Unsupervised learning: Anomaly detection
Part II: Functional data

Pavlo Mozharovskyi

LTCI, Telecom Paris, Institut Polytechnique de Paris

Parcours Data Science BPCE

Paris, the 13th of June 2023

Contents

Contents

Functional data framework

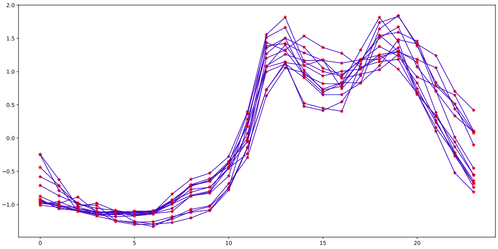
- Let $\boldsymbol{F} = \{\boldsymbol{F}(t) \in \mathbb{R}^d,\ t \in [0,1]\}$ be a random variable that takes its values in a (multivariate) functional space.

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

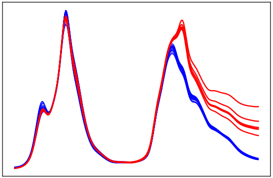
- Let $\boldsymbol{F} = \{\boldsymbol{F}(t) \in \mathbb{R}^d,\ t \in [0, 1]\}$ be a random variable that takes its values in a (multivariate) functional space.
- In practice, we only have access to the realization of $\boldsymbol{F}$ at a finite number of arguments/times, $\boldsymbol{f} = \{\boldsymbol{f}(t_1),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



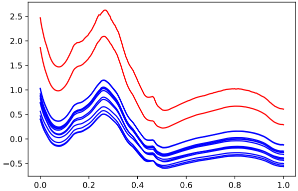
Taxonomy of functional anomalies (Hubert *et al.*, 2015)

A non-complete taxonomy of functional abnormalities:

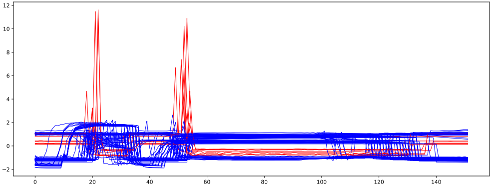
Shape anomalies



Shift anomalies



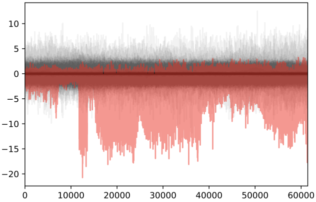
Isolated anomalies

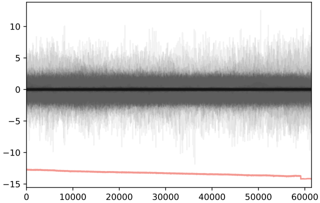


Taxonomy of functional anomalies (Airbus data)

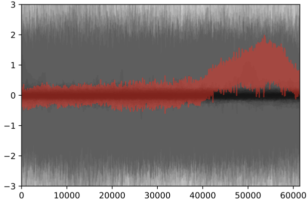
A non-complete taxonomy of functional abnormalities:

Magnitude (=location, shift) anomalies

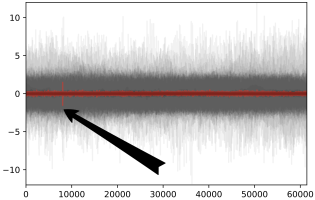




Shape anomalies

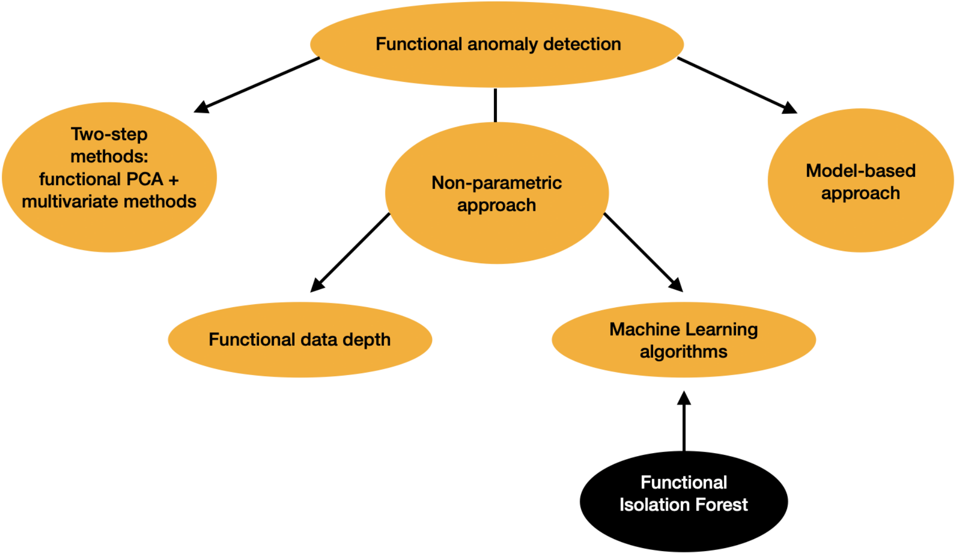


Isolated anomalies



Contents

FIF in the context of FAD contributions



Contents

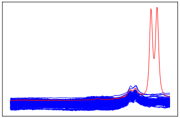
Functional Isolation Forest

- $X_1, \ldots, X_n$ are random variables in Hilbert space $\mathcal{H}$ and $\mathcal{D} \subset \mathcal{H}$.
- This ensemble learning algorithm builds a collection of binary tree based on a recursive and randomized tree-structured partitioning procedure.

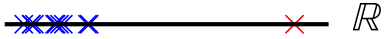
Functional Isolation Forest

- $X_1, \ldots, X_n$ are random variables in Hilbert space $\mathcal{H}$ and $\mathcal{D} \subset \mathcal{H}$.
- This ensemble learning algorithm builds a collection of binary tree based on a recursive and randomized tree-structured partitioning procedure.

Step 1:



Draw $\mathbf{d}$ from $\boldsymbol{\nu} \in \mathcal{P}(\mathcal{D})$ $\longrightarrow$

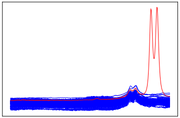


$\{\langle X_i, \mathbf{d} \rangle_{\mathcal{H}}, \ \ i \leq n\}$

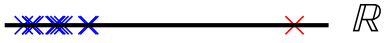
Functional Isolation Forest

- $X_1, \ldots, X_n$ are random variables in Hilbert space $\mathcal{H}$ and $\mathcal{D} \subset \mathcal{H}$.
- This ensemble learning algorithm builds a collection of binary tree based on a recursive and randomized tree-structured partitioning procedure.

Step 1:

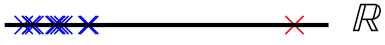


Draw $\mathbf{d}$ from $\boldsymbol{\nu} \in \mathcal{P}(\mathcal{D})$ $\longrightarrow$



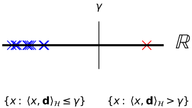
$\{\langle X_i, \mathbf{d} \rangle_{\mathcal{H}}, \ \ i \leq n\}$

Step 2:



$\{\langle X_i, \mathbf{d} \rangle_{\mathcal{H}}, \ \ i \leq n\}$

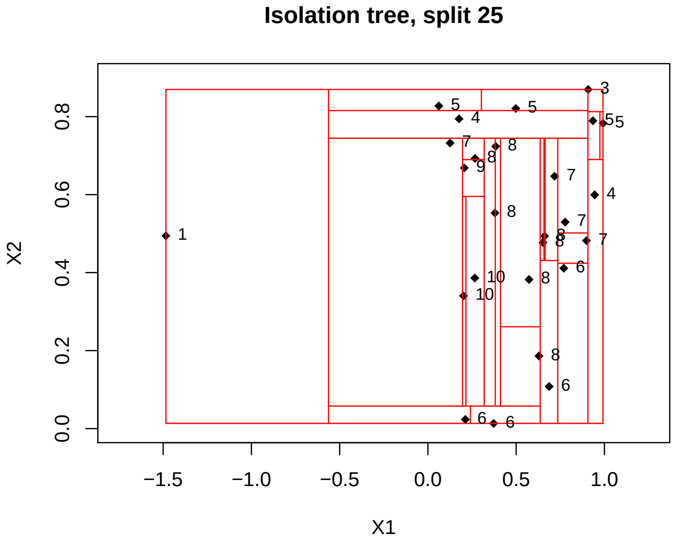
Draw uniformly γ $\longrightarrow$



- The trick: an anomaly should be isolated faster than normal data.

