distribution P and $\alpha \in [0,1]$, the level sets $D_\alpha(P)$ form a family of **depth-trimmed** of **central regions**.
- The innermost region arises at some depth $\alpha_{\max} \leq 1$, which depends on the depth notion D and distribution P. Then $D_\alpha(X)$ is the set of **deepest points**.
- Central regions describe distribution w.r.t. location, dispersion, and shape.
- Properties of central regions, for any α:
  - Due to **D1** and **D2** $D_\alpha(X)$ is **affine equivariant**: $D_\alpha(AX+b) = AD_\alpha(X) + b$ for any $p \times p$ non-singular matrix A and any $b \in \mathbb{R}^p$;

Central regions

- For given distribution P and $\alpha \in [0, 1]$, the level sets $D_\alpha(P)$ form a family of **depth-trimmed** of **central regions**.
- The innermost region arises at some depth $\alpha_{\max} \leq 1$, which depends on the depth notion D and distribution P. Then $D_\alpha(X)$ is the set of **deepest points**.
- Central regions describe distribution w.r.t. location, dispersion, and shape.
- Properties of central regions, for any α:
  - Due to **D1** and **D2** $D_\alpha(X)$ is **affine equivariant**: $D_\alpha(AX + b) = AD_\alpha(X) + b$ for any $p \times p$ non-singular matrix A and any $b \in \mathbb{R}^p$;
  - Due to **D3** $D_\alpha(X)$ is bounded;

Central regions

- For given distribution P and $\alpha \in [0,1]$, the level sets $D_\alpha(P)$ form a family of **depth-trimmed** of **central regions**.
- The innermost region arises at some depth $\alpha_{\mathsf{max}} \leq 1$, which depends on the depth notion D and distribution P. Then $D_\alpha(X)$ is the set of **deepest points**.
- Central regions describe distribution w.r.t. location, dispersion, and shape.
- Properties of central regions, for any α:
  - Due to **D1** and **D2** $D_\alpha(X)$ is **affine equivariant**: $D_\alpha(AX + b) = AD_\alpha(X) + b$ for any $p \times p$ non-singular matrix A and any $b \in \mathbb{R}^p$;
  - Due to **D3** $D_\alpha(X)$ is bounded;
  - Due to **D4** $D_\alpha(X)$-s are nested: if $\alpha \geq \beta$, then $D_\alpha(X) \subseteq D_\beta(X)$, and star-shaped;

Central regions

- For given distribution P and $\alpha \in [0, 1]$, the level sets $D_\alpha(P)$ form a family of **depth-trimmed** of **central regions**.
- The innermost region arises at some depth $\alpha_{\text{max}} \leq 1$, which depends on the depth notion D and distribution P. Then $D_\alpha(X)$ is the set of **deepest points**.
- Central regions describe distribution w.r.t. location, dispersion, and shape.
- Properties of central regions, for any α:
  - Due to **D1** and **D2** $D_\alpha(X)$ is **affine equivariant**: $D_\alpha(AX + b) = AD_\alpha(X) + b$ for any $p \times p$ non-singular matrix A and any $b \in \mathbb{R}^p$;
  - Due to **D3** $D_\alpha(X)$ is bounded;
  - Due to **D4** $D_\alpha(X)$-s are nested: if $\alpha \geq \beta$, then $D_\alpha(X) \subseteq D_\beta(X)$, and star-shaped; due to **D4con** $D_\alpha(X)$ is in addition convex;

Central regions

- For given distribution P and $\alpha \in [0,1]$, the level sets $D_\alpha(P)$ form a family of **depth-trimmed** of **central regions**.
- The innermost region arises at some depth $\alpha_{\text{max}} \leq 1$, which depends on the depth notion D and distribution P. Then $D_\alpha(X)$ is the set of **deepest points**.
- Central regions describe distribution w.r.t. location, dispersion, and shape.
- Properties of central regions, for any α:
  - Due to **D1** and **D2** $D_\alpha(X)$ is **affine equivariant**: $D_\alpha(AX + b) = AD_\alpha(X) + b$ for any $p \times p$ non-singular matrix A and any $b \in \mathbb{R}^p$;
  - Due to **D3** $D_\alpha(X)$ is bounded;
  - Due to **D4** $D_\alpha(X)$-s are nested: if $\alpha \geq \beta$, then $D_\alpha(X) \subseteq D_\beta(X)$, and star-shaped; due to **D4con** $D_\alpha(X)$ is in addition convex;
  - Due to **D5** $D_\alpha(X)$ is closed.

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:

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

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:

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

Properties:

Depth:	**Regions:**
▸ Affine invariant;	Affine equivariant;

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:

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

Properties:

Depth:	**Regions:**
▶ Affine invariant;	Affine equivariant;
▶ Vanishing at infinity;	Bounded;

Tukey-trimmed regions

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

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

Properties:

Depth:	**Regions:**
▸ Affine invariant;	Affine equivariant;
▸ Vanishing at infinity;	Bounded;
▸ Monotone w.r.t. deepest point;	Nested;

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:

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

Properties:

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

Tukey-trimmed regions

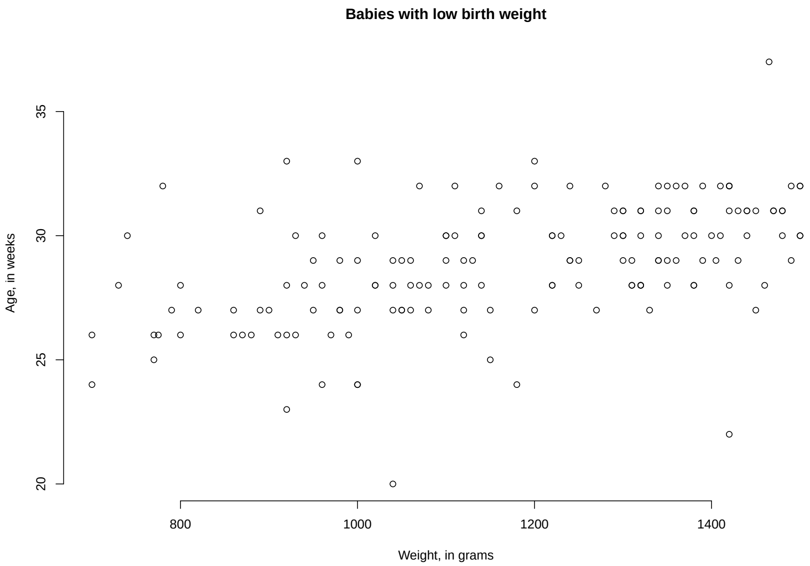
Tukey depth defines a family of (depth-)trimmed (central) regions $D_\tau^T(X)$, the upper-level sets of the depth function:

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

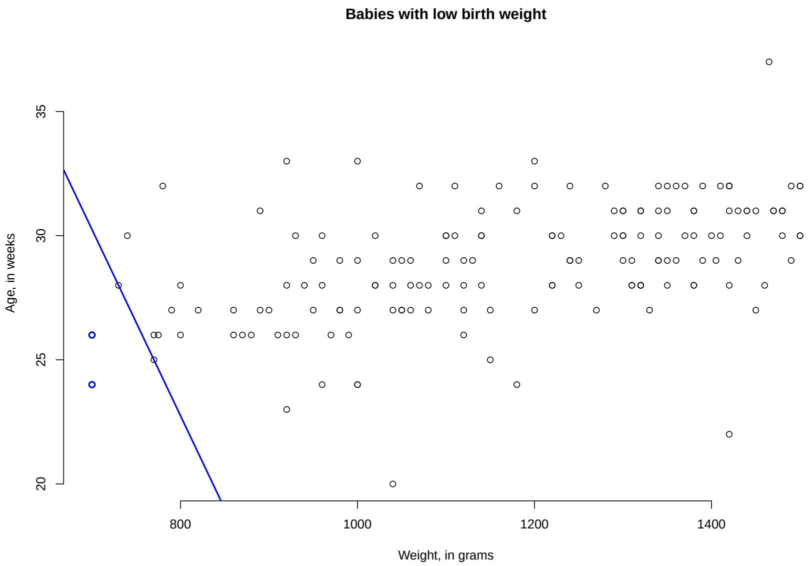
Properties:

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

Tukey (=halfspace, location) depth-trimmed regions

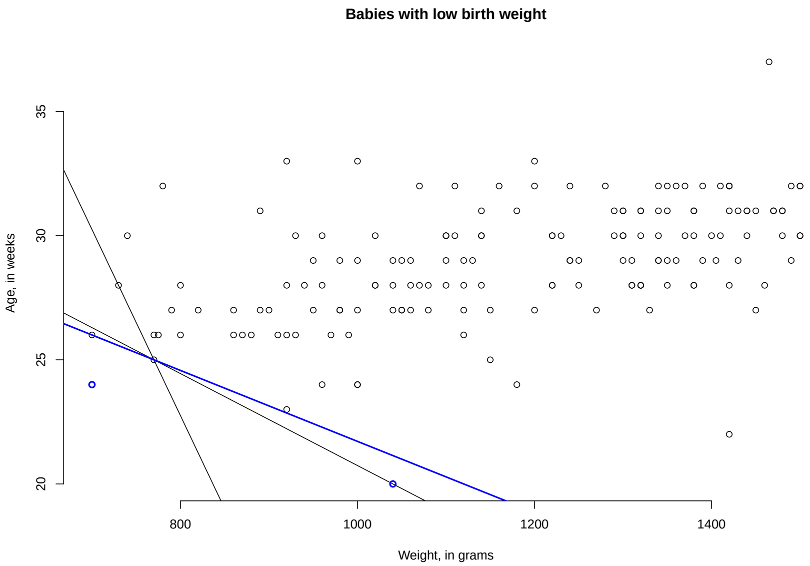


Tukey (=halfspace, location) depth-trimmed regions

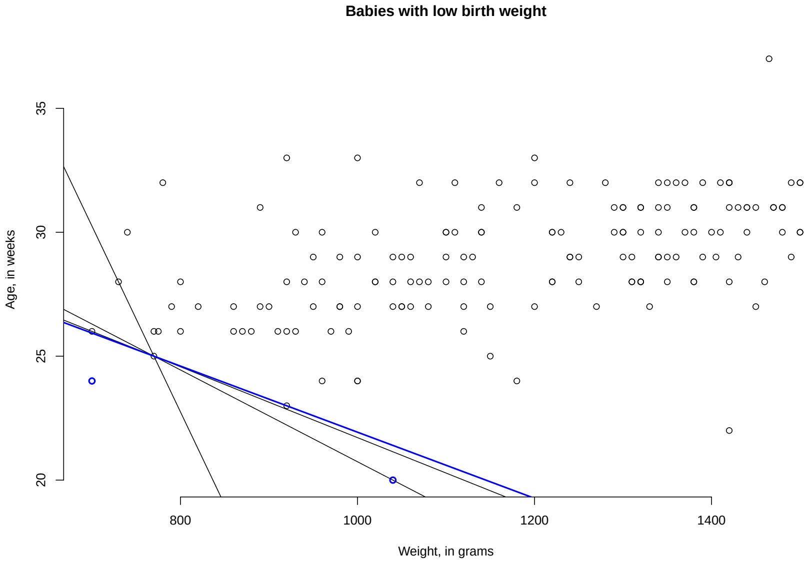


Tukey (=halfspace, location) depth-trimmed regions

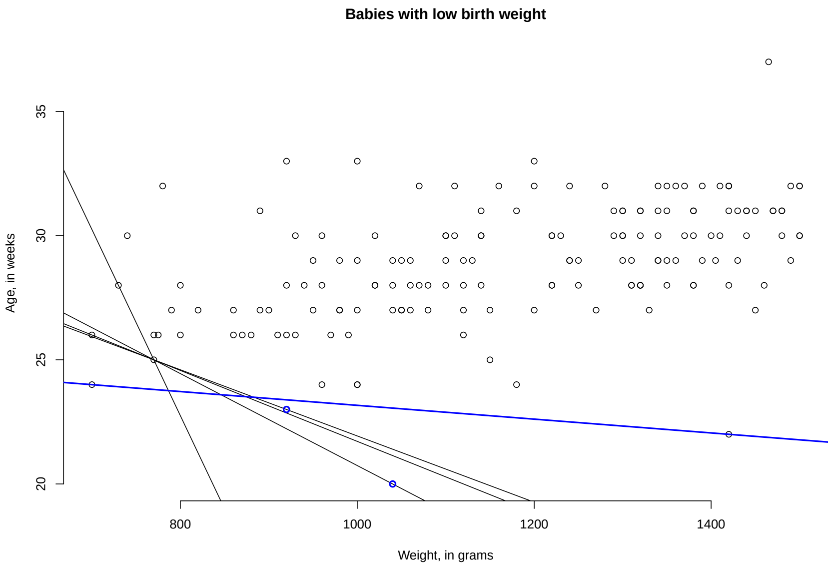
Tukey (=halfspace, location) depth-trimmed regions



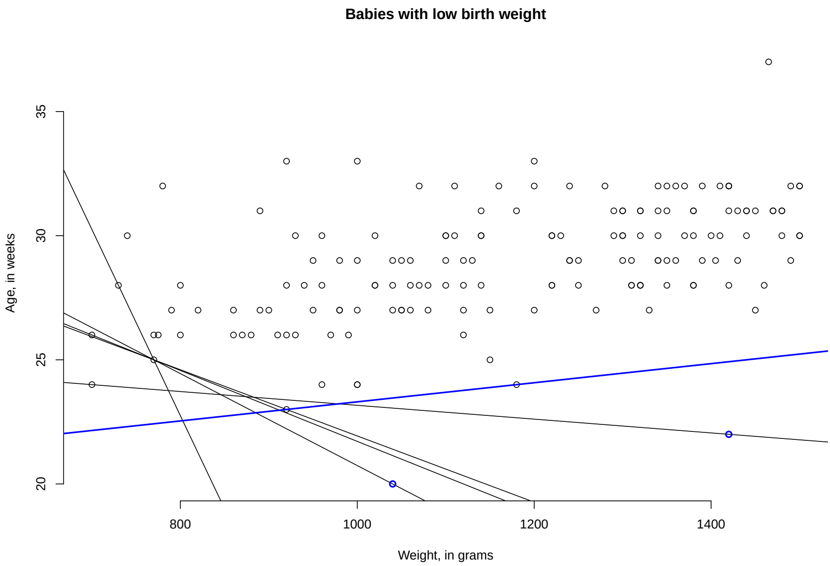
Tukey (=halfspace, location) depth-trimmed regions



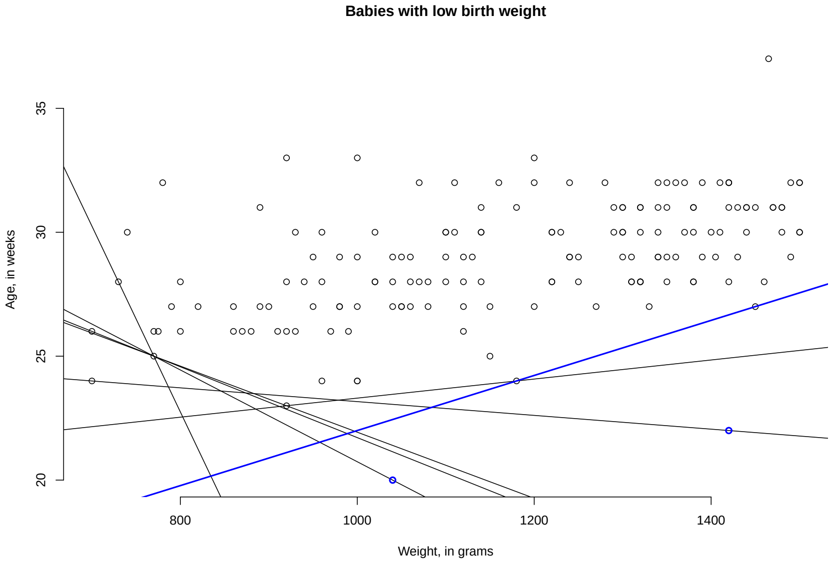
Tukey (=halfspace, location) depth-trimmed regions



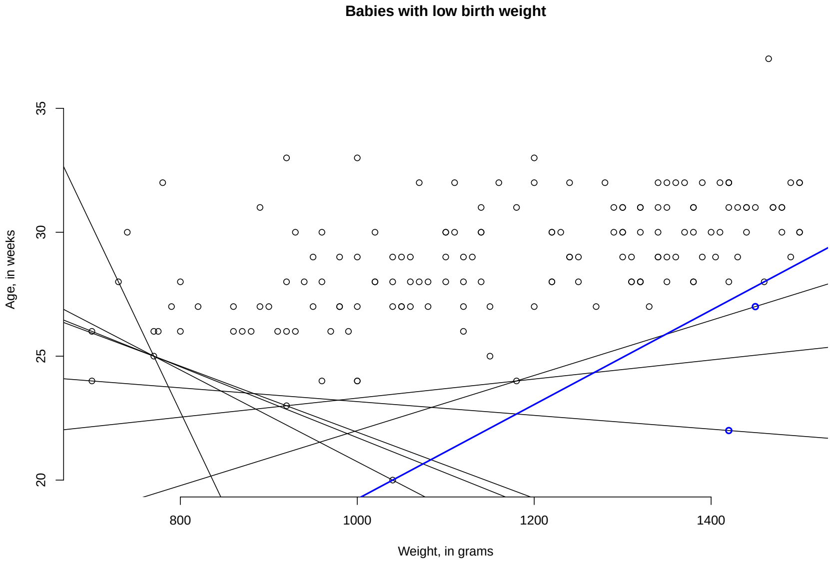
Tukey (=halfspace, location) depth-trimmed regions



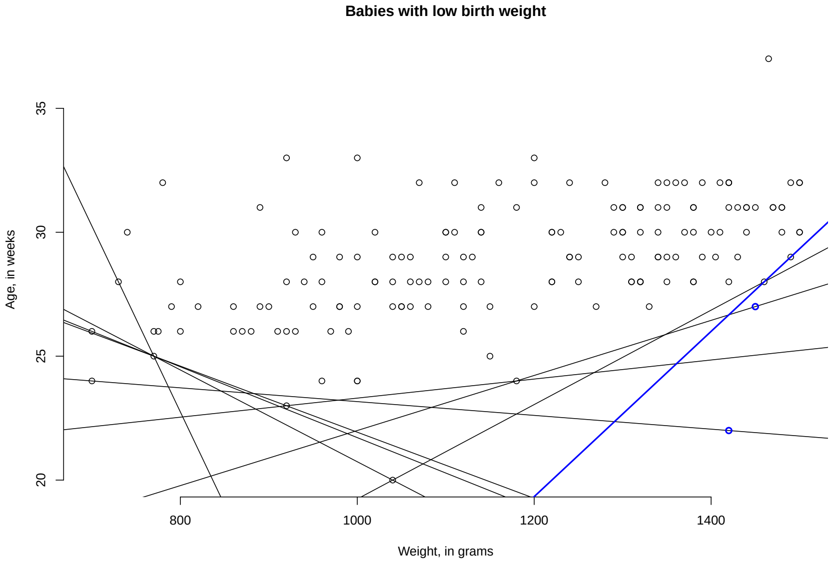
Tukey (=halfspace, location) depth-trimmed regions



