Functional Anomaly Detection and Robust Estimation

PhD Defense, Guillaume Staerman

LTCI, Télécom Paris, Institut Polytechnique de Paris

April 12th, 2022

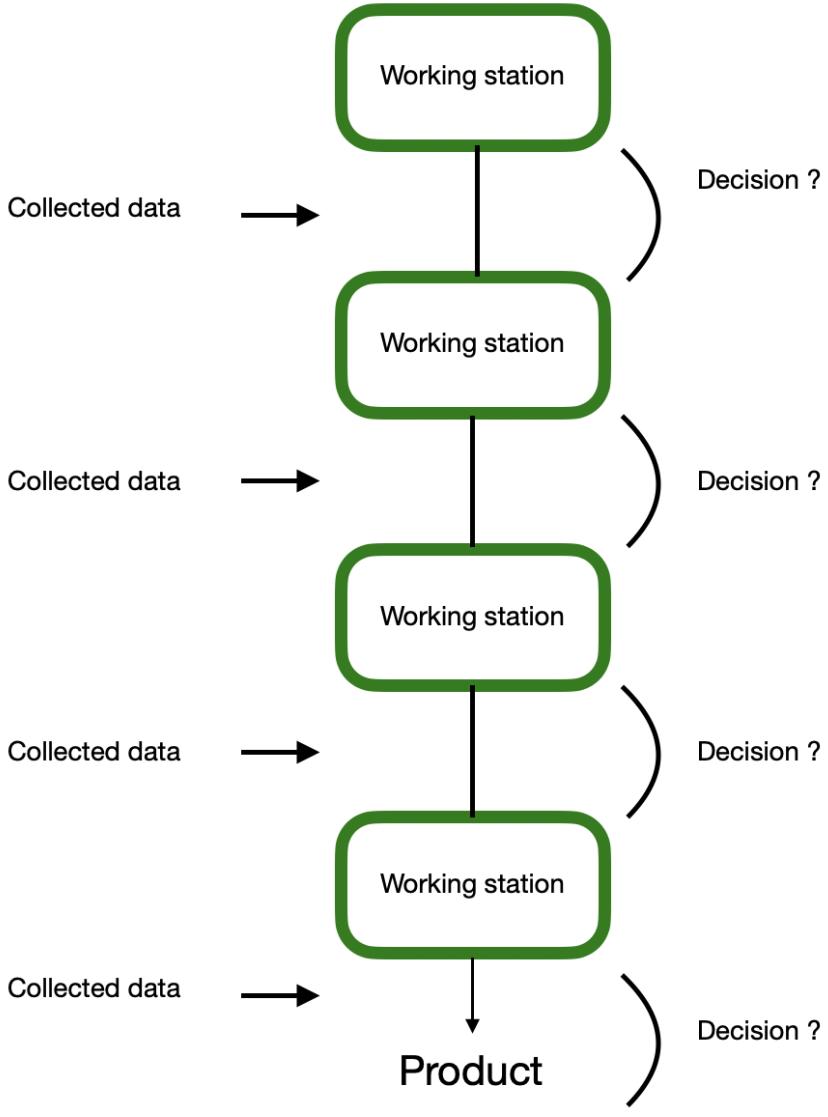
Motivation: increasing amount of data

- IoT
- Distributed platforms
- Industry 4.0

Real-world data:

- incomplete, inconsistent, duplicate, noisy, **contaminated**

Motivation: contaminated data and production line

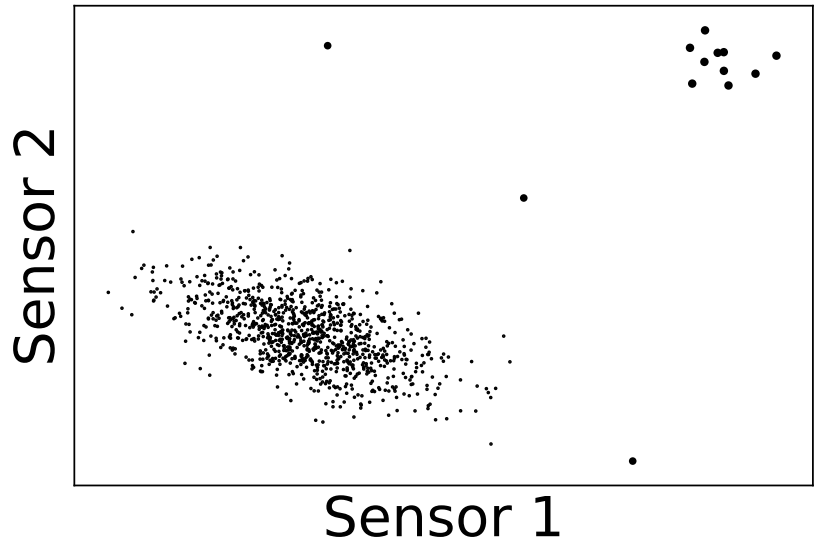


Robust Learning

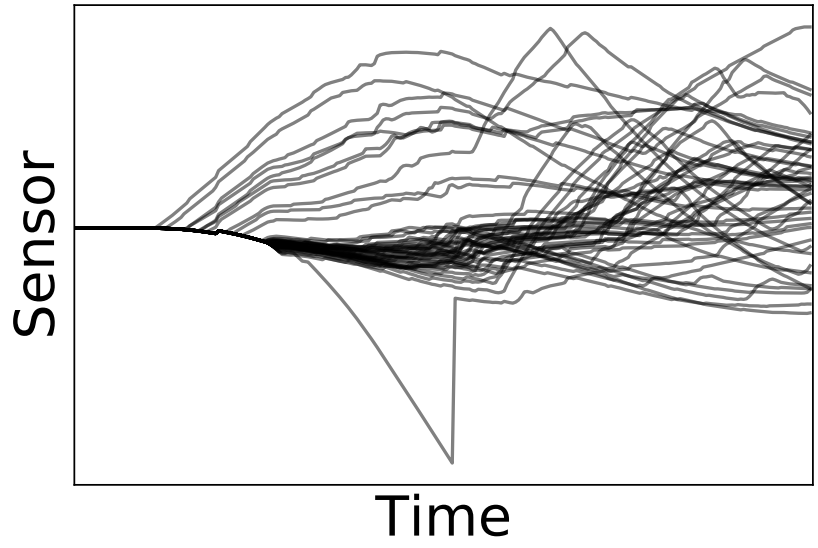
Anomaly Detection

Motivation: contaminated data

Multivariate data



Functional data



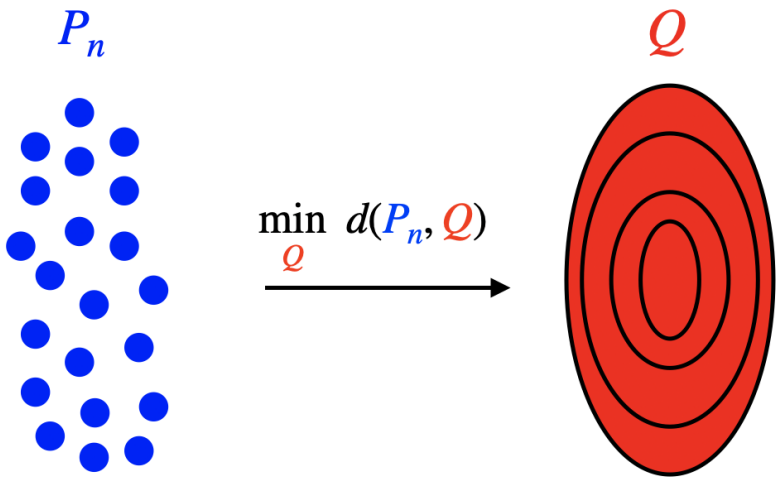
Outline

- PART I. Robustness and probability metrics
  - **Median-of-Means and Wasserstein distance**
  - **Robust discrepancy measures based on Data Depth**
  - The AI-IRW depth
- PART II. Functional anomaly detection
  - **The ACH depth**
  - **Functional Isolation Forest**
  - Benchmark of recent FAD methods on aeronautics data

Robustness and Probability metrics

Metrics between Probability Distributions

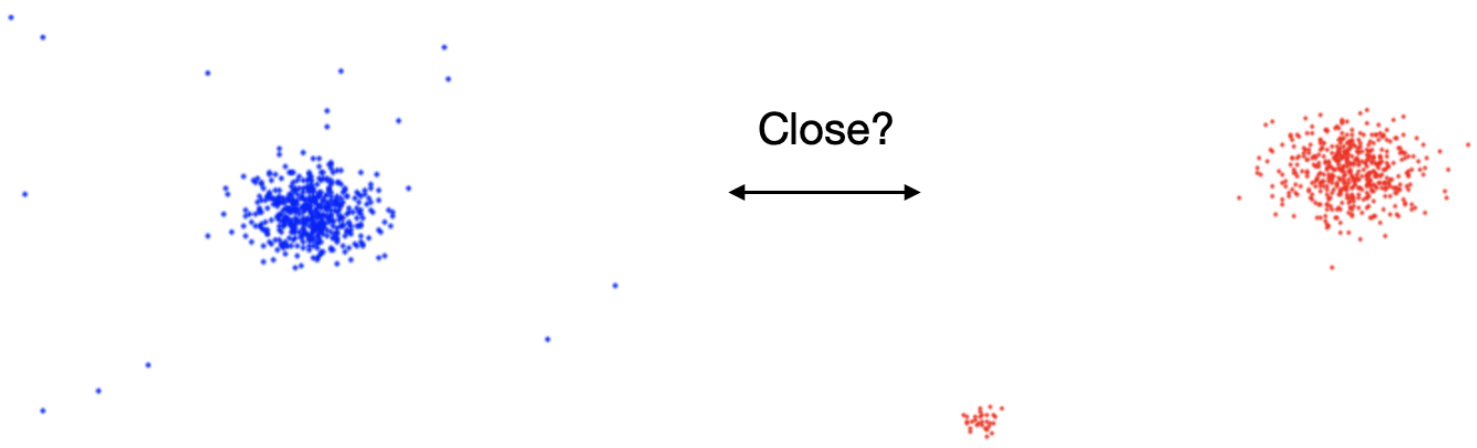
Many applications: generative modeling [GPAM+14], text evaluation [KSKW15], domain adaptation [CFT14], variational inference [LHJ19]



Many metrics: Kullback-Leibler divergence, f-divergences [Csi63], Maximum Mean Discrepancy [GBR+07], Wasserstein distance [Kan42]

Robustness and Probability Metrics

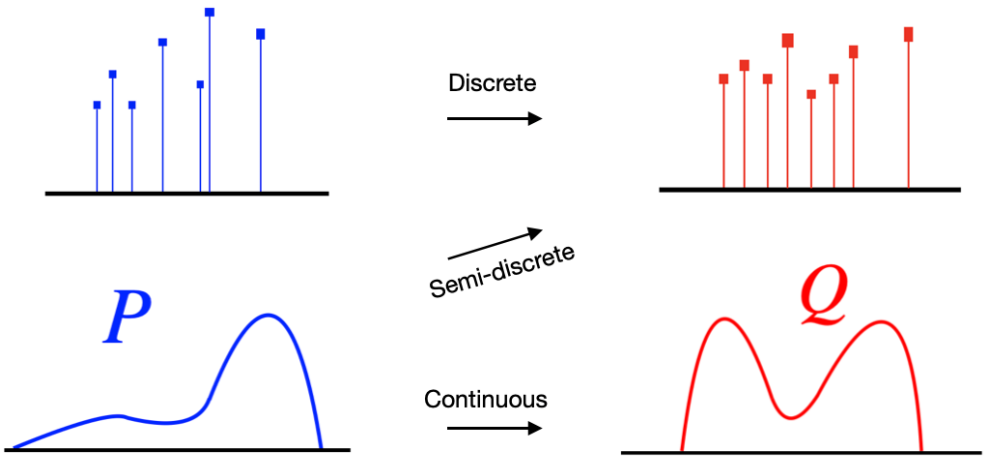
$\tilde{P} = (1-\epsilon)P + \epsilon P_A$ $\qquad$ $\tilde{Q} = (1-\epsilon)Q + \epsilon Q_A$



Goal

To design metrics d s.t. $\quad d(\tilde{P}, \tilde{Q}) \approx d(P, Q)$

Wasserstein distance [Kantorovich, 1942]



- Transport P to Q minimizing a cost function.
- The minimal cost of transportation defines the Wasserstein distance

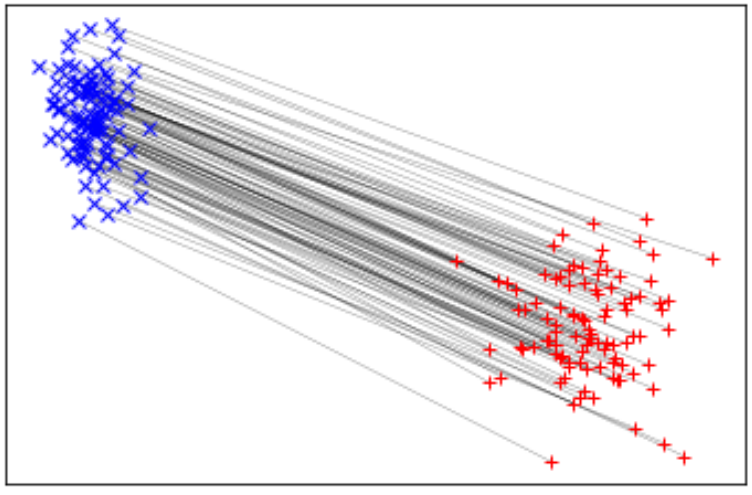
Let $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$ such that $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$.

$$\mathcal{W}_p^p(P, Q) = \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y)^p d\pi(x, y)$$

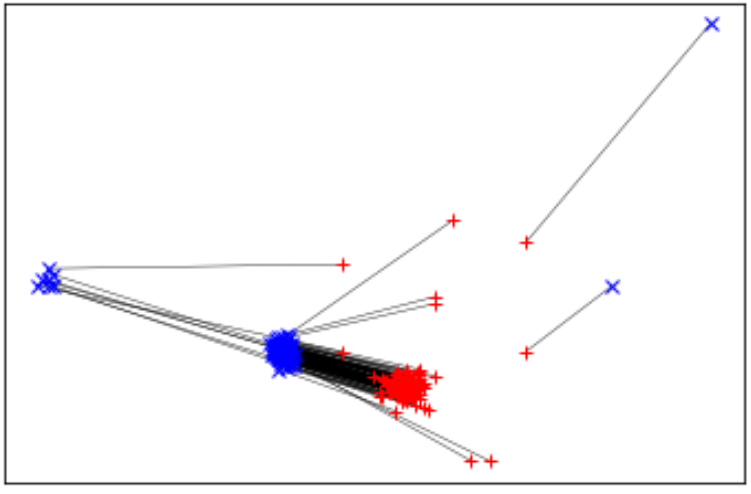
with $\Pi(P, Q) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \int \pi(x, y)dy = P(x); \int \pi(x, y)dx = Q(y)\}$.