Functional Isolation Forest

Illustration: Isolation tree



Children node construction in a functional isolation tree

If a node (j, k) is non terminal, it is split in three steps as follows:

1. Choose a Split function $\mathbf{d}$ according to the probability distribution $\boldsymbol{\nu}$ on $\mathcal{D}$.

Children node construction in a functional isolation tree

If a node (j, k) is non terminal, it is split in three steps as follows:

1. Choose a Split function $\mathbf{d}$ according to the probability distribution $\boldsymbol{\nu}$ on $\mathcal{D}$.
2. Choose randomly and uniformly a Split value γ in the interval

$$\left[\min_{\mathbf{x}\in\mathcal{S}_{j,k}} \langle \mathbf{x}, \mathbf{d} \rangle_{\mathcal{H}}, \max_{\mathbf{x}\in\mathcal{S}_{j,k}} \langle \mathbf{x}, \mathbf{d} \rangle_{\mathcal{H}}\right],$$

Children node construction in a functional isolation tree

If a node (j, k) is non terminal, it is split in three steps as follows:

1. Choose a Split function $\mathbf{d}$ according to the probability distribution $\boldsymbol{\nu}$ on $\mathcal{D}$.
2. Choose randomly and uniformly a Split value γ in the interval

$$\left[\min_{\mathbf{x}\in\mathcal{S}_{j,k}} \langle \mathbf{x}, \mathbf{d}\rangle_{\mathcal{H}}, \max_{\mathbf{x}\in\mathcal{S}_{j,k}} \langle \mathbf{x}, \mathbf{d}\rangle_{\mathcal{H}}\right],$$

3. Form the children subsets

$$\begin{aligned} \mathcal{C}_{j+1,2k} &= \mathcal{C}_{j,k} \cap \{\mathbf{x} \in \mathcal{H} : \langle \mathbf{x}, \mathbf{d}\rangle_{\mathcal{H}} \leq \gamma\}, \\ \mathcal{C}_{j+1,2k+1} &= \mathcal{C}_{j,k} \cap \{\mathbf{x} \in \mathcal{H} : \langle \mathbf{x}, \mathbf{d}\rangle_{\mathcal{H}} > \gamma\}. \end{aligned}$$

as well as the children training datasets

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

Children node construction in a functional isolation tree

If a node (j, k) is non terminal, it is split in three steps as follows:

1. Choose a Split function $\mathbf{d}$ according to the probability distribution $\boldsymbol{\nu}$ on $\mathcal{D}$.
2. Choose randomly and uniformly a Split value γ in the interval

$$\left[\min_{\mathbf{x}\in\mathcal{S}_{j,k}} \langle \mathbf{x}, \mathbf{d}\rangle_{\mathcal{H}}, \max_{\mathbf{x}\in\mathcal{S}_{j,k}} \langle \mathbf{x}, \mathbf{d}\rangle_{\mathcal{H}}\right],$$

3. Form the children subsets

$$\begin{aligned} \mathcal{C}_{j+1,2k} &= \mathcal{C}_{j,k} \cap \{\mathbf{x} \in \mathcal{H} :\ \langle \mathbf{x}, \mathbf{d}\rangle_{\mathcal{H}} \leq \gamma\}, \\ \mathcal{C}_{j+1,2k+1} &= \mathcal{C}_{j,k} \cap \{\mathbf{x} \in \mathcal{H} :\ \langle \mathbf{x}, \mathbf{d}\rangle_{\mathcal{H}} > \gamma\}. \end{aligned}$$

   as well as the children training datasets

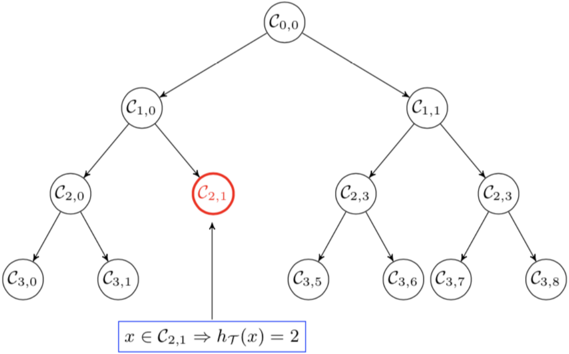
$$\mathcal{S}_{j+1,2k} = \mathcal{S}_{j,k} \cap \mathcal{C}_{j+1,2k} \text{ and } \mathcal{S}_{j+1,2k+1} = \mathcal{S}_{j,k} \cap \mathcal{C}_{j+1,2k+1}.$$

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

Anomaly score prediction

- One may then define the piecewise constant function $h_\tau : \mathcal{H} \to \mathbb{N}$ by: $\forall \boldsymbol{x} \in \mathcal{H}$,

  $h_\tau(\boldsymbol{x}) = j$ if and only if $x \in \mathcal{C}_{j,k}$ and $\mathcal{C}_{j,k}$ is associated to a terminal n



Anomaly score prediction

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

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

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

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

Anomaly score prediction

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

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

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

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

Contents

Parameters of FIF

- Classical parameters of Isolation Forest :
  - number of trees,
  - size of the subsample,
  - height limit.

- New parameters due to the functional setup :

  1. The dictionary $\mathcal{D}$.
  2. The probability measure $\boldsymbol{\nu}$.
  3. The scalar product $\langle ., . \rangle_{\mathcal{H}}$.

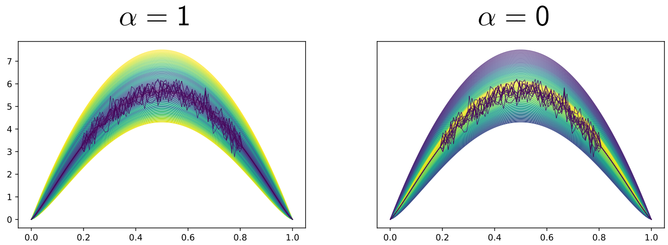
The role of the scalar product

- Compromise between both location and shape :

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

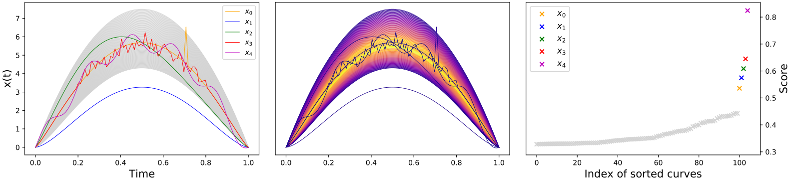
Example on a toy dataset :

- 90 curves defined by $\mathbf{x}(t) = 30(1 - t)^q t^q$ with q equispaced in $[1, 1.4]$,
- 10 *abnormal* curves defined by $\mathbf{x}(t) = 30(1 - t)^{1.2} t^{1.2}$ noised by $\varepsilon \sim \mathcal{N}(0, 0.3^2)$ on the interval $[0.2, 0.8]$.



Ability to detect a variety of anomalies

- Sobolev inner product: $\langle ., . \rangle_{W_{1,2}}$.
- Gaussian wavelets dictionary
  $\mathbf{d}_{\theta,\sigma}(t) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}} \left(1 - \left(\frac{t-\theta}{\sigma}\right)^2\right) \exp\left(\frac{-(t-\theta)^2}{2\sigma^2}\right)$.
- Uniform measure $\boldsymbol{\nu}$.



Contents

Performance on real datasets (1)

- FIF with 4 setups (Dictionary+scalar product):
  - Dyadic indicator (DI)+L_2
  - Cosine (Cos)+L_2
  - Cosine (Cos)+Sobolev
  - Dataset itself (Self)+L_2

  Competitors:

- *Isolation Forest*,*Local Outlier Factor*, *One-class SVM* after dimension reduction by FPCA.
- fHD_{RP} : Random projection method with functional Halspace depth.
- fSDO : Functional Stahel-Donoho Outlyingness.

Performance on real datasets (2)

Methods :	DI_{L_2}	Cos_{Sob}	Cos_{L_2}	$Self_{L_2}$	IF	LOF	OCSVM	fHD_{RP}	fSDO
Chinatown	0.93	0.82	0.74	0.77	0.69	0.68	0.70	0.76	0.98
Coffee	0.76	0.87	0.73	0.77	0.60	0.51	0.59	0.74	0.67
ECGFiveDays	0.78	0.75	0.81	0.56	0.81	0.89	0.90	0.60	0.81
ECG200	0.86	0.88	0.88	0.87	0.80	0.80	0.79	0.85	0.86
Handoutlines	0.73	0.76	0.73	0.72	0.68	0.61	0.71	0.73	0.76
SonyRobotAI1	0.89	0.80	0.85	0.83	0.79	0.69	0.74	0.83	0.94
SonyRobotAI2	0.77	0.75	0.79	0.92	0.86	0.78	0.80	0.86	0.81
StarLightCurves	0.82	0.81	0.76	0.86	0.76	0.72	0.77	0.77	0.85
TwoLeadECG	0.71	0.61	0.61	0.56	0.71	0.63	0.71	0.65	0.69
Yoga	0.62	0.54	0.60	0.58	0.57	0.52	0.59	0.55	0.55
EOGHorizontal	0.72	0.76	0.81	0.74	0.70	0.69	0.74	0.73	0.75
CinECGTorso	0.70	0.92	0.86	0.43	0.51	0.46	0.41	0.64	0.80
ECG5000	0.93	0.98	0.98	0.95	0.96	0.93	0.95	0.91	0.93

Table: AUC of different anomaly detection methods calculated on the test set. Bold numbers correspond to the best result.

Contents

Extension to multivariate functional data

FIF can be easily extended to the multivariate functional data, *i.e.* when the quantity of interest lies in $\mathbb{R}^d$ for each moment of time:

$$x : [0,1] \longrightarrow \mathbb{R}^d$$
$$t \longmapsto \Big((x^1(t), \; \ldots, \; x^d(t) \Big)$$

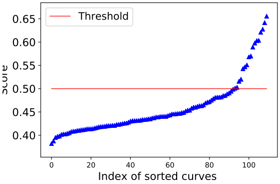
- Coordinate-wise sum of the d corresponding scalar products:

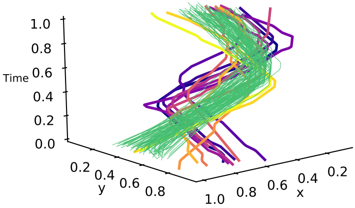
$$\langle \mathbf{f}, \mathbf{g} \rangle_{L_2^{\otimes d}} := \sum_{i=1}^{d} \langle f^{(i)}, g^{(i)} \rangle_{L_2}$$

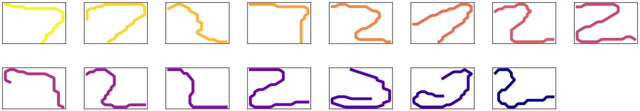
- Dictionaries : Composed by univariate function on each axis, multivariate wavelets, multivariate Brownian motion ...

Example with MNIST dataset

We extract the digits' contours and obtain bivariate functional curves from MNIST dataset. Each digit is transformed into a curve in $(L_2([0,1]) \times L_2([0,1]))$ using length parametrization on $[0,1]$.







Connection to data depth and supervised classification

- One may define a functional depth by $D_{FIF}(x;\mathcal{S}) = 1 - s_n(x;\mathcal{S})$.

Assume that we have a training classification dataset of q classes $\mathcal{S} = \mathcal{S}^1 \cup \ldots \cup \mathcal{S}^q$.