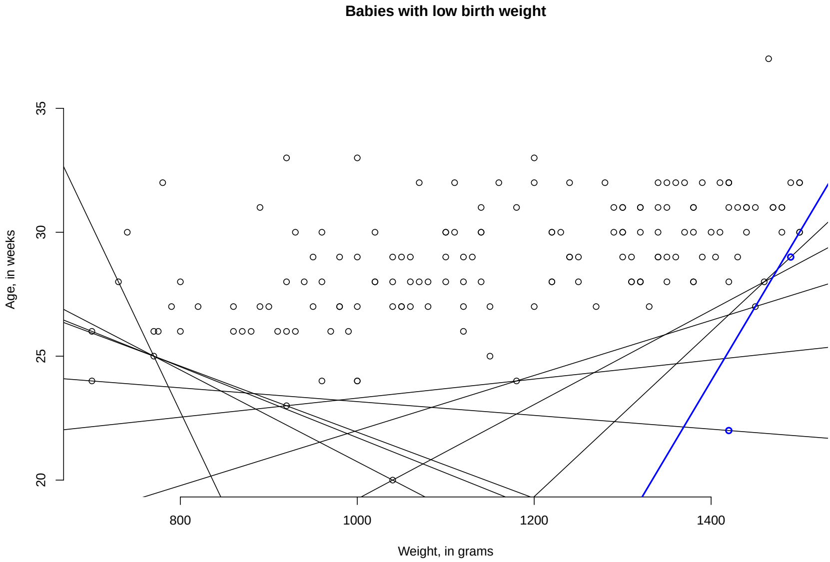
Tukey (=halfspace, location) depth-trimmed regions



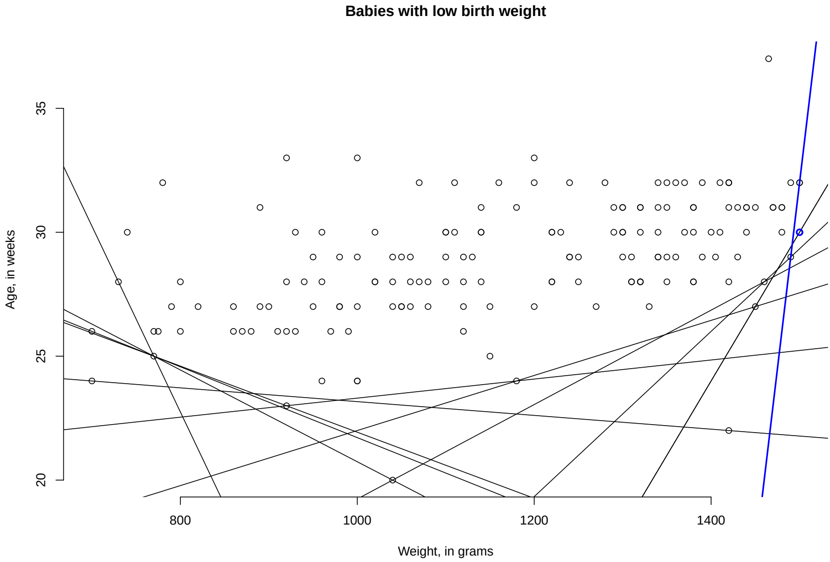
Tukey (=halfspace, location) depth-trimmed regions



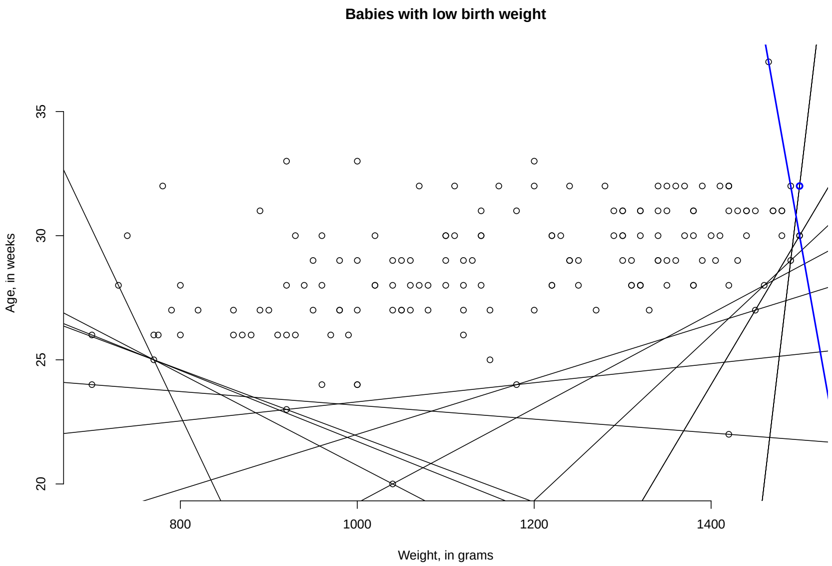
Tukey (=halfspace, location) depth-trimmed regions



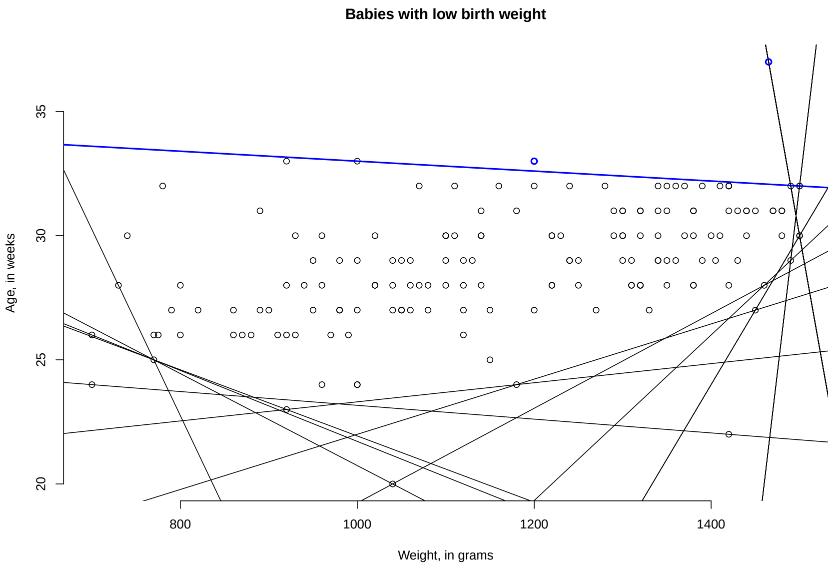
Tukey (=halfspace, location) depth-trimmed regions



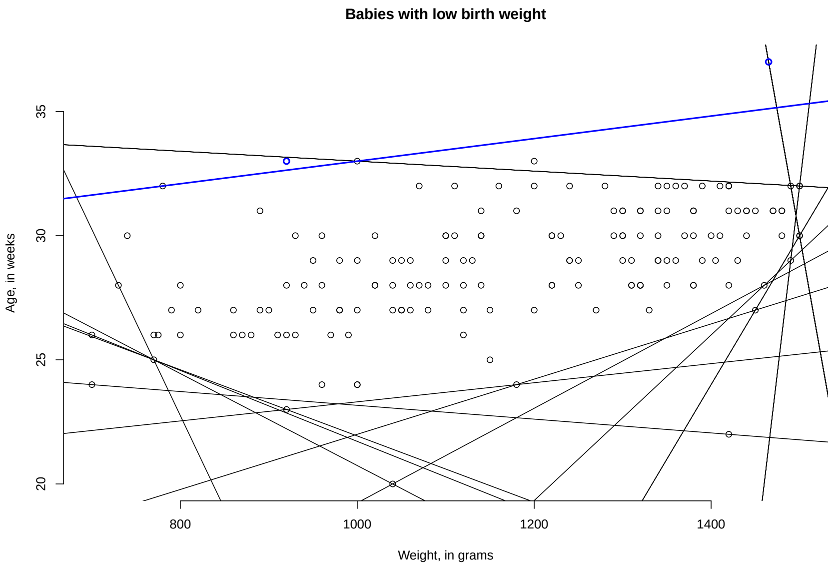
Tukey (=halfspace, location) depth-trimmed regions



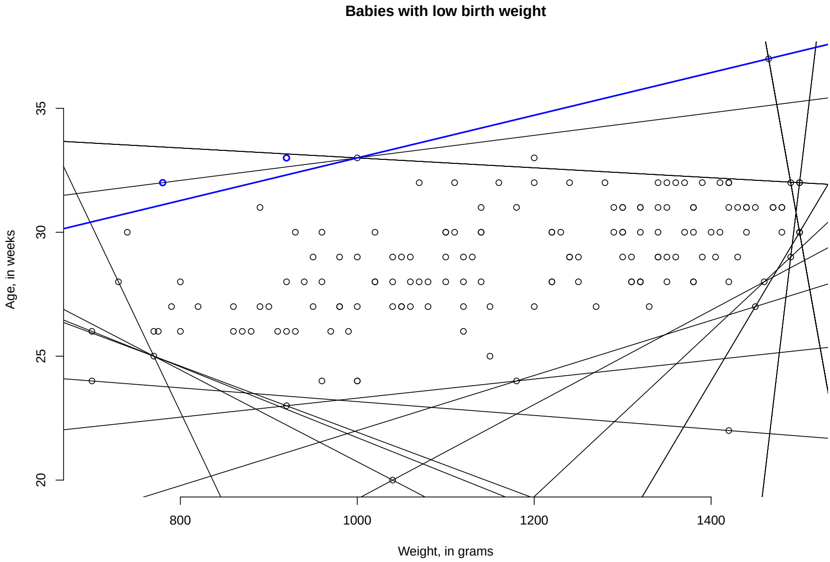
Tukey (=halfspace, location) depth-trimmed regions



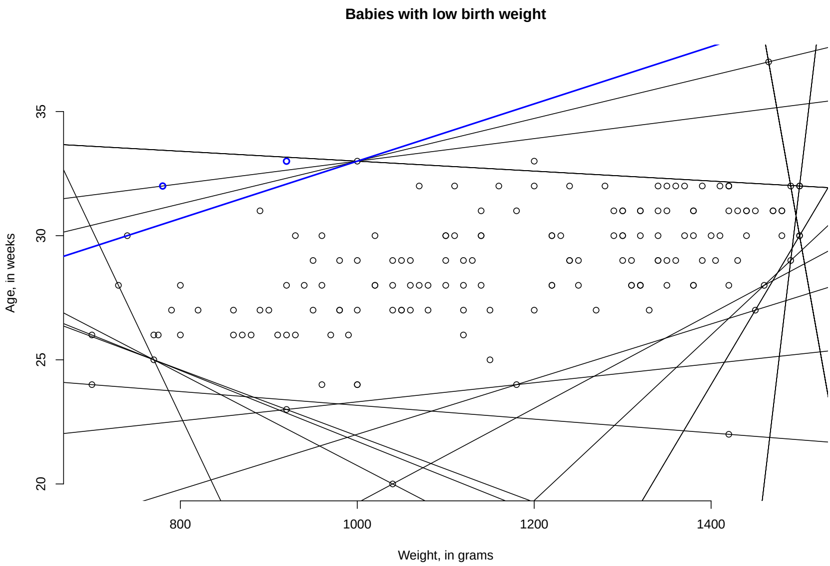
Tukey (=halfspace, location) depth-trimmed regions



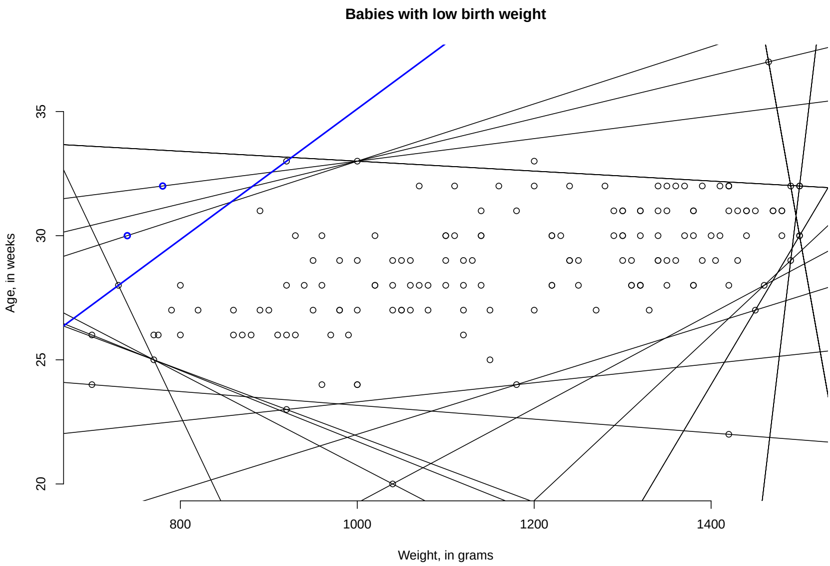
Tukey (=halfspace, location) depth-trimmed regions



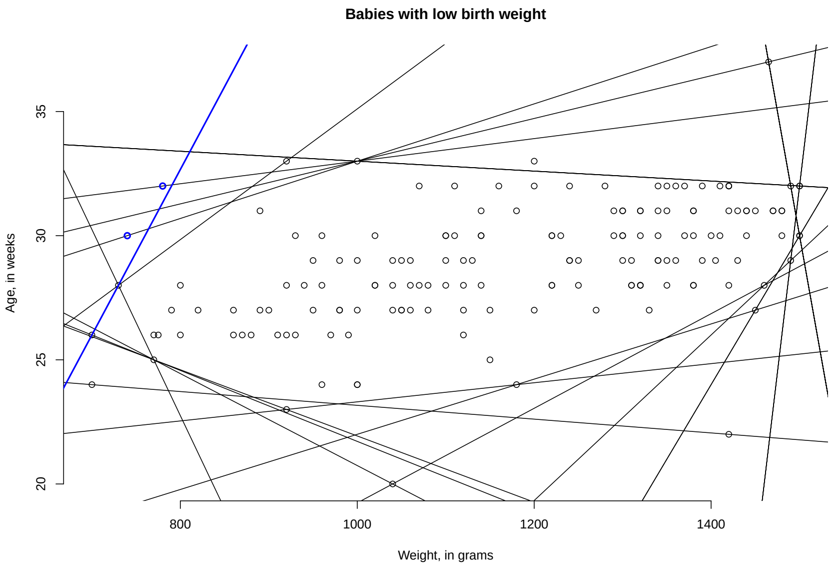
Tukey (=halfspace, location) depth-trimmed regions



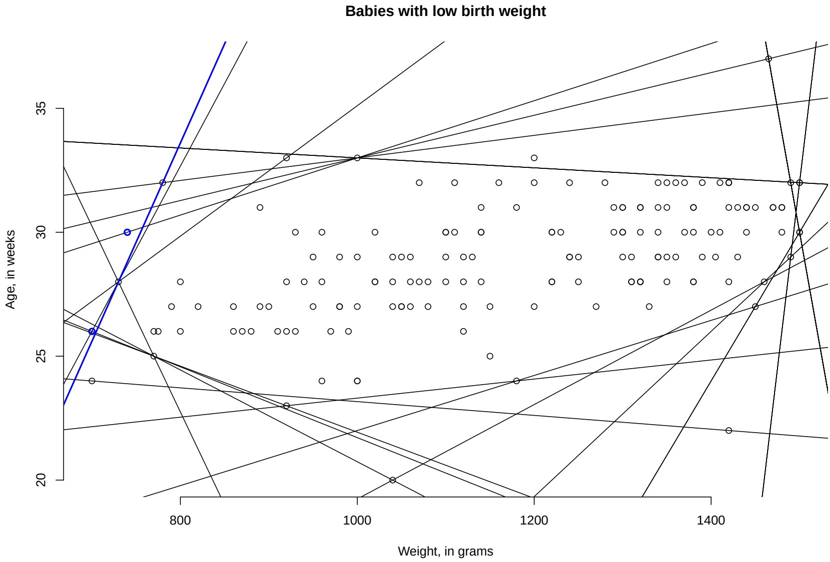
Tukey (=halfspace, location) depth-trimmed regions



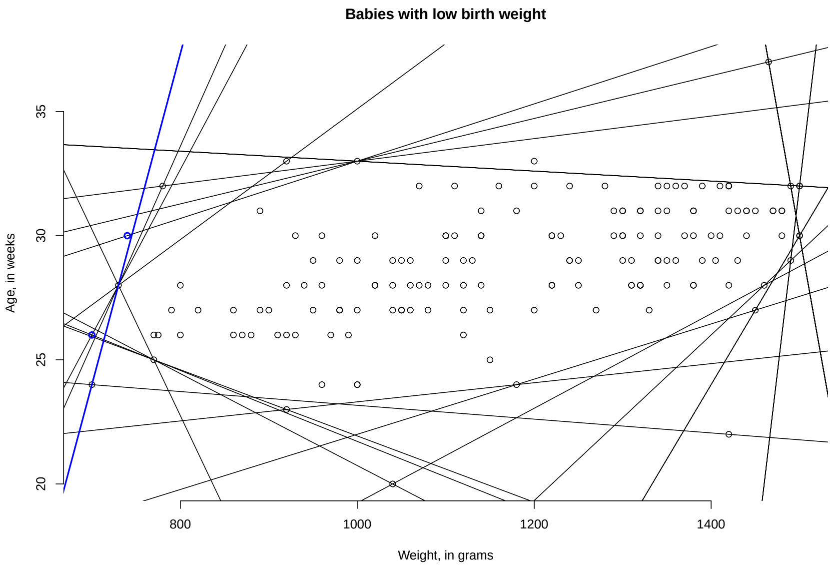
Tukey (=halfspace, location) depth-trimmed regions



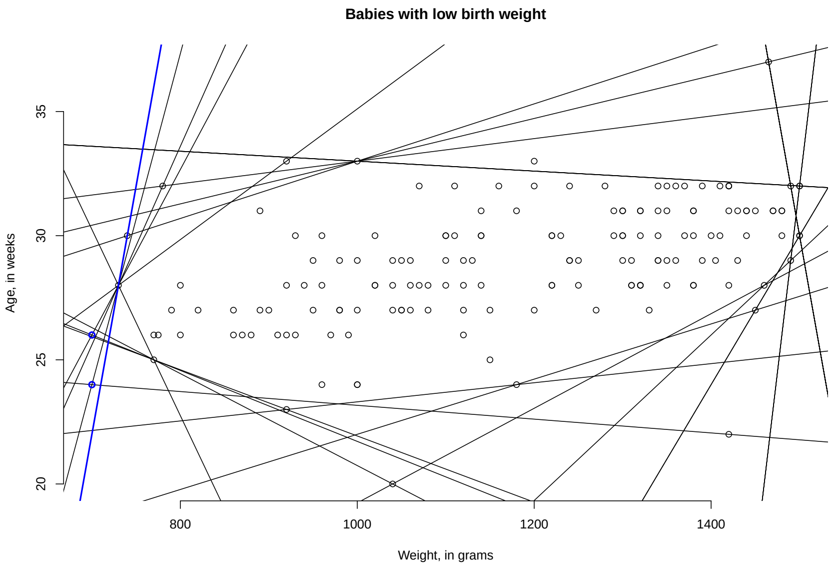
Tukey (=halfspace, location) depth-trimmed regions