Wasserstein and Outliers

Without outliers



With outliers



Relaxation of mass conservation:

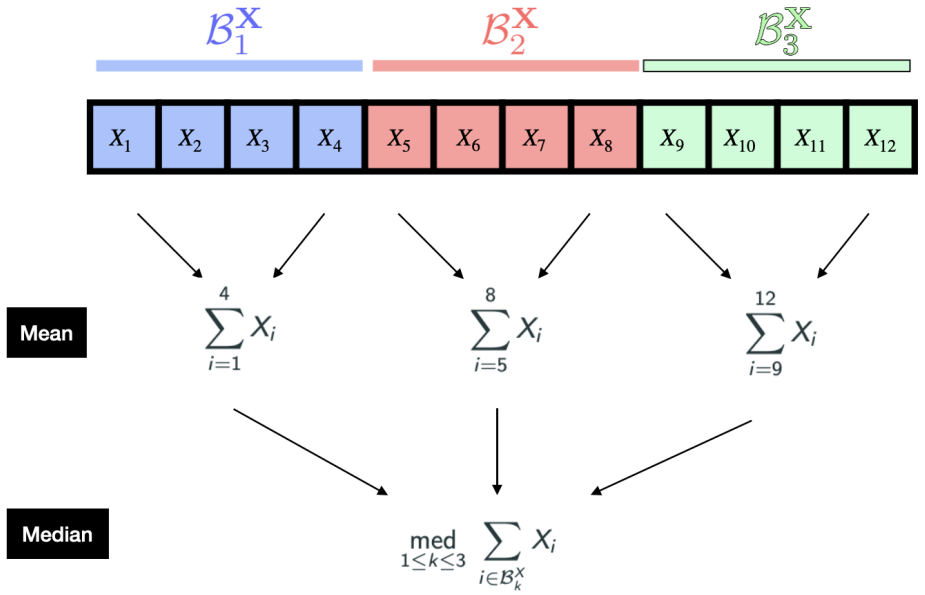
- Metrics based on unbalanced optimal transport [BCF20, NCG21, MGS+21]

When OT meets MoM: robust estimation of Wasserstein distance

Median-of-Means (MoM) estimator

$X_1, \dots, X_n$ an i.i.d. sample from X following $P \in \mathcal{P}(\mathcal{X})$ and $g : \mathcal{X} \to \mathbb{R}$.
Consider $\mathcal{B}_1^X, \dots, \mathcal{B}_K^X$ a partition of $\{1, \dots, n\}$ with the same size $B = n/K$ and $1 \leq K \leq n$.

$$\mathrm{MoM}_X[g] = \underset{1 \leq k \leq K}{\mathrm{med}} \left\{ \frac{1}{B} \sum_{i \in \mathcal{B}_k^X} g(X_i) \right\}$$



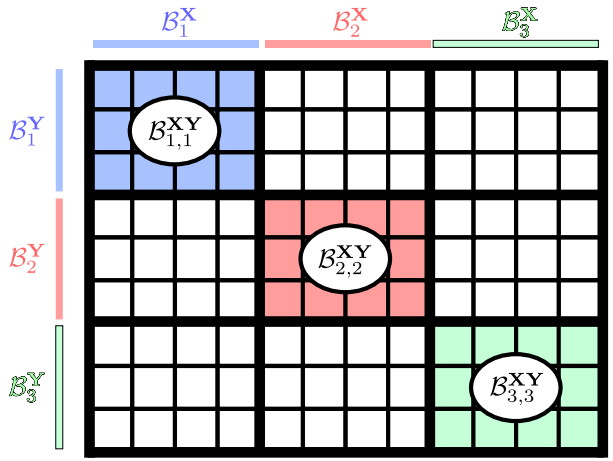
- Attractive robustness properties [Cat12, DLLO16, Dep20, LL20, LM19]
- Successful applications in machine learning tasks such as classification [BJL15, LLM20], bandits [BCBL13] or mean embedding [LSML19]

Median-of-U-statistics (MoU)

$X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ be two samples from X and Y following $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$. Let $1 \leq K_X \leq n$ and $1 \leq K_Y \leq m$, $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ and consider $\mathcal{B}_1^{\mathbf{X}}, \dots, \mathcal{B}_{K_X}^{\mathbf{X}}$ and $\mathcal{B}_1^{\mathbf{Y}}, \dots, \mathcal{B}_{K_Y}^{\mathbf{Y}}$ be partitions of $\{1, \dots, n\}$ and $\{1, \dots, m\}$ with size $B_X = n/K_X$ and $B_Y = m/K_Y$. Let $\mathcal{B}_{i,j}^{XY} = \mathcal{B}_i^X \mathcal{B}_j^{Y\top}$.

$$\mathbb{E}[h(X, Y)]$$

$$\mathrm{MoU}_{\mathbf{XY}}^{\mathrm{diag}}[h] = \underset{1 \leq k \leq K}{\mathrm{med}} \left\{ \frac{1}{B_X B_Y} \sum_{(i,j) \in \mathcal{B}_{k,k}^{XY}} h(X_i, Y_j) \right\}$$



- Attractive robustness properties [JL16, LCB19, LSC21]
- Still few applications! Ranking? Similarity learning?

MoM/MoU based estimators of the Wasserstein distance

The dual Kantorovich-Rubinstein formulation of the Wasserstein distance is defined as

$$\begin{aligned}\mathcal{W}_1(P, Q) &= \sup_{\Phi \in \mathcal{B}_L} \mathbb{E}_P\left[\Phi(X)\right] - \mathbb{E}_Q\left[\Phi(Y)\right], \\ &= \sup_{\Phi \in \mathcal{B}_L} \mathbb{E}_{(P \otimes Q)}\left[h_\Phi(X, Y)\right],\end{aligned}$$

where $h_\Phi(x, y) = \Phi(x) - \Phi(y)$ and $\mathcal{B}_L$ the unit ball of the Lipschitz functions space.

MoM/MoU based estimators of the Wasserstein distance

The dual Kantorovich-Rubinstein formulation of the Wasserstein distance is defined as

$$\begin{aligned}\mathcal{W}_1(P, Q) &= \sup_{\Phi \in \mathcal{B}_L} \mathbb{E}_P\left[\Phi(X)\right] - \mathbb{E}_Q\left[\Phi(Y)\right], \\ &= \sup_{\Phi \in \mathcal{B}_L} \mathbb{E}_{(P \otimes Q)}\left[h_\Phi(X, Y)\right],\end{aligned}$$

where $h_\Phi(x, y) = \Phi(x) - \Phi(y)$ and $\mathcal{B}_L$ the unit ball of the Lipschitz functions space.

We define MoM/MoU based estimators of the Wasserstein distance:

$$\mathcal{W}_{\mathrm{MoM}}(P_n, Q_m) = \sup_{\Phi \in \mathcal{B}_L} \{\mathrm{MoM}_X[\Phi] - \mathrm{MoM}_Y[\Phi]\},$$

$$\mathcal{W}_{\mathrm{MoU-diag}}(P_n, Q_m) = \sup_{\Phi \in \mathcal{B}_L} \{\mathrm{MoU}^{\mathrm{diag}}_{XY}[h_\Phi]\}.$$

Contamination Setting and Assumptions

- $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$: polluted with $\tau_X, \tau_Y < 1/2$ outliers.

- Inliers: i.i.d. in a compact set $\mathcal{K} \subset \mathbb{R}^d$, no assumption on outliers.

Contamination Setting and Assumptions

- $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$: polluted with $\tau_X, \tau_Y < 1/2$ outliers.

- Inliers: i.i.d. in a compact set $\mathcal{K} \subset \mathbb{R}^d$, no assumption on outliers.

Asymptotically, the size of blocks $B_X \to +\infty$:

- $B_X = n/K_X \Longrightarrow K_X = o(n)$
- if $\tau_X = \text{constant} \Longrightarrow n_{\mathcal{O}}^X = O(n)$

$\Big\}$ Problem: $K_X > 2n_{\mathcal{O}}^X$

There exist $C \geq 1$ and $0 \leq \alpha < 1$ such that $\tau_X \leq Cn^{\alpha-1}$ and $\tau_Y \leq Cm^{\alpha-1}$.

Theoretical Guarantees [Staerman et al., 2020.]

Choosing $K_X = \lceil \sqrt{2\tau_{\mathbf{X}}}\ n \rceil$ with $\tau_X < 1/2$, it holds:

$$\mathcal{W}_{\mathrm{MoM}}(\tilde{P}_n, P) \xrightarrow{a.s} 0,$$

and

$$\mathbb{E}\left[\mathcal{W}_{\mathrm{MoM}}(\tilde{P}_n, P)\right] \ \leq\ \kappa(\tau_{\mathbf{X}})\ n^{-1/(d+2)}.$$

If $\tau_{\mathbf{X}} + \tau_{\mathbf{Y}} < 1/2$ and $n = m$, then choosing $K_X = K_Y = \lceil \sqrt{2(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}})}\ n \rceil$, it holds:

$$\left|\mathcal{W}_{\mathrm{MoU-diag}}(\tilde{P}_n, \tilde{Q}_m) - \mathcal{W}(P, Q)\right| \xrightarrow{a.s} 0,$$

and

$$\mathbb{E}\left|\mathcal{W}_{\mathrm{MoU-diag}}(\tilde{P}_n, \tilde{Q}_m) - \mathcal{W}(P, Q)\right| \leq\ \kappa'(\tau_X + \tau_Y)\ n^{-1/(d+2)}.$$

Application: MoM-WGAN (1/2)

Assume now that $X_1, \ldots, X_n$ is an input sample and $Z_1, \ldots, Z_m$ is a generated sample from a latent space. Let Φ be the discriminant NN and Ψ be the generator NN.

Wasserstein GAN (WGAN) formulation:

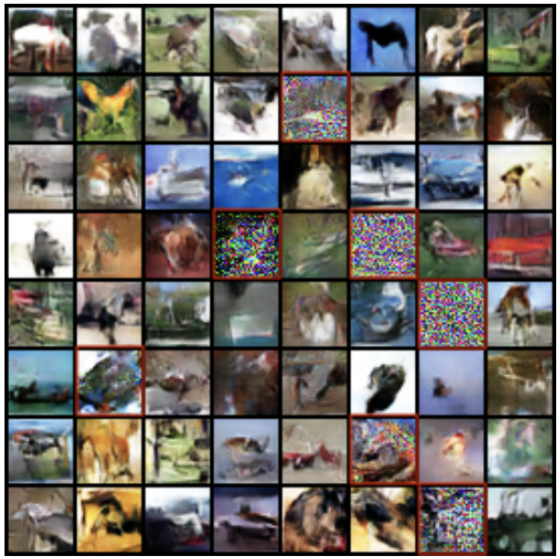
$$\min_{\theta} \max_{w} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Phi_w(X_i) - \frac{1}{m} \sum_{j=1}^{m} \Phi_w(\Psi_\theta(Z_j)) \right\}$$

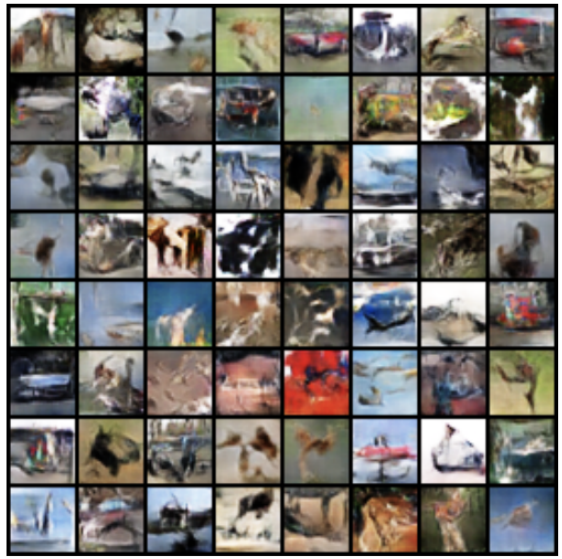
Median-of-Means Wasserstein GAN (MoM-WGAN) formulation:

$$\min_{\theta} \max_{w} \left\{ \mathrm{MoM}_{\mathbf{X}}[\Phi_w] - \frac{1}{m} \sum_{j=1}^{m} \Phi_w(\Psi_\theta(Z_j)) \right\}$$

Application: MoM-WGAN (2/2)

WGAN





MoM-WGAN

Further contributions

- Algorithms to compute $\mathcal{W}_{\mathrm{MoM}}, \mathcal{W}_{\mathrm{MoU-diag}}$
- Convergence of the algorithms
- Robustness of the proposed approaches

Open source code using `Pytorch` Python library available at

https://github.com/GuillaumeStaermanML/MoM-Wasserstein.

Discrepancy measure based on Data Depth

From Wasserstein to Robustness

If $P \in \mathcal{P}(\mathbb{R})$ and $Q \in \mathcal{P}(\mathbb{R})$,

$$W_p^p(P, Q) = \int_0^1 |F_P^{-1}(q) - F_Q^{-1}(q)|^p \, dq.$$

From Wasserstein to Robustness

If $P \in \mathcal{P}(\mathbb{R})$ and $Q \in \mathcal{P}(\mathbb{R})$,

$$W_p^p(P, Q) = \int_0^1 |F_P^{-1}(q) - F_Q^{-1}(q)|^p \, dq.$$

If $P \in \mathcal{P}(\mathbb{R})$ and $Q \in \mathcal{P}(\mathbb{R})$,

$$\int_\varepsilon^{1-\varepsilon} |F_P^{-1}(q) - F_Q^{-1}(q)|^p \, dq.$$

From Wasserstein to Robustness

If $P \in \mathcal{P}(\mathbb{R})$ and $Q \in \mathcal{P}(\mathbb{R})$,

$$W_p^p(P, Q) = \int_0^1 |F_P^{-1}(q) - F_Q^{-1}(q)|^p \, dq.$$

If $P \in \mathcal{P}(\mathbb{R})$ and $Q \in \mathcal{P}(\mathbb{R})$,

$$\int_\varepsilon^{1-\varepsilon} |F_P^{-1}(q) - F_Q^{-1}(q)|^p \, dq.$$

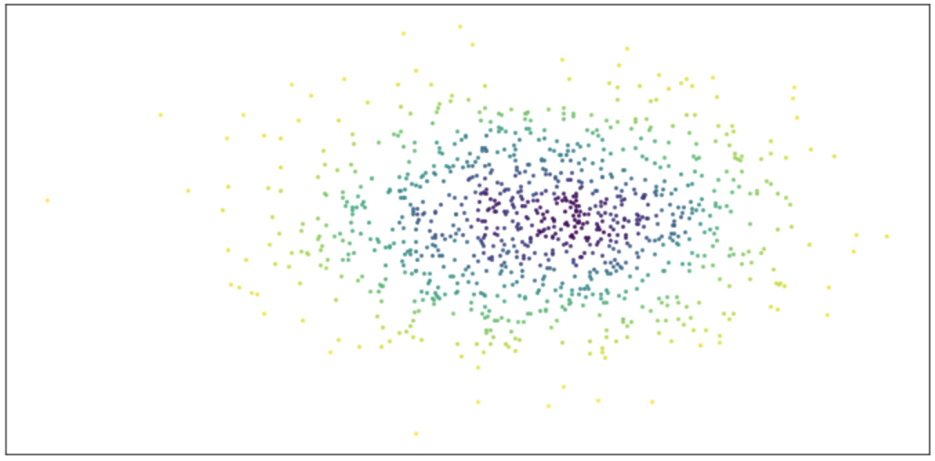
How to extend the notion of order to $\mathbb{R}^d$ when $d > 1$?

- Component-wise ranks [Hay02, Hod55]
- Spatial ranks [GG29, Sca33, Hal48]
- Depth-based ranks [Tuk75, Liu90]

Statistical Data Depth [Tukey, 1975]

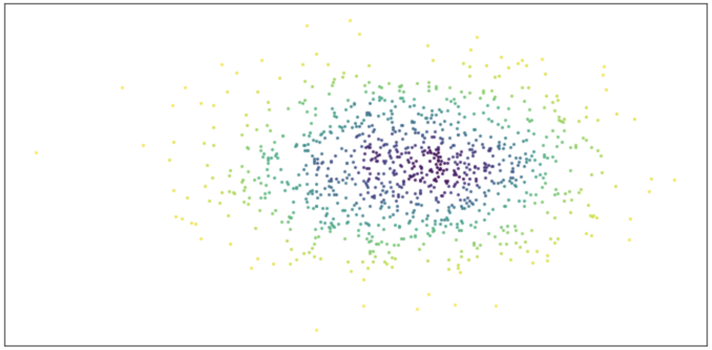
A **data depth** is a non parametric statistic which measures the centrality of any element of a general space $\mathbb{R}^d$ w.r.t. a probability distribution $P \in \mathcal{P}(\mathbb{R}^d)$. Precisely, it is a function

$$\begin{aligned} D : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) &\longrightarrow [0,1] \\ (x, P) &\longmapsto D(x, P) \end{aligned}$$



Applications: anomaly detection [Ser06]; clustering [Jö04]; classification [LPR06, LCAL12, LMM14]; missing data imputation [MJH20]

Properties [Zuo & Serfling, 2000]



(**D1**) affine-invariant: $D(Ax + b, P_{AX+b}) = D(x, P_X)$ for any $d \times d$ non-singular matrix A and any $b \in \mathbb{R}^d$.

(**D2**) maximal at symmetry center: $D(z, P) = \sup_{x \in \mathbb{R}^d} D(x, P)$ holds for any $P \in \mathcal{P}(\mathbb{R}^d)$ having symmetry center z.

(**D3**) monotone on rays: for any z with $D(z, P) = \sup_{x \in \mathbb{R}^d} D(x, P)$, any $x \in \mathbb{R}^d$, and any $0 \leq \alpha \leq 1$ it holds: $D(x, P) \leq D(z + \alpha(x - z)|P)$.

(**D4**) vanishing at infinity: $\lim_{||x|| \to \infty} D(x, P) = 0$.