- Low dimensional representation based on depth-based map can be defined by

$$\mathbf{x} \mapsto \phi(\mathbf{x}) = \big(D_{FIF}(\mathbf{x};\mathcal{S}^1), \ldots, D_{FIF}(\mathbf{x};\mathcal{S}^q)\big) \in [0,1]^q \,.$$

- One may define a DD-plot classifier by using a classifier on the low dimension representation of the functional dataset.

Example of depth map on MNIST dataset

$\mathcal{S}$ is constructed by taking 100 digits from class 1, 100 from class 5 and 100 from class 7.

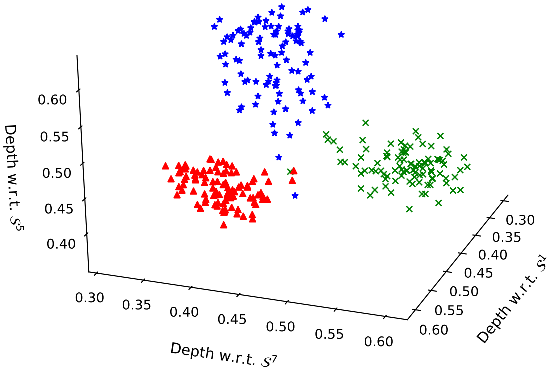


Figure: Depth space embedding of the three digits (1, 5 and 7) of the MNIST dataset.

Some remarks on FIF

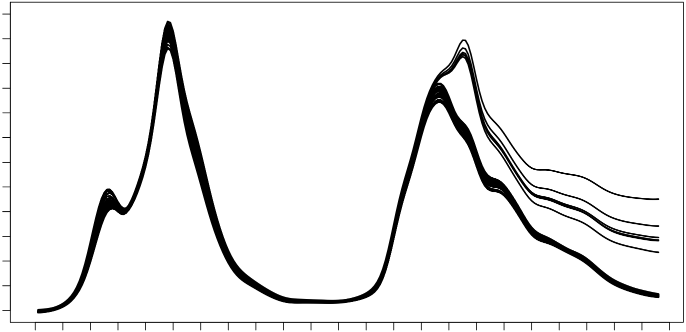
- New anomaly detection algorithm for functional data:
  - Great flexibility but dictionaries (and scalar product) are tricky to choose in an unsupervised setting.
  - Low complexity and memory requierement.
- Lack of theoretical garanties!

Staerman, G., Mozharovskyi, P., Clémençon, S., and d'Alché-Buc, F. **Functional Isolation Forest.** *ACML 2019.*

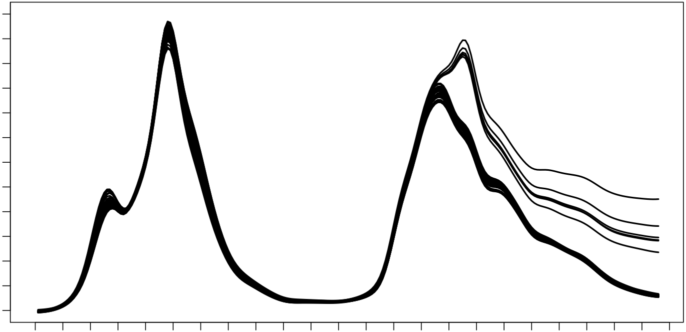
All codes are available at:
`https://github.com/guillaumestaermanML/FIF`.

Contents

Detection of (multivariate) functional anomalies



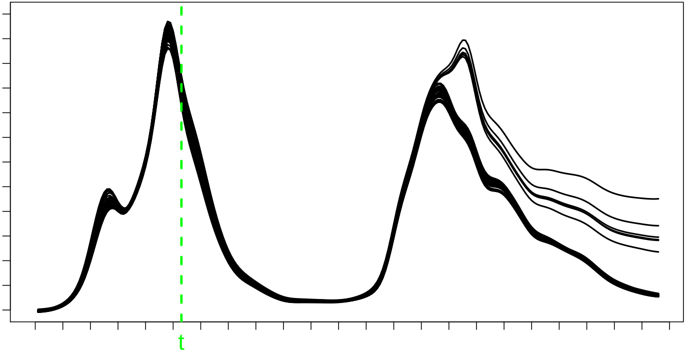
Detection of (multivariate) functional anomalies



- **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

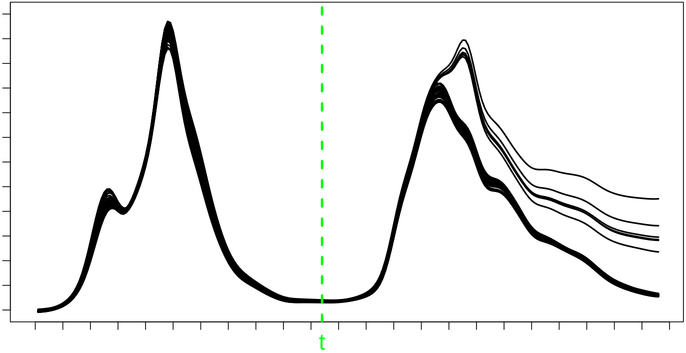


- **Functional depth** of $\boldsymbol{f}$ w.r.t. $\mathcal{F} = \{\boldsymbol{f}_i\}_{i=1}^n$:

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

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

Detection of (multivariate) functional anomalies

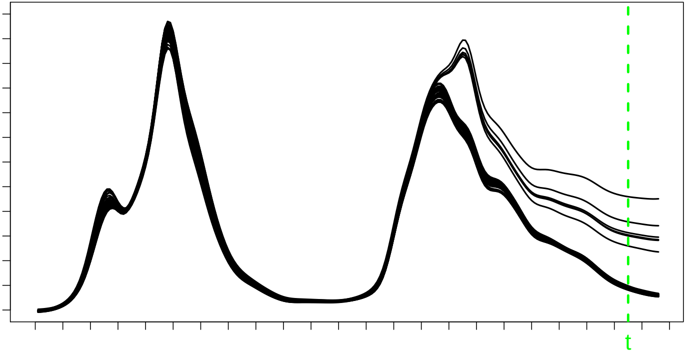


- **Functional depth** of $\boldsymbol{f}$ w.r.t. $\mathcal{F} = \{\boldsymbol{f}_i\}_{i=1}^n$:

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

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

Detection of (multivariate) functional anomalies

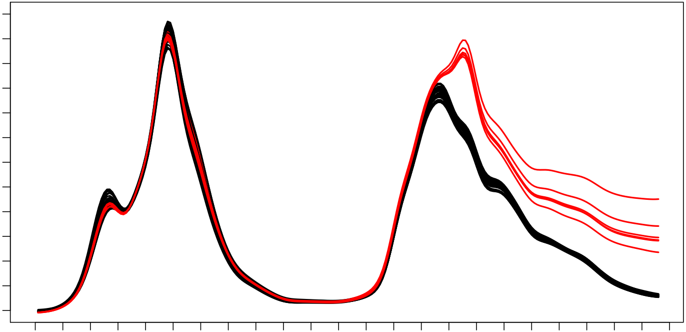


- **Functional depth** of $\boldsymbol{f}$ w.r.t. $\mathcal{F} = \{\boldsymbol{f}_i\}_{i=1}^n$:

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

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

Detection of (multivariate) functional anomalies

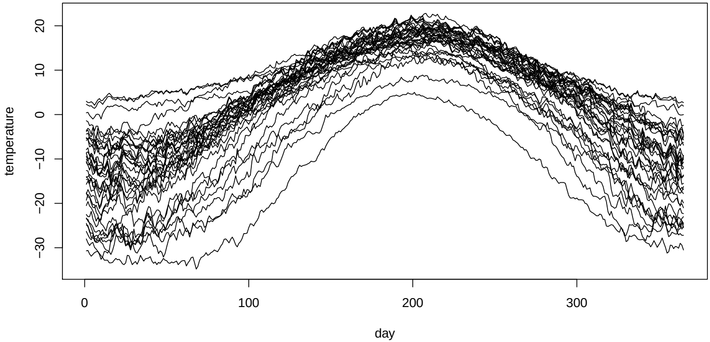


- **Functional depth** of $\boldsymbol{f}$ w.r.t. $\mathcal{F} = \{\boldsymbol{f}_i\}_{i=1}^n$:

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

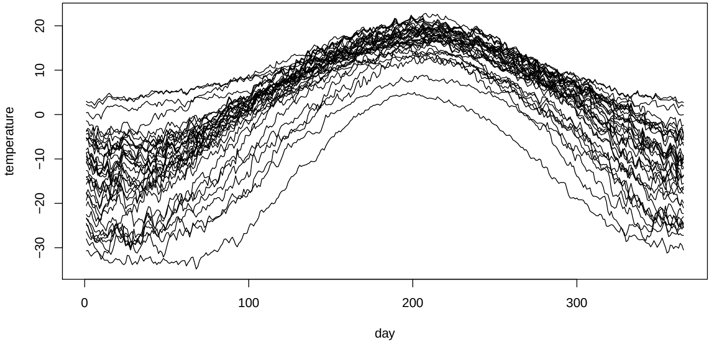
  where $D^d(\cdot|\cdot)$ is a multivariate data depth, as defined above.
- Label $\boldsymbol{f}$ as anomaly if $D(\boldsymbol{f}|\mathcal{F}) < \min(D)$.