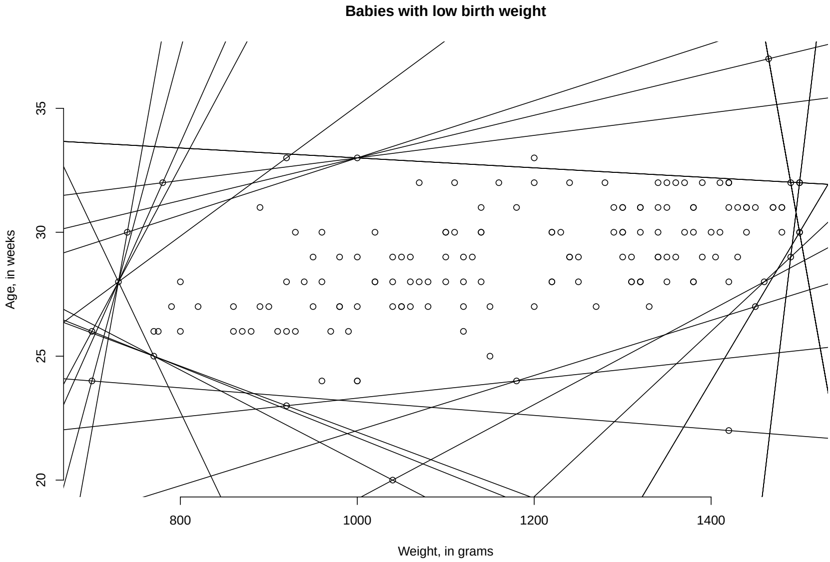
Tukey (=halfspace, location) depth-trimmed regions



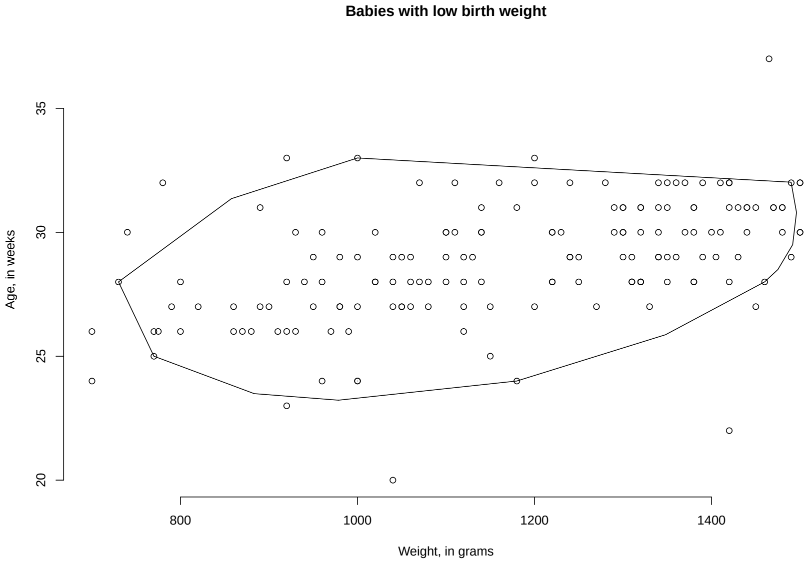
Tukey (=halfspace, location) depth-trimmed regions



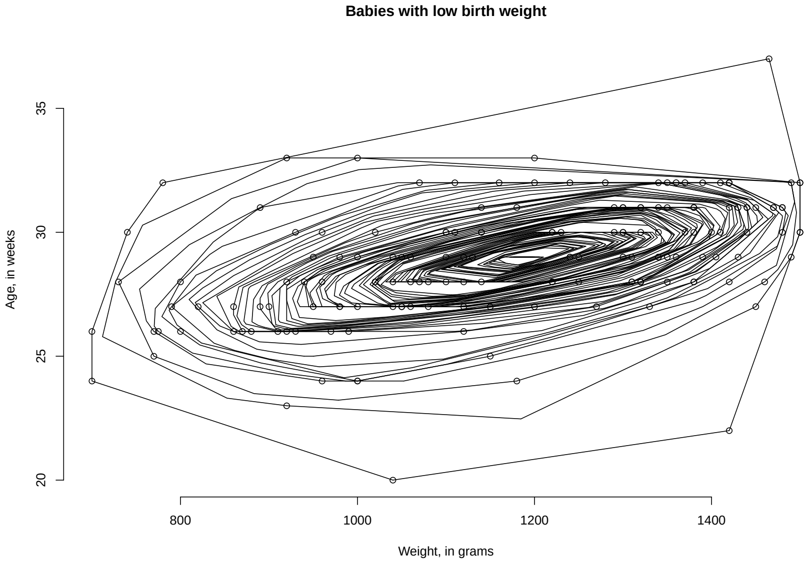
Tukey (=halfspace, location) depth-trimmed regions



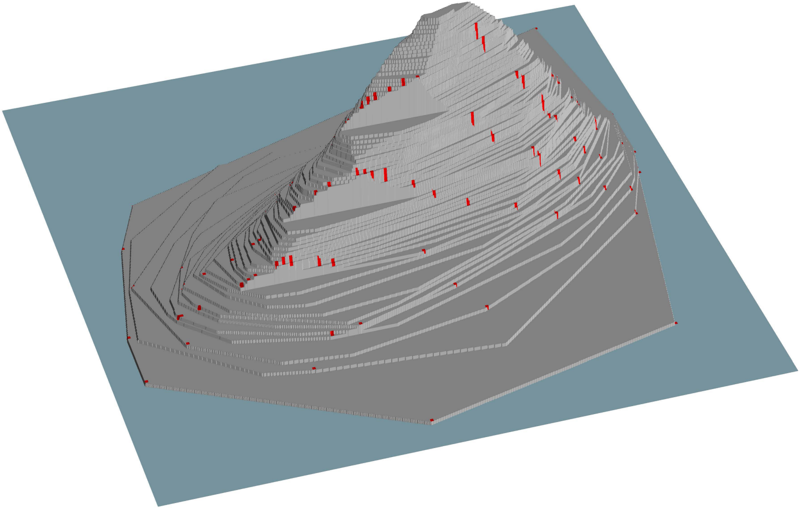
Tukey (=halfspace, location) depth-trimmed regions



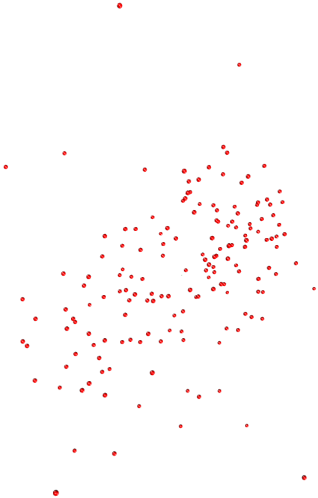
Tukey (=halfspace, location) depth-trimmed regions



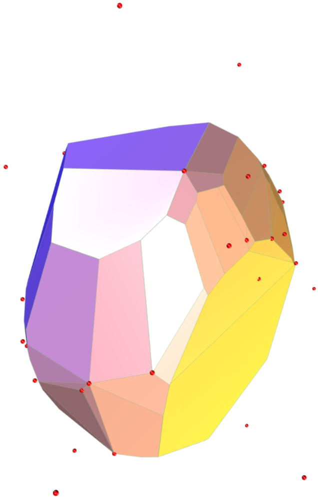
Tukey (=halfspace, location) data depth



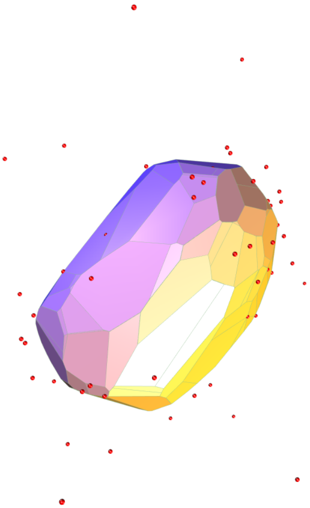
Tukey (=halfspace, location) depth region



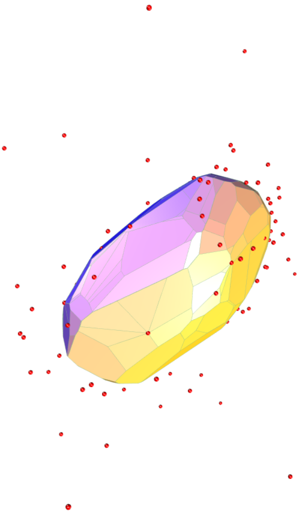
Tukey (=halfspace, location) depth region: $\tau = 2/161$



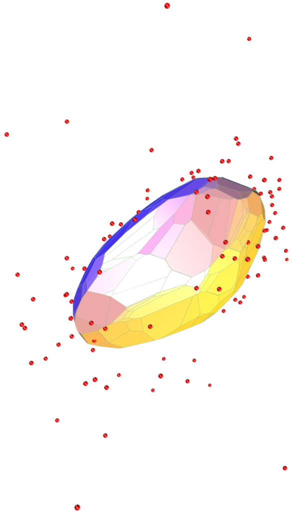
Tukey (=halfspace, location) depth region: $\tau = 5/161$



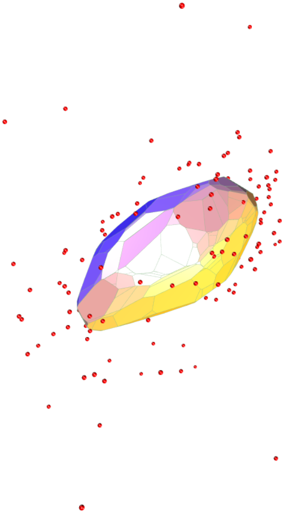
Tukey (=halfspace, location) depth region: $\tau = 9/161$



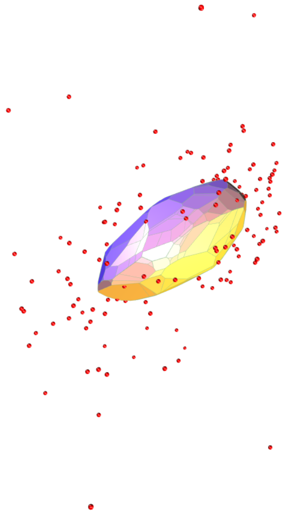
Tukey (=halfspace, location) depth region: τ =13/161



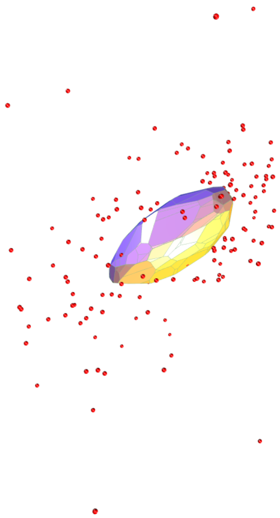
Tukey (=halfspace, location) depth region: $\tau = 17/161$



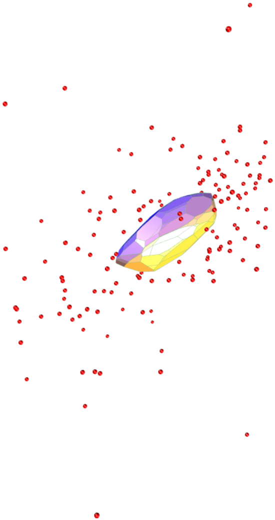
Tukey (=halfspace, location) depth region: $\tau = 25/161$



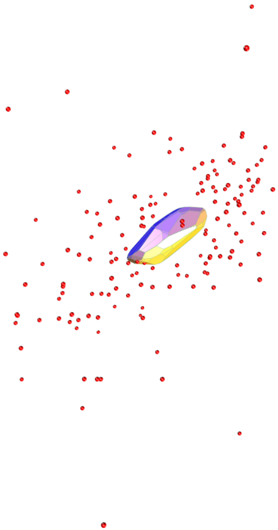
Tukey (=halfspace, location) depth region: $\tau = 33/161$



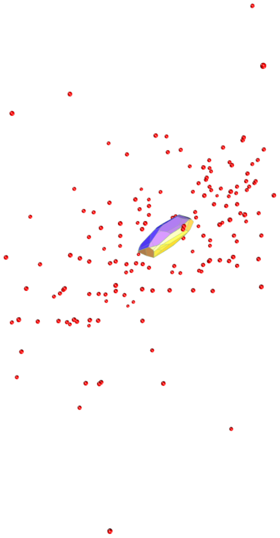
Tukey (=halfspace, location) depth region: $\tau = 41/161$



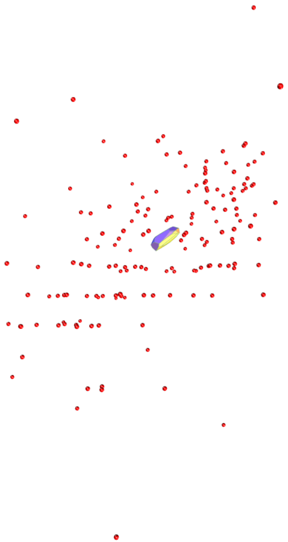
Tukey (=halfspace, location) depth region: $\tau = 49/161$



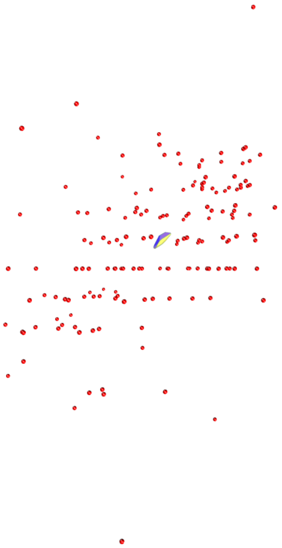
Tukey (=halfspace, location) depth region: $\tau = 57/161$



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



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



Contents

Mahalanobis depth (Mahalanobis, 1936)

- **Mahalanobis depth** is defined as:

$$D^{Mah}(\boldsymbol{x}|X) = \frac{1}{1 + (\delta^{Mah})^2(\boldsymbol{x}|X)},$$

based on Mahalanobis distance:

$$(\delta^{Mah})^2(\boldsymbol{x}|X) = (\boldsymbol{x} - \boldsymbol{\mu}_X)^T \boldsymbol{\Sigma}_X^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_X).$$

Mahalanobis depth (Mahalanobis, 1936)

- **Mahalanobis depth** is defined as:

$$D^{Mah}(\boldsymbol{x}|X) = \frac{1}{1 + (\delta^{Mah})^2(\boldsymbol{x}|X)},$$

based on Mahalanobis distance:

$$(\delta^{Mah})^2(\boldsymbol{x}|X) = (\boldsymbol{x} - \boldsymbol{\mu}_X)^T \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_X).$$

- In the empirical version, $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ are substituted by suitable estimates:
  - moment estimates;
  - robust estimates such as **minimum volume ellipsoid** or **minimum covariance determinant** (MCD).

Mahalanobis depth (Mahalanobis, 1936)

- **Mahalanobis depth** is defined as:

$$D^{Mah}(\boldsymbol{x}|X) = \frac{1}{1 + (\delta^{Mah})^2(\boldsymbol{x}|X)},$$

based on Mahalanobis distance:

$$(\delta^{Mah})^2(\boldsymbol{x}|X) = (\boldsymbol{x} - \boldsymbol{\mu}_X)^T \boldsymbol{\Sigma}_X^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_X).$$

- In the empirical version, $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ are substituted by suitable estimates:
  - moment estimates;
  - robust estimates such as **minimum volume ellipsoid** or **minimum covariance determinant** (MCD).
- Properties:
  - satisfies **D1 – D5** and **D4con**, is continuous;

Mahalanobis depth (Mahalanobis, 1936)

- **Mahalanobis depth** is defined as:

$$D^{Mah}(\boldsymbol{x}|X) = \frac{1}{1 + (\delta^{Mah})^2(\boldsymbol{x}|X)},$$

based on Mahalanobis distance:

$$(\delta^{Mah})^2(\boldsymbol{x}|X) = (\boldsymbol{x} - \boldsymbol{\mu}_X)^T \boldsymbol{\Sigma}_X^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_X).$$

- In the empirical version, $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ are substituted by suitable estimates:
  - moment estimates;
  - robust estimates such as **minimum volume ellipsoid** or **minimum covariance determinant** (MCD).
- Properties:
  - satisfies **D1 – D5** and **D4con**, is continuous;
  - being defined by $d(d+1)$ parameters, can be seen as a **parametric depth**;

Mahalanobis depth (Mahalanobis, 1936)

- **Mahalanobis depth** is defined as:

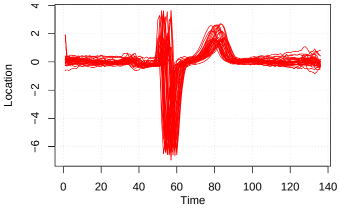
$$D^{Mah}(\boldsymbol{x}|X) = \frac{1}{1 + (\delta^{Mah})^2(\boldsymbol{x}|X)} \,,$$

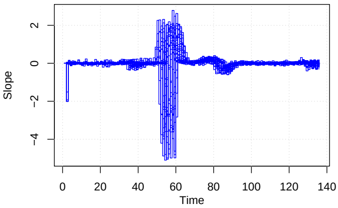
based on Mahalanobis distance:

$$(\delta^{Mah})^2(\boldsymbol{x}|X) = (\boldsymbol{x} - \boldsymbol{\mu}_X)^T \boldsymbol{\Sigma}_X^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_X)\,.$$

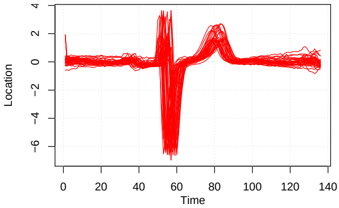
- In the empirical version, $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ are substituted by suitable estimates:
  - moment estimates;
  - robust estimates such as **minimum volume ellipsoid** or **minimum covariance determinant** (MCD).
- Properties:
  - satisfies **D1 – D5** and **D4con**, is continuous;
  - being defined by $d(d+1)$ parameters, can be seen as a **parametric depth**;
  - by a single elliptical contour characterizes a multivariate **normal distribution** or one within an affine **family of** non-degenerate **elliptical distributions**.

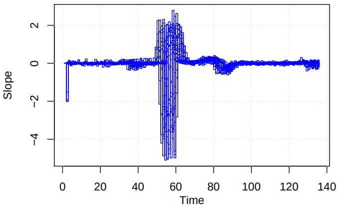
ECG five days data





ECG five days data

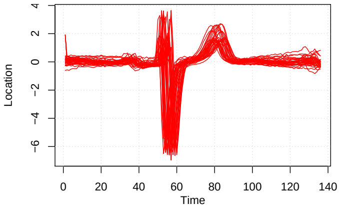


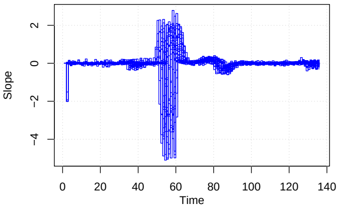


$$\hat{f}_i \mapsto \boldsymbol{x}_i = \left[\int_0^T \hat{f}_i(t)dt, \int_0^T \hat{f}_i'(t)dt \right],$$

with $\hat{f}_i(t)$ being the function obtained by connecting the points $(t_{ij},\ f_i(t_{ij})), j = 1, \ldots, N_i$ with line segments, $\hat{f}_i'(t)$ its derivative.