Tukey (=halfspace, location) depth

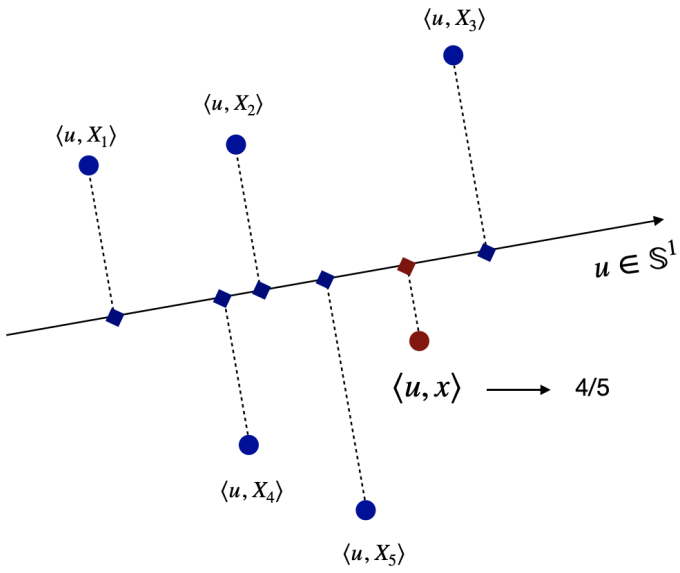
For any $x \in \mathbb{R}^d$ and $P \in \mathcal{P}(\mathbb{R}^d)$, the Halfspace depth is defined as

$$D_{\mathrm{T}}(x, P) = \inf_{u \in \mathbb{S}^{d-1}} \mathbb{P}\left(\langle u, X \rangle \leq \langle u, x \rangle\right)$$

- $X_1, \ldots, X_n$ i.i.d. sample from P

- $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$

$$\widehat{D}_{\mathrm{T},n}(x, P_n) = \min_{u \in \mathbb{S}^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left\{\langle u, X_i \rangle \leq \langle u, x \rangle\right\}$$

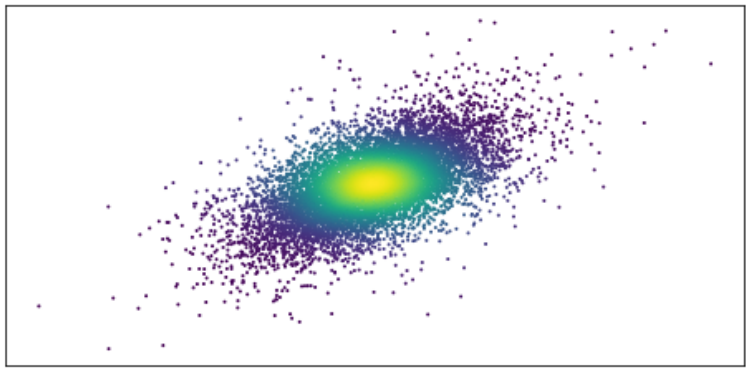


Affine-Invariant Integrated Rank Weighted (AI-IRW) Depth [Staerman et al., 2021]

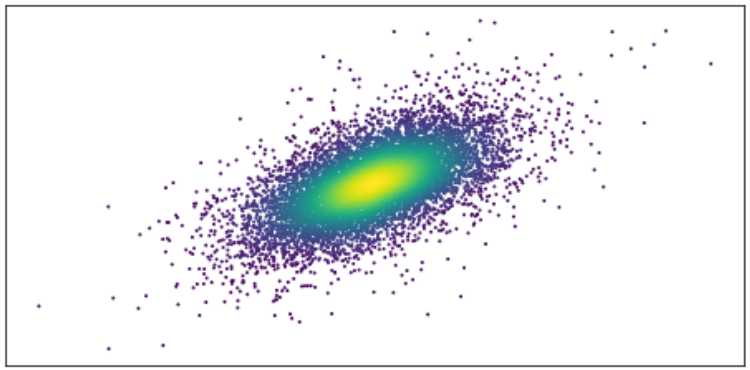
The AI-IRW depth of $x \in \mathbb{R}^d$ w.r.t. $P \in \mathcal{P}(\mathbb{R}^d)$ is given by:

$$D_{\text{AI-IRW}}(x, P) = \mathbb{E}\left[\min\left(F_V(\langle V, x\rangle), 1 - F_V(\langle V, x\rangle)\right)\right],$$

where $V = \Sigma^{-\top/2}U/||\Sigma^{-\top/2}U||$ and U being uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$.



IRW



AI-IRW

Concentration results for AI-IRW

Generating a number $n_{\mathrm{proj}} \geq 1$ i.i.d. random directions $U_1, \ldots, U_m$, copies of the generic r.v. U and independent from the original data $\mathcal{D}_n = \{X_1, \ldots, X_n\}$: $\forall x \in \mathbb{R}^d$,

$$\tilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}(x, \mathcal{D}_n) = \frac{1}{n_{\mathrm{proj}}} \sum_{j=1}^{n_{\mathrm{proj}}} \min\left\{ \widehat{F}_{\hat{V}_j}\left(\langle \hat{V}_j,\ x\rangle\right), 1 - \widehat{F}_{\hat{V}_j}\left(\langle \hat{V}_j,\ x\rangle\right)\right\},$$

where, for all $j \in \{1,\ \ldots, n_{\mathrm{proj}}\}$ and $t \in \mathbb{R}$, we set

$$\hat{V}_j = \hat{\Sigma}^{-\top/2} U_j / ||\hat{\Sigma}^{-\top/2} U_j|| \text{ and } \widehat{F}_{\hat{V}_j}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\left\{\langle \hat{V}_j, X_i\rangle \leq t\right\}}.$$

$$\sup_{x \in\ \mathcal{B}_r} |\tilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}(x, \mathcal{D}_n) - D_{\text{AI-IRW}}(x, P)| \leq \text{Rates}\ (n, n_{\mathrm{proj}}, d)$$

Further contributions

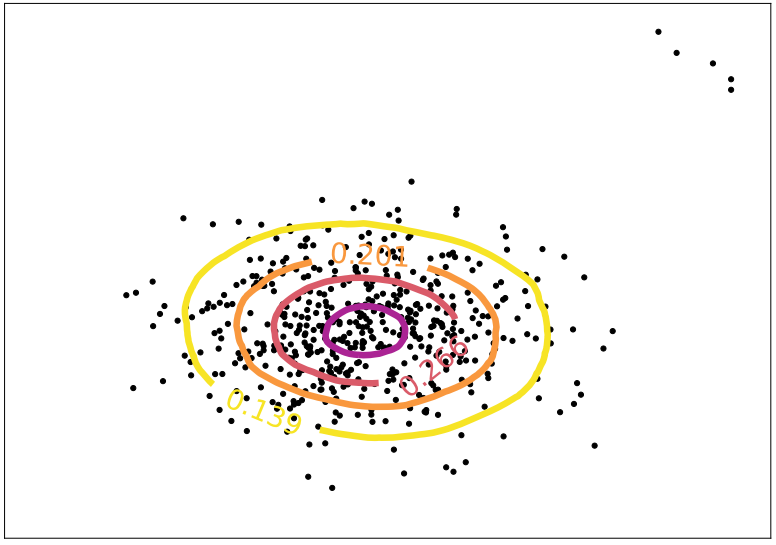
Non-asymptotic results	✓	for IRW and AI-IRW
Fast approximation	✓	with Monte-Carlo scheme
Robustness	✓	with MCD estimator
Performance	✓	in anomaly detection
Open-source code	✓	in Python

From quantile regions to depth-trimmed regions

For any $\alpha \in [0,1]$, $\quad D_P^\alpha = \{x,\ D(x,P) \geq \alpha\}$

- Affine-equivariant (due to **D1**)
- Nested (due to **D3**)
- Bounded (due to **D4**)
- Closed (if D is upper-continuous)
- Convexity (if D is quasi-concave)

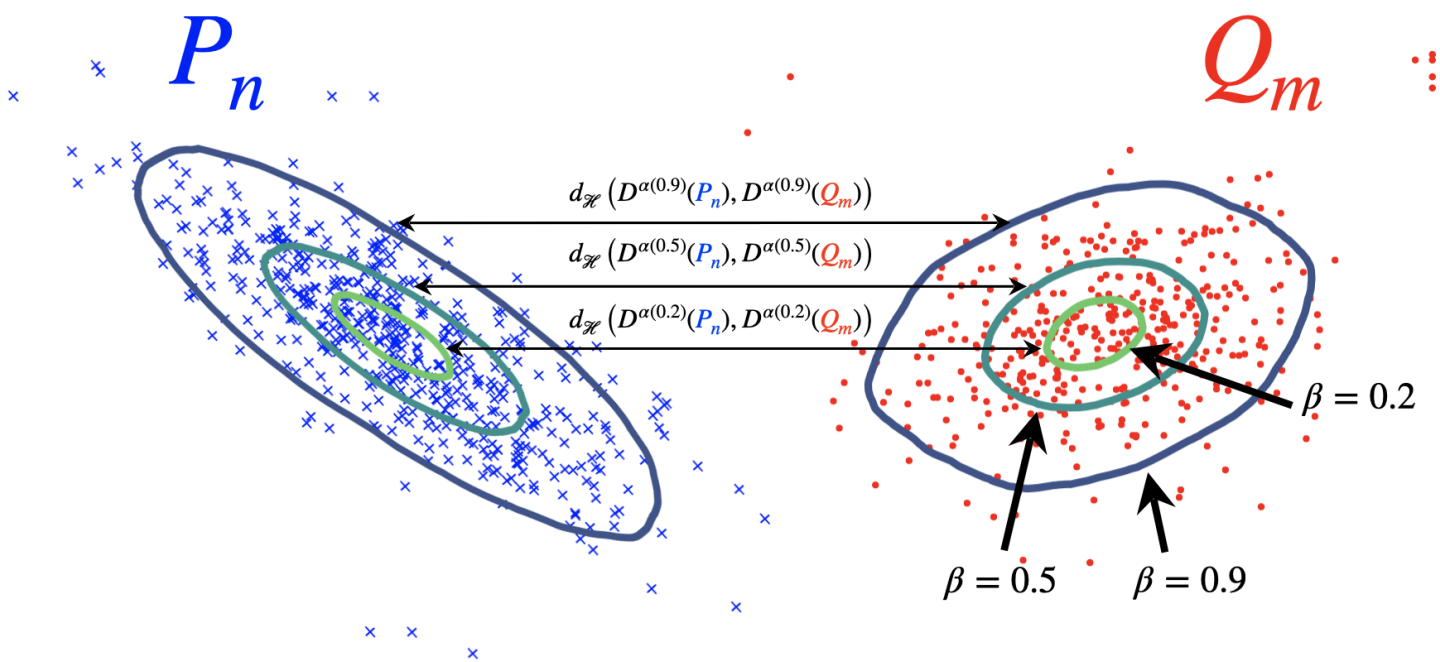
AI-IRW contours



A Pseudo-Metric based on Depth-Trimmed Regions [Staerman et al., 2021]

Let $\varepsilon \in (0, 1], p \in [1, \infty)$. The depth-trimmed regions ($DR_{p,\varepsilon}$) discrepancy measure between $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$ is:

$$DR^{p}_{p,\varepsilon}(P, Q) = \int_{0}^{1-\varepsilon} d_{\mathcal{H}} \left(D_P^{\alpha(\beta)}, D_Q^{\alpha(\beta)} \right)^p d\beta.$$



Properties

Pseudo-metric

- $DR_{p,\varepsilon}(P, Q) = DR_{p,\varepsilon}(Q, P)$
- $DR_{p,\varepsilon}(P, Q) \leq DR_{p,\varepsilon}(P, R) + DR_{p,\varepsilon}(R, Q)$
- $P = Q \implies DR_{p,\varepsilon}(P, Q) = 0$
- $DR_{p,\varepsilon}(P, Q) = 0 \not\Rightarrow P = Q$

Isometry invariance

- $DR_{p,\varepsilon}(A_\sharp P, A_\sharp Q) = DR_{p,\varepsilon}(P, Q)$ for any $AA^\top = I_d$

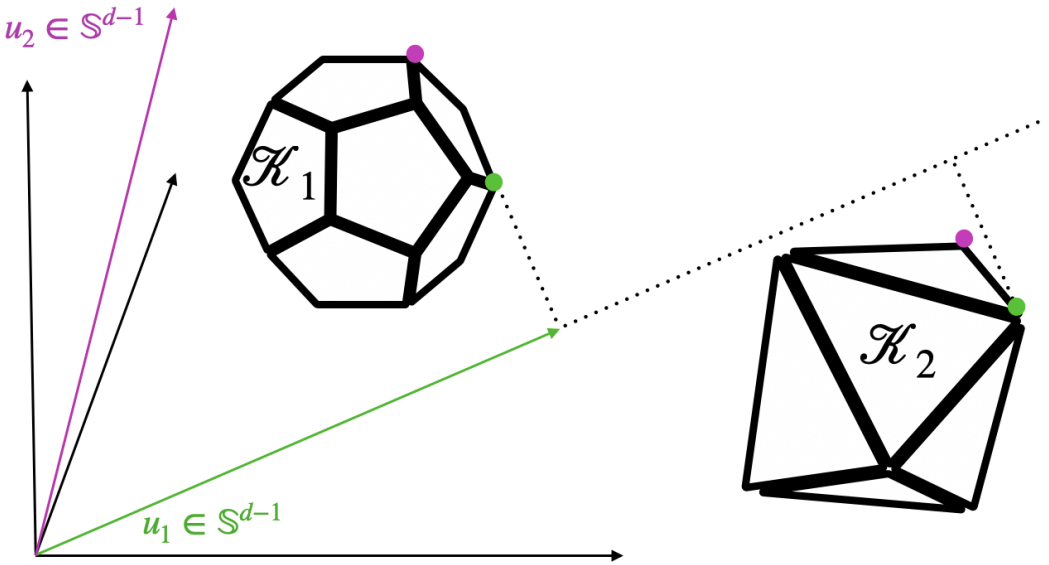
Robustness

- Breakdown point of $DR_{p,\varepsilon}$ is greater than or equal to that of $D_P^\varepsilon \vee D_Q^\varepsilon$

Computationally efficient approximation

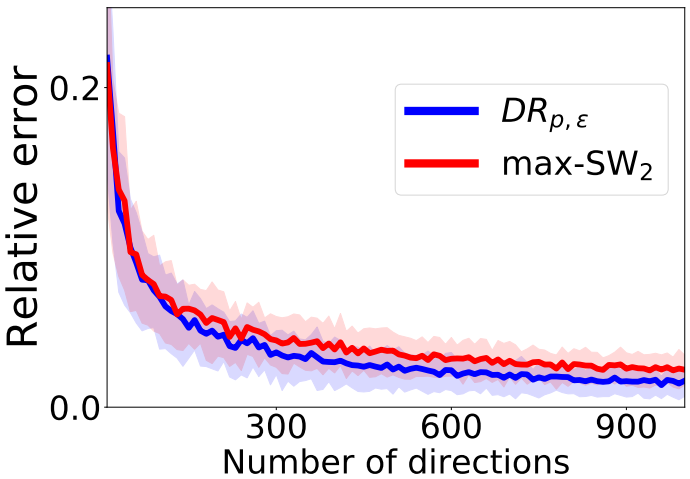
For any convex data depth, we have

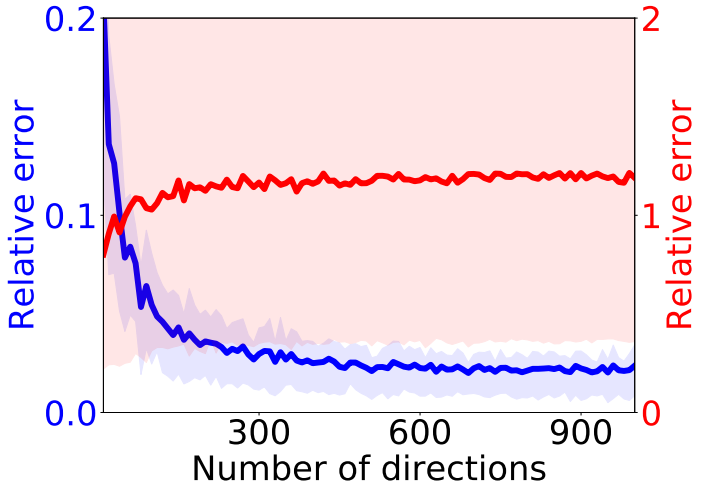
$$d_{\mathcal{H}}(D_P^{\alpha(\beta)}, D_Q^{\alpha(\beta)}) = \sup_{u \in \mathbb{S}^{d-1}} \left| h_{D_P^{\alpha(\beta)}}(u) - h_{D_Q^{\alpha(\beta)}}(u) \right|,$$

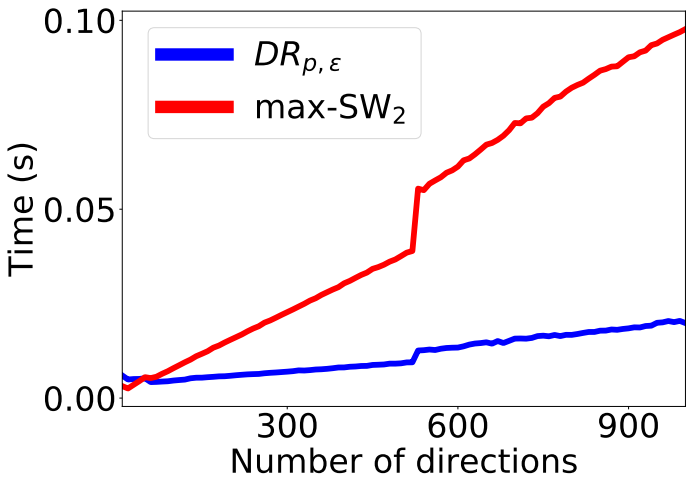


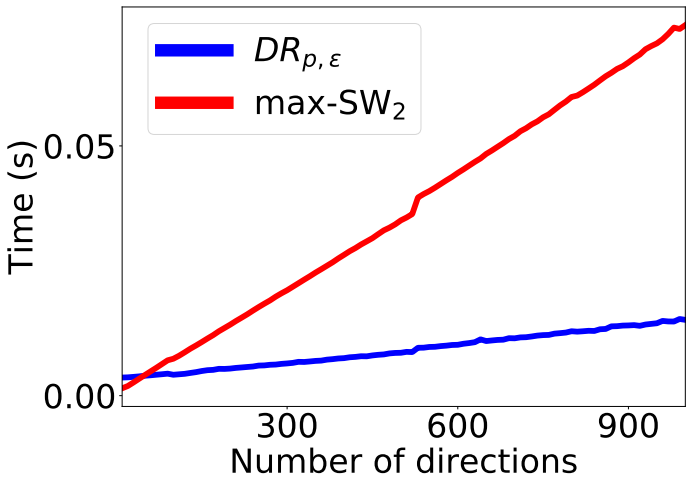
- $h_{D_P^{\alpha(\beta)}}(u)$ approximated by $\sup\{\langle u, X_i \rangle, \quad X_i \in D_P^{\alpha(\beta)}\}$
- Approximation of the supremum with n_{proj} finite unit sphere vectors
- Computation in $O(n_{\text{proj}} nd)$

Approximation quality and computation time









(Robust) Clustering on images

Bag of pixels: images represented as probability distributions in $\mathbb{R}^3$

$\longrightarrow$ (x-axis, y-axis, intensity)

	Fashion-MNIST		Cont. Fashion-MNIST	
	NMI	ARI	NMI	ARI
$DR_{p,\varepsilon}$	**0.58**	**0.43**	**0.55**	**0.42**
Wass	0.50	0.35	0.48	0.30
Sliced-Wass	0.55	0.39	0.47	0.33
MMD	0.54	0.37	0.50	0.36
Euclidian	0.50	0.32	0.48	0.30

Table 1: Spectral clustering performances.