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

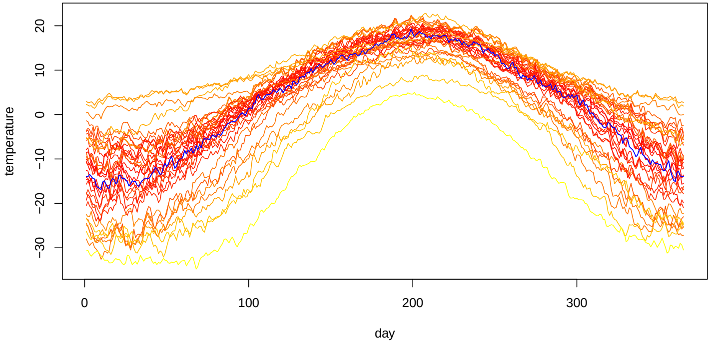


Let $\boldsymbol{F}$ be a stochastic process with continuous paths defined on $[0,1]$, and $\boldsymbol{f}$ its realization. Then:

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

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

Integrated depth for functional data

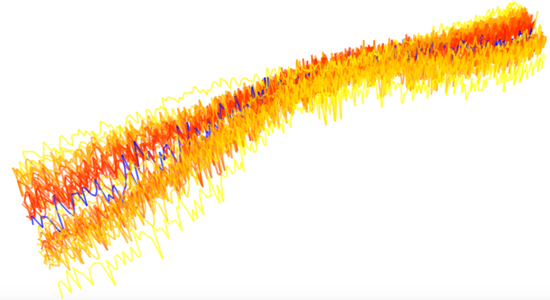


Let $\boldsymbol{F}$ be a stochastic process with continuous paths defined on $[0,1]$, and $\boldsymbol{f}$ its realization. Then:

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

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:

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

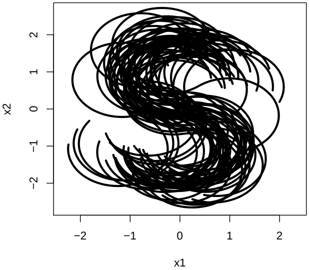
$$w(t) = w_\alpha\big(t, \boldsymbol{F}(t)\big) = \frac{\mathrm{vol}\big\{D_\alpha\big(\boldsymbol{F}(t)\big)\big\}}{\int_0^1 \mathrm{vol}\big\{D_\alpha\big(\boldsymbol{F}(u)\big)\big\}du}.$$

see Claeskens, Hubert, Slaets, Vakili, 2014.

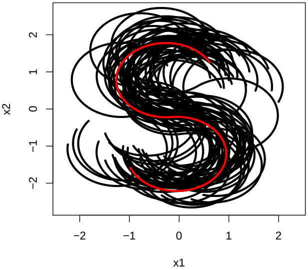
Contents

Contents

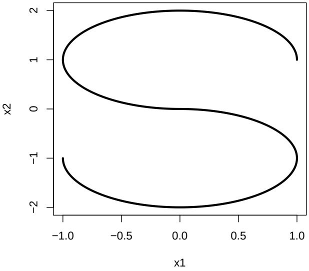
Functional depth: Motivation 1



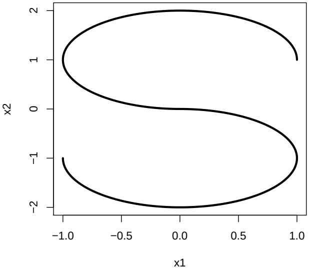
Functional depth: Motivation 1



Functional depth: Motivation 1



Functional depth: Motivation 1



Regard the following different parametrizations of a curve:

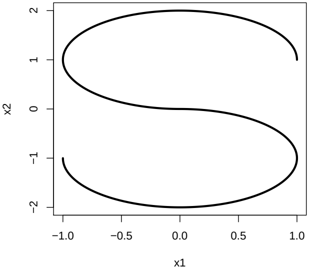
Parametrization A:

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

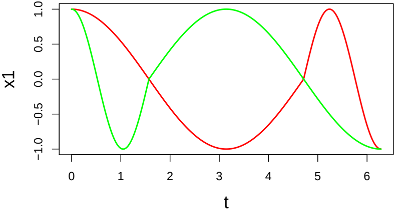
Parametrization B:

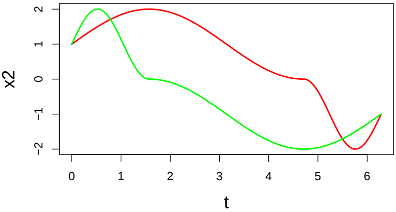
$$x_1(t) = -\big(cos(3t)+1\big)\mathbb{1}\{t < \tfrac{\pi}{2}\} - \big(cos(t+\pi)+1\big)\mathbb{1}\{t \geq \tfrac{\pi}{2}\}+1$$
$$x_2(t) = \quad \big(sin(3t)\ +1\big)\mathbb{1}\{t < \tfrac{\pi}{2}\} - \big(sin(t+\pi)\ +1\big)\mathbb{1}\{t \geq \tfrac{\pi}{2}\}$$

Functional depth: Motivation 1



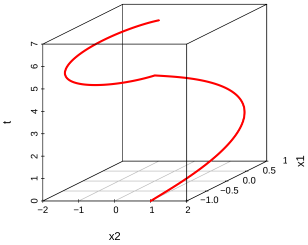
Regard the following different parametrizations of a curve:



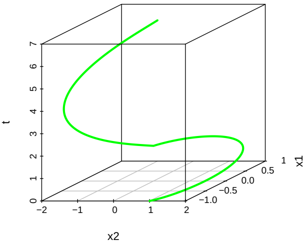


Functional depth: Motivation 1

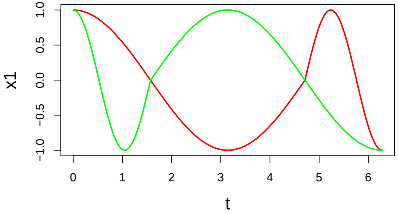
Parametrization A

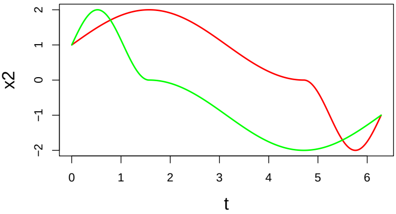


Parametrization B



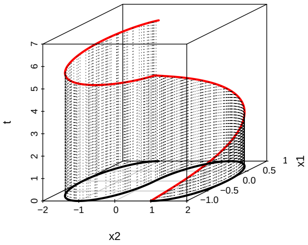
Parametrization:



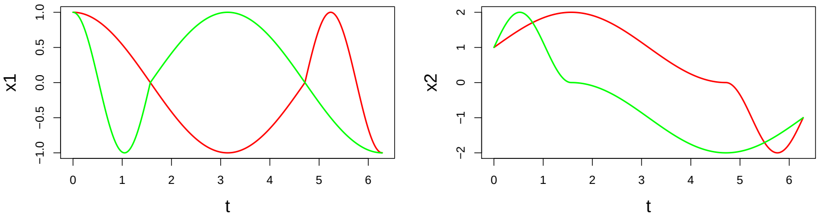


Functional depth: Motivation 1

Parametrization A Parametrization B



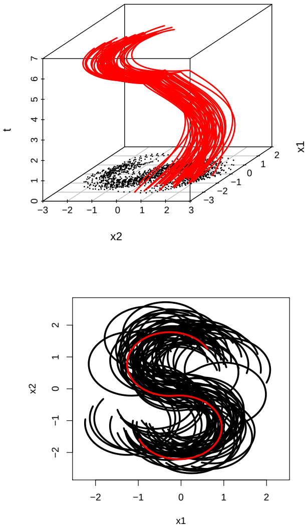
Parametrization:



Functional depth: Motivation 1

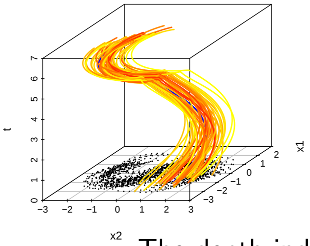
Parametrization A

Parametrization B

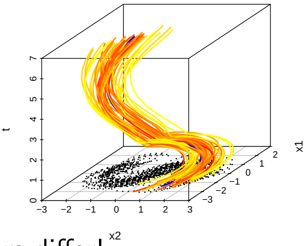


Functional depth: Motivation 1

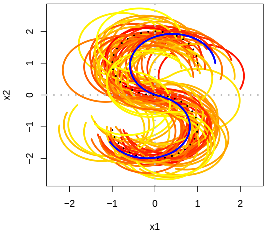
Parametrization A

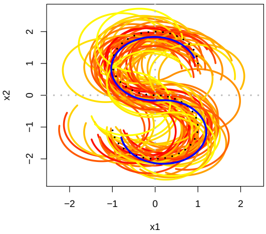


Parametrization B



The depth-induced orders differ!

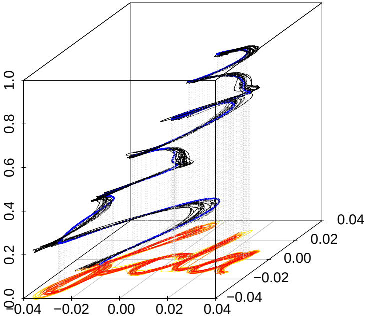




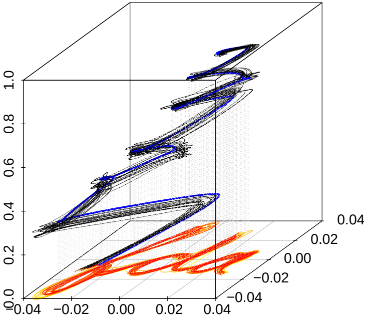
Functional depth: Motivation 2

Functional halfspace depth for the FDA-data

Parametrization by time



Parametrization by length



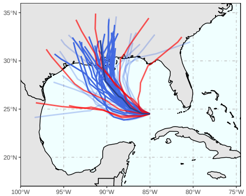
Depth-induced ranking for parametrizations by time and by length:

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

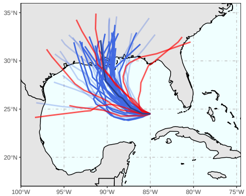
Functional depth: Motivation 3

Simulated hurricane tracks: **curve boxplot**

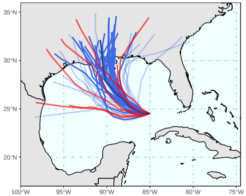
MFH depth – par. time



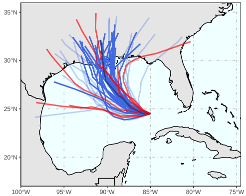
MFH depth - par. length



mSB depth – par. time



mSB depth – par. length



Contents

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 $\gamma_1, \gamma_2 : \beta_1 \circ \gamma_1 = \beta_2 \circ \gamma_2$.