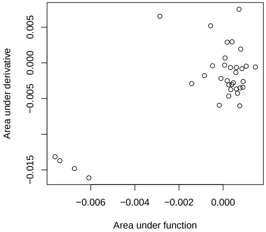
ECG five days data



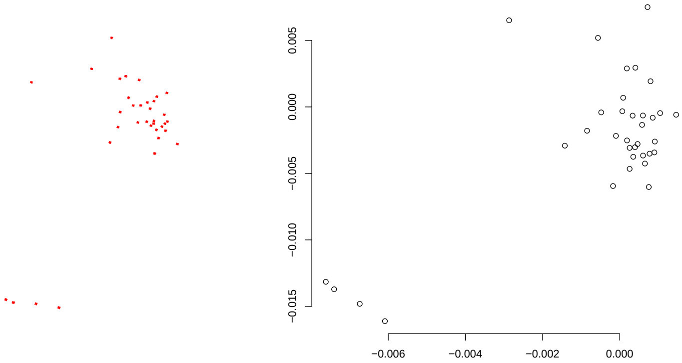


$$\hat{f}_i \mapsto \boldsymbol{x}_i = \left[\int_0^T \hat{f}_i(t)dt, \int_0^T \hat{f}_i'(t)dt \right],$$

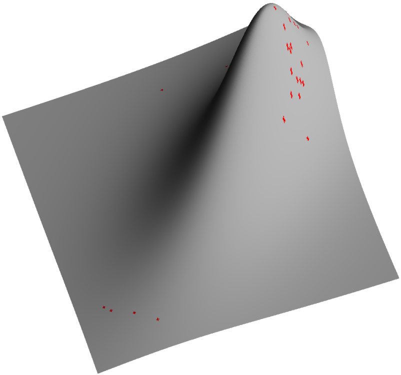
with $\hat{f}_i(t)$ being the function obtained by connecting the points $(t_{ij}, f_i(t_{ij})), j = 1, \ldots, N_i$ with line segments, $\hat{f}_i'(t)$ its derivative.

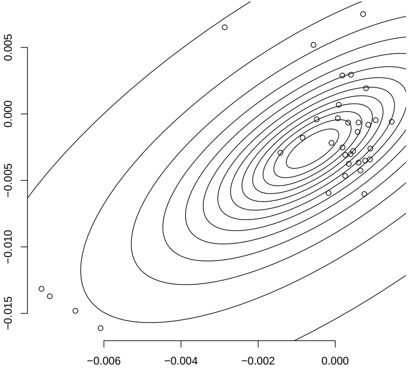


Mahalanobis depth (Mahalanobis, 1936)

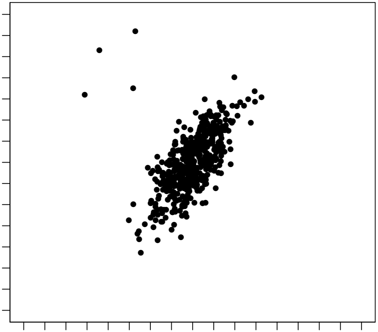


Mahalanobis depth (Mahalanobis, 1936)

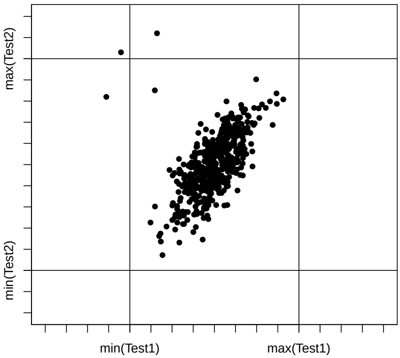




Multivariate anomaly detection: an example

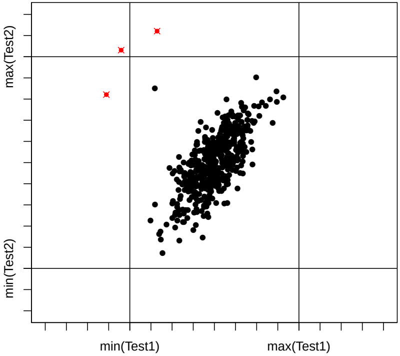


Multivariate anomaly detection: an example



- Checking for **minimum** and **maximum** in each test result.

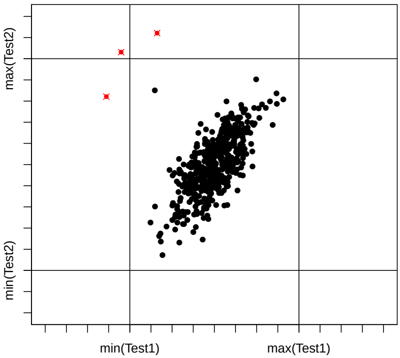
Multivariate anomaly detection: an example



- Checking for **minimum** and **maximum** in each test result.
- Label observation $\boldsymbol{x}$ as anomaly if:

$$\boldsymbol{x} \notin [\min(\text{Test1}),\max(\text{Test1})] \times [\min(\text{Test2}),\max(\text{Test2})] \,.$$

Multivariate anomaly detection: an example

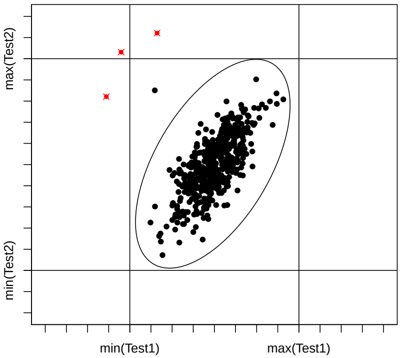


- Checking for **minimum** and **maximum** in each test result.
- Label observation $\boldsymbol{x}$ as anomaly if:

$$\boldsymbol{x} \notin [\min(\text{Test1}),\max(\text{Test1})] \times [\min(\text{Test2}),\max(\text{Test2})] \,.$$

- **Not all** anomalies can be detected.

Multivariate anomaly detection: an example

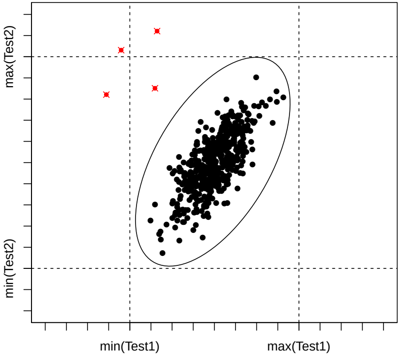


- **Mahalanobis distance** of an observation $\boldsymbol{x} \in \mathbb{R}^2$ (from the mean) is defined as follows:

$$d_{Mah}(\boldsymbol{x}|\boldsymbol{X}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu}$ is the **mean** and $\boldsymbol{\Sigma}$ is the **covariance** matrix.

Multivariate anomaly detection: an example



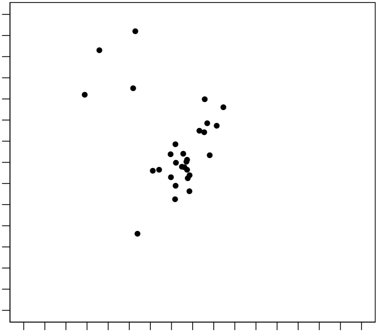
- **Mahalanobis distance** of an observation $\boldsymbol{x} \in \mathbb{R}^2$ (from the mean) is defined as follows:

$$d_{Mah}(\boldsymbol{x}|\boldsymbol{X}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}),$$

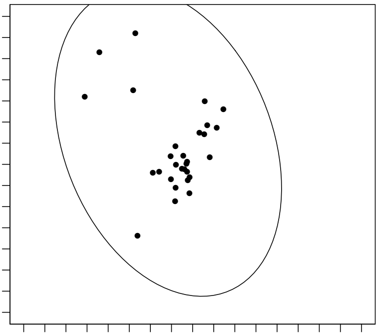
where $\boldsymbol{\mu}$ is the **mean** and $\boldsymbol{\Sigma}$ is the **covariance** matrix.

- Label $\boldsymbol{x}$ as anomaly $d_{Mah}(\boldsymbol{x}|\boldsymbol{X}) > \max(d_{Mah})$.

Multivariate anomaly detection: robustness

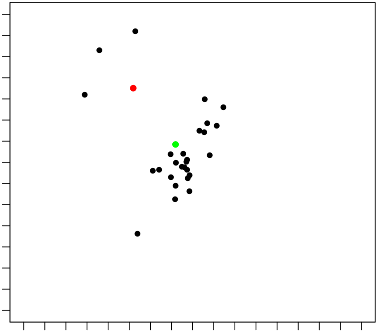


Multivariate anomaly detection: robustness



- Mahalanobis distance (moment estimators) **not robust**.

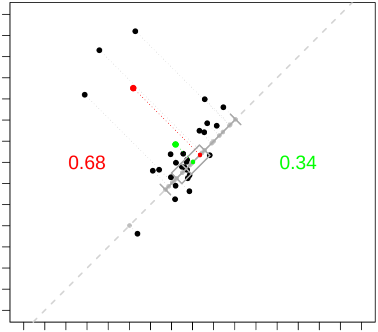
Multivariate anomaly detection: robustness



- Mahalanobis distance (moment estimators) **not robust**.
- **Stahel-Donoho outlyingness** of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top \boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})} .$$

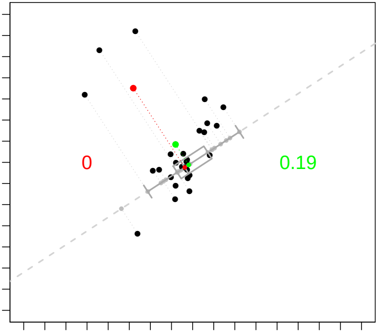
Multivariate anomaly detection: robustness



- Mahalanobis distance (moment estimators) **not robust**.
- **Stahel-Donoho outlyingness** of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top \boldsymbol{u} - \mathrm{med}(\boldsymbol{X}\boldsymbol{u})|}{\mathrm{MAD}(\boldsymbol{X}\boldsymbol{u})}\,.$$

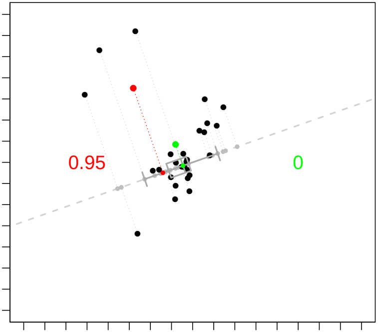
Multivariate anomaly detection: robustness



- Mahalanobis distance (moment estimators) **not robust**.
- **Stahel-Donoho outlyingness** of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top \boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})}.$$

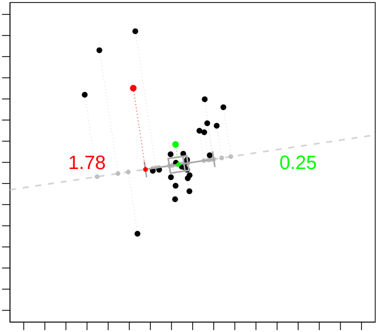
Multivariate anomaly detection: robustness



- Mahalanobis distance (moment estimators) **not robust**.
- **Stahel-Donoho outlyingness** of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u} \in \mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top \boldsymbol{u} - \mathrm{med}(\boldsymbol{X}\boldsymbol{u})|}{\mathrm{MAD}(\boldsymbol{X}\boldsymbol{u})} .$$

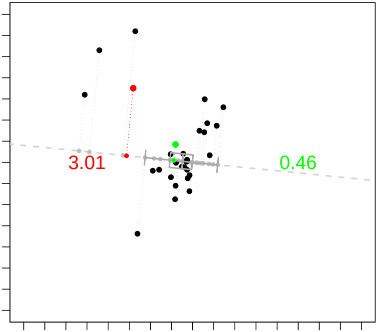
Multivariate anomaly detection: robustness



- Mahalanobis distance (moment estimators) **not robust**.
- **Stahel-Donoho outlyingness** of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top \boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})}\,.$$

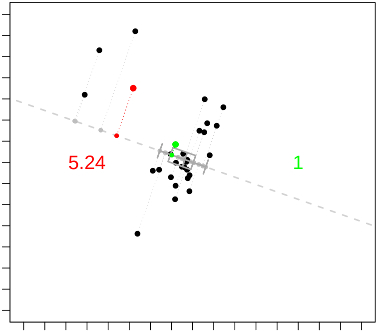
Multivariate anomaly detection: robustness



- Mahalanobis distance (moment estimators) **not robust**.
- **Stahel-Donoho outlyingness** of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top\boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})} .$$

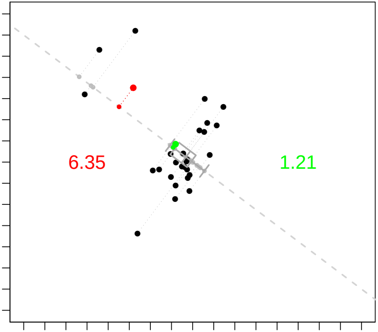
Multivariate anomaly detection: robustness



- Mahalanobis distance (moment estimators) **not robust**.
- **Stahel-Donoho outlyingness** of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top \boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})} .$$

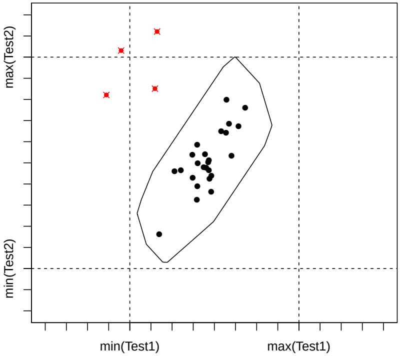
Multivariate anomaly detection: robustness



- Mahalanobis distance (moment estimators) **not robust**.
- **Stahel-Donoho outlyingness** of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top\boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})} .$$

Multivariate anomaly detection: robustness

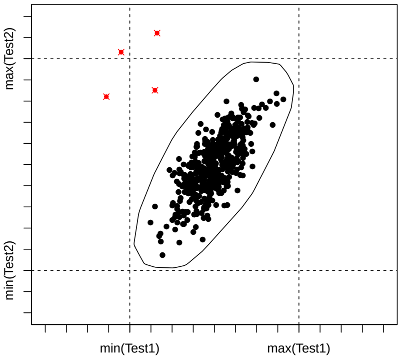


- Mahalanobis distance (moment estimators) **not robust**.
- **Stahel-Donoho outlyingness** of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u}\in\mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top \boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})} \,.$$

- Label $\boldsymbol{x}$ as anomaly if $O_{SD}(\boldsymbol{x}|\boldsymbol{X}) > \max(O_{SD})$.

Multivariate anomaly detection: robustness



- Mahalanobis distance (moment estimators) **not robust**.
- **Stahel-Donoho outlyingness** of $\boldsymbol{x}$ w.r.t. $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$:

$$O_{SD}(\boldsymbol{x}|\boldsymbol{X}) = \max_{\boldsymbol{u} \in \mathcal{S}^{d-1}} \frac{|\boldsymbol{x}^\top \boldsymbol{u} - \text{med}(\boldsymbol{X}\boldsymbol{u})|}{\text{MAD}(\boldsymbol{X}\boldsymbol{u})} .$$

- Label $\boldsymbol{x}$ as anomaly if $O_{SD}(\boldsymbol{x}|\boldsymbol{X}) > \max(O_{SD})$.

Projection depth (Zuo & Serfling, 2000)

According to Zuo & Serfling (2000), **projection depth** is defined as:

$$D^{prj}(\boldsymbol{x}|X) = \frac{1}{1 + O_{SD}(\boldsymbol{x}|X)},$$

Projection depth (Zuo & Serfling, 2000)

According to Zuo & Serfling (2000), **projection depth** is defined as:

$$D^{prj}(\boldsymbol{x}|X) = \frac{1}{1 + O_{SD}(\boldsymbol{x}|X)},$$

where

$$O_{SD}(\boldsymbol{x}|X) = \sup_{\boldsymbol{r} \in S^{d-1}} \frac{|X^T\boldsymbol{r} - \mathrm{med}(X^T\boldsymbol{r})|}{\mathrm{MAD}(X^T\boldsymbol{r})}$$

is the **projected outlyingness** (Stahel, 1981; Donoho, 1982), $\mathrm{med}(Y)$ and $\mathrm{MAD}(Y) = \mathrm{med}\left(\left|Y - \mathrm{med}(Y)\right|\right)$ are the univariate median and median absolute deviation from the median, respectively.

Projection depth (Zuo & Serfling, 2000)

According to Zuo & Serfling (2000), **projection depth** is defined as:

$$D^{prj}(\boldsymbol{x}|X) = \frac{1}{1 + O_{SD}(\boldsymbol{x}|X)},$$

where

$$O_{SD}(\boldsymbol{x}|X) = \sup_{\boldsymbol{r} \in S^{d-1}} \frac{|X^T\boldsymbol{r} - \text{med}(X^T\boldsymbol{r})|}{\text{MAD}(X^T\boldsymbol{r})}$$

is the **projected outlyingness** (Stahel, 1981; Donoho, 1982), $\text{med}(Y)$ and $\text{MAD}(Y) = \text{med}\big(\big|Y - \text{med}(Y)\big|\big)$ are the univariate median and median absolute deviation from the median, respectively.

Properties:

- Satisfies **D1** – **D5** and **D4con**, is continuous;

Projection depth (Zuo & Serfling, 2000)

According to Zuo & Serfling (2000), **projection depth** is defined as:

$$D^{prj}(\boldsymbol{x}|X) = \frac{1}{1 + O_{SD}(\boldsymbol{x}|X)},$$

where

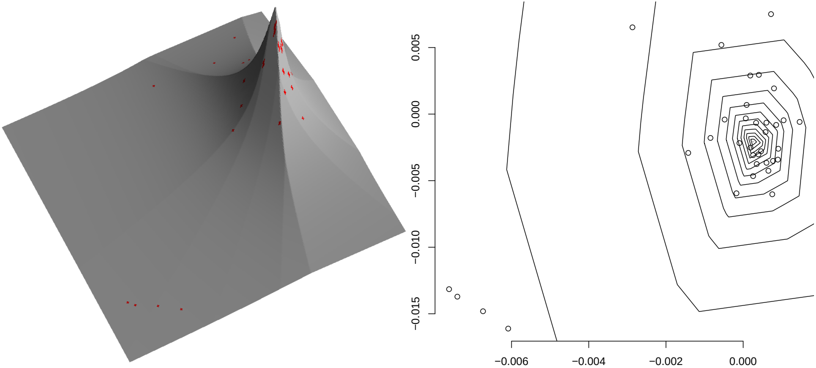
$$O_{SD}(\boldsymbol{x}|X) = \sup_{\boldsymbol{r} \in S^{d-1}} \frac{|X^T\boldsymbol{r} - \text{med}(X^T\boldsymbol{r})|}{\text{MAD}(X^T\boldsymbol{r})}$$

is the **projected outlyingness** (Stahel, 1981; Donoho, 1982), $\text{med}(Y)$ and $\text{MAD}(Y) = \text{med}\left(\left|Y - \text{med}(Y)\right|\right)$ are the univariate median and median absolute deviation from the median, respectively.

Properties:

- Satisfies **D1 – D5** and **D4con**, is continuous;
- its **median** has asymptotic breakdown point of 0.5.