Functional Anomaly Detection

Unsupervised Anomaly Detection

Goal: Learn a score function $s : \mathcal{X} \longrightarrow \mathbb{R}$

Multivariate methods

- One-Class SVM (OCSVM) [SPST+01]
- Isolation Forest (IF) [LTZ08, LTZ12]
- Local Outlier Factor (LOF) [BKNS00]
- Histogram-based Outlier Score [GD12]
- k-NN approaches [ZS09, SRH11]
- Autoencoders [AC15, ZP17]
- Minimum volume set [EM92, SN06]
- Data depth [Tuk75, Liu90, DG92]

Functional methods

- Functional data depth [FM01, LPR09, LPR11, NRB16]
- Functional outlyingness measures [HRS15, RRH18, DG19]

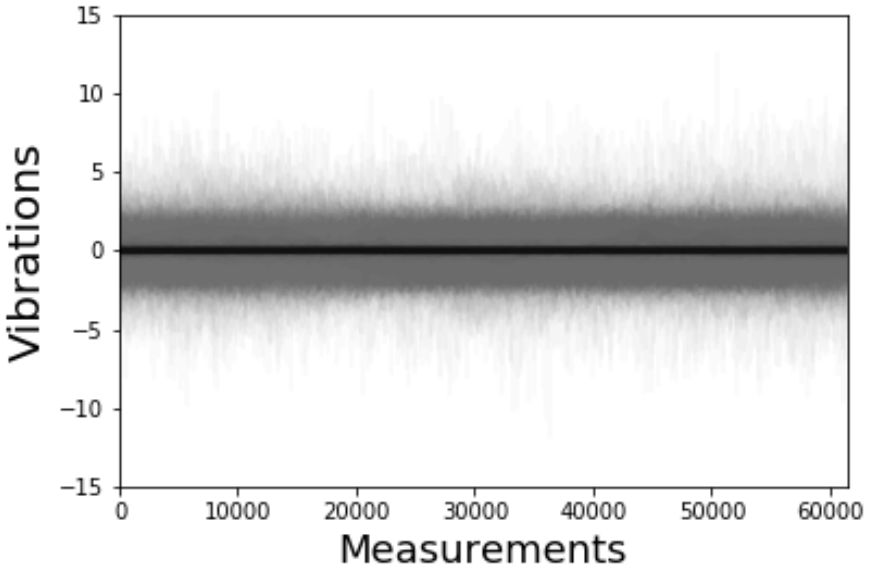
- Machine Learning?

Functional Data Framework

- Let $\mathbf{X} = \{\mathbf{X}(t) \in \mathbb{R}, \quad t \in [0,1]\}$ be a functional random variable that takes its values in a functional space $\mathcal{F}$
- Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be an i.i.d. sample from $\mathbf{X}$. In practice, we only have access to partial observations of the sample:
  $\mathcal{S} = \{\mathbf{X}_i(t_1), \ldots, \mathbf{X}_i(t_p), \quad 1 \leq i \leq n\}$ with $0 \leq t_1 < \ldots < t_p \leq 1$
- First step: reconstruct a functional object from time-series either by interpolation or basis decomposition

---

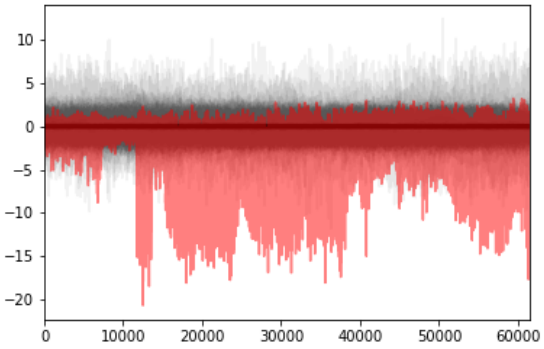
Example with Airbus data: one-minute-sequences at 1024Hz of accelerometer data measured on test helicopters, i.e. $p = 61440$.

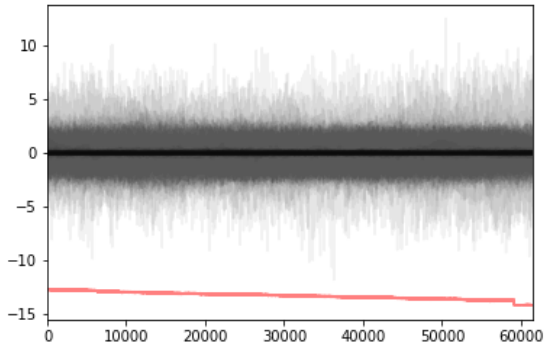


Challenges of Functional Anomaly Detection

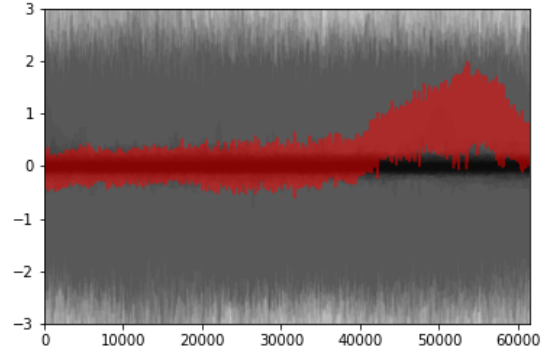
A taxonomy of functional abnormalities [HRS15]:

Magnitude (=location, shift) anomalies

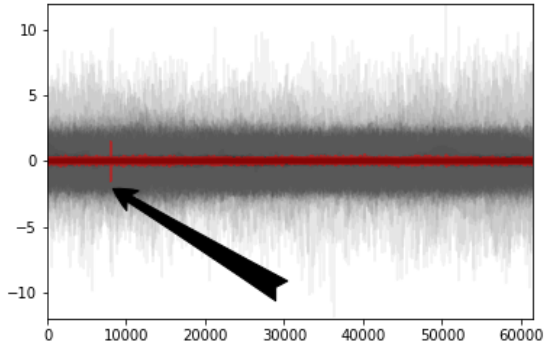




Shape anomalies



Isolated anomalies

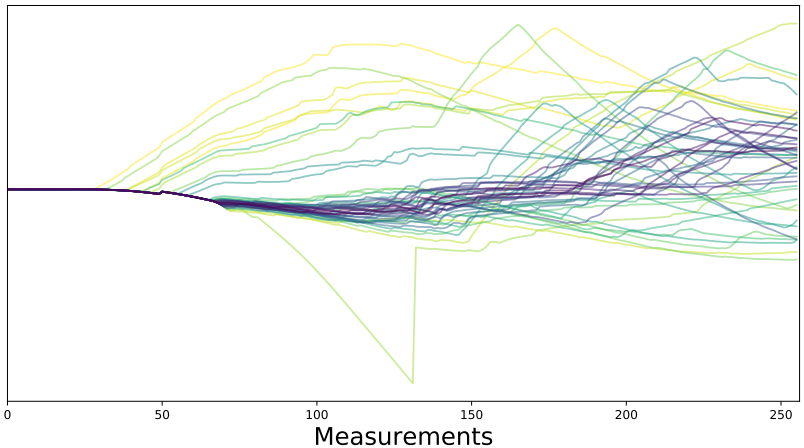


The ACH depth

Functional Data Depth

A **functional data depth** is a non-parametric statistic which measures the centrality of $\mathbf{x} \in \mathcal{F}$ w.r.t. a probability distribution $\mathbf{P} \in \mathcal{P}(\mathcal{F})$:

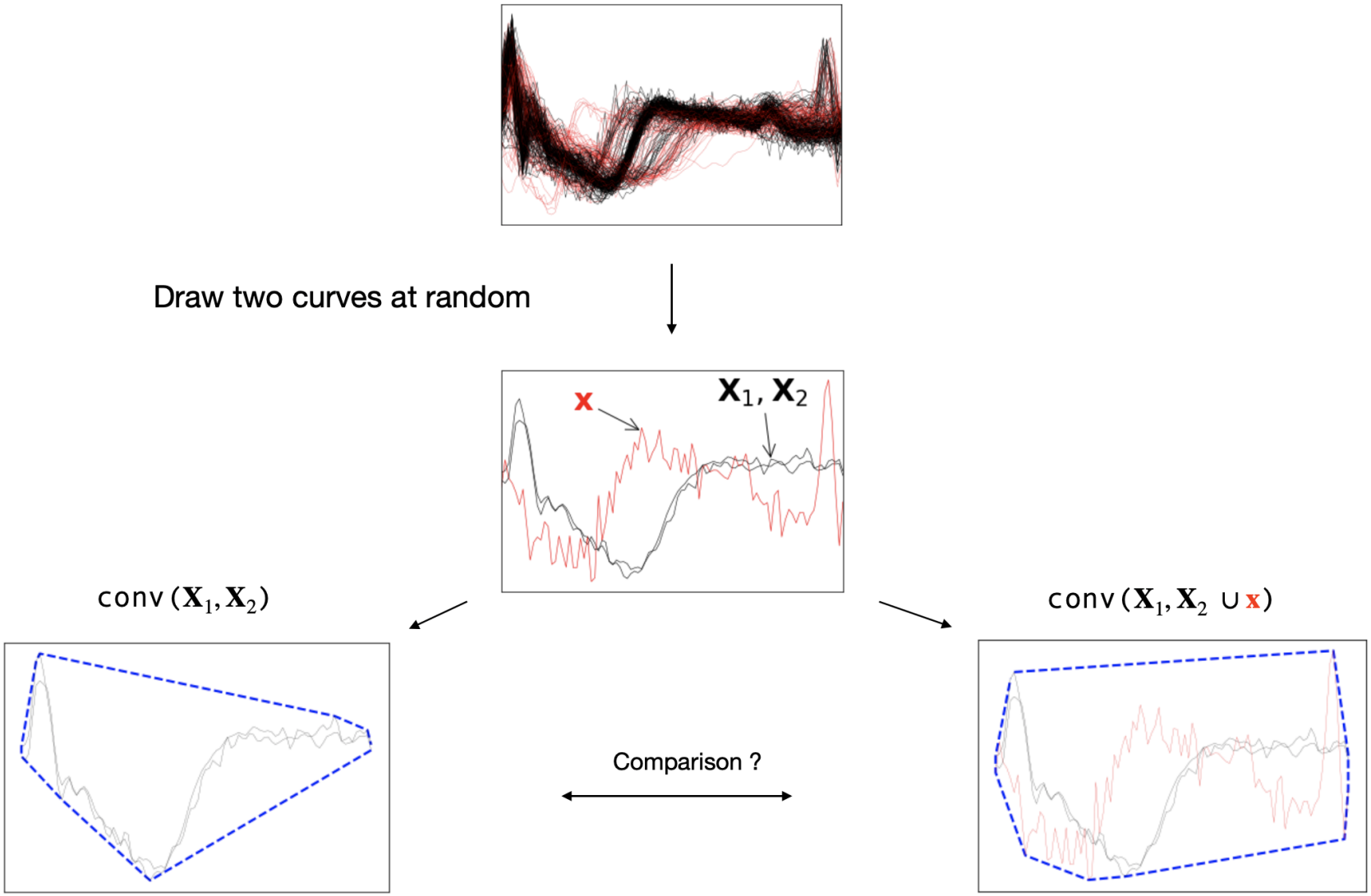
$$\begin{aligned} FD : \mathcal{F} \times \mathcal{P}(\mathcal{F}) &\longrightarrow [0,1] \\ (\mathbf{x}, \mathbf{P}) &\longmapsto FD(\mathbf{x}, \mathbf{P}) \end{aligned}$$



Examples: integrated depth [FM01], band depth [LPR09], local band depth [AR13], half-region depth [LPR11]

- Focusing on shape anomalies: [AGR14, KR16, NGH17, HTLS21]
- Focusing on isolated anomalies?

The Area of the Convex Hull (ACH) of Sampled Curves (1/2) [Staerman et al., 2020]



The Area of the Convex Hull (ACH) of Sampled Curves (2/2)

Definition

Let $J \geq 1$ be a fixed integer. The ACH depth of degree J is the function $FD_J : \mathcal{C}(\mathcal{T}) \times \mathcal{P}(\mathcal{C}(\mathcal{T})) \to [0,1]$ defined by: $\forall \mathbf{x} \in \mathcal{C}(\mathcal{T})$,

$$FD_J(\mathbf{x}, \mathbf{P}) = \mathbb{E}\left[\frac{\lambda\left(conv\left(graph\left(\{\mathbf{X}_1, \ldots, \mathbf{X}_J\}\right)\right)\right)}{\lambda\left(conv\left(graph\left(\{\mathbf{X}_1, \ldots, \mathbf{X}_J\} \cup \{\mathbf{x}\}\right)\right)\right)}\right],$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_J$ are i.i.d. r.v.'s drawn from $\mathbf{P} \in \mathcal{P}(\mathcal{C}(\mathcal{T}))$.

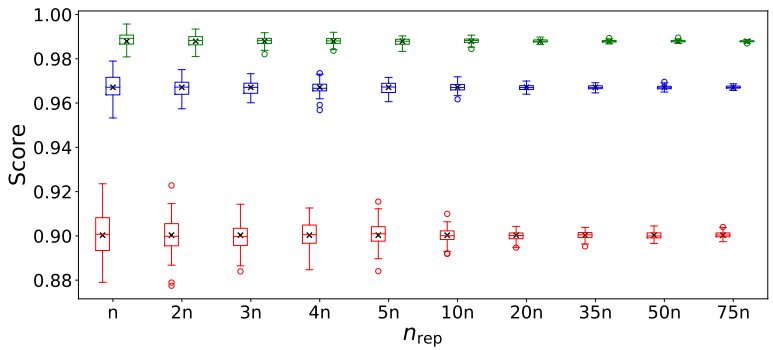
When $n \geq J$, an unbiased statistical estimation of $FD_J(., \mathbf{P})$ can be obtained by computing the symmetric U-statistic of degree J:

$$FD_{J,n}(\mathbf{x}) = \underbrace{\frac{1}{\binom{n}{J}} \sum_{1 \leq i_1 < \ldots < i_J \leq n}}_{\text{Approximation with } n_{\text{rep}} \text{ combinations?}} \frac{\lambda\left(conv\left(graph\left(\{\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_J}\}\right)\right)\right)}{\lambda\left(conv\left(graph\left(\{\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_J}, \mathbf{x}\}\right)\right)\right)}.$$

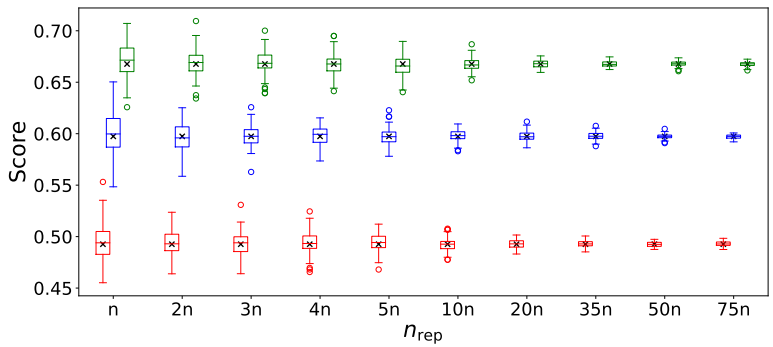
Choice of n_{rep}

Parameter n_{rep} reflects the trade-off between statistical accuracy and computational time

$FD(\mathbf{x}_0)$



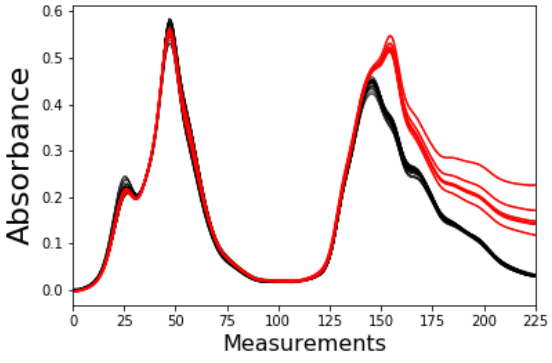
$FD(\mathbf{x}_3)$



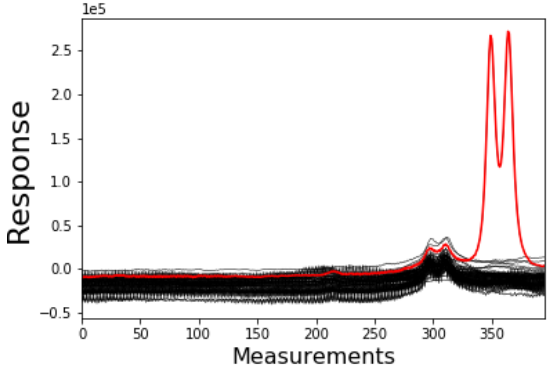
- The variance of each depth decreases sufficiently fast to obtain good approximation
- For example, for $n = 100$ and $J = 3$, we have $\binom{n}{J} = 161700$. $n_{\text{rep}} = 2000$ leads to very good approximation