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*:

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

Associated distribution and the sampling scheme

- n

Associated distribution and the sampling scheme

- n
- 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

- n
- 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

- n
- 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

- n
- 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:

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

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

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:

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

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:

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

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

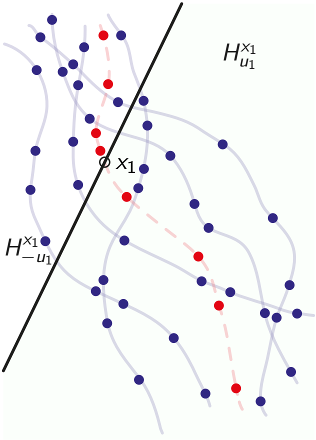
Definition

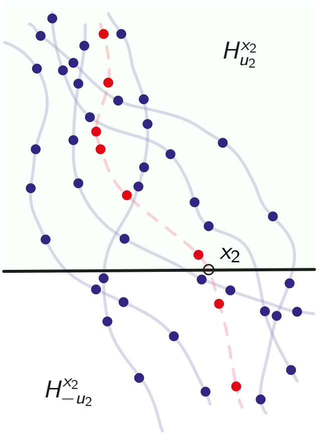
The **sample Tukey curve depth** of $\mathcal{C} \in \mathfrak{B}$ w.r.t. $\mathcal{X}_1, ..., \mathcal{X}_n$ is:

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

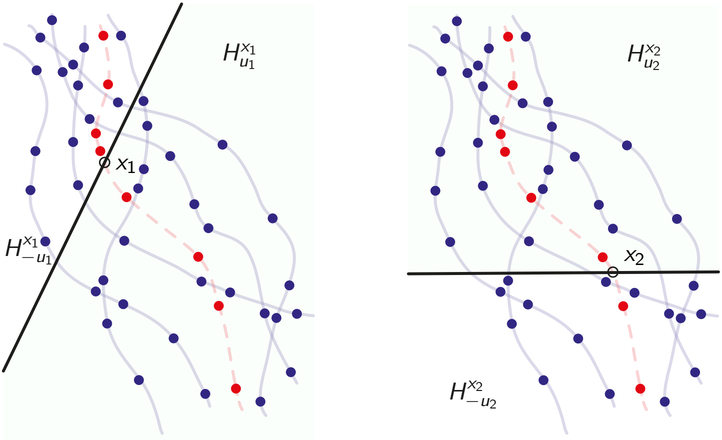
where $Q_n = (\mu_{\mathcal{X}_1} + \cdots + \mu_{\mathcal{X}_n})/n$.

Data depth for an unparametrized curve: intuition





Data depth for an unparametrized curve: intuition



Traditional reasoning:

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

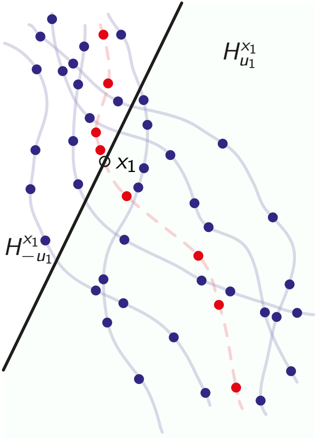
$\widehat{Q}_P(H_{-u_1}^{x_1}) = \frac{15}{40}$, $\widehat{\mu}_C(H_{-u_1}^{x_1}) = \frac{4}{8}$

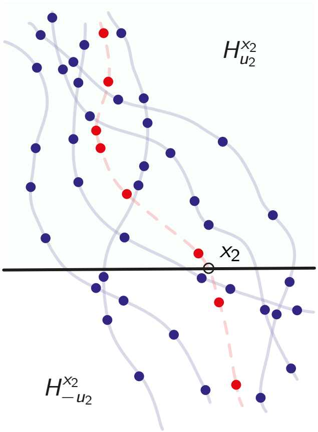
Curve-based reasoning:

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

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

Data depth for an unparametrized curve: intuition





Traditional reasoning:

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

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

Curve-based reasoning:

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

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

Data depth for an unparametrized curve: 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) $\mathbb{X}_{n,m} = \{X_{i,j},\ i = 1, ..., n,\ j = 1, ..., m\}$

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

- 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. $\mathcal{X}_1, \ldots, \mathcal{X}_n$ is defined as:

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

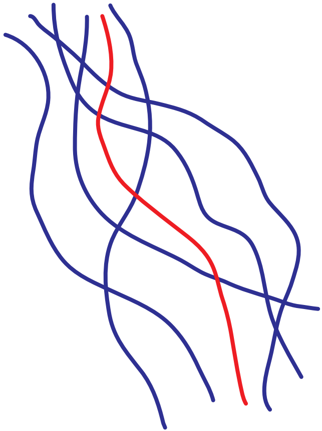
with the depth of an arbitrary point $\boldsymbol{x} \in \mathbb{R}^d$ w.r.t. $\widehat{Q}_{n,m}$ being:

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

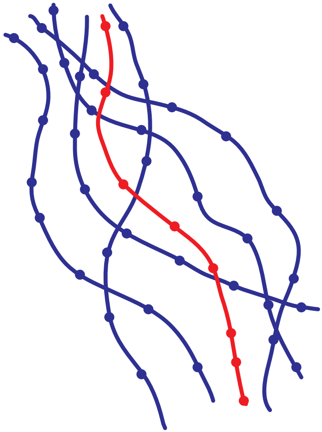
and $\frac{0}{0} = 0$ as before.

Contents

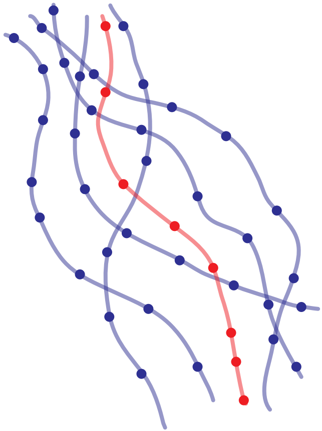
Calculation of the Tukey curve depth



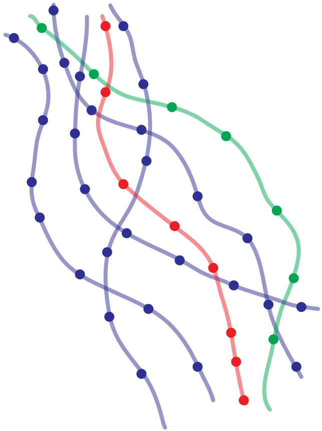
Calculation of the Tukey curve depth



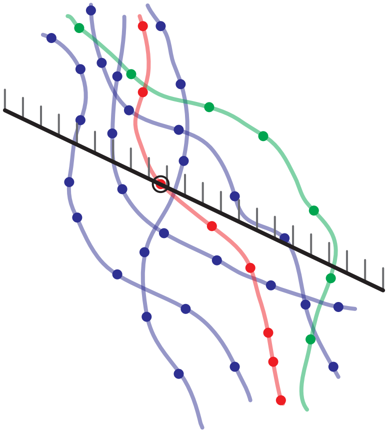
Calculation of the Tukey curve depth



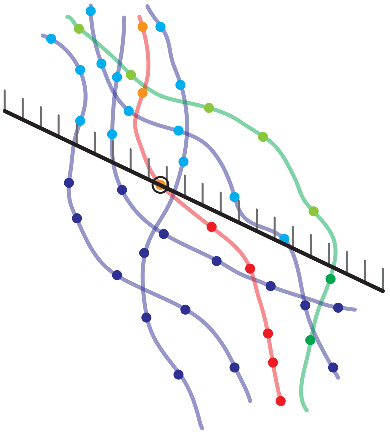
Calculation of the Tukey curve depth



Calculation of the Tukey curve depth

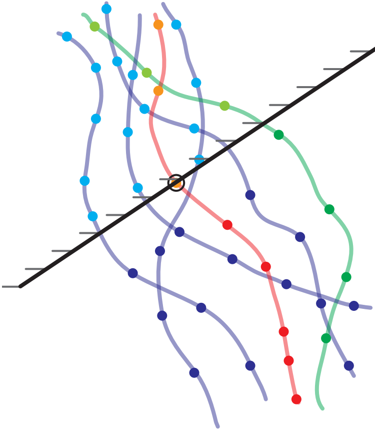


Calculation of the Tukey curve depth



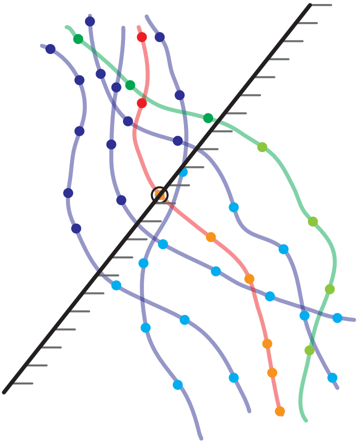
$$D(\mathbb{Y}_{2,c}|Q_m, \mathcal{H}_{m,b}) = \frac{\frac{1}{5}\left(\frac{5}{7}+\frac{3}{8}+\frac{6}{8}+\frac{2}{7}+\frac{3}{6}\right)}{\frac{2}{8}} = 2.1$$

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: consistency

Theorem

Let $\mathcal{C} \in \mathfrak{B}$ be a rectifiable curve, and let P be a probability measure in the space of curves such that $P \in \mathcal{P}$. Let (Δ_m) be a decreasing sequence of positive numbers such that (Δ_m) and $(\sqrt{\frac{\log(m)}{m}}/\Delta_m^2)$ converges to zero when $m \to \infty$.