Projection depth (Zuo & Serfling, 2000)



Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

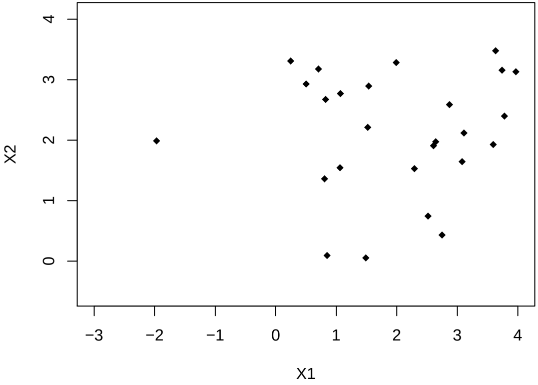
Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[\frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|} \right] \right\| \quad \text{with} \quad \frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|} = 0 \quad \text{if} \quad \boldsymbol{x} - X = \mathbf{0}\,.$$

Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

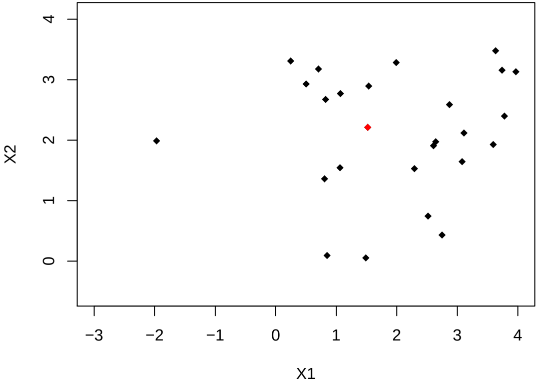
$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[\frac{\boldsymbol{x}-X}{\|\boldsymbol{x}-X\|}\right] \right\| \quad \text{with} \quad \frac{\boldsymbol{x}-X}{\|\boldsymbol{x}-X\|} = 0 \quad \text{if} \quad \boldsymbol{x}-X = \mathbf{0}\,.$$



Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

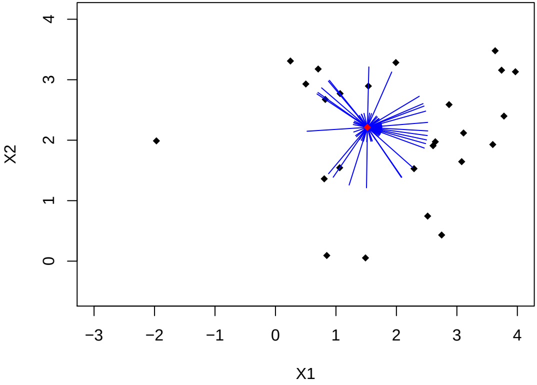
$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[\frac{\boldsymbol{x}-X}{\|\boldsymbol{x}-X\|}\right]\right\| \quad \text{with} \quad \frac{\boldsymbol{x}-X}{\|\boldsymbol{x}-X\|} = 0 \quad \text{if} \quad \boldsymbol{x}-X = \boldsymbol{0}\,.$$



Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

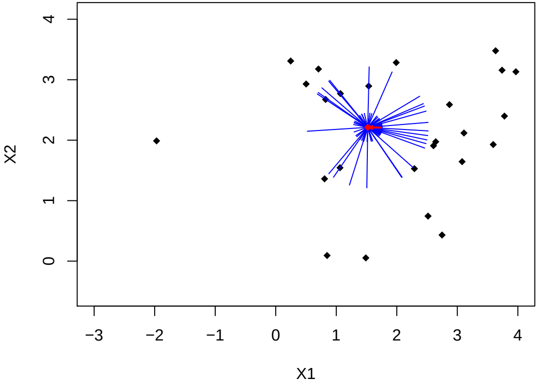
$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[\frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|} \right] \right\| \quad \text{with} \quad \frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|} = 0 \quad \text{if} \quad \boldsymbol{x} - X = \boldsymbol{0}\,.$$



Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

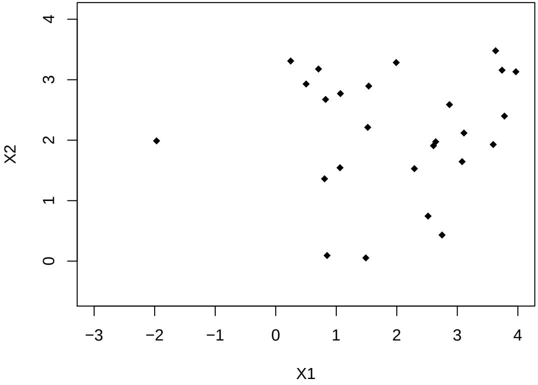
$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[\frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|}\right] \right\| \quad \text{with} \quad \frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|} = 0 \quad \text{if} \quad \boldsymbol{x} - X = \mathbf{0}\,.$$



Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

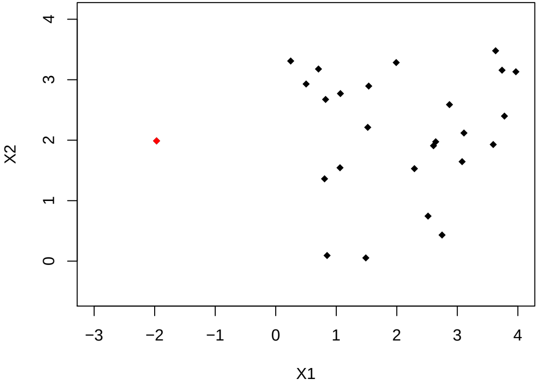
$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[\frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|} \right] \right\| \quad \text{with} \quad \frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|} = 0 \quad \text{if} \quad \boldsymbol{x} - X = \mathbf{0}\,.$$



Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

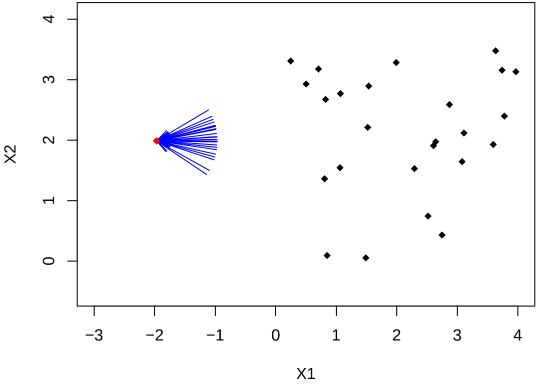
$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[\frac{\boldsymbol{x}-X}{\|\boldsymbol{x}-X\|}\right]\right\| \quad \text{with} \quad \frac{\boldsymbol{x}-X}{\|\boldsymbol{x}-X\|} = 0 \quad \text{if} \quad \boldsymbol{x}-X = \mathbf{0}\,.$$



Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

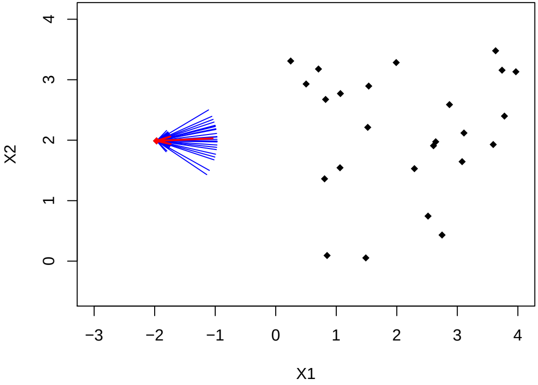
$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[\frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|}\right] \right\| \quad \text{with} \quad \frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|} = 0 \quad \text{if} \quad \boldsymbol{x} - X = \mathbf{0}\,.$$



Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[\frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|} \right] \right\| \quad \text{with} \quad \frac{\boldsymbol{x} - X}{\|\boldsymbol{x} - X\|} = 0 \quad \text{if} \quad \boldsymbol{x} - X = \boldsymbol{0}\,.$$



Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[v\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - X) \right) \right] \right\| ,$$

with

$$v(\boldsymbol{y}) = \begin{cases} \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} & \text{if } \boldsymbol{y} \neq \boldsymbol{0} , \\ \boldsymbol{0} & \text{if } \boldsymbol{y} = \boldsymbol{0} . \end{cases}$$

Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[v\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - X) \right) \right] \right\| ,$$

with

$$v(\boldsymbol{y}) = \begin{cases} \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} & \text{if } \boldsymbol{y} \neq \boldsymbol{0} , \\ \boldsymbol{0} & \text{if } \boldsymbol{y} = \boldsymbol{0} . \end{cases}$$

Properties:

- satisfies **D1 – D5**, but not **D4con**, is continuous;

Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[v\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - X)\right)\right] \right\| ,$$

with

$$v(\boldsymbol{y}) = \begin{cases} \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} & \text{if } \boldsymbol{y} \neq \boldsymbol{0}\,, \\ \boldsymbol{0} & \text{if } \boldsymbol{y} = \boldsymbol{0}\,. \end{cases}$$

Properties:

- satisfies **D1 – D5**, but not **D4con**, is continuous;
- if $\boldsymbol{\Sigma}$ is orthogonal, satisfies **D2iso** only;

Spatial depth (Vardi & Zhang, 2000; Serfling 2002)

Exploiting the idea of spatial quantiles of Chaudhuri (1996) and Koltchinskii (1997), Vardi & Zhang (2000) and Serflig (2002) formulate the **spatial depth** (also L_1-depth) as:

$$D^{spt}(\boldsymbol{x}|X) = 1 - \left\| \mathbb{E}\left[v\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - X) \right) \right] \right\|,$$

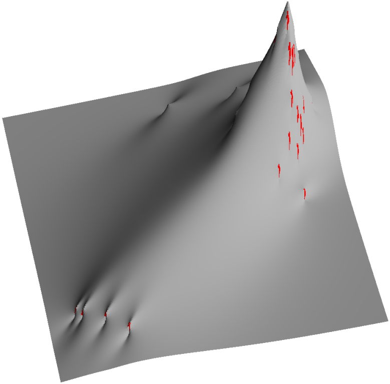
with

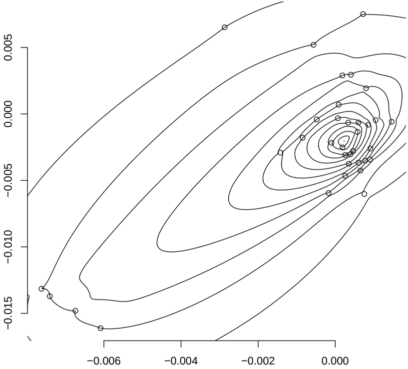
$$v(\boldsymbol{y}) = \begin{cases} \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} & \text{if } \boldsymbol{y} \neq \boldsymbol{0}, \\ \boldsymbol{0} & \text{if } \boldsymbol{y} = \boldsymbol{0}. \end{cases}$$

Properties:

- satisfies **D1 – D5**, but not **D4con**, is continuous;
- if $\boldsymbol{\Sigma}$ is orthogonal, satisfies **D2iso** only;
- with **D2iso** its maximum (say $\boldsymbol{x}^*$) is referred to as **spatial median**, a multivariate location estimator having asymptotic breakdown point of 0.5.

Spatial depth (Vardi & Zhang, 2000; Serfling 2002)





Contents

Functional data framework

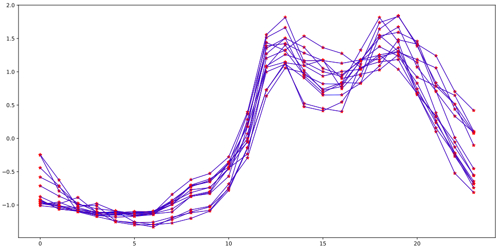
- Let $\boldsymbol{F} = \{\boldsymbol{F}(t) \in \mathbb{R}^d,\ t \in [0,1]\}$ be a random variable that takes its values in a (multivariate) functional space.

Functional data framework

- Let $\boldsymbol{F} = \{\boldsymbol{F}(t) \in \mathbb{R}^d,\ t \in [0,1]\}$ be a random variable that takes its values in a (multivariate) functional space.
- In practice, we only have access to the realization of $\boldsymbol{F}$ at a finite number of arguments/times, $\boldsymbol{f} = \{\boldsymbol{f}(t_1), ..., \boldsymbol{f}(t_p)\}$ such that $0 \leq t_1 < \cdots < t_p \leq 1$.

Functional data framework

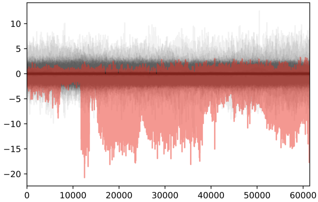
- Let $\boldsymbol{F} = \{\boldsymbol{F}(t) \in \mathbb{R}^d,\ t \in [0,1]\}$ be a random variable that takes its values in a (multivariate) functional space.
- In practice, we only have access to the realization of $\boldsymbol{F}$ at a finite number of arguments/times, $\boldsymbol{f} = \{\boldsymbol{f}(t_1), ..., \boldsymbol{f}(t_p)\}$ such that $0 \leq t_1 < \cdots < t_p \leq 1$.
- The first step: reconstruct **functional object** from partial observations (time-series) with interpolation or basis decomposition.

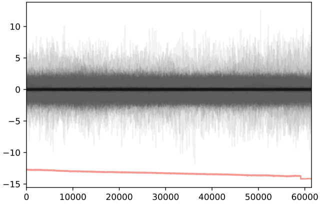


Taxonomy of functional anomalies

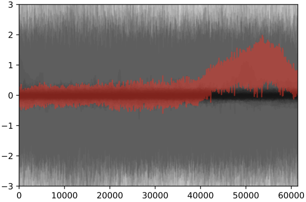
A non-complete taxonomy of functional abnormalities:

Magnitude (=location, shift) anomalies

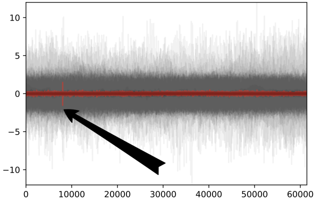




Shape anomalies

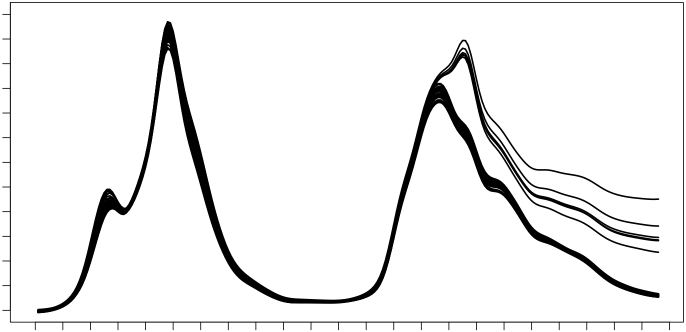


Isolated anomalies

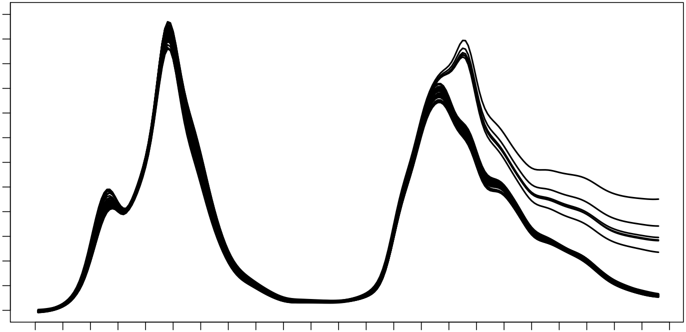


Contents

Detection of (multivariate) functional anomalies



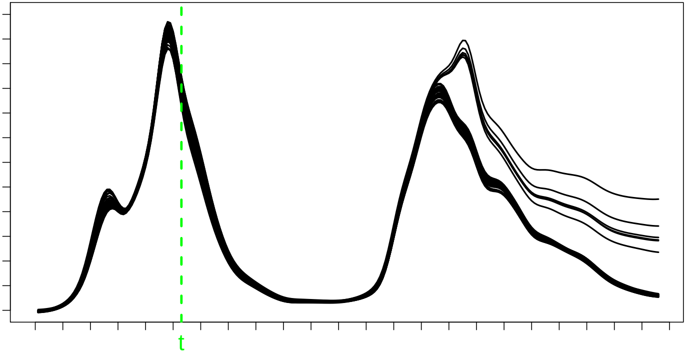
Detection of (multivariate) functional anomalies



- **Functional depth** of $\boldsymbol{f}$ w.r.t. $\mathcal{F} = \{\boldsymbol{f}_i\}_{i=1}^{n}$:

$$D(\boldsymbol{f}|\mathcal{F}) = \int_{t_{\min}}^{t^{\max}} D^1\big(\boldsymbol{f}(t)|\mathcal{F}(t)\big)\, dt\,,$$

Detection of (multivariate) functional anomalies

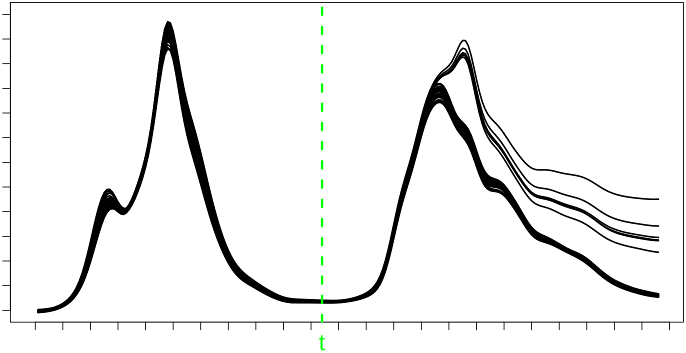


- **Functional depth** of $\boldsymbol{f}$ w.r.t. $\mathcal{F} = \{\boldsymbol{f}_i\}_{i=1}^n$:

$$D(\boldsymbol{f}|\mathcal{F}) = \int_{t_{\text{min}}}^{t^{\text{max}}} D^1\big(\boldsymbol{f}(t)|\{\boldsymbol{f}_1(t), ..., \boldsymbol{f}_n(t)\}\big)\, dt\,,$$

where $D^d(\cdot|\cdot)$ is a multivariate data depth, as defined above.

Detection of (multivariate) functional anomalies

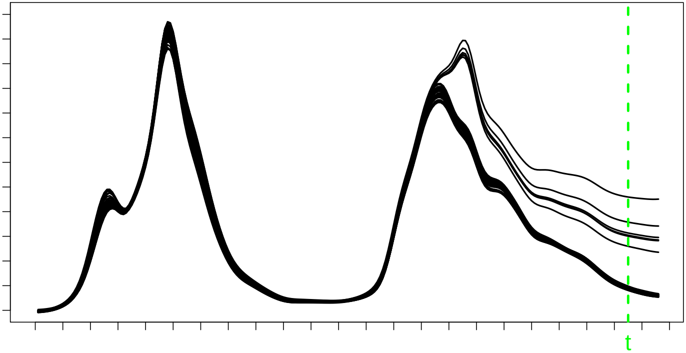


- **Functional depth** of $\boldsymbol{f}$ w.r.t. $\mathcal{F} = \{\boldsymbol{f}_i\}_{i=1}^n$:

$$D(\boldsymbol{f}|\mathcal{F}) = \int_{t_{\min}}^{t^{\max}} D^1\big(\boldsymbol{f}(t)|\{\boldsymbol{f}_1(t), ..., \boldsymbol{f}_n(t)\}\big)\, dt\,,$$

where $D^d(\cdot|\cdot)$ is a multivariate data depth, as defined above.

Detection of (multivariate) functional anomalies

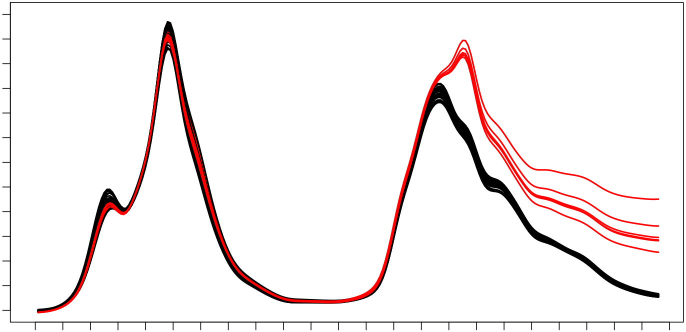


- **Functional depth** of $\boldsymbol{f}$ w.r.t. $\mathcal{F} = \{\boldsymbol{f}_i\}_{i=1}^{n}$:

$$D(\boldsymbol{f}|\mathcal{F}) = \int_{t_{\min}}^{t^{\max}} D^1\big(\boldsymbol{f}(t)|\{\boldsymbol{f}_1(t), \dots, \boldsymbol{f}_n(t)\}\big)\, dt\,,$$

where $D^d(\cdot|\cdot)$ is a multivariate data depth, as defined above.