Application to Anomaly Detection

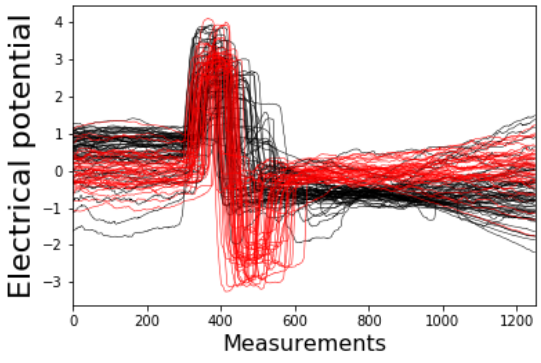
Octane



Wine



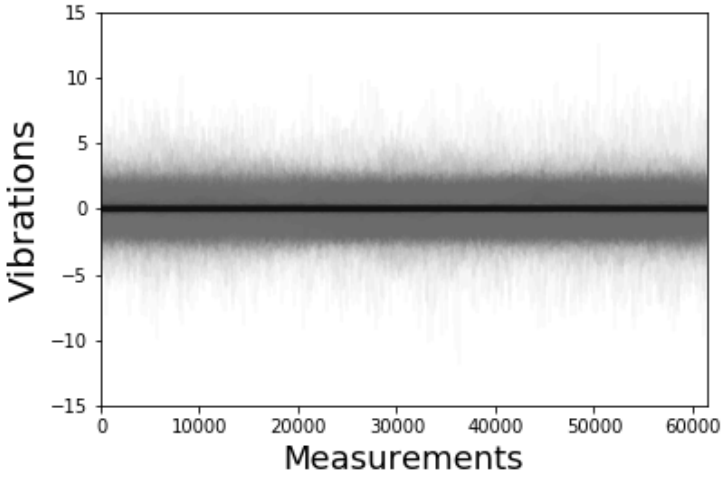
EOG

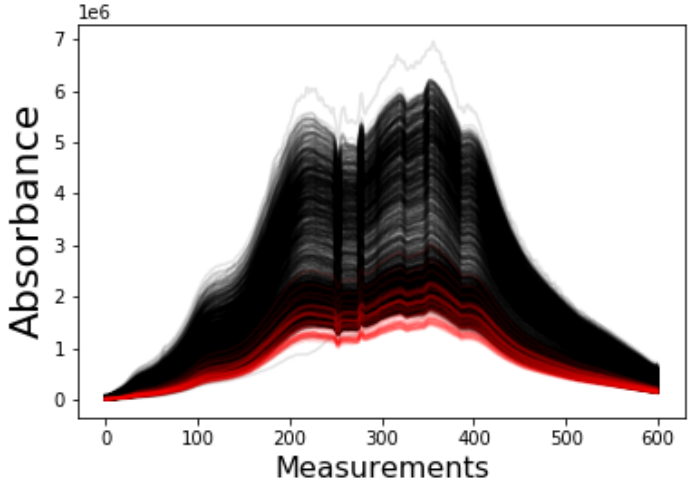


	ACHD	FSDO	FT	IF	OC
Octane	**1**	0.5	0.33	0.5	0.5
Wine	**1**	0	0	0	**1**
EOG	**0.73**	0.55	0.48	0.63	0.6

Proportion of detected anomalies.

Limitations





Crossed curves and salient points

Further contributions

- Investigation of functional depth properties
- Consistency of the sample version and the approximation
- Robustness of returned ranks

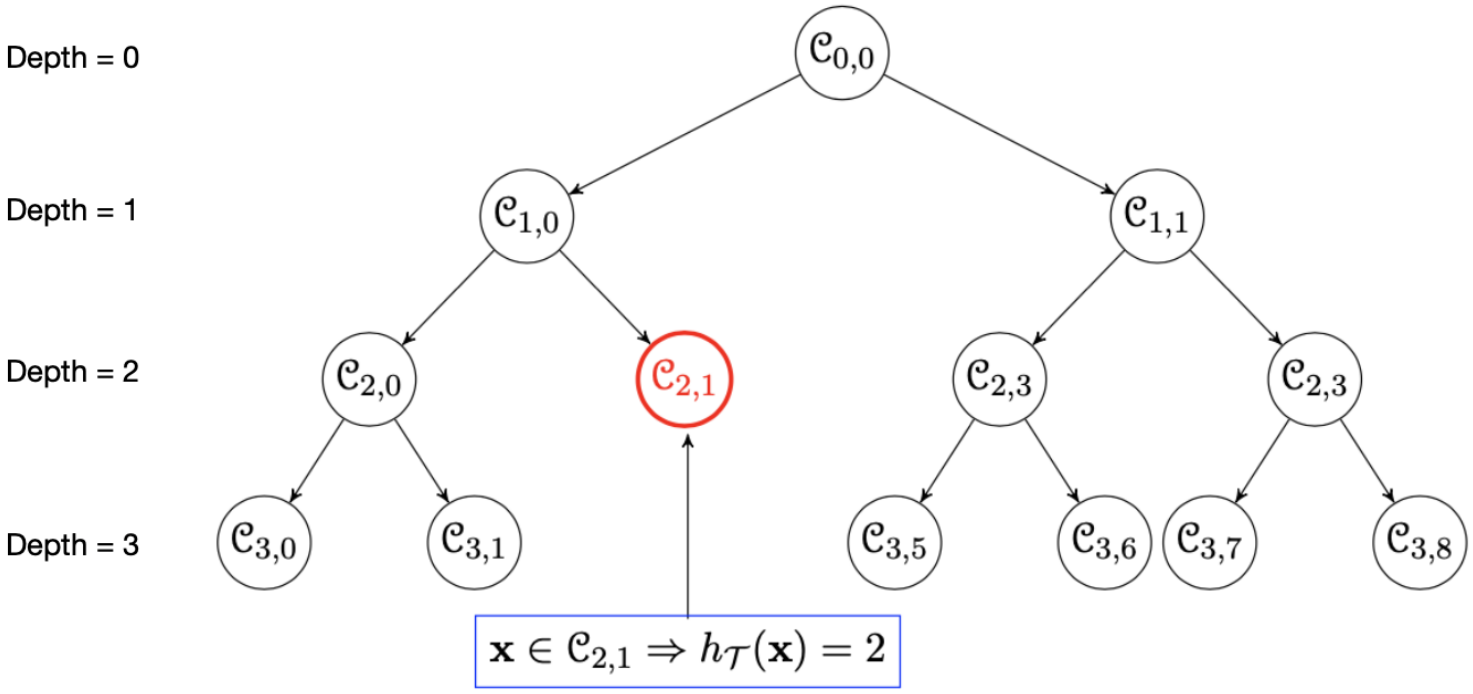
Cython/C++ implementation is available at

https://github.com/GuillaumeStaermanML/ACHD.

Functional Isolation Forest

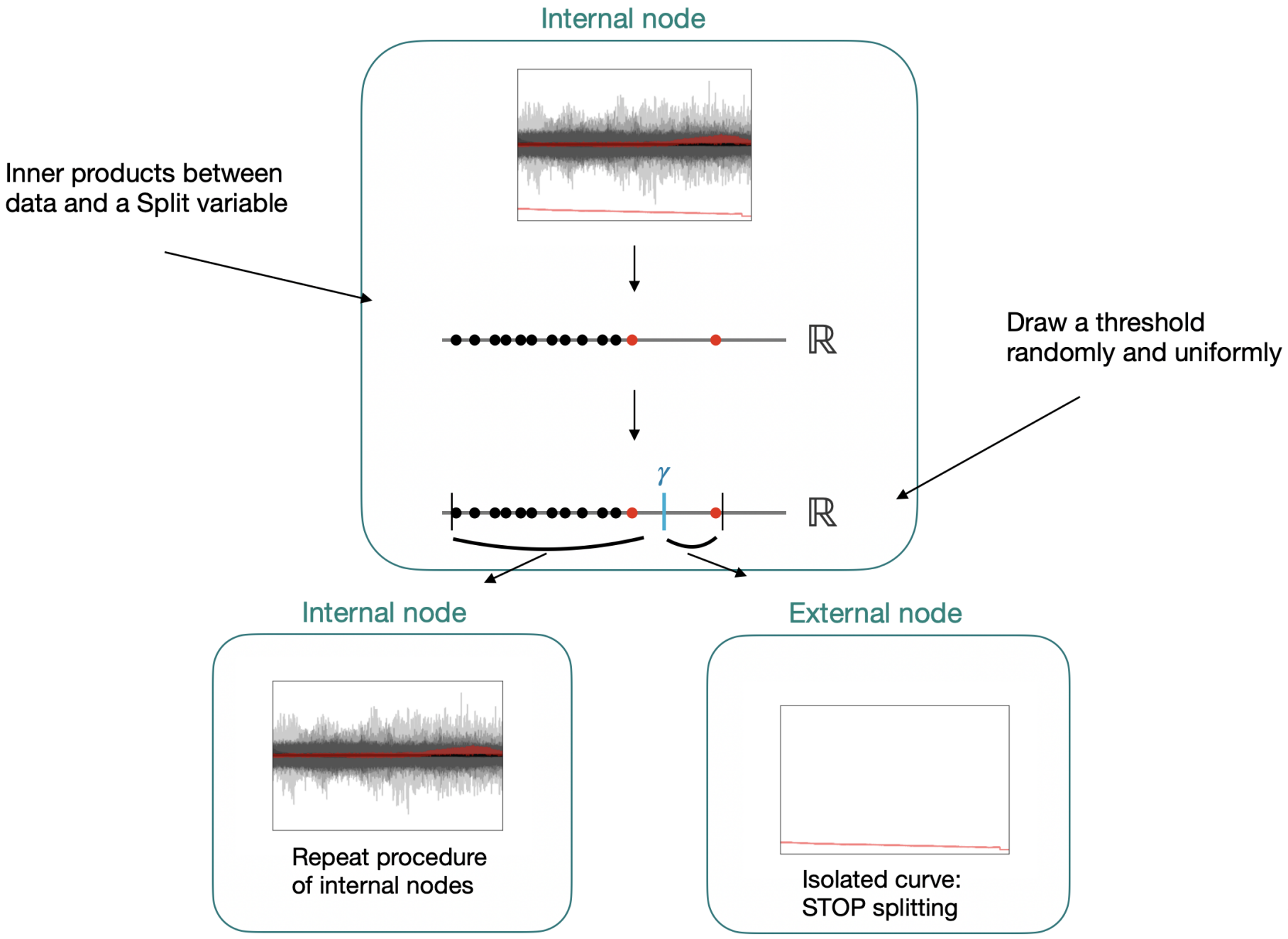
Functional Isolation Forest (FIF) [Staerman et al., 2019]

Extension of Isolation Forest [LTZ08] to the functional case



- A splitting rule: randomly and uniformly with one-dimensional projections
- A stopping rule: stop when each cell contains a single observation

Children Node Construction in a Functional Isolation Tree (1/2)



Children Node Construction in a Functional Isolation Tree (2/2)

$\mathcal{S}_{0,0} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ is a sample of r.v. in a Hilbert space $\mathcal{H} \triangleq \mathcal{C}_{0,0}$ and $\mathcal{D} \subset \mathcal{H}$.

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

1. Choose a Split variable $\mathbf{d}$ according to the probability distribution $\boldsymbol{Q}$ 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

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

Parameters of FIF

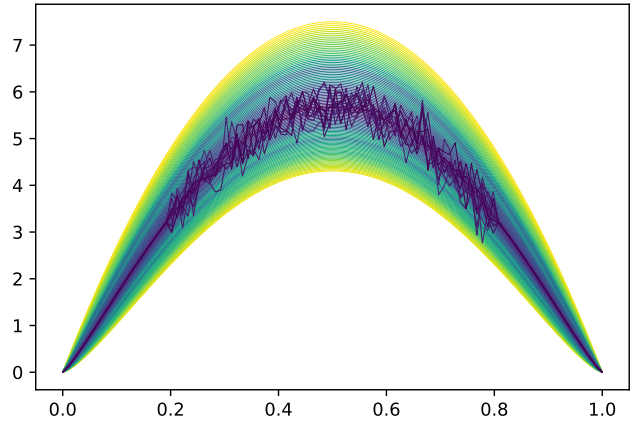
- Classical parameters of Isolation Forest: the number of trees, the size of the subsample and the height limit
- New parameters due to the functional setup:
  1. Dictionary $\mathcal{D}$
  2. Probability measure $\boldsymbol{\nu}$
  3. Inner product $\langle ., . \rangle_{\mathcal{H}}$

Example of inner product:

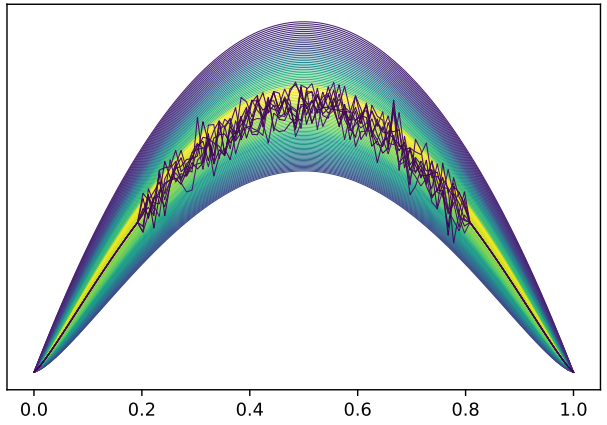
- Compromise between both location and shape:

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

$\alpha = 1$



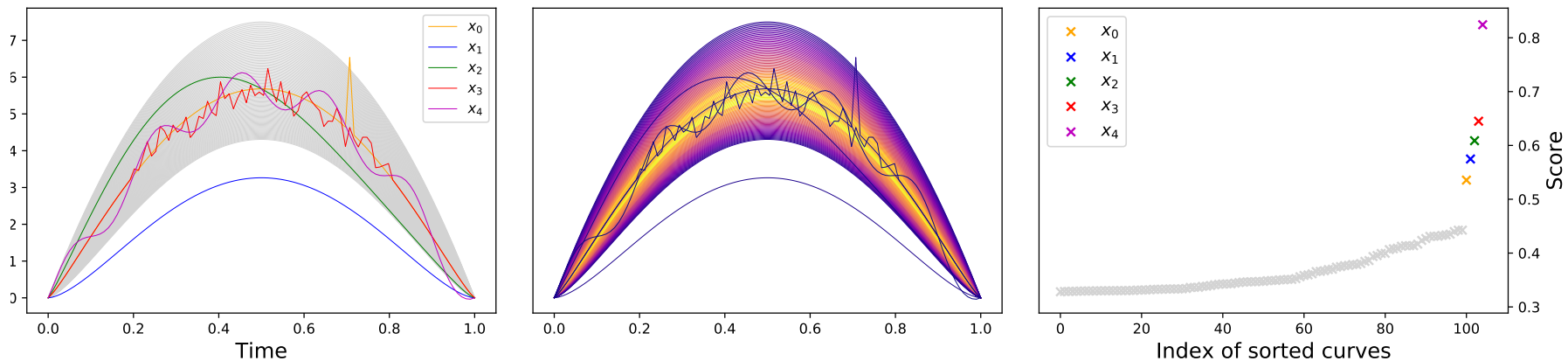
$\alpha = 0$



Ability to detect a Variety of Anomalies

- Sobolev inner product: $\langle ., . \rangle_{W_{1,2}}$ and uniform measure $\boldsymbol{Q}$
- Gaussian wavelets dictionary: $\mathcal{D} \triangleq \{\mathbf{d}_{\theta,\sigma}, \quad \theta \in [-4, 4], \sigma \in [0.2, 1]\}$

  with $\mathbf{d}_{\theta,\sigma}(t) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}} \left(1 - \left(\frac{t-\theta}{\sigma}\right)^2\right) \exp\left(\frac{-(t-\theta)^2}{2\sigma^2}\right)$



Performance on Real Datasets

13 real-world dataset from UCR Time Series Classification Archive[1]

Methods :	DI_{L_2}	Cos_{Sob}	Cos_{L_2}	$Self_{L_2}$	IF	LOF	OCSVM	FT	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 2: AUROC of different anomaly detection methods calculated on the test set.

[1] Chen, Yanping and Keogh, Eamonn and Hu, Bing and Begum, Nurjahan and Bagnall, Anthony and Mueen, Abdullah and Batista, Gustavo. The UCR Time Series Classification Archive, 2015. www.cs.ucr.edu/~eamonn/time_series_data/

Contributions

About robustness:

- MoM-based robust estimation of the 1-Wasserstein distance with theoretical guarantees and applications to WGANs
- Depth-regions based pseudo-metric extending the one-dimensional formula of the Wasserstein closed-form
- the AI-IRW depth, finite sample analysis of AI-IRW/IRW

About functional anomaly detection:

- A functional depth, the ACH depth, capable of detecting isolated anomalies
- Functional Isolation Forest: flexible algorithm capable of detecting a wide variety of anomalies

All approaches are implemented in Cython/C++ for FAD methods and in Python for multivariate ones and are available at https://github.com/GuillaumeStaermanML.