Data depth for an unparametrized curve: consistency

Theorem

Let $\mathcal{C} \in \mathfrak{B}$ be a rectifiable curve, and let P be a probability measure in the space of curves such that $P \in \mathcal{P}$. Let (Δ_m) be a decreasing sequence of positive numbers such that (Δ_m) and $(\sqrt{\frac{\log(m)}{m}}/\Delta_m^2)$ converges to zero when $m \to \infty$.

Then:

- *the Monte Carlo approximation $\widehat{D}_{n,m,\Delta_m}(\mathcal{C}|\mathcal{X}_1, ..., \mathcal{X}_n)$ converges in probability to $D(\mathcal{C}|\mathcal{X}_1, \ldots, \mathcal{X}_n)$ as $m \to \infty$;*

Data depth for an unparametrized curve: consistency

Theorem

Let $\mathcal{C} \in \mathfrak{B}$ be a rectifiable curve, and let P be a probability measure in the space of curves such that $P \in \mathcal{P}$. Let (Δ_m) be a decreasing sequence of positive numbers such that (Δ_m) and $(\sqrt{\frac{\log(m)}{m}}/\Delta_m^2)$ converges to zero when $m \to \infty$.

Then:

- *the Monte Carlo approximation $\widehat{D}_{n,m,\Delta_m}(\mathcal{C}|\mathcal{X}_1, ..., \mathcal{X}_n)$ converges in probability to $D(\mathcal{C}|\mathcal{X}_1, \ldots, \mathcal{X}_n)$ as $m \to \infty$;*
- *the Monte Carlo approximation $\widehat{D}_{n,m,\Delta_m}(\mathcal{C}|\mathcal{X}_1, ..., \mathcal{X}_n)$ converges in probability to $D(\mathcal{C}|P)$ as $m, n \to \infty$;*

Data depth for an unparametrized curve: consistency

Theorem

Let $\mathcal{C} \in \mathfrak{B}$ be a rectifiable curve, and let P be a probability measure in the space of curves such that $P \in \mathcal{P}$. Let (Δ_m) be a decreasing sequence of positive numbers such that (Δ_m) and $(\sqrt{\frac{\log(m)}{m}}/\Delta_m^2)$ converges to zero when $m \to \infty$.

Then:

- *the Monte Carlo approximation $\widehat{D}_{n,m,\Delta_m}(\mathcal{C}|\mathcal{X}_1, ..., \mathcal{X}_n)$ converges in probability to $D(\mathcal{C}|\mathcal{X}_1, \ldots, \mathcal{X}_n)$ as $m \to \infty$;*
- *the Monte Carlo approximation $\widehat{D}_{n,m,\Delta_m}(\mathcal{C}|\mathcal{X}_1, ..., \mathcal{X}_n)$ converges in probability to $D(\mathcal{C}|P)$ as $m, n \to \infty$;*
- *the sample Tukey curve depth $D(\mathcal{C}|\mathcal{X}_1, ..., \mathcal{X}_n)$ converges in probability to $D(\mathcal{C}|P)$ as $n \to \infty$.*

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

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

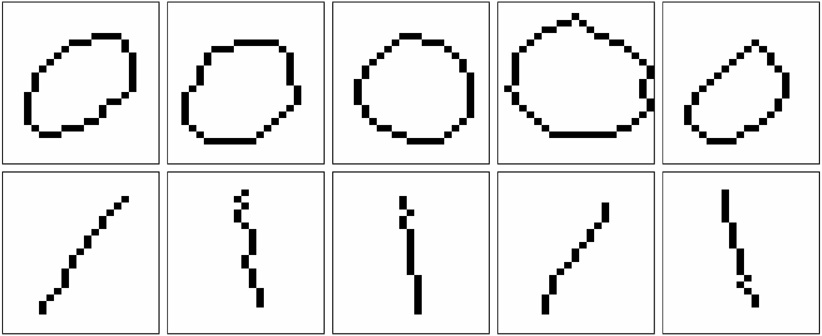
- **Vanishing at infinity:**

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

Contents

Binary supervised classification: MNIST (“0” *vs* “1”)

Some examples:



Given: training sample $\mathcal{S}_0 = \{\mathcal{C}_1, ..., \mathcal{C}_m\}$ stemming from P_0 and $\mathcal{S}_1 = \{\mathcal{C}_{m+1}, ..., \mathcal{C}_{m+n}\}$ stemming from P_1 in $\mathfrak{B}$.

Find: classifier $g : \mathfrak{B} \to \{0, 1\}$ best separating P_0 and P_1.

Binary supervised classification: MNIST (“0” *vs* “1”)

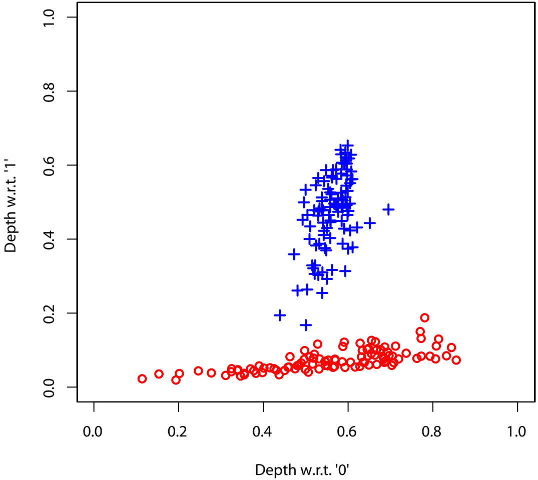
Consider *DD*-plot (Li, Cuesta-Albertos, Liu ’12):

$$\boldsymbol{Z} = \{\boldsymbol{z}_i \,:\, \boldsymbol{z}_i = \big(D(\mathcal{C}_i|Q_{P_0}), D(\mathcal{C}_i|Q_{P_1})\big),\, i = 1, ..., m+n\}\,.$$

Binary supervised classification: MNIST (“0” *vs* “1”)

Consider *DD*-plot (Li, Cuesta-Albertos, Liu ’12):

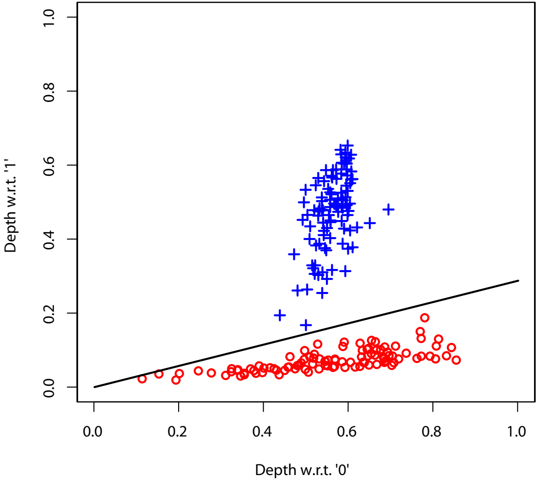
$$\boldsymbol{Z} = \{\boldsymbol{z}_i \, : \, \boldsymbol{z}_i = \big(D(\mathcal{C}_i|Q_{P_0}), D(\mathcal{C}_i|Q_{P_1})\big), \; i = 1, \ldots, m+n\} \, .$$



Binary supervised classification: MNIST (“0” *vs* “1”)

Consider *DD*-plot (Li, Cuesta-Albertos, Liu ’12):

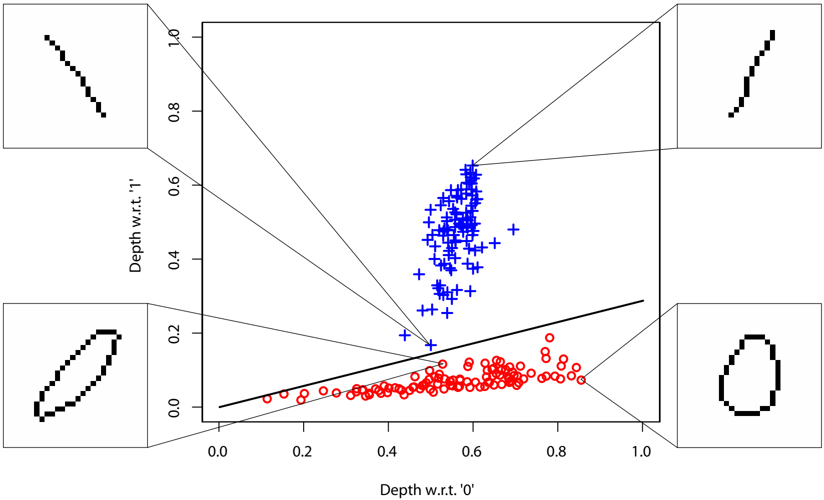
$$\boldsymbol{Z} = \{\boldsymbol{z}_i \,:\, \boldsymbol{z}_i = \big(D(\mathcal{C}_i|Q_{P_0}), D(\mathcal{C}_i|Q_{P_1})\big),\; i = 1, \ldots, m+n\}\,.$$



Binary supervised classification: MNIST (“0” *vs* “1”)

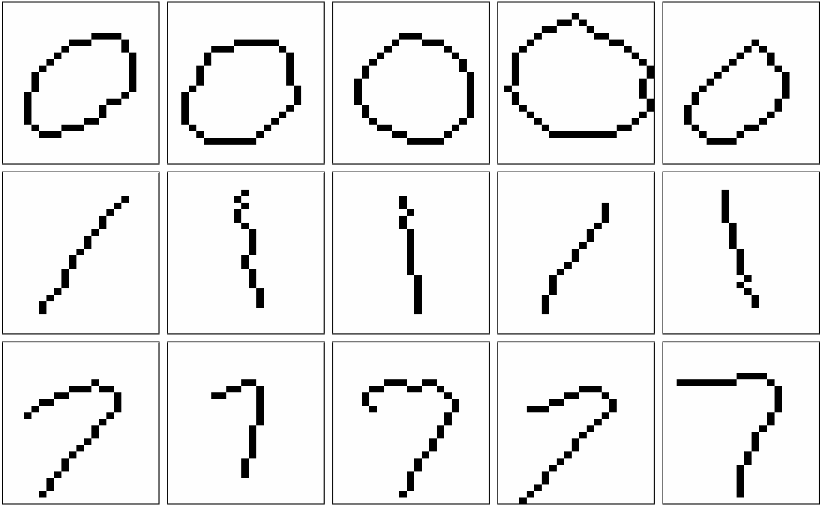
Consider *DD*-plot (Li, Cuesta-Albertos, Liu ’12):

$$\boldsymbol{Z} = \{\boldsymbol{z}_i \,:\, \boldsymbol{z}_i = \big(D(\mathcal{C}_i|Q_{P_0}), D(\mathcal{C}_i|Q_{P_1})\big),\ i = 1, ..., m+n\}\,.$$



Unsupervised classification: MNIST (“0”, “1”, and “7”)

Some examples:



Task: Find reasonable grouping with data depth (Jörnsten ’04).