Detection of (multivariate) functional anomalies

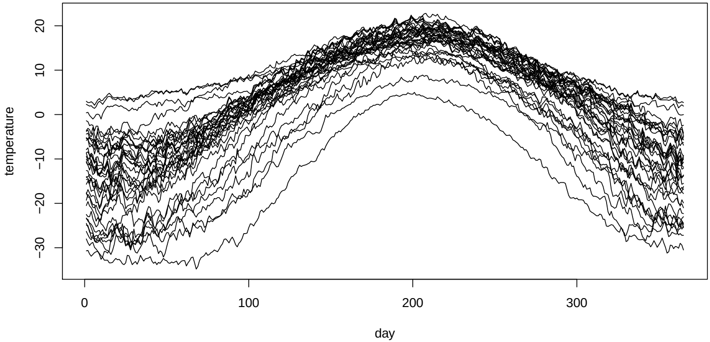


- **Functional depth** of $\boldsymbol{f}$ w.r.t. $\mathcal{F} = \{\boldsymbol{f}_i\}_{i=1}^n$:

$$D(\boldsymbol{f}|\mathcal{F}) = \int_{t_{\text{min}}}^{t^{\text{max}}} D^1\big(\boldsymbol{f}(t)|\{\boldsymbol{f}_1(t), \dots, \boldsymbol{f}_n(t)\}\big)\, dt\,,$$

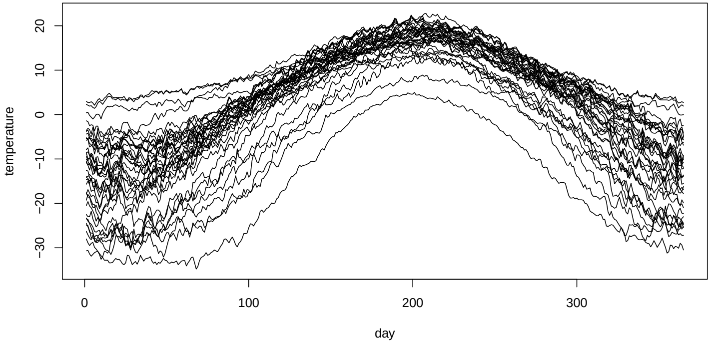
  where $D^d(\cdot|\cdot)$ is a multivariate data depth, as defined above.
- Label $\boldsymbol{f}$ as anomaly if $D(\boldsymbol{f}|\mathcal{F}) < \min(D)$.

Integrated depth for functional data



Let $\boldsymbol{F}$ be a stochastic process with continuous paths defined on $[0, 1]$, and $\boldsymbol{f}$ its realization.

Integrated depth for functional data

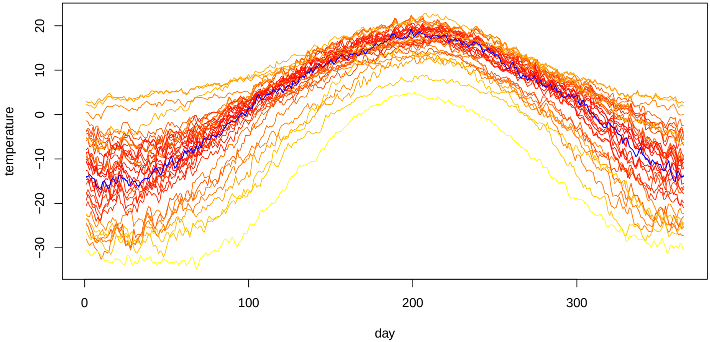


Let $\boldsymbol{F}$ be a stochastic process with continuous paths defined on $[0, 1]$, and $\boldsymbol{f}$ its realization. Then:

$$D(\boldsymbol{f}|\boldsymbol{F}) = \int_0^1 D\big(\boldsymbol{f}(t)|\boldsymbol{F}(t)\big)\,dt.$$

see Fraiman, Muniz, 2001; also López-Pintado, Romo, 2011.

Integrated depth for functional data

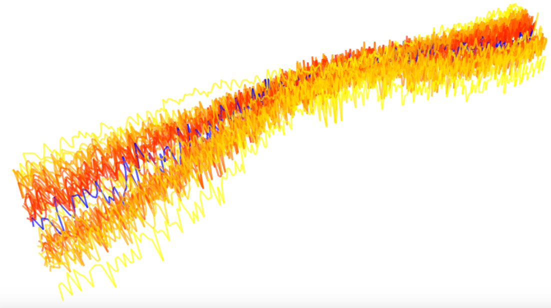


Let $\boldsymbol{F}$ be a stochastic process with continuous paths defined on $[0, 1]$, and $\boldsymbol{f}$ its realization. Then:

$$D(\boldsymbol{f}|\boldsymbol{F}) = \int_0^1 \min\{F_{\boldsymbol{F}(t)}\big(\boldsymbol{f}(t)\big), 1 - F_{\boldsymbol{F}(t)}\big(\boldsymbol{f}(t)^-\big)\}dt.$$

see Fraiman, Muniz, 2001; also López-Pintado, Romo, 2011.

Multivariate functional halfspace depth



Let $\boldsymbol{F}$ be a d-real-valued stochastic process with continuous paths defined on $[0, 1]$, and $\boldsymbol{f}$ its realization. Then:

$$MFD(\boldsymbol{f}|\boldsymbol{F}) = \int_0^1 D\big(\boldsymbol{f}(t)|\boldsymbol{F}(t)\big) \cdot w(t)dt,$$

$$w(t) = w_\alpha\big(t, \boldsymbol{F}(t)\big) = \frac{\text{vol}\big\{D_\alpha\big(\boldsymbol{F}(t)\big)\big\}}{\int_0^1 \text{vol}\big\{D_\alpha\big(\boldsymbol{F}(u)\big)\big\}du}.$$

see Claeskens, Hubert, Slaets, Vakili, 2014.

Contents

Functional isolation forest
(Staerman, M., Clémençon, d'Alché-Buc; 2019)

- **Functional isolation forest** is an adaptation of the multivariate isolation forest algorithm
(see also Hariri, Carrasco Kind, Brunner, 2018).

Functional isolation forest (Staerman, M., Clémençon, d'Alché-Buc; 2019)

- **Functional isolation forest** is an adaptation of the multivariate isolation forest algorithm (see also Hariri, Carrasco Kind, Brunner, 2018).
- **No predefined basis** to project on.

Functional isolation forest (Staerman, M., Clémençon, d'Alché-Buc; 2019)

- **Functional isolation forest** is an adaptation of the multivariate isolation forest algorithm (see also Hariri, Carrasco Kind, Brunner, 2018).
- **No predefined basis** to project on.
- Drawing a direction on a unit sphere as in (Hariri *et al.*, 2018) is no longer possible due to the excessive richness of $\mathcal{H}$.

Functional isolation forest (Staerman, M., Clémençon, d'Alché-Buc; 2019)

- **Functional isolation forest** is an adaptation of the multivariate isolation forest algorithm (see also Hariri, Carrasco Kind, Brunner, 2018).
- **No predefined basis** to project on.
- Drawing a direction on a unit sphere as in (Hariri *et al.*, 2018) is no longer possible due to the excessive richness of $\mathcal{H}$.
- Project the observations $\boldsymbol{f}$ on (random) **elements $\boldsymbol{d}$ of a dictionary** $\mathcal{D} \in \mathcal{H}$ chosen to be rich enough to explore different properties of data and well appropriate to be sampled in a representative manner.

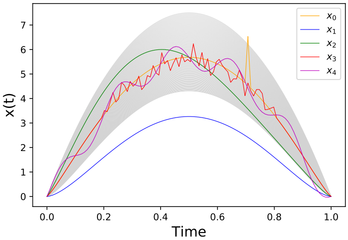
Functional isolation forest (Staerman, M., Clémençon, d'Alché-Buc; 2019)

- **Functional isolation forest** is an adaptation of the multivariate isolation forest algorithm (see also Hariri, Carrasco Kind, Brunner, 2018).
- **No predefined basis** to project on.
- Drawing a direction on a unit sphere as in (Hariri *et al.*, 2018) is no longer possible due to the excessive richness of $\mathcal{H}$.
- Project the observations $\boldsymbol{f}$ on (random) **elements $\boldsymbol{d}$ of a dictionary** $\mathcal{D} \in \mathcal{H}$ chosen to be rich enough to explore different properties of data and well appropriate to be sampled in a representative manner.
- To account for both location and shape anomalies, we suggest the following **scalar product** that provides a compromise between the both (for $\lambda = 0.5$, Sobolev $W_{1,2}$ scalar product):

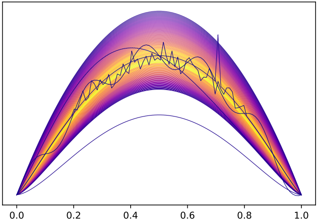
$$\langle \mathbf{f}, \mathbf{d} \rangle := \lambda \times \frac{\langle \mathbf{f}, \mathbf{d} \rangle_{L_2}}{||\mathbf{f}||\,||\mathbf{d}||} + (1-\lambda) \times \frac{\langle \mathbf{f}', \mathbf{d}' \rangle_{L_2}}{||\mathbf{f}'||\,||\mathbf{d}'||}, \quad \lambda \in [0,1]\,.$$

Functional isolation forest: an example

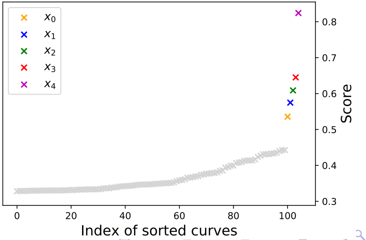
Functional **data set** with **anomalies**



Color-indicated anomaly score

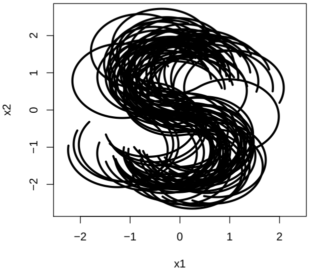


Anomaly score

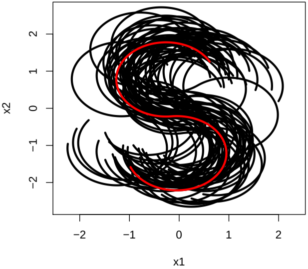


Contents

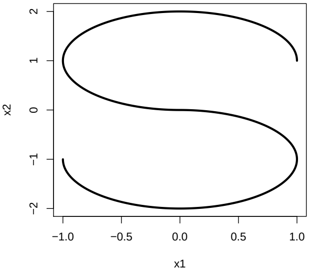
Functional depth: Motivation 1



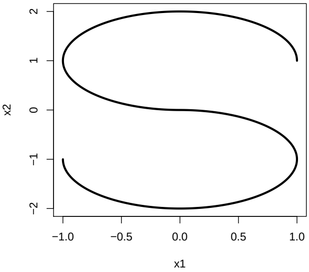
Functional depth: Motivation 1



Functional depth: Motivation 1



Functional depth: Motivation 1



Regard the following different parametrizations of a curve:

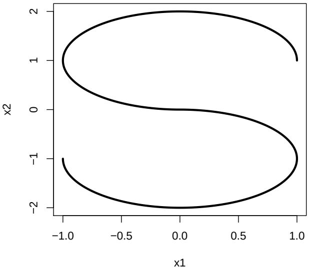
Parametrization A:

$$x_1(t) = -\big(\cos(t)+1\big)\mathbb{1}\{t < \tfrac{3\pi}{2}\} - \big(\cos(3t-3\pi)\ +1\big)\mathbb{1}\{t \geq \tfrac{3\pi}{2}\}+1$$
$$x_2(t) = \quad \big(sin(t)+1\big)\mathbb{1}\{t < \tfrac{3\pi}{2}\} - \big(sin(3t-3\pi)+1\big)\mathbb{1}\{t \geq \tfrac{3\pi}{2}\}$$

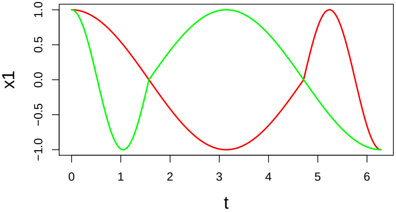
Parametrization B:

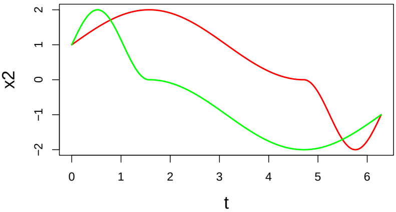
$$x_1(t) = -\big(cos(3t)+1\big)\mathbb{1}\{t < \tfrac{\pi}{2}\} - \big(cos(t+\pi)+1\big)\mathbb{1}\{t \geq \tfrac{\pi}{2}\}+1$$
$$x_2(t) = \quad \big(sin(3t)\ +1\big)\mathbb{1}\{t < \tfrac{\pi}{2}\} - \big(sin(t+\pi)\ +1\big)\mathbb{1}\{t \geq \tfrac{\pi}{2}\}$$

Functional depth: Motivation 1

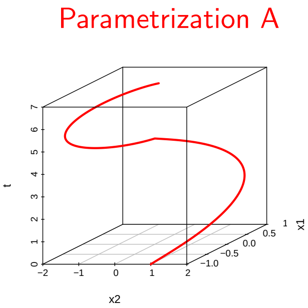


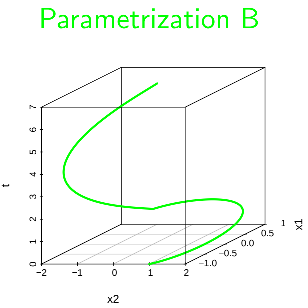
Regard the following different parametrizations of a curve:



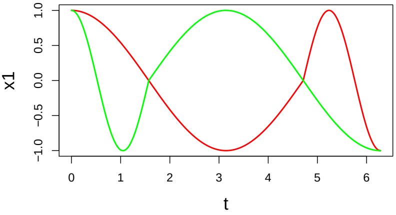


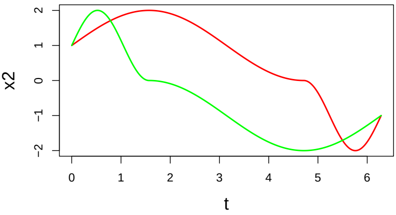
Functional depth: Motivation 1





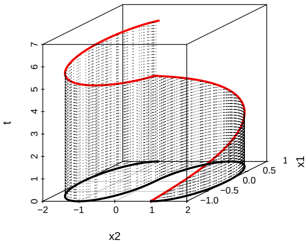
Parametrization:



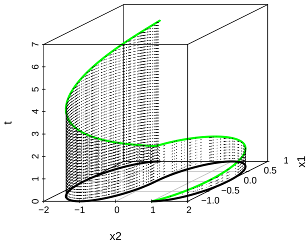


Functional depth: Motivation 1

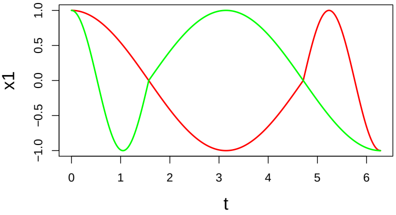
Parametrization A

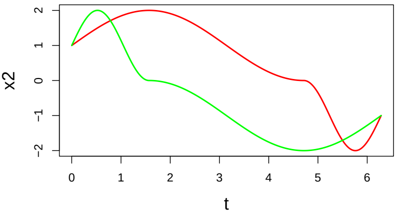


Parametrization B

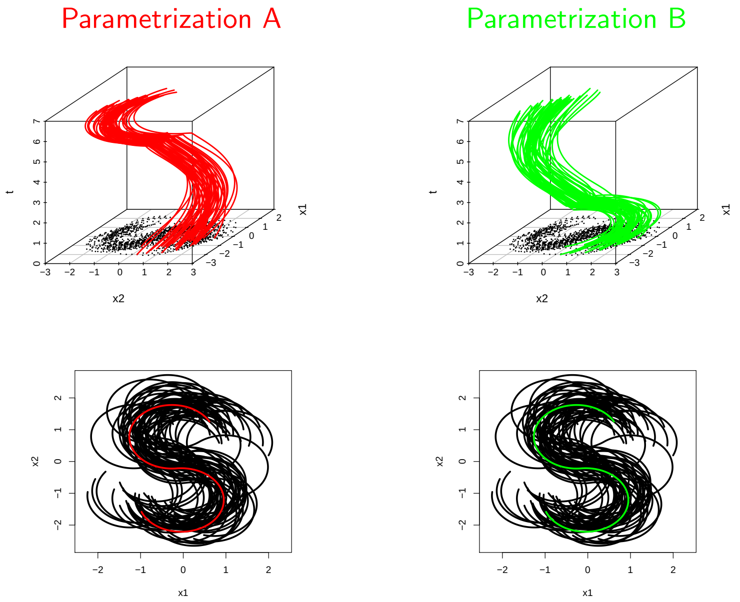


Parametrization:



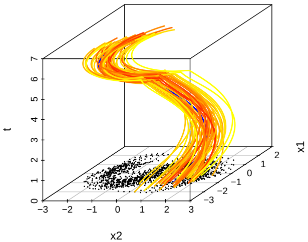


Functional depth: Motivation 1

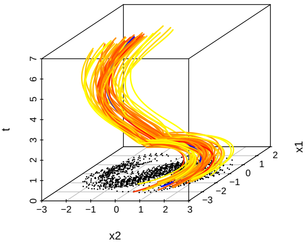


Functional depth: Motivation 1

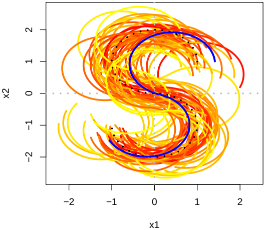
Parametrization A

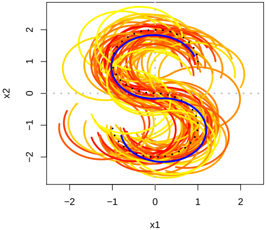


Parametrization B



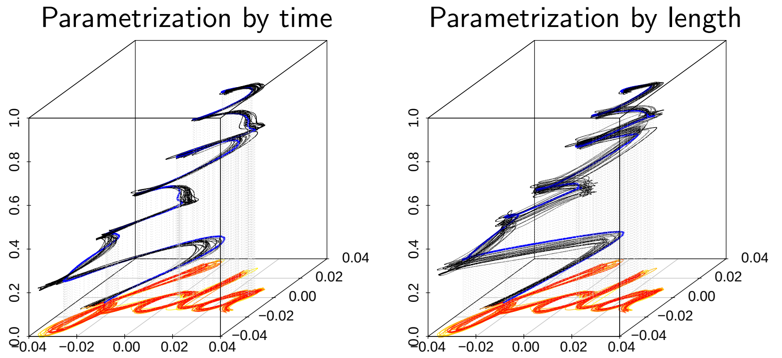
The depth-induced orders differ!





Functional depth: Motivation 2

Functional halfspace depth for the FDA-data



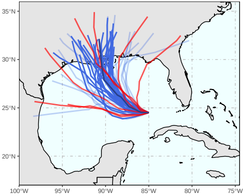
Depth-induced ranking for parametrizations by time and by length:

Time	2	3	13	12	4	8	**1**	17	11	9	7	19	15	20	18	16	14	5	6	10
Length	6	3	16	14	5	7	**13**	11	1	17	2	19	8	20	12	18	15	4	9	10

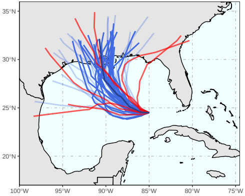
Functional depth: Motivation 3

Simulated hurricane tracks: **curve boxplot**

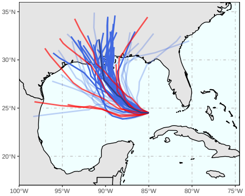
MFH depth – par. time



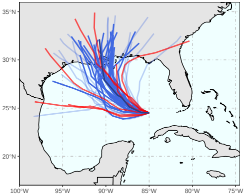
MFH depth - par. length



mSB depth – par. time



mSB depth – par. length



The space of curves

- Let $(\mathbb{R}^d, |\cdot|_2)$ be the Euclidean space.