Perspectives

Local perspectives on $DR_{p,\varepsilon}$:

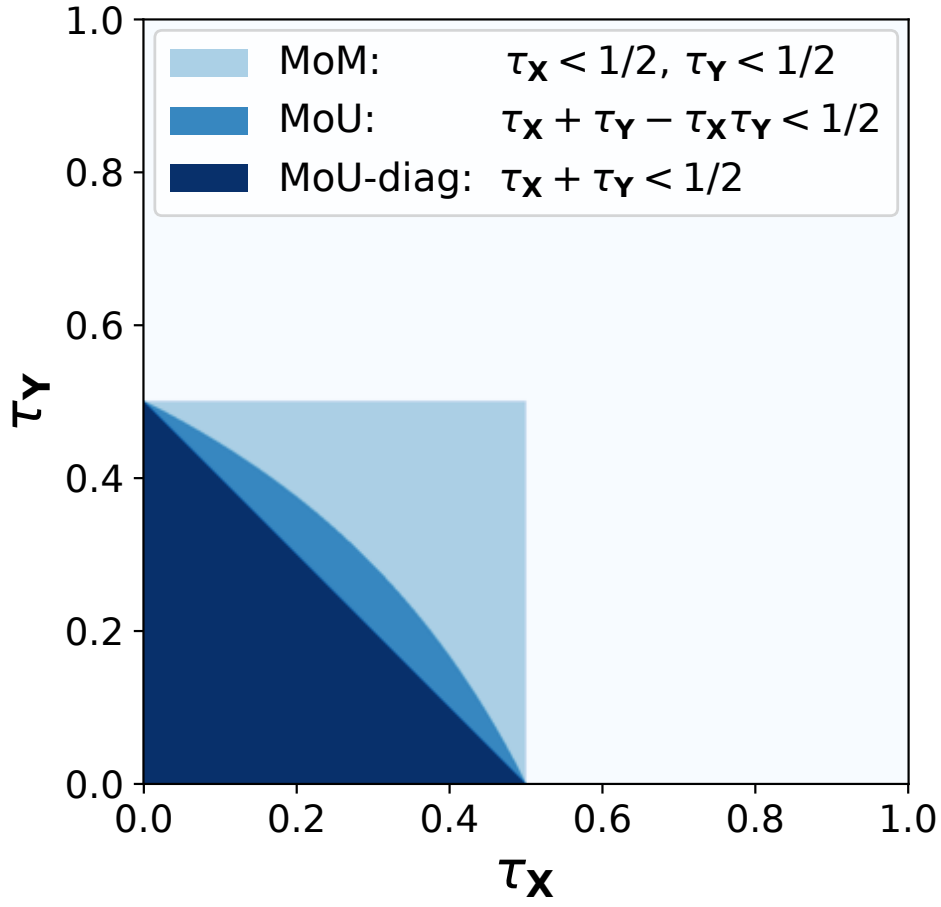
- Consistency, finite-sample analysis
- Statistical accuracy of the approximation
- Approximation for any data depth
- Replacing the Hausdorff distance

Global perspectives: reliable neural networks

- Adversarial attacks detection
- Out-of-Distribution detection
- Defense methods against adversarial attacks

Appendix of MoM-Wasserstein

Admitted proportion of outliers



Algorithm for $\mathcal{W}_{\mathrm{MoM}}$

Algorithm 8.1 Approximation of $\mathcal{W}_{\mathrm{MoM}}(P_n, Q_m)$.

input : ς, the learning rate. c, the clipping parameter. w_0, the initial weights. K_X, K_Y the number of blocks for $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$.

for $t = 0, \ldots, n_{iter}$ **do**

Sample K_X disjoint blocks $\mathcal{B}_1^X, \ldots, \mathcal{B}_{K_X}^X$ and K_Y disjoint blocks $\mathcal{B}_1^Y, \ldots, \mathcal{B}_{K_Y}^Y$ from a sampling scheme and find median blocks $\mathcal{B}_{med}^X$ and $\mathcal{B}_{med}^Y$

$$G_w \leftarrow \left\lfloor K_X/n \right\rfloor \sum_{i \in \mathcal{B}_{med}^X} \nabla_w \Psi_w(X_i) - \left\lfloor K_Y/m \right\rfloor \sum_{j \in \mathcal{B}_{med}^Y} \nabla_w \Psi_w(Y_j))$$

$w \leftarrow w + \varsigma \times \mathrm{RMSProp}(w, G_w)$

$w \leftarrow \mathrm{clip}(w, -c, c)$

return w, $\widetilde{\mathcal{W}}_{\mathrm{MoM}}$, Ψ_w.

Algorithm for $\mathcal{W}_{\mathrm{MoU-diag}}$

Algorithm 8.2 Computation of $\mathcal{W}_{\mathrm{MoU-diag}}(P_n, Q_m)$.

input : ς, the learning rate. c, the clipping parameter. w_0 the initial weights. K_X, K_Y the number of blocks for $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$.

for $t = 0, \ldots, n_{iter}$ **do**

Sample $K = K_X \wedge K_Y$ disjoint blocks $\mathcal{B}^{XY}_{1,1}, \mathcal{B}^{XY}_{2,2}, \ldots, \mathcal{B}^{XY}_{k,k}, \ldots \mathcal{B}^{XY}_{K,K}$ from a sampling scheme and find the median block $\mathcal{B}^{XY}_{med}$

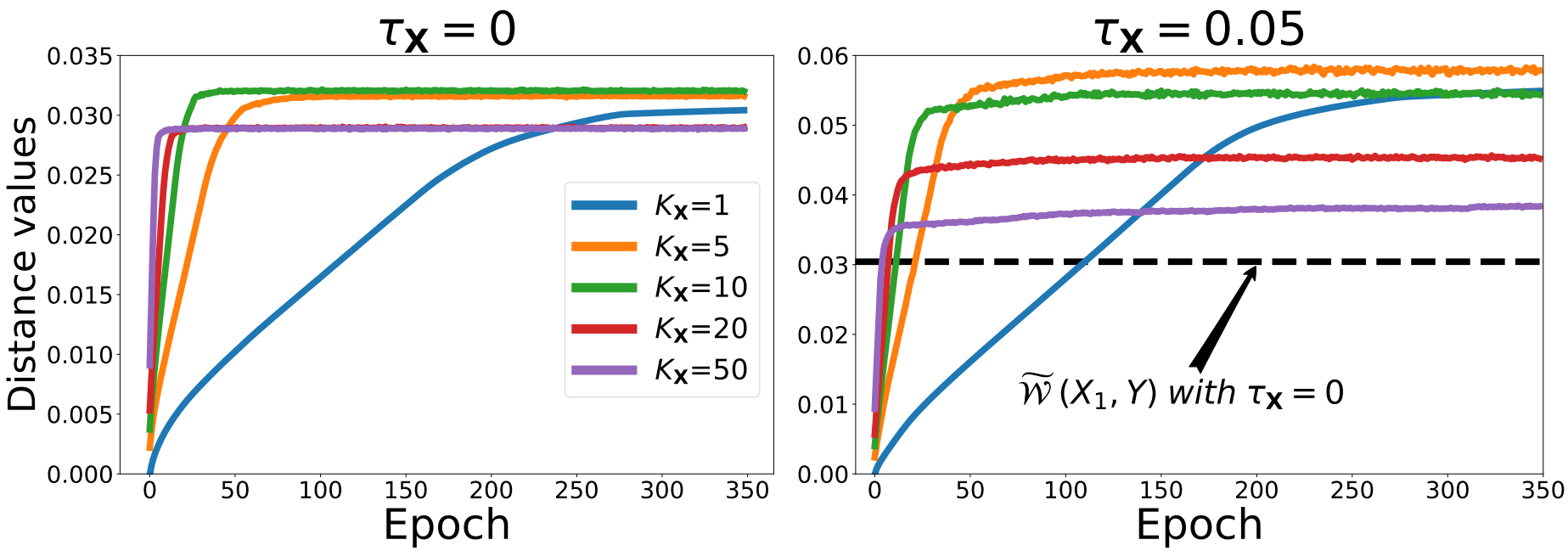
$$G_w \longleftarrow \Big\lfloor K/n \Big\rfloor \sum_{(i,j)\in\mathcal{B}^{XY}_{med}} \nabla_w \Big[\Psi_w(X_i) - \Psi_w(Y_j)\Big]$$

$w \leftarrow w + \varsigma \times \mathrm{RMSProp}(w, G_w)$

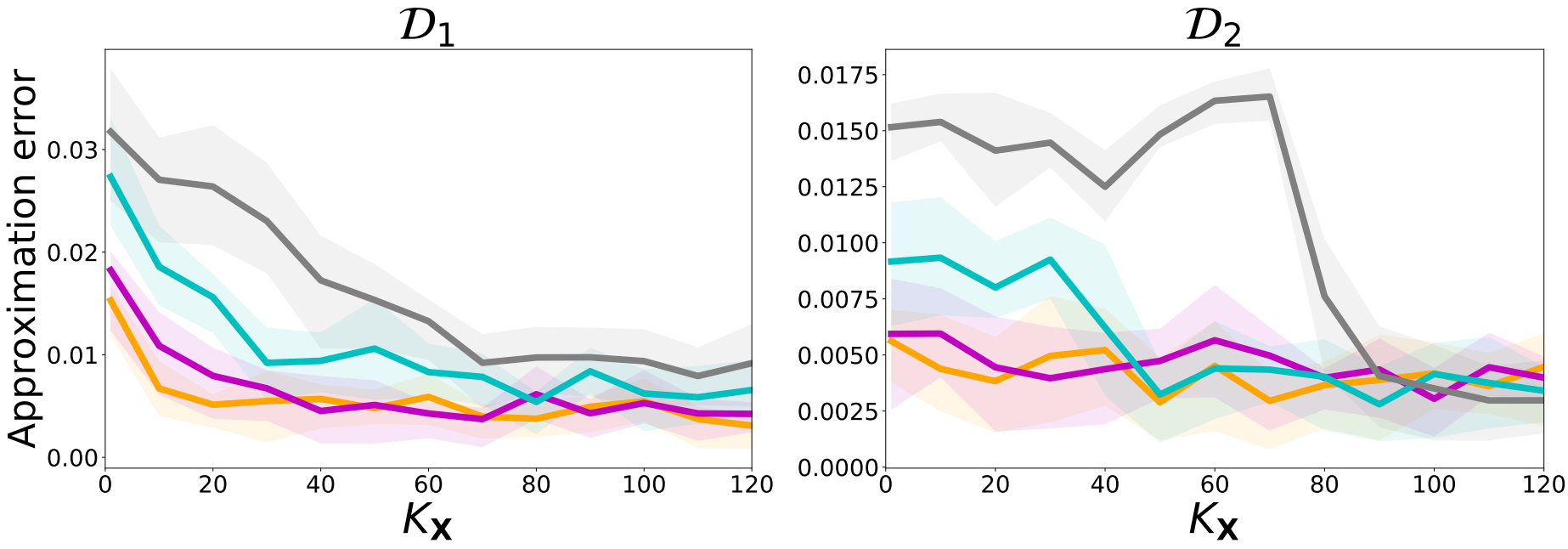
$w \leftarrow \mathrm{clip}(w, -c, c)$

return w, $\widetilde{\mathcal{W}}_{\mathrm{MoU-diag}}$, Ψ_w.

Convergence



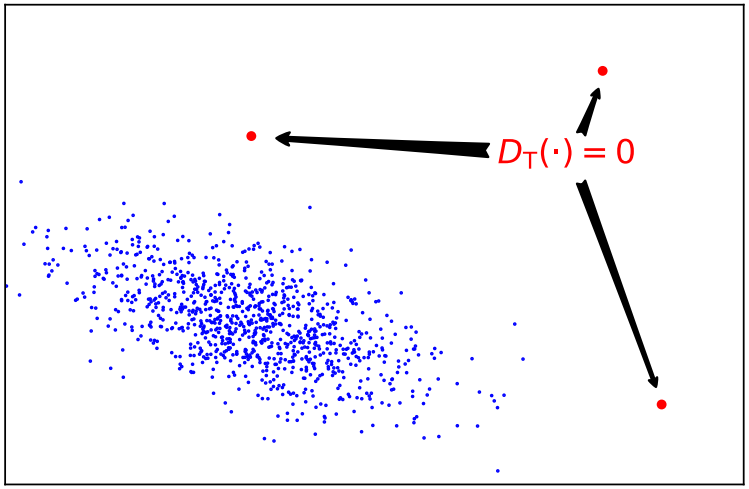
Robustness w.r.t. the number of blocks



Appendix of AI-IRW

Limitations of the Halfspace Depth

Halfspace depth is zero beyond the convex hull of the data.



How to compute the *minimum* over the unit sphere $\mathbb{S}^{d-1}$?

- Exact computation: computational complexity $O\left(n^{d-1}\log(n)\right)$ [DM16].
- Monte-Carlo approximation: (statistical) asymptotic rates in $O\left((\log(n_{proj})/n_{proj})^{\frac{1}{d-1}}\right)$ [DMN21].

Integrated Rank-Weighted (IRW) depth [Ramsay et al., 2019]

The IRW depth of $x \in \mathbb{R}^d$ w.r.t. $P \in \mathcal{P}(\mathbb{R}^d)$ is given by:

$$D_{\text{IRW}}(x, P) = \int_{\mathbb{S}^{d-1}} \min\left(F_u(\langle u, x\rangle), 1 - F_u(\langle u, x\rangle)\right)\ \omega_{d-1}(du)$$

where ω_{d-1} is the spherical probability measure on $\mathbb{S}^{d-1}$ and F_u is the cdf of the random variable $\langle u, X\rangle$, *i.e.* $F_u(t) = \mathbb{P}\left(\langle u, X\rangle \leq t\right)$.

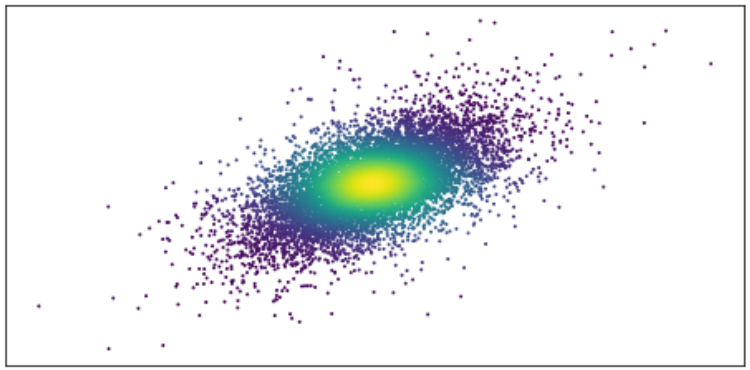
$\Rightarrow$ D_{IRW} satisfies (**D2**, **D3**, **D4**) but not **D1**!

Affine-Invariant Integrated Rank Weighted (AI-IRW) Depth [Staerman et al., 2021]

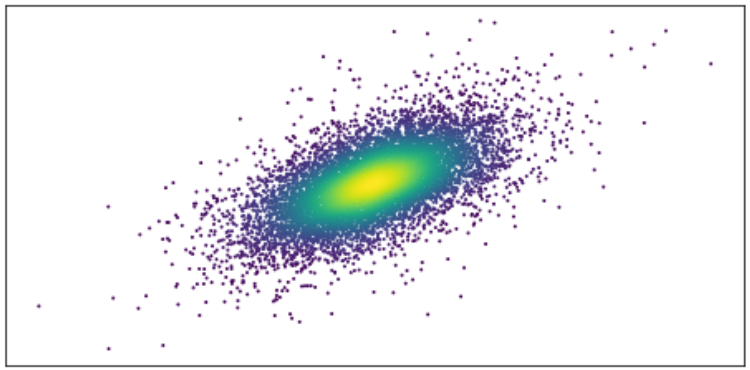
The AI-IRW depth of $x \in \mathbb{R}^d$ w.r.t. $P \in \mathcal{P}(\mathbb{R}^d)$ is given by:

$$D_{\text{AI-IRW}}(x, P) = \mathbb{E}\left[\min\left(F_V(\langle V, x\rangle), 1 - F_V(\langle V, x\rangle)\right)\right],$$

where $V = \Sigma^{-\top/2}U/||\Sigma^{-\top/2}U||$ and U being uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$.