Unsupervised classification: MNIST (“0”, “1”, and “7”)

Depth-based clustering (Jörnsten ’04):

Unsupervised classification: MNIST ("0", "1", and "7")

Depth-based clustering (Jörnsten '04):

- Let $\{\mathcal{C}_1, ..., \mathcal{C}_{\sum_j n_j}\}$ be the observed sample and let I_j, $j = 1, ..., J$ denote the corresponding partitioning into J clusters (indices of observations belonging to each cluster j) with $\cup_j I_j = \{1, ..., \sum_j n_j\}$ and $I_{j_1} \cap I_{j_2} = \emptyset$ for all $j_1 \neq j_2$.

Unsupervised classification: MNIST ("0", "1", and "7")

Depth-based clustering (Jörnsten '04):

- Let $\{\mathcal{C}_1, ..., \mathcal{C}_{\sum_j n_j}\}$ be the observed sample and let I_j, $j = 1, ..., J$ denote the corresponding partitioning into J clusters (indices of observations belonging to each cluster j) with $\cup_j I_j = \{1, ..., \sum_j n_j\}$ and $I_{j_1} \cap I_{j_2} = \emptyset$ for all $j_1 \neq j_2$.
- Define the **silhouette width** of an observation i belonging to cluster j as

$$Sil_i^j = \frac{\bar{d}_i^{-j} - \bar{d}_i^j}{\max\{\bar{d}_i^{-j}\,,\, \bar{d}_i^j\}}\,,$$

  where $\bar{d}_i^j = \frac{1}{\#I_j - 1} \sum_{i' \in I_j\,,\, i' \neq i} d_{\mathfrak{B}}(\mathcal{C}_i, \mathcal{C}_{i'})$ and $\bar{d}_i^{-j} \in \text{argmin}_{j' \neq j} \frac{1}{\#I_{j'}} \sum_{i' \in I_{j'}} d_{\mathfrak{B}}(\mathcal{C}_i, \mathcal{C}_{i'})$ are average distances to the observations in its own cluster and in the closest among foreign clusters respectively.

Unsupervised classification: MNIST (“0”, “1”, and “7”)

Depth-based clustering (Jörnsten ’04):

- Let $\{\mathcal{C}_1, \ldots, \mathcal{C}_{\sum_j n_j}\}$ be the observed sample and let I_j, $j = 1, \ldots, J$ denote the corresponding partitioning into J clusters (indices of observations belonging to each cluster j) with $\cup_j I_j = \{1, \ldots, \sum_j n_j\}$ and $I_{j_1} \cap I_{j_2} = \emptyset$ for all $j_1 \neq j_2$.
- Define the **silhouette width** of an observation i belonging to cluster j as

$$Sil_i^j = \frac{\bar{d}_i^{-j} - \bar{d}_i^j}{\max\{\bar{d}_i^{-j}\,,\,\bar{d}_i^j\}}\,,$$

  where $\bar{d}_i^j = \frac{1}{\#I_j - 1}\sum_{i' \in I_j\,,\,i' \neq i} d_{\mathfrak{B}}(\mathcal{C}_i, \mathcal{C}_{i'})$ and $\bar{d}_i^{-j} \in \mathrm{argmin}_{j' \neq j}\,\frac{1}{\#I_{j'}}\sum_{i' \in I_{j'}} d_{\mathfrak{B}}(\mathcal{C}_i, \mathcal{C}_{i'})$ are average distances to the observations in its own cluster and in the closest among foreign clusters respectively.
- The **relative depth** is defined as

$$ReD_i^j = D(\mathcal{C}_i | \{\mathcal{C}_{i'}\}_{i' \in I_j}) - \max_{j' \neq j} D(\mathcal{C}_i | \{\mathcal{C}_{i'}\}_{i' \in I_{j'}})\,.$$

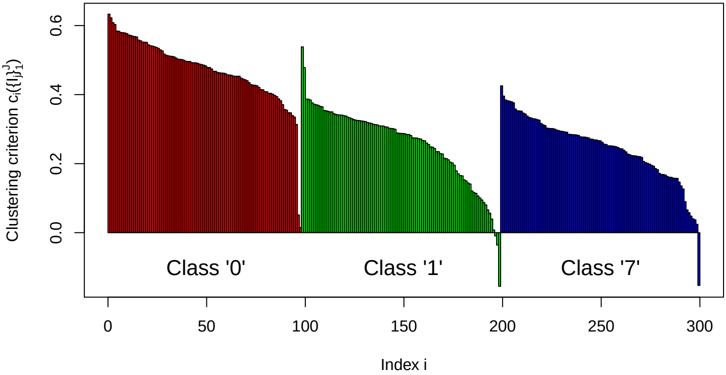
Unsupervised classification: MNIST (“0”, “1”, and “7”)

Clustering criterion:

$$C(\{I_j\}_1^J) = \frac{1}{\sum_j n_j} \sum_{j=1}^{J} \sum_{i \in I_j} c_i(\{I_j\}_1^J),$$

with the **observation-wise clustering criterion:**

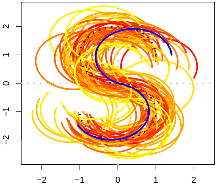
$$c_i(\{I_j\}_1^J) = (1-\lambda) Sil_i^j + \lambda ReD_i^j.$$



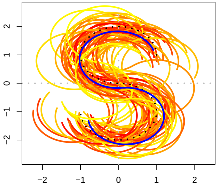
Comparison with functional depth: Example 1

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

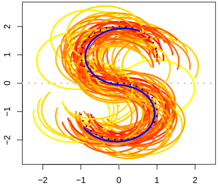
MFHD – time



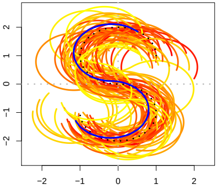
MFHD - length



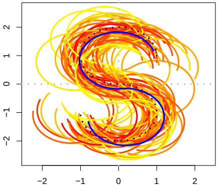
Curve depth



mSBD – time



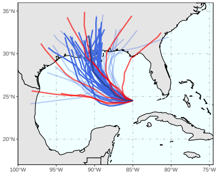
mSBD – length



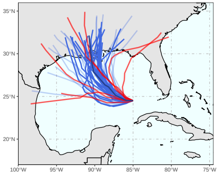
Comparison with functional depth: Example 2

Simulated hurricane tracks: **curve boxplot**

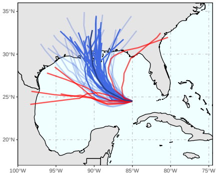
MFHD – time



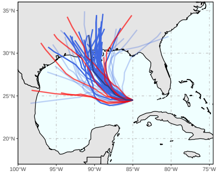
MFHD - length



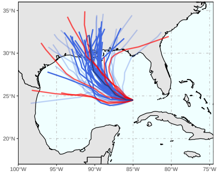
Curve depth



mSBD – time

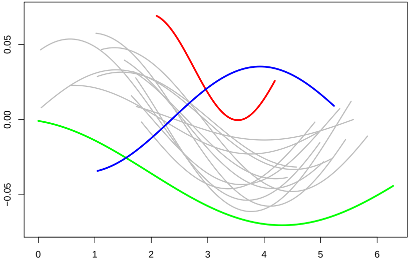


mSBD – length

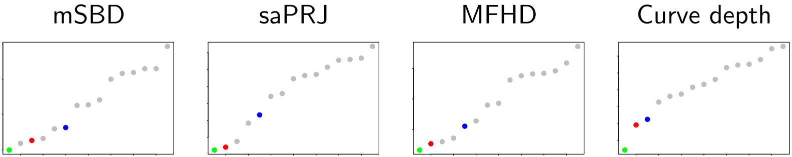


Comparison with functional depth: Anomaly detection 1

Data set 1 with introduced anomalies:

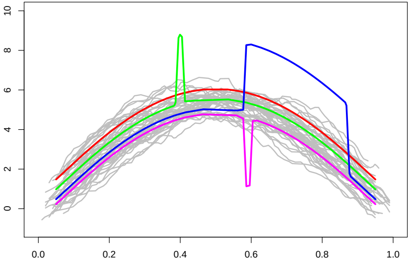


Ordered depth values:



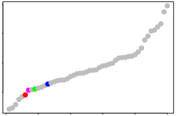
Comparison with functional depth: Anomaly detection 2

Data set 2 with introduced anomalies:

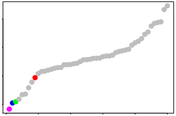


Ordered depth values:

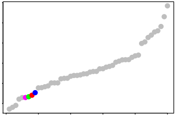
mSBD



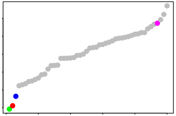
saPRJ



MFHD



Curve depth



Contents

Application: brain imaging – OATS data

- The **Older Australian Twins Study** (OATS) includes *diffusion tensor magnetic resonance images* (DTI) of 34 twin pairs: 11 dizygotic (DZ) and 23 monozygotic (MZ).