The space of curves

- Let $(\mathbb{R}^d, |\cdot|_2)$ be the Euclidean space.
- A *parametrized curve* $\beta : [0,1] \to \mathbb{R}^d$ is a continuous map.
  A reparametrization $\gamma : [0,1] \to [0,1]$ is increasing continuous function: $\gamma(0) = 0$ and $\gamma(1) = 1$.

The space of curves

- Let $(\mathbb{R}^d, |\cdot|_2)$ be the Euclidean space.
- A *parametrized curve* $\beta : [0,1] \to \mathbb{R}^d$ is a continuous map.
  A reparametrization $\gamma : [0,1] \to [0,1]$ is increasing continuous function: $\gamma(0) = 0$ and $\gamma(1) = 1$.
- Two parametrized curves β_1, β_2 are equivalent if and only if there exist two reparametrizations γ_1, γ_2 : $\beta_1 \circ \gamma_1 = \beta_2 \circ \gamma_2$.

The space of curves

- Let $(\mathbb{R}^d, |\cdot|_2)$ be the Euclidean space.
- A *parametrized curve* $\beta : [0,1] \to \mathbb{R}^d$ is a continuous map. A reparametrization $\gamma : [0,1] \to [0,1]$ is increasing continuous function: $\gamma(0) = 0$ and $\gamma(1) = 1$.
- Two parametrized curves β_1, β_2 are equivalent if and only if there exist two reparametrizations γ_1, γ_2 : $\beta_1 \circ \gamma_1 = \beta_2 \circ \gamma_2$.
- An *unparametrized curve*, noted $\mathcal{C} := \mathcal{C}_\beta$, is defined as the equivalence class of β up to the above equivalence relation. The *space of unparametrized curves* is then defined as

$$\mathfrak{B} = \{\mathcal{C}_\beta : \beta \in \mathcal{C}([0,1], \mathbb{R}^d)\}.$$

The space of curves

- Let $(\mathbb{R}^d, |\cdot|_2)$ be the Euclidean space.
- A *parametrized curve* $\beta : [0,1] \to \mathbb{R}^d$ is a continuous map. A reparametrization $\gamma : [0,1] \to [0,1]$ is increasing continuous function: $\gamma(0) = 0$ and $\gamma(1) = 1$.
- Two parametrized curves β_1, β_2 are equivalent if and only if there exist two reparametrizations γ_1, γ_2 : $\beta_1 \circ \gamma_1 = \beta_2 \circ \gamma_2$.
- An *unparametrized curve*, noted $\mathcal{C} := \mathcal{C}_\beta$, is defined as the equivalence class of β up to the above equivalence relation. The *space of unparametrized curves* is then defined as

$$\mathfrak{B} = \{\mathcal{C}_\beta : \beta \in \mathcal{C}([0,1], \mathbb{R}^d)\}.$$

- We endow $\mathfrak{B}$ with the Fréchet *metric*:

$$d_{\mathfrak{B}}\left(\mathcal{C}_1, \mathcal{C}_2\right) = \inf\left\{\|\beta_1 - \beta_2\|_\infty, \beta_1 \in \mathcal{C}_1,\ \beta_2 \in \mathcal{C}_2\right\}, \quad \mathcal{C}_1, \mathcal{C}_2 \in \mathfrak{B}\,.$$

Associated distribution and the sampling scheme

- Let $\mathcal{C}$ be an unparameterized curve. The *length of* $\mathcal{C}$:

$$L(\mathcal{C}) = \sup_{\tau} \left\{ \sum_{i=1}^{N} |\beta(\tau_i) - \beta(\tau_{i-1})|_2 \ : \tau \text{ is a partition of } [0,1] \right\},$$

for all $\beta \in \mathcal{C}$.

Associated distribution and the sampling scheme

- Let $\mathcal{C}$ be an unparameterized curve. The *length of* $\mathcal{C}$:

$$L(\mathcal{C}) = \sup_{\tau} \left\{ \sum_{i=1}^{N} |\beta(\tau_i) - \beta(\tau_{i-1})|_2 \ : \tau \text{ is a partition of } [0,1] \right\},$$

for all $\beta \in \mathcal{C}$.

- An unparametrized curve $\mathcal{C}$ is called *rectifiable* if $L(\mathcal{C})$ is finite. The length $L : \mathfrak{B} \to \mathbb{R} + \cup \{\infty\}$ is measurable:

$$\mathcal{P} = \left\{ P \text{ prob. measure on } (\mathfrak{B}, d_{\mathfrak{B}}) \ : \ P(\{\mathcal{C} \in \mathfrak{B}; 0 < L(\mathcal{C}) < \infty\}) = 1 \right\}.$$

Associated distribution and the sampling scheme

- Let $\mathcal{C}$ be an unparameterized curve. The *length of* $\mathcal{C}$:

$$L(\mathcal{C}) = \sup_{\tau} \left\{ \sum_{i=1}^{N} |\beta(\tau_i) - \beta(\tau_{i-1})|_2 \ : \tau \text{ is a partition of } [0,1] \right\},$$

for all $\beta \in \mathcal{C}$.

- An unparametrized curve $\mathcal{C}$ is called *rectifiable* if $L(\mathcal{C})$ is finite. The length $L : \mathfrak{B} \to \mathbb{R} + \cup\{\infty\}$ is measurable:

$$\mathcal{P} = \left\{ P \text{ prob. measure on } (\mathfrak{B}, d_{\mathfrak{B}}) \ : \ P(\{\mathcal{C} \in \mathfrak{B}; 0 < L(\mathcal{C}) < \infty\}) = 1 \right\}.$$

- Let $\mathcal{X}$ be a random element of $\mathfrak{B}$ stemming from distribution $P \in \mathcal{P}$.

Associated distribution and the sampling scheme

- Let $\mathcal{C}$ be an unparameterized curve. The *length of* $\mathcal{C}$:

$$L(\mathcal{C}) = \sup_{\tau} \left\{ \sum_{i=1}^{N} |\beta(\tau_i) - \beta(\tau_{i-1})|_2 \ : \tau \text{ is a partition of } [0,1] \right\},$$

  for all $\beta \in \mathcal{C}$.
- An unparametrized curve $\mathcal{C}$ is called *rectifiable* if $L(\mathcal{C})$ is finite. The length $L : \mathfrak{B} \to \mathbb{R} + \cup\{\infty\}$ is measurable:

$$\mathcal{P} = \left\{ P \text{ prob. measure on } (\mathfrak{B}, d_{\mathfrak{B}}) \ : \ P(\{\mathcal{C} \in \mathfrak{B}; 0 < L(\mathcal{C}) < \infty\}) = 1 \right\}.$$

- Let $\mathcal{X}$ be a random element of $\mathfrak{B}$ stemming from distribution $P \in \mathcal{P}$.
- We derive the probability distribution Q_P on $\mathbb{R}^d$ as follows: if $X \sim Q_P$, then distribution of $X \,|\, \mathcal{X} = \mathcal{C}$ is the (uniform on $\mathcal{C}$) probability distribution $\mu_{\mathcal{C}}$:

$$\mu_{\mathcal{C}}(A) = \int_{\mathcal{C}} \mathbb{1}_A(x) dx.$$

Associated distribution and the sampling scheme

The statistical model:

$$\mathcal{X}_1, \ldots, \mathcal{X}_n \text{ i.i.d. from } P.$$

For Monte-Carlo estimation, we can consider the following **sampling scheme**:

$$\begin{cases} \mathcal{X}_1, \ldots, \mathcal{X}_n \text{ i.i.d. from } P, \\ \text{for all } i = 1, \ldots, n \\ \qquad X_{i,1}, \ldots, X_{i,m} \text{ i.i.d. from } \mu_{\mathcal{X}_i}. \end{cases}$$

Data depth for an unparametrized curve

Definition

The **Tukey curve depth** of $\mathcal{C} \in \mathfrak{B}$ w.r.t. Q_P is defined as:

$$D(\mathcal{C}|Q_P) = \int_{\mathcal{C}} D(\boldsymbol{x}|Q_P, \mu_{\mathcal{C}}) d\mu_{\mathcal{C}}(\boldsymbol{x}),$$

Data depth for an unparametrized curve

Definition

The **Tukey curve depth** of $\mathcal{C} \in \mathfrak{B}$ w.r.t. Q_P is defined as:

$$D(\mathcal{C}|Q_P) = \int_{\mathcal{C}} D(\boldsymbol{x}|Q_P, \mu_{\mathcal{C}}) d\mu_{\mathcal{C}}(\boldsymbol{x}),$$

where the depth $D(\boldsymbol{x}|Q_P, \mu_{\mathcal{C}})$ of an arbitrary point $\boldsymbol{x} \in \mathcal{C}$ w.r.t. the distribution Q_P is defined as:

$$D(\boldsymbol{x}|Q_P, \mu_{\mathcal{C}}) = \inf\left\{\frac{Q_P(H)}{\mu_{\mathcal{C}}(H)} : H \text{ closed half-space} \subset \mathbb{R}^d, \boldsymbol{x} \in \partial H\right\},$$

where convention $\frac{0}{0} = 0$ is adopted.

Data depth for an unparametrized curve

Definition

The **Tukey curve depth** of $\mathcal{C} \in \mathfrak{B}$ w.r.t. Q_P is defined as:

$$D(\mathcal{C}|Q_P) = \int_{\mathcal{C}} D(\boldsymbol{x}|Q_P, \mu_{\mathcal{C}}) d\mu_{\mathcal{C}}(\boldsymbol{x}) \,,$$

where the depth $D(\boldsymbol{x}|Q_P, \mu_{\mathcal{C}})$ of an arbitrary point $\boldsymbol{x} \in \mathcal{C}$ w.r.t. the distribution Q_P is defined as:

$$D(\boldsymbol{x}|Q_P, \mu_{\mathcal{C}}) = \inf\Big\{\frac{Q_P(H)}{\mu_{\mathcal{C}}(H)} \;:\; H \text{ closed half-space} \subset \mathbb{R}^d \,, \boldsymbol{x} \in \partial H\Big\},$$

where convention $\frac{0}{0} = 0$ is adopted.

Definition

The **sample Tukey curve depth** of $\mathcal{C} \in \mathfrak{B}$ w.r.t. $\mathcal{X}_1, ..., \mathcal{X}_n$ is:

$$D(\mathcal{C}|\mathcal{X}_1, \dots, \mathcal{X}_n) = \int_{\mathcal{C}} D(\boldsymbol{x}|Q_n, \mu_{\mathcal{C}}) d\mu_{\mathcal{C}}(\boldsymbol{x}) \,,$$

Data depth for an unparametrized curve

Definition

The **Tukey curve depth** of $\mathcal{C} \in \mathfrak{B}$ w.r.t. Q_P is defined as:

$$D(\mathcal{C}|Q_P) = \int_{\mathcal{C}} D(\boldsymbol{x}|Q_P, \mu_{\mathcal{C}}) d\mu_{\mathcal{C}}(\boldsymbol{x}),$$

where the depth $D(\boldsymbol{x}|Q_P, \mu_{\mathcal{C}})$ of an arbitrary point $\boldsymbol{x} \in \mathcal{C}$ w.r.t. the distribution Q_P is defined as:

$$D(\boldsymbol{x}|Q_P, \mu_{\mathcal{C}}) = \inf\left\{\frac{Q_P(H)}{\mu_{\mathcal{C}}(H)} : H \text{ closed half-space} \subset \mathbb{R}^d, \boldsymbol{x} \in \partial H\right\},$$

where convention $\frac{0}{0} = 0$ is adopted.

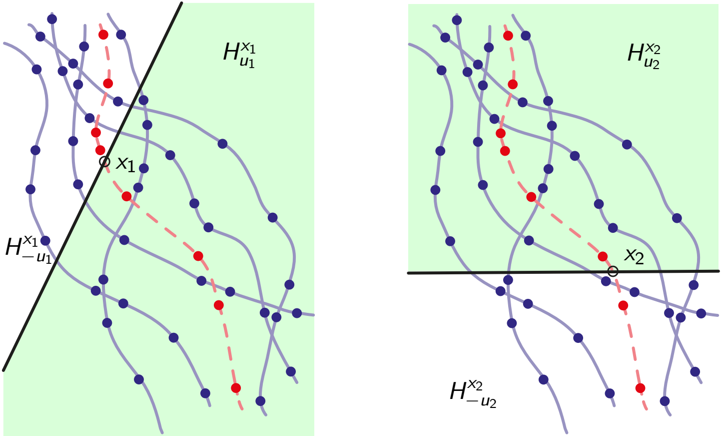
Definition

The **sample Tukey curve depth** of $\mathcal{C} \in \mathfrak{B}$ w.r.t. $\mathcal{X}_1, ..., \mathcal{X}_n$ is:

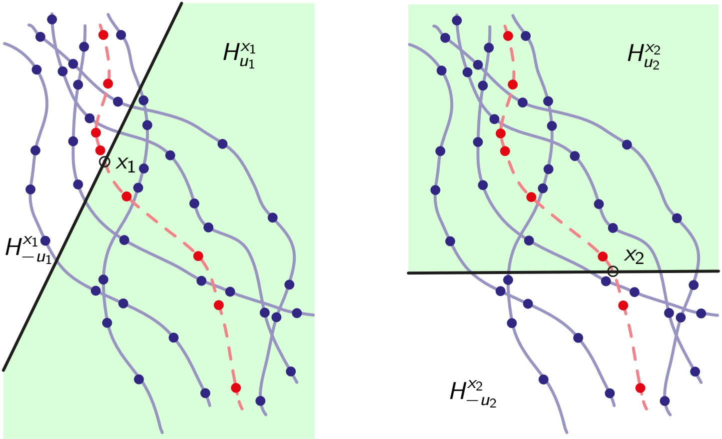
$$D(\mathcal{C}|\mathcal{X}_1, \ldots, \mathcal{X}_n) = \int_{\mathcal{C}} D(\boldsymbol{x}|Q_n, \mu_{\mathcal{C}}) d\mu_{\mathcal{C}}(\boldsymbol{x}),$$

where $Q_n = (\mu_{\mathcal{X}_1} + \cdots + \mu_{\mathcal{X}_n})/n$.

Data depth for an unparametrized curve: intuition



Data depth for an unparametrized curve: intuition



Traditional reasoning:

$\widehat{Q}_P(H^{x_1}_{u_1}) = \frac{25}{40}$, $\widehat{\mu}_{\mathcal{C}}(H^{x_1}_{u_1}) = \frac{4}{8}$

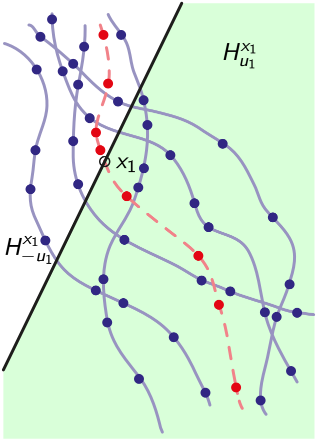
$\widehat{Q}_P(H^{x_1}_{-u_1}) = \frac{15}{40}$, $\widehat{\mu}_{\mathcal{C}}(H^{x_1}_{-u_1}) = \frac{4}{8}$

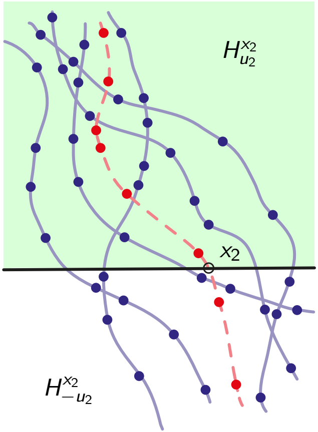
Curve-based reasoning:

$\widehat{Q}_P(H^{x_2}_{u_2}) = \frac{25}{40}$, $\widehat{\mu}_{\mathcal{C}}(H^{x_2}_{u_2}) = \frac{6}{8}$

$\widehat{Q}_P(H^{x_2}_{-u_2}) = \frac{15}{40}$, $\widehat{\mu}_{\mathcal{C}}(H^{x_2}_{-u_2}) = \frac{2}{8}$

Data depth for an unparametrized curve: intuition





Traditional reasoning:

$\widehat{Q}_P(H^{x_1}_{u_1}) = \frac{25}{40}$, $\widehat{\mu}_C(H^{x_1}_{u_1}) = \frac{4}{8}$

$\boldsymbol{\widehat{Q}_P(H^{x_1}_{-u_1}) = \frac{15}{40}}$, $\boldsymbol{\widehat{\mu}_C(H^{x_1}_{-u_1}) = \frac{4}{8}}$

Curve-based reasoning:

$\boldsymbol{\widehat{Q}_P(H^{x_2}_{u_2}) = \frac{25}{40}}$, $\boldsymbol{\widehat{\mu}_C(H^{x_2}_{u_2}) = \frac{6}{8}}$

$\widehat{Q}_P(H^{x_2}_{-u_2}) = \frac{15}{40}$, $\widehat{\mu}_C(H^{x_2}_{-u_2}) = \frac{2}{8}$

Data depth for an unparametrized curve: empirical version

- Let a chosen curve consist of two (independently drawn on $\mathcal{C}$) parts $\mathbb{Y}_{1,m} = (Y_{1,1}, \ldots, Y_{1,m})$ and $\mathbb{Y}_{2,m} = (Y_{2,1}, \ldots, Y_{2,m})$ with empirical distribution

$$\widehat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{Y_{1,i}} \,,$$

where $\delta_{\boldsymbol{x}}$ is the Dirac measure in $\boldsymbol{x} \in \mathbb{R}^d$.

Data depth for an unparametrized curve: empirical version

- Let a chosen curve consist of two (independently drawn on $\mathcal{C}$) parts $\mathbb{Y}_{1,m} = (Y_{1,1}, \ldots, Y_{1,m})$ and $\mathbb{Y}_{2,m} = (Y_{2,1}, \ldots, Y_{2,m})$ with empirical distribution

$$\widehat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{Y_{1,i}} ,$$

  where $\delta_{\boldsymbol{x}}$ is the Dirac measure in $\boldsymbol{x} \in \mathbb{R}^d$.
- Let $\widehat{Q}_{n,m}$ be the empirical distribution (observed sample) $\mathbb{X}_{n,m} = \{X_{i,j},\ i = 1, ..., n,\ j = 1, ..., m\}$

$$\widehat{Q}_{n,m} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{X_{i,j}} .$$

Data depth for an unparametrized curve: empirical version

- Let a chosen curve consist of two (independently drawn on $\mathcal{C}$) parts $\mathbb{Y}_{1,m} = (Y_{1,1}, \ldots, Y_{1,m})$ and $\mathbb{Y}_{2,m} = (Y_{2,1}, \ldots, Y_{2,m})$ with empirical distribution

$$\widehat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{Y_{1,i}} ,$$

  where $\delta_{\boldsymbol{x}}$ is the Dirac measure in $\boldsymbol{x} \in \mathbb{R}^d$.
- Let $\widehat{Q}_{n,m}$ be the empirical distribution (observed sample) $\mathbb{X}_{n,m} = \{X_{i,j},\ i = 1, \ldots, n,\ j = 1, \ldots, m\}$

$$\widehat{Q}_{n,m} = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{X_{i,j}} .$$

- To compute the sample Tukey curve depth, a Monte Carlo approximation is used.