IRW



AI-IRW

Inference of AI-IRW

Generating a number $n_{\text{proj}} \geq 1$ i.i.d. random directions $U_1, \ldots, U_m$, copies of the generic r.v. U and independent from the original data $\mathcal{D}_n = \{X_1, \ldots, X_n\}$: $\forall x \in \mathbb{R}^d$,

$$\tilde{D}^{\text{MC}}_{\text{AI-IRW}}(x, \mathcal{D}_n) = \frac{1}{n_{\text{proj}}} \sum_{j=1}^{n_{\text{proj}}} \min\left\{ \widehat{F}_{\hat{V}_j}\left(\langle \hat{V}_j,\ x\rangle\right), 1 - \widehat{F}_{\hat{V}_j}\left(\langle \hat{V}_j,\ x\rangle\right)\right\},$$

where, for all $j \in \{1,\ \ldots, n_{\text{proj}}\}$ and $t \in \mathbb{R}$, we set

$$\hat{V}_j = \hat{\Sigma}^{-\top/2} U_j / ||\hat{\Sigma}^{-\top/2} U_j|| \text{ and } \widehat{F}_{\hat{V}_j}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\langle \hat{V}_j, X_i\rangle \leq t\}}.$$

Goal: Finite-Sample Analysis

Control the deviations between the approximation of the empirical depth $\tilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}(x, \mathcal{D}_n)$ and $D_{\text{AI-IRW}}(x, P)$ such that for any finite n and n_{proj}, with high probability we have:

$$\underbrace{\sup_{x \in ?}}_{?} |\tilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}(x, \mathcal{D}_n) - D_{\text{AI-IRW}}(x, P)| \leq \text{Rates}\ (n, n_{\mathrm{proj}}, d)$$

Existing literature:

- Non-asymptotic analysis of the halfspace depth [SW86, BF17].
- Asymptotic rates for the MC approximation of the Halfspace depth [NDM20].

Sketch of Proof

$$|\tilde{D}^{\text{MC}}_{\text{AI-IRW}}(x, \mathcal{D}_n) - D_{\text{AI-IRW}}(x, P)| \leq \underbrace{\left|\hat{D}_{\text{AI-IRW}}(x, \mathcal{D}_n) - D_{\text{AI-IRW}}(x, P)\right|}_{\text{Sample estimation}} + \underbrace{\left|D^{\text{MC}}_{\text{AI-IRW}}(x, P) - D_{\text{AI-IRW}}(x, P)\right|}_{\text{Monte-Carlo approximation}}$$

Sample estimation:

- Step 1. Linearization through Lipschitz Ass in proj.
- Step 2. Controlling halfspace deviations.
- Step 3. Control of $||\widehat{\Sigma}^{-1/2} - \Sigma^{-1/2}||$
  (a) Eigenvectors by Davis-Kahan Theorem
  (b) Eigenvalues by Weyl's inequality

MC approximation:

- Step 1. Linearization through radial Lipschitz Ass.
- Step 2. Chaining technique using the fact that the covering number of $\mathcal{B}_r$ is finite.

Assumptions (1/2)

Assume that $\phi_P : (u, x) \in \mathbb{S}^{d-1} \times \mathbb{R}^d \mapsto \mathbb{P}\{\langle u, X \rangle \leq \langle u, x \rangle\}$.

Uniform radial Lipschitz condition

For all $(u, v) \in \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$, there exists $L_R < +\infty$ such that

$$\sup_{x \in \mathbb{R}^d} |\phi_P(u, x) - \phi_P(v, x)| \leq L_R ||u - v||.$$

Uniform Lipschitz condition in projection

For all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, there exists $L_p < +\infty$ such that

$$\sup_{u \in \mathbb{S}^{d-1}} |\phi_P(u, x) - \phi_P(u, y)| \leq L_{pr} ||x - y||.$$

- $L_R = M\ V_{d,r}$ and $L_{pr} = M\ V_{d-1,r}$ when P has M-bounded density

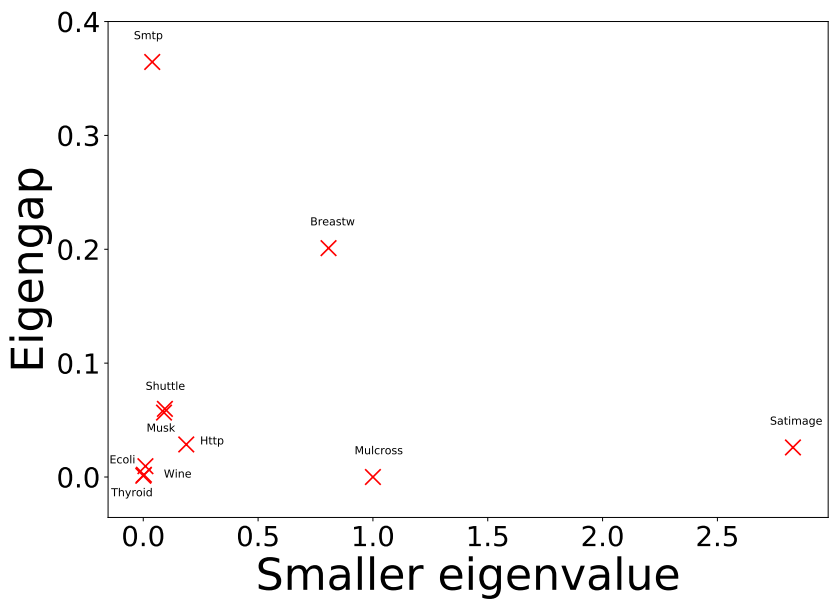
Assumptions (2/2)

Minimum eigenvalue

There exists $\varepsilon > 0$ such that: $\forall j \in \{1, \ldots, d\}, \varepsilon \leq \sigma_j$.

Non-zero eigengap

Assume $\gamma > 0$ s.t. $\gamma = \min\{\sigma_{(i)} - \sigma_{(i+1)}, \ 1 \leq j \leq d\}$, where $\sigma_{(1)} \geq \ldots \geq \sigma_{(d)}$ are sorted Σ's eigenvalues.

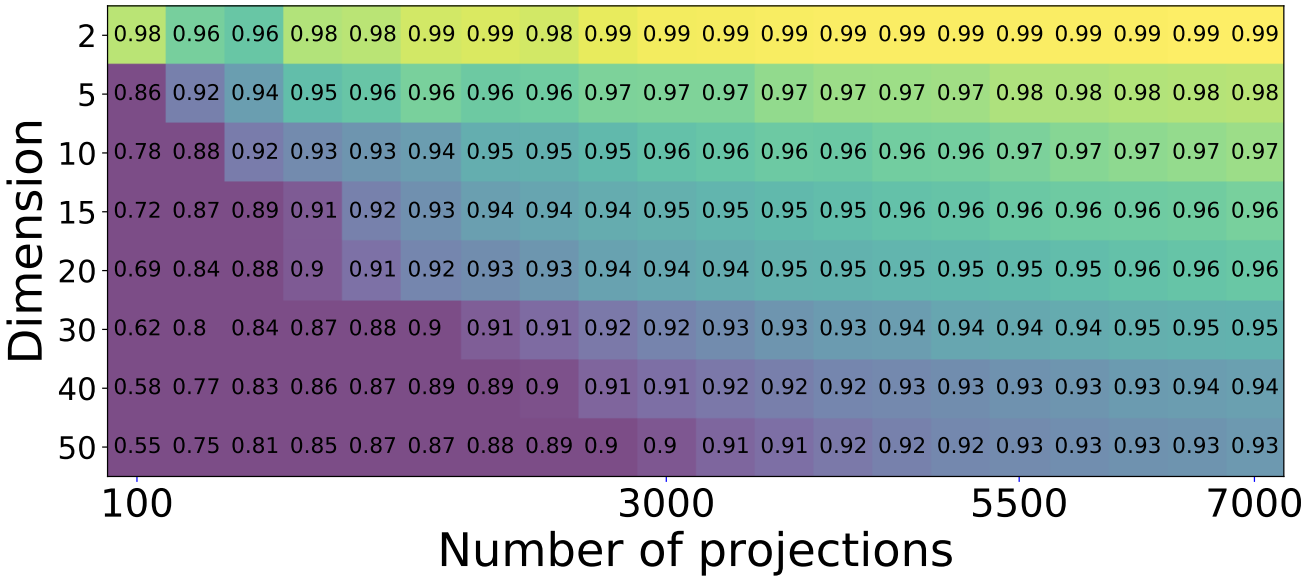


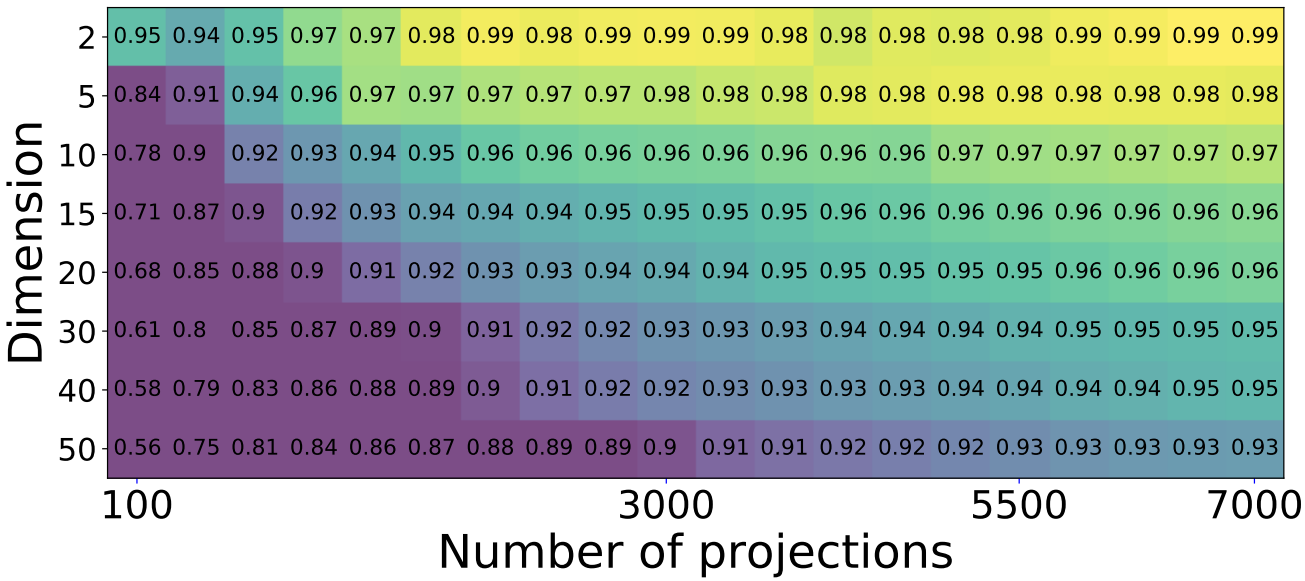
Theorem [Staerman et al., 2021]

For any $\delta \in (c_1 e^{-c_2 n}, 1)$, we have with probability at least $1-\delta$:

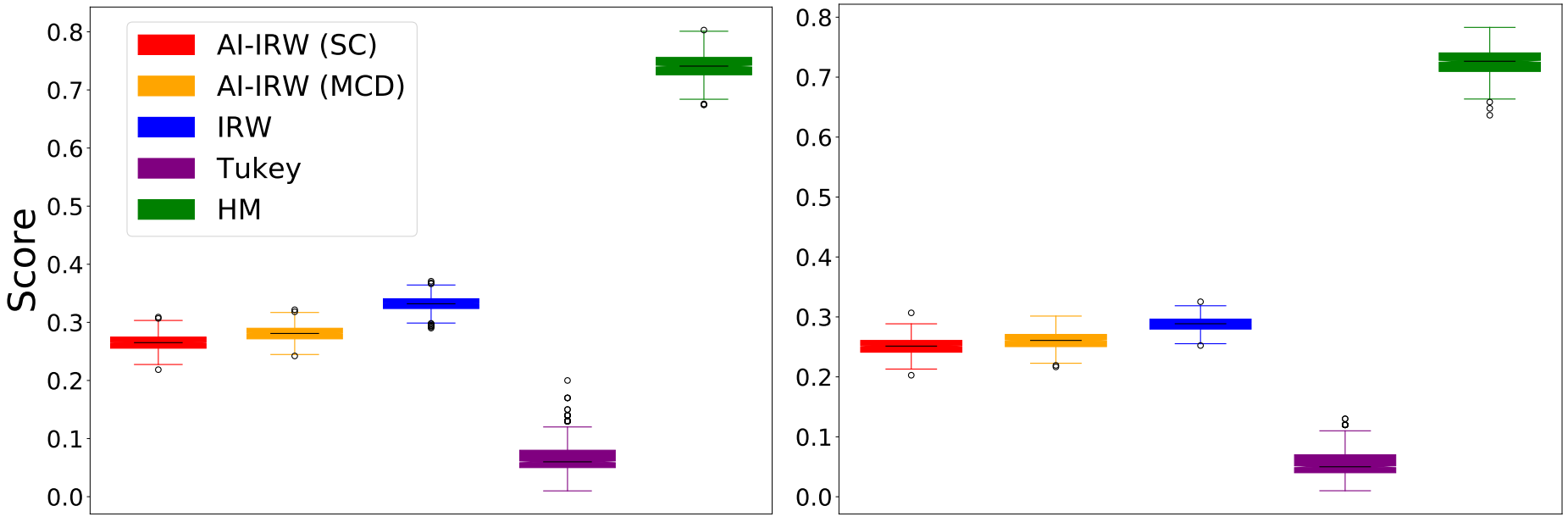
$$\sup_{x \in \mathcal{B}_r} \left| \tilde{D}^{\mathrm{MC}}_{\text{AI-IRW}}(x, \mathcal{D}_n) - D_{\text{AI-IRW}}(x, P) \right| \leq \sqrt{\frac{128 \log(c_3 n/\delta)}{9n}} + \frac{c_4 L_R}{\varepsilon \wedge \gamma} \max_{s=1,2} \left(\frac{d + \log(2/\delta)}{n} \right)^{1/s} + \frac{4 L_{pr}}{3 n_{\mathrm{proj}}} + 2\sqrt{\frac{d \log(3 r n_{\mathrm{proj}}) + \log(6/\delta)}{18 n_{\mathrm{proj}}}}.$$

Approximation quality

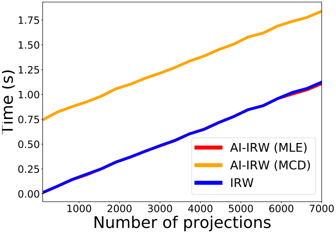


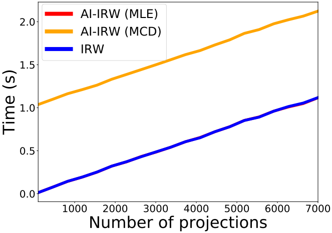


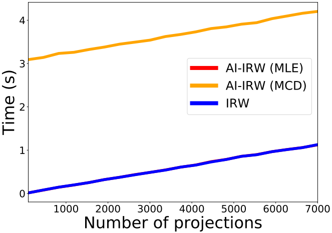
Variance of the score

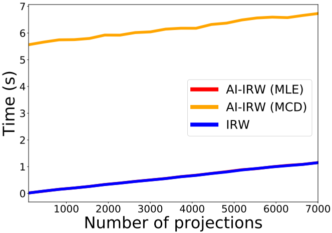


Computation time

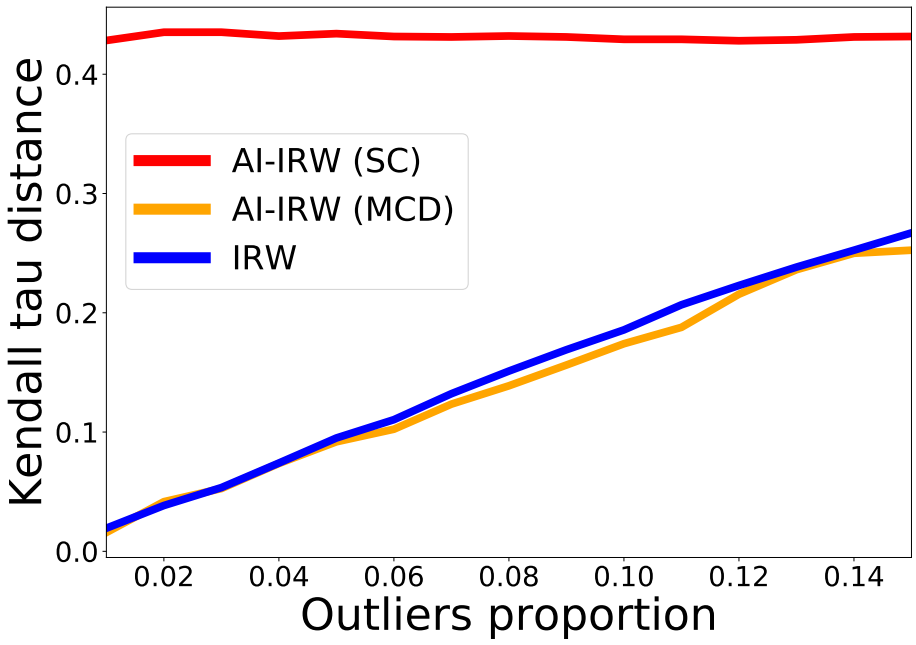


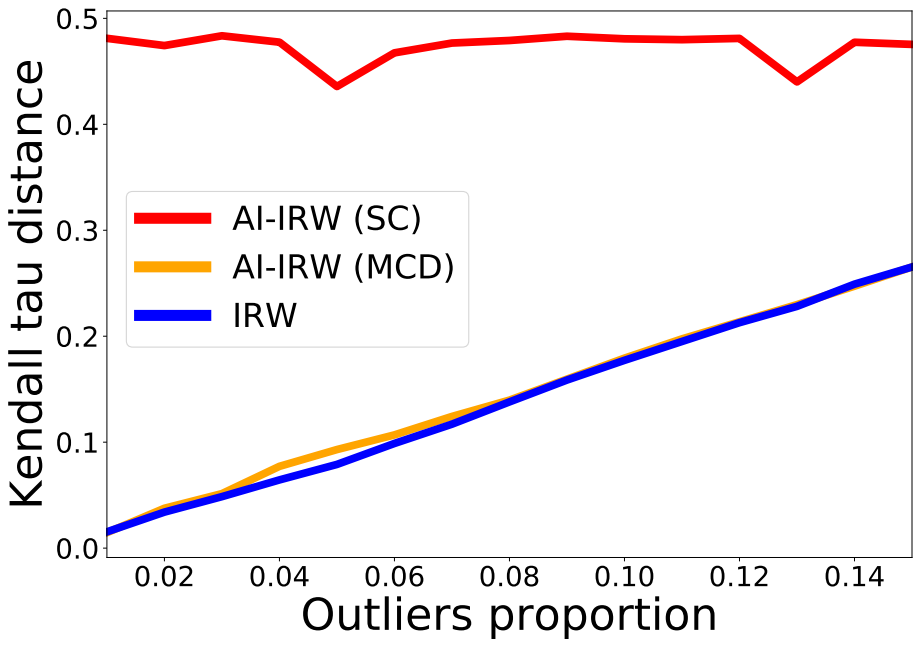






Robustness


Kendall tau distance
0.0
0.1
0.2
0.3
0.4
AI-IRW (SC)
AI-IRW (MCD)
IRW
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Outliers proportion