Application: brain imaging – OATS data

- The **Older Australian Twins Study** (OATS) includes *diffusion tensor magnetic resonance images* (DTI) of 34 twin pairs: 11 dizygotic (DZ) and 23 monozygotic (MZ).
- For each individual, **1000 fiber tracts** connecting the motor cortex with the brain stem extracted for each hemisphere.

Application: brain imaging – OATS data

- The **Older Australian Twins Study** (OATS) includes *diffusion tensor magnetic resonance images* (DTI) of 34 twin pairs: 11 dizygotic (DZ) and 23 monozygotic (MZ).
- For each individual, **1000 fiber tracts** connecting the motor cortex with the brain stem extracted for each hemisphere.

Questions to answer:

- **Information compression** for better understanding of brain functioning.

Application: brain imaging – OATS data

- The **Older Australian Twins Study** (OATS) includes *diffusion tensor magnetic resonance images* (DTI) of 34 twin pairs: 11 dizygotic (DZ) and 23 monozygotic (MZ).
- For each individual, **1000 fiber tracts** connecting the motor cortex with the brain stem extracted for each hemisphere.

Questions to answer:

- **Information compression** for better understanding of brain functioning.
- **Outlier detection** for indication of wrongly tracked fibers.

Application: brain imaging – OATS data

- The **Older Australian Twins Study** (OATS) includes *diffusion tensor magnetic resonance images* (DTI) of 34 twin pairs: 11 dizygotic (DZ) and 23 monozygotic (MZ).
- For each individual, **1000 fiber tracts** connecting the motor cortex with the brain stem extracted for each hemisphere.

Questions to answer:

- **Information compression** for better understanding of brain functioning.
- **Outlier detection** for indication of wrongly tracked fibers.
- **Curve registration** for aligning data from different individuals before further analysis.

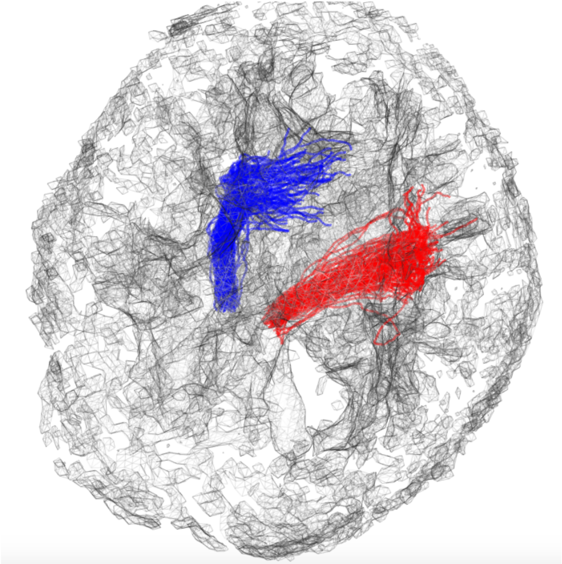
Application: brain imaging – OATS data

- The **Older Australian Twins Study** (OATS) includes *diffusion tensor magnetic resonance images* (DTI) of 34 twin pairs: 11 dizygotic (DZ) and 23 monozygotic (MZ).
- For each individual, **1000 fiber tracts** connecting the motor cortex with the brain stem extracted for each hemisphere.

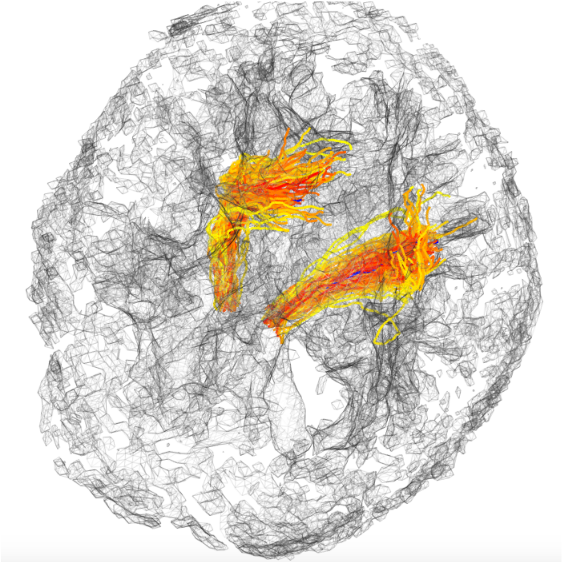
Questions to answer:

- **Information compression** for better understanding of brain functioning.
- **Outlier detection** for indication of wrongly tracked fibers.
- **Curve registration** for aligning data from different individuals before further analysis.
- Studying **genetic dependency** (DZ *vs.* MZ) for identifying disease causes.

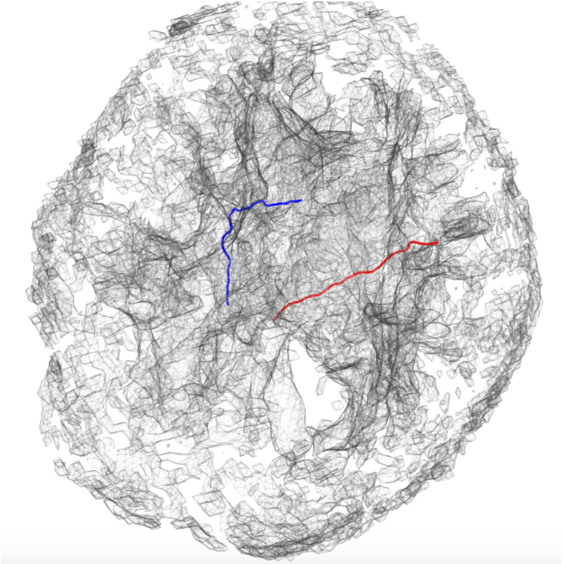
Application: brain imaging



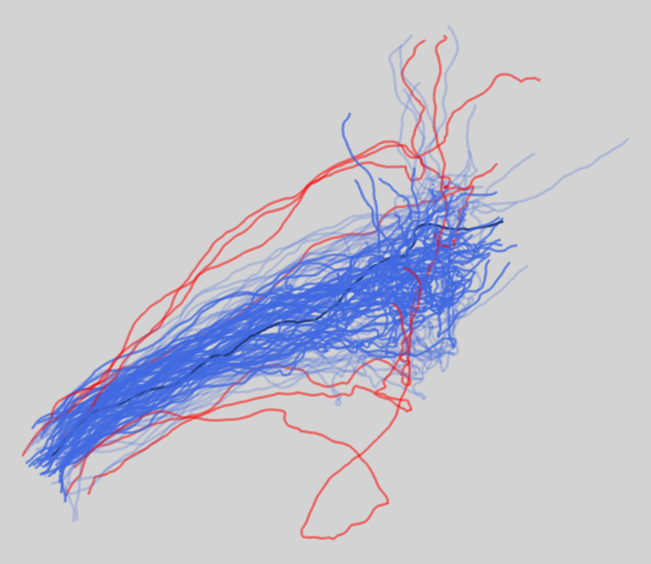
Application: brain imaging – depth-based ordering



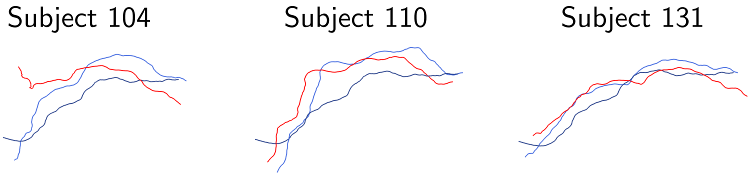
Application: brain imaging – information compression



Application: brain imaging, right stem – outlier detection

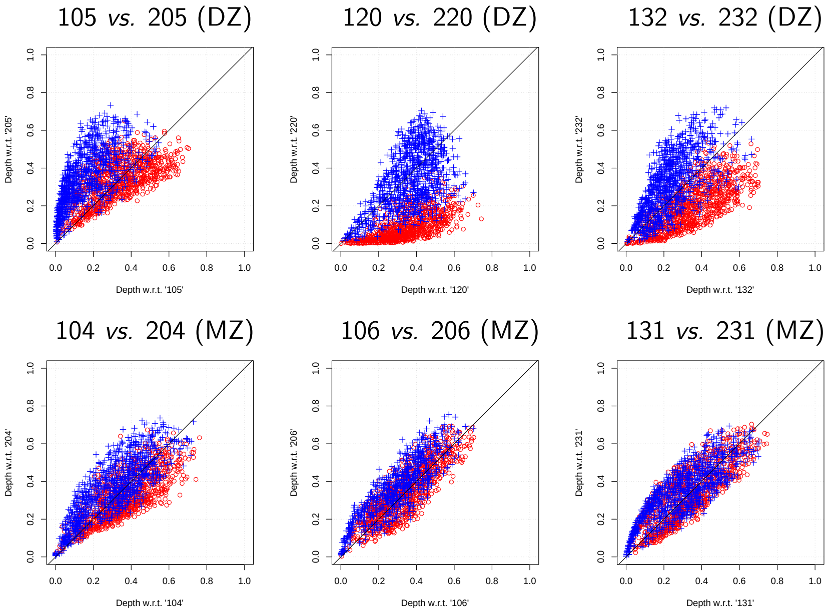


Application: brain imaging, right stem – registration



- The red and the dark blue curves are respectively the **deepest** curves before registration of the respective subject and subject **235**, the subject whose deepest curve is the **deepest of all**.
- We **bring** the red curve **as close as possible** (in terms of the *distance*) to the black curve. The transformed curve (after registration) is the light blue curve.
- **Distances** from each curve to the deepest one (dark blue) before (red) and after (light blue) registration are 10.271 and 3.245 (for subject 104), 4.539 and 3.395 (for subject 110), 3.329 and 2.084 (for subject 131), respectively.

Application: brain imaging, right stem – twins comparison



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 (part II)

Notebooks:

- anomdet_simulation1.Rmd,
- anomdet_hurricanes.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.

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.

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.