Data depth for an unparametrized curve: empirical version

- Let H be a closed halfspace in $\mathbb{R}^d$ and $\mathcal{H}_{\Delta}^{n,m}$ denote a collection of such halfspaces such that for all $H \in \mathcal{H}_{\Delta}^{n,m}$ either $\widehat{Q}_{n,m}(H) = 0$ or $\widehat{\mu}_m(H) > \Delta$, almost surely, for $\Delta \in (0, \frac{1}{2})$.

Data depth for an unparametrized curve: empirical version

- Let H be a closed halfspace in $\mathbb{R}^d$ and $\mathcal{H}_\Delta^{n,m}$ denote a collection of such halfspaces such that for all $H \in \mathcal{H}_\Delta^{n,m}$ either $\widehat{Q}_{n,m}(H) = 0$ or $\widehat{\mu}_m(H) > \Delta$, almost surely, for $\Delta \in (0, \frac{1}{2})$.

Definition

The **Monte Carlo approximation** of the **Tukey curve depth** of $\mathcal{C}$ w.r.t. $\mathcal{X}_1, ..., \mathcal{X}_n$ is defined as:

$$\widehat{D}_{n,m,\Delta}(\mathcal{C}|\mathcal{X}_1, ..., \mathcal{X}_n) = \frac{1}{m}\sum_{i=1}^{m} \widehat{D}(Y_{2,i}|\widehat{Q}_{n,m}, \widehat{\mu}_m, \mathcal{H}_\Delta^{n,m}),$$

Data depth for an unparametrized curve: empirical version

- Let H be a closed halfspace in $\mathbb{R}^d$ and $\mathcal{H}_{\Delta}^{n,m}$ denote a collection of such halfspaces such that for all $H \in \mathcal{H}_{\Delta}^{n,m}$ either $\widehat{Q}_{n,m}(H) = 0$ or $\widehat{\mu}_m(H) > \Delta$, almost surely, for $\Delta \in (0, \frac{1}{2})$.

Definition

The **Monte Carlo approximation** of the **Tukey curve depth** of $\mathcal{C}$ w.r.t. $\mathcal{X}_1, ..., \mathcal{X}_n$ is defined as:

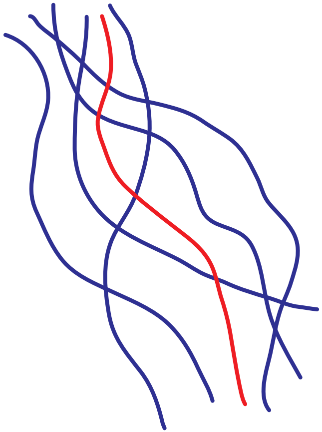
$$\widehat{D}_{n,m,\Delta}(\mathcal{C}|\mathcal{X}_1, ..., \mathcal{X}_n) = \frac{1}{m} \sum_{i=1}^{m} \widehat{D}(Y_{2,i}|\widehat{Q}_{n,m}, \widehat{\mu}_m, \mathcal{H}_{\Delta}^{n,m}) \,,$$

with the depth of an arbitrary point $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. $\widehat{Q}_{n,m}$ being:

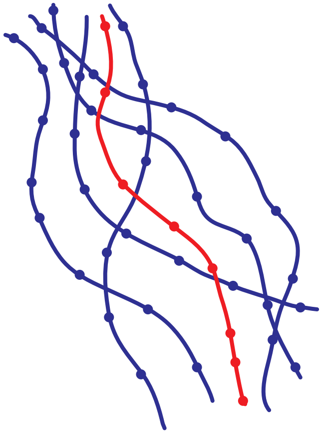
$$\widehat{D}(\boldsymbol{x}|\widehat{Q}_{n,m}, \widehat{\mu}_m, \mathcal{H}_{\Delta}^{n,m}) \;=\; \inf\Big\{\frac{\widehat{Q}_{n,m}(H)}{\widehat{\mu}_m(H)} \;:\; H \in \mathcal{H}_{\Delta}^{n,m}, \, \boldsymbol{x} \in \partial H\Big\}$$

and $\frac{0}{0} = 0$ as before.

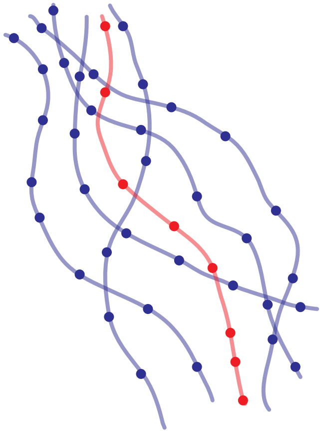
Calculation of the Tukey curve depth



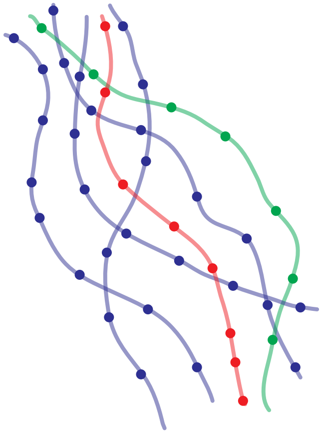
Calculation of the Tukey curve depth



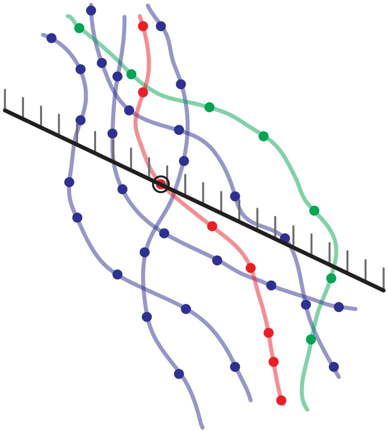
Calculation of the Tukey curve depth



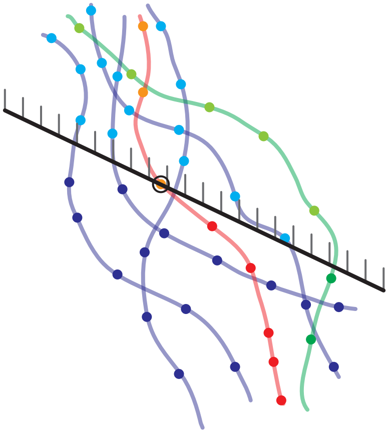
Calculation of the Tukey curve depth



Calculation of the Tukey curve depth

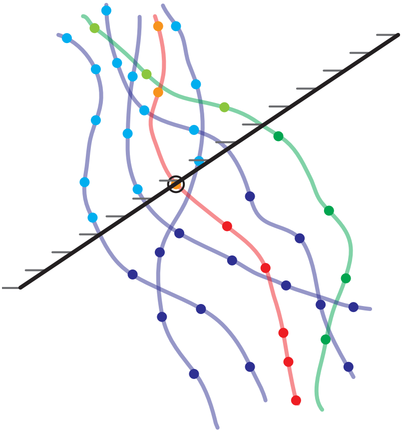


Calculation of the Tukey curve depth



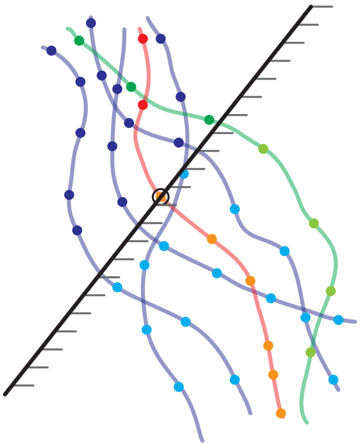
$$D(\mathbb{Y}_{2,c}|Q_m, \mathcal{H}_{m,b}) = \frac{\frac{1}{5}\left(\frac{5}{7}+\frac{3}{8}+\frac{6}{8}+\frac{2}{7}+\frac{3}{6}\right)}{\frac{2}{8}} = 2.1$$

Calculation of the Tukey curve depth



$$D(\mathbb{Y}_{2,c}|Q_m, \mathcal{H}_{m,b}) = \frac{\frac{1}{5}\left(\frac{3}{7}+\frac{5}{8}+\frac{4}{8}+\frac{3}{7}+\frac{3}{6}\right)}{\frac{2}{8}} = 1.9857$$

Calculation of the Tukey curve depth



$$D(\mathbb{Y}_{2,c}|Q_m, \mathcal{H}_{m,b}) = \frac{\frac{1}{5}\left(\frac{4}{7}+\frac{3}{8}+\frac{4}{8}+\frac{4}{7}+\frac{4}{6}\right)}{\frac{5}{8}} = 0.7159$$

Data depth for an unparametrized curve: properties

Restrict to $\mathfrak{B}_\ell$, the subset of unparametrized curves of positive length bounded by $\ell > 0$. Then the Tukey curve depth satisfies the following properties:

- **Nonnegativity and boundedness by one:**

$$D(\mathcal{C}|Q_P) \in [0, 1] \,.$$

Data depth for an unparametrized curve: properties

Restrict to $\mathfrak{B}_\ell$, the subset of unparametrized curves of positive length bounded by $\ell > 0$. Then the Tukey curve depth satisfies the following properties:

- **Nonnegativity and boundedness by one:**

$$D(\mathcal{C}|Q_P) \in [0,1]\,.$$

- **Similarity invariance:** Let $f : \mathbb{R}^d \to \mathbb{R}^d f$ be a similarity, i.e. there exists an orthogonal matrix A, a factor $r > 0$ and a vector $\boldsymbol{b} \in \mathbb{R}^d$ such that for all $\boldsymbol{x} \in \mathbb{R}^d$, $f(\boldsymbol{x}) = rA\boldsymbol{x} + \boldsymbol{b}$. In particular for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^d$, $|f(\boldsymbol{x}) - f(\boldsymbol{y})|_2 = r|\boldsymbol{x} - \boldsymbol{y}|_2$. Denote by P_f the distribution of the image through f of a stochastic process having a distribution P. Then

$$D(f \circ \mathcal{C}|Q_{P_f}) = D(\mathcal{C}|Q_P)\,.$$

Data depth for an unparametrized curve: properties

Restrict to $\mathfrak{B}_\ell$, the subset of unparametrized curves of positive length bounded by $\ell > 0$. Then the Tukey curve depth satisfies the following properties:

- **Nonnegativity and boundedness by one:**

$$D(\mathcal{C}|Q_P) \in [0,1]\,.$$

- **Similarity invariance:** Let $f : \mathbb{R}^d \to \mathbb{R}^d f$ be a similarity, i.e. there exists an orthogonal matrix A, a factor $r > 0$ and a vector $\boldsymbol{b} \in \mathbb{R}^d$ such that for all $\boldsymbol{x} \in \mathbb{R}^d$, $f(\boldsymbol{x}) = rA\boldsymbol{x} + \boldsymbol{b}$. In particular for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^d$, $|f(\boldsymbol{x}) - f(\boldsymbol{y})|_2 = r|\boldsymbol{x} - \boldsymbol{y}|_2$. Denote by P_f the distribution of the image through f of a stochastic process having a distribution P. Then

$$D(f \circ \mathcal{C}|Q_{P_f}) = D(\mathcal{C}|Q_P)\,.$$

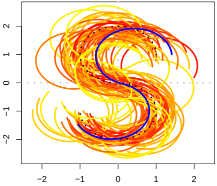
- **Vanishing at infinity:**

$$\lim_{d_{\mathbb{G}}(\mathcal{C},\mathbf{0}) \to \infty, \mathcal{C} \in \mathfrak{B}_\ell} D(\mathcal{C}, Q_P) = \inf_{\mathcal{C} \in \mathfrak{B}_\ell} D(\mathcal{C}, Q_P) = 0\,.$$

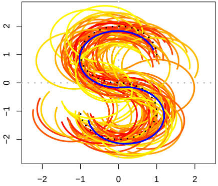
Comparison with functional depth: Example 1

Simulated S letters: **depth-induced ranking**

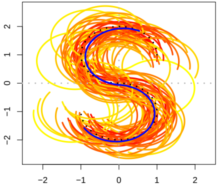
MFHD – time



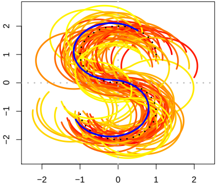
MFHD - length



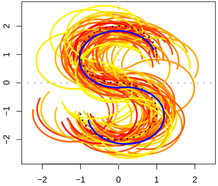
Curve depth



mSBD – time



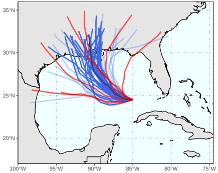
mSBD – length



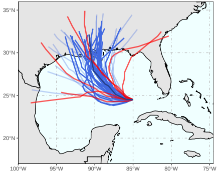
Comparison with functional depth: Example 2

Simulated hurricane tracks: **curve boxplot**

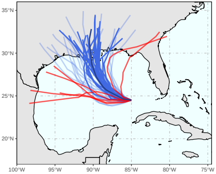
MFHD – time



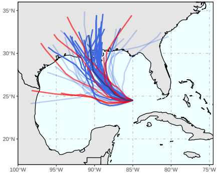
MFHD - length



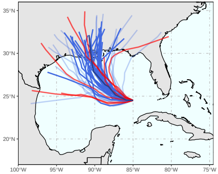
Curve depth



mSBD – time

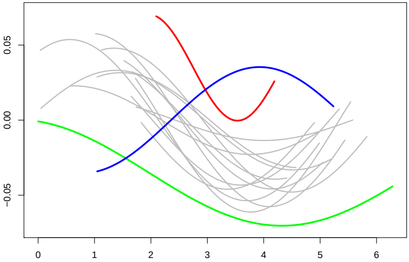


mSBD – length



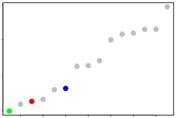
Comparison with functional depth: Anomaly detection 1

Data set 1 with introduced anomalies:

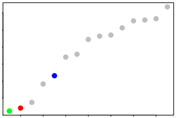


Ordered depth values:

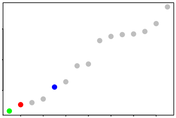
mSBD



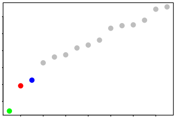
saPRJ



MFHD

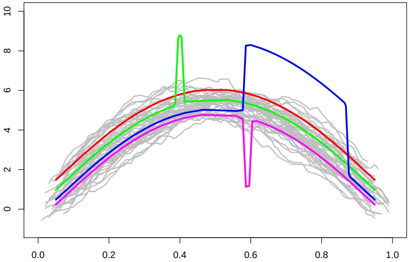


Curve depth



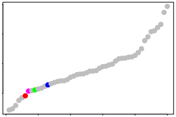
Comparison with functional depth: Anomaly detection 2

Data set 2 with introduced anomalies:

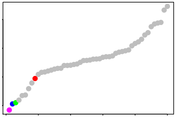


Ordered depth values:

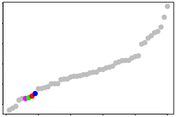
mSBD



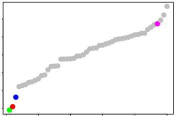
saPRJ



MFHD



Curve depth



Contents

Thank you for attention! (and a short list of literature)

- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. ACM *Computing Surveys (CSUR)*, 41(3):15, 1–58.
- Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. (2000). LOF: Identifying density-based local outliers. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 29, 93–104.
- Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), 1443–1471.
- Liu, F.T., Ting, K.M., and Zhou, Z. (2008). Isolation forest. In: *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 413–422.
- Mosler, K. (2013). Depth statistics. In: *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*, 17-34.
- Hubert, M., Rousseeuw, P.J., and Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2), 177-202.

Practical session

Notebooks:

- anomdet_simulation1.Rmd,
- anomdet_hurricanes.Rmd,
- anomdet_brainimaging.Rmd,
- anomdet_cars.ipynb,
- anomdet_airbus.ipynb.

Data sets:

- carsanom.csv: Data set on anomaly detection for cars.
- airbus_data.csv: Data set from Airbus.
- hurdat2-1851-2019-052520.txt: Historical hurricane data.
- 101_1_dwi_fa.nii: Anatomical brain volume data.
- 101_1_dwi.voxelcoordsL.txt: Left brain fiber's bundle.
- 101_1_dwi.voxelcoordsR.txt: Right brain fiber's bundle.

Supplementary scripts:

- depth_routines.py: Routines for data depth calculation.
- FIF.py: Implementation of the functional isolation forest.
- depth_routines.R: Routines for curves' parametrization.
- DTI.R: Routines for input of brain imaging data.

Literature (mentioned in the tutorial) (1)

- Boser, B.E., Guyon, I., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, Pittsburgh, ACM, 5, 144–152.
- Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. (2000). LOF: Identifying density-based local outliers. In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 29, 93–104.
- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. ACM *Computing Surveys (CSUR)*, 41(3):15, 1–58.
- Chaudhuri P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91, 862–872.
- Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, 109(505), 411-423.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.
- Donoho D. (1982). *Breakdown Properties of Multivariate Location Estimators.* Ph.D. thesis, Harvard University.
- Donoho D.L., Gasko M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20, 1803–1827.

Literature (mentioned in the tutorial) (2)

- Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *TEST*, 10, 419-440.
- Hariri, S., Carrasco Kind, M., and Brunner, R.J. (2018). Extended isolation forest. `arXiv:1811.02141`.
- Hubert, M., Rousseeuw, P.J., and Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2), 177-202.
- Koltchinskii V. (1997). M-estimation, convexity and quantiles. *The Annals of Statistics*, 25, 435–477.
- Koshevoy G., Mosler K. (1997). Zonoid trimming for multivariate distributions. *The Annals of Statistics*, 25, 1998–2017.
- Liu R.Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18, 405–414.
- Liu, Z. and Modarres, R. (2011). Lens data depth and median. *Journal of Nonparametric Statistics*, 23, 1063–1074.
- Liu, F.T., Ting, K.M., and Zhou, Z. (2008). Isolation forest. In: *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 413–422.

Literature (mentioned in the tutorial) (3)

- López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486), 718–734.
- Mahalanobis P.C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 12, 49–55.
- Markou, M. and Singh, S. (2003). Novelty detection: a review - Part 1: Statistical approaches. *Signal Processing*, 83(12), 2481–2497.
- Markou, M. and Singh, S. (2003). Novelty detection: a review - Part 2: Neural network based approaches. *Signal Processing*, 83(12), 2499–2521.
- Miljković, D. (2010). Review of novelty detection methods. *The 33rd International Convention MIPRO*, Opatija, 593–598.
- Mosler, K. (2013). Depth statistics. In: *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*, 17-34.
- Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, 1, 327–332.
- Pimentel, M.A.F., Clifton, D.A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99, 215–249.
- Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), 1443–1471.

Literature (mentioned in the tutorial) (4)

- Serfling, R. (2002). A depth function and a scale curve based on spatial quantiles. In: *Statistical Data Analysis Based on the L_1-Norm and Related Methods*m Birkhäser, Basel, 25-38.
- Stahel W. (1981). *Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators (In German).* Ph.D. thesis, Swiss Federal Institute of Technology in Zurich.
- Tukey J.W. (1975). Mathematics and the picturing of data. In: *Proceedings of the International Congress of Mathematicians*, volume 2, Canadian Mathematical Congress, 523–531.
- Vapnik, V. and Chervonenkis, A. (1974). *Theory of Pattern Recognition* (in Russian). Nauka, Moscow.
- Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portraits. *Avtomatika i Telemekhanika*, 24, 774–780.
- Vardi Y., Zhang C. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 1423–1426.
- Zuo Y., Serfling R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28, 461–482.