Kendall tau distance
0.0
0.1
0.2
0.3
0.4
0.5
AI-IRW (SC)
AI-IRW (MCD)
IRW
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Outliers proportion

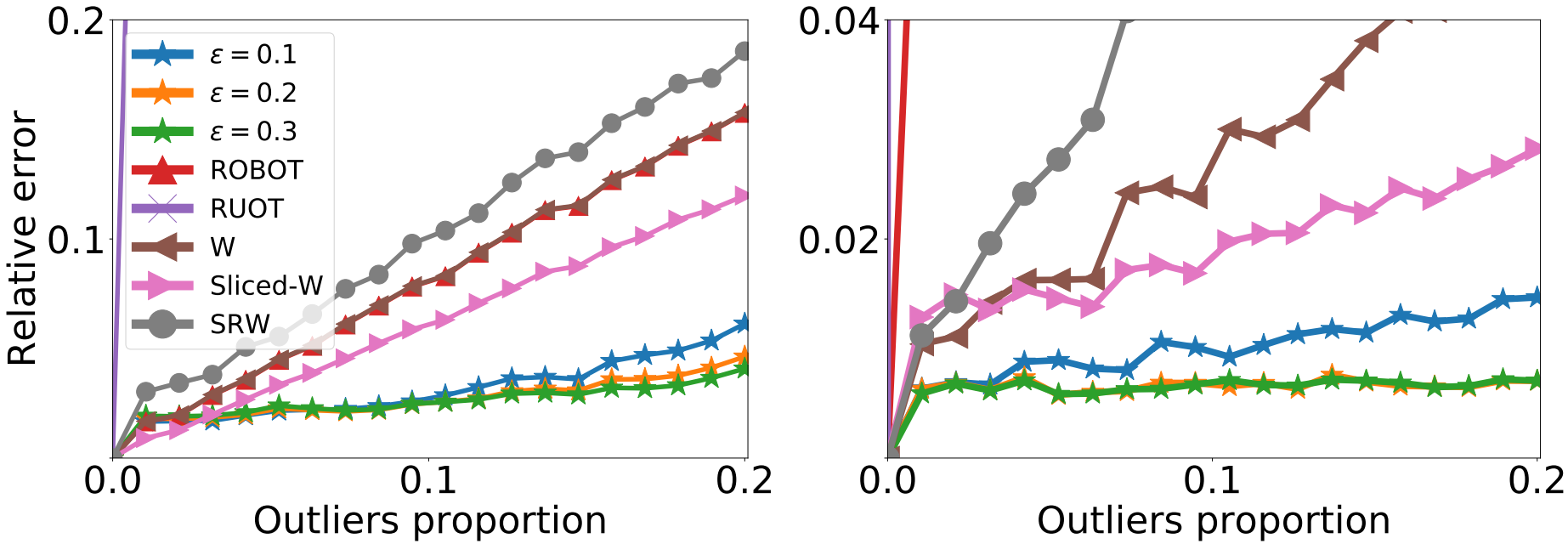
Anomaly detection

	AI-IRW	IRW	HM	T	IF	AE
Ecoli	0.85	0.83	**0.88**	0.68	0.77	0.64
Shuttle	0.99	0.99	0.99	0.86	0.99	0.99
Mulcross	**1**	0.98	**1**	0.87	0.96	**1**
Thyroid	**0.98**	0.80	0.84	0.92	0.97	0.97
Wine	0.96	0.96	**0.99**	0.71	0.8	0.72
Http	**1**	0.95	0.97	0.99	**1**	**1**
Smtp	**0.96**	0.77	0.74	0.85	0.90	0.82
Breastw	0.97	0.97	**0.99**	0.84	**0.99**	0.91
Musk	**1**	0.84	0.97	0.77	**1**	**1**
Satimage	**0.99**	0.96	0.98	0.95	**0.99**	0.98

Table 3: AUROCs of benchmarked anomaly detection methods.

Appendix
Depth-Trimmed-Regions
Pseudo-Metric

Robustness


0.2
0.1
Relative error
$\varepsilon = 0.1$
$\varepsilon = 0.2$
$\varepsilon = 0.3$
ROBOT
RUOT
W
Sliced-W
SRW
0.0
0.1
0.2
Outliers proportion
0.04
0.02
0.0
0.1
0.2
Outliers proportion

Appendix ACH

Properties of the ACH depth

- **Non-degeneracy**.
- **scale-translation invariance** with $a, b \in \mathbb{R}$.
- **Vanishing at infinity**.
- **Continuity in** x.
- **Uniform continuity in** P.
- **Consistency:** With probability one,

$$\sup_{x \in \mathcal{C}([0,1])} |FD_{J,n}(x) - FD_J(x, P)| \to 0.$$

- **Approximation consistency:** Suppose that, as $n \to \infty$,

$$\delta = \max_{1 \le i \le n} \max_{2 \le k \le p_i} \left\{ t_{k+1}^{(i)} - t_k^{(i)} \right\} \to 0.$$

As $n \to \infty$, we have, with probability one,

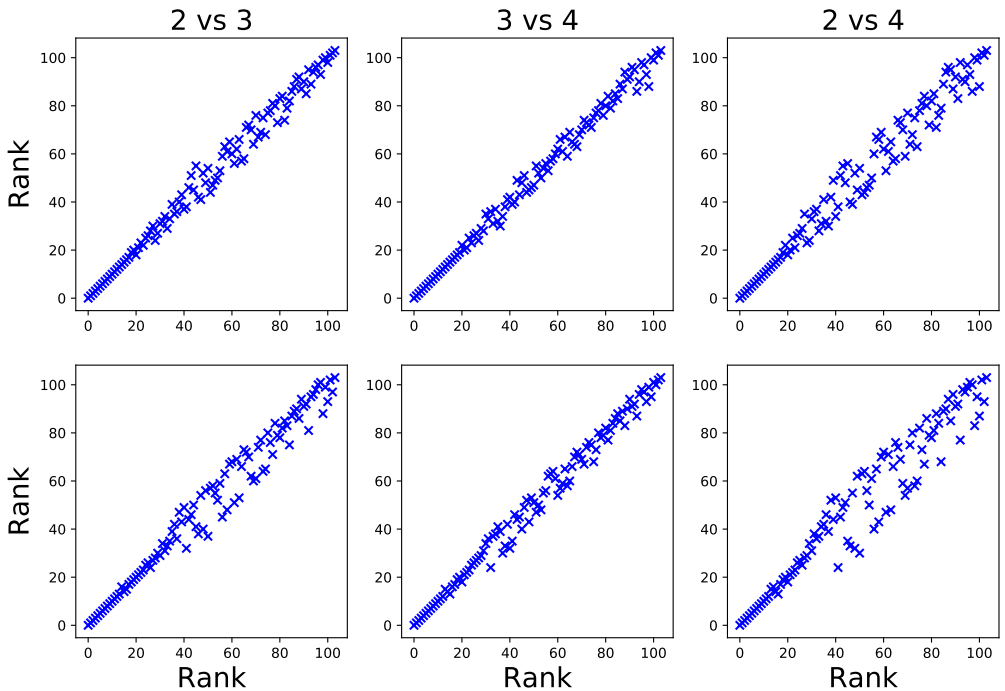
$$\sup_{x \in \mathcal{C}([0,1])} \left| FD'_{J,n}(x) - FD_J(x, P) \right| \to 0.$$

Choice of J

- Not an easy question !
- The use of an average version $\overline{FD}_J$ defined by: $\forall \mathbf{x} \in \mathcal{C}([0,1])$,

$$\overline{FD}_J(\mathbf{x}, \mathbf{P}) = \frac{1}{J} \sum_{j=2}^{J} FD_j(\mathbf{x}, \mathbf{P}).$$

is possible due to the approximation version, but the benefits of the averaging are still not clear.



Robustness (1/1)

The stability of the returned rank is studied adding a percentage $\alpha \in [0, 30]$ of anomalies separately:

- Location for dataset (a), Isolated and Shape for dataset (b).
- The performance is evaluated using Kendall's tau distance.

We compare the robustness of ACH depth with the three following anomaly detection/functional depth methods:

- Functional Stahel Donoho Outlyingness (Hubert et al., 2015)
- Functional Tukey Depth (Claeskens et al., 2014)
- Functional Isolation Forest (Staerman et al., 2019)

Robustness (2/2)

$d_\tau(\sigma_0, \sigma_\alpha)(\times 10^{-2})$							
	α	0	5	10	15	25	30
ACH	Location	0	**0.6**	**1.3**	**2.2**	**4.3**	**5.2**
	Isolated	0	**0.3**	**1.3**	**0.9**	**1.6**	**2.4**
	Shape	0	**0.9**	**2**	**2.6**	**4.2**	**4.7**
FSDO	Location	0	3.6	7.3	10	16	20
	Isolated	0	0.8	3.6	3.2	7.2	9.4
	Shape	0	1.6	2.9	4.2	6.6	7.4
FT	Location	0	5.1	9.5	13	20	23
	Isolated	0	0.7	2.7	2.7	5.9	7.2
	Shape	0	1.7	2.9	4.3	6.6	7.7
FIF	Location	0	7	8.2	7.3	7.3	8.9
	Isolated	0	9.3	12	11	10	12
	Shape	0	7.4	7.9	10	14	14

Appendix FIF

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:

$$\begin{aligned} x : [0,1] &\longrightarrow \mathbb{R}^d \\ t &\longmapsto \left((x^1(t), \ldots, x^d(t)\right) \end{aligned}$$

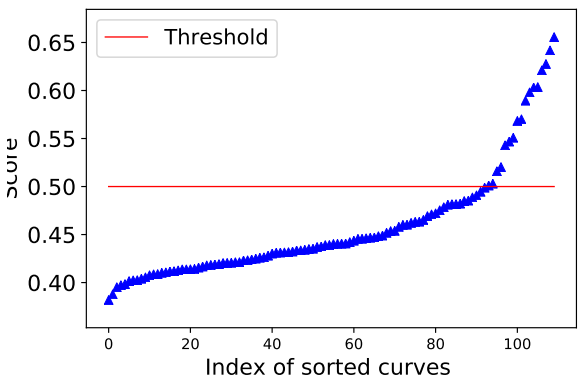
- Coordinate-wise sum of the d corresponding scalar products:

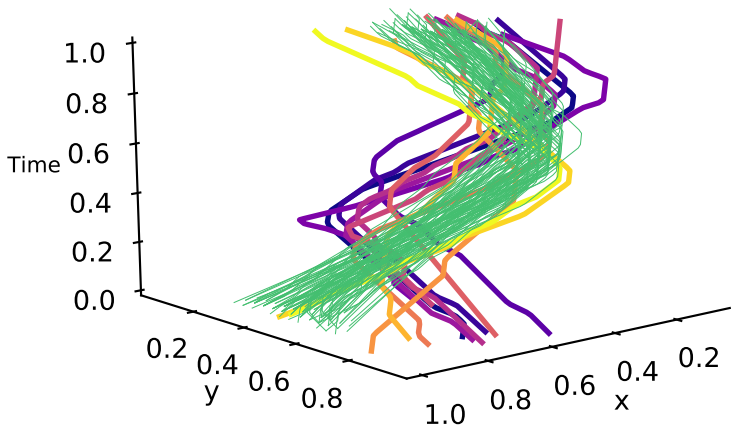
$$\langle \mathbf{f}, \mathbf{g} \rangle_{L_2^{\otimes d}} := \sum_{i=1}^{d} \langle f^{(i)}, g^{(i)} \rangle_{L_2}$$

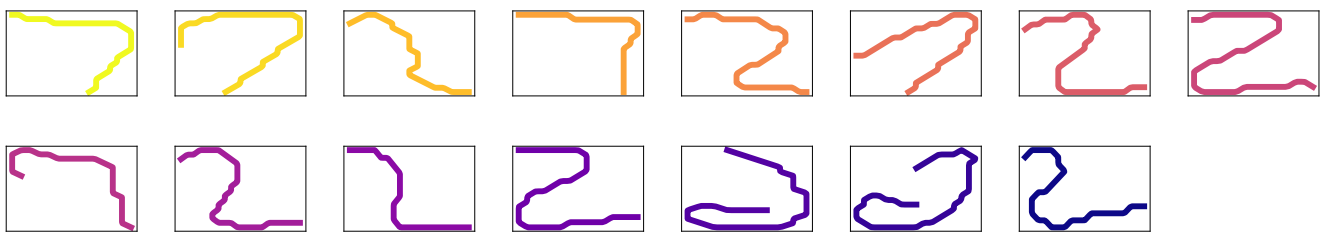
- Dictionaries : Composed by univariate function on each axis, multivariate wavelets, multivariate Brownian motion ...

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 $[0,1]$.







References i


Jinwon An and Sungzoon Cho.
Variational autoencoder based anomaly detection using reconstruction probability.
Special Lecture on IE, 2(1):1–18, 2015.


Ana Arribas-Gil and Juan Romo.
Shape outlier detection and visualization for functional data: the outliergram.
Biostatistics, 15(4):603–619, 2014.


Claudio Agostinelli and Mario Romanazzi.
Ordering curves by data depth.
Studies in Classification, Data Analysis, and Knowledge Organization, pages 1–8, 2013.

References ii


Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi.
Bandits with heavy tail.
IEEE Transactions on Information Theory, 59(11):7711–7717, 2013.


Yogesh Balaji, Rama Chellappa, and Soheil Feizi.
Robust optimal transport with applications in generative modeling and domain adaptation.
In *Advances in Neural Information Processing Systems*, volume 33, pages 12934–12944, 2020.


Michael A. Burr and Robert J. Fabrizio.
Uniform convergence rates for halfspace depth.
Statistics & Probability Letters, 124:33–40, 2017.


Christian Brownlees, Emilien Joly, and Gábor Lugosi.
Empirical risk minimization for heavy-tailed losses.
The Annals of Statistics, 43(6):2507–2536, 2015.

References iii

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander.

Lof: identifying density-based local outliers.

In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, volume 29, pages 93–104, 2000.

Olivier Catoni.

Challenging the empirical mean and empirical variance: a deviation study.

In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185, 2012.

References iv


Nicolas Courty, Rémi Flamary, and Devis Tuia.
Domain adaptation with regularized optimal transport.
In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 274–289, 2014.


Imre Csiszàr.
Eine informationstheoretische ungleichung und ihre anwendung auf den bewis der ergodizität von markhoffschen kette.
Magyer Tud. Akad. Mat. Kutato Int. Koezl, 8:85–108, 1963.


Jules Depersin.
Robust subgaussian estimation with vc-dimension.
arXiv preprint arXiv:2004.11734, 2020.

References v


David L. Donoho and Miriam Gasko.
Breakdown properties of location estimates based on half space depth and projected outlyingness.
The Annals of Statistics, 20:1803–1827, 1992.


Wenlin Dai and Marc Genton.
Multivariate functional data visualization and outlier detection.
Journal of Computational and Graphical Statistics, 27(4):923–934, 2019.


Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira.
Sub-gaussian mean estimators.
The Annals of Statistics, 44(6):2695–2725, 2016.


Rainer Dyckerhoff and Pavlo Mozharovskyi.
Exact computation of the halfspace depth.
Computational Statistics & Data Analysis, 98:19–30, 2016.

References vi


Rainer Dyckerhoff, Pavlo Mozharovskyi, and Stanislav Nagy.
Approximate computation of projection depths.
Computational Statistics & Data Analysis, 157(C):107166, 2021.


John H. J. Einmahl and David M. Mason.
Generalized Quantile Processes.
The Annals of Statistics, 20(2):1062–1078, 1992.


Ricardo Fraiman and Graciela Muniz.
Trimmed means for functional data.
Test, 10(2):419–440, 2001.


Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola.
A kernel method for the two-sample-problem.
In *Advances in Neural Information Processing Systems*, volume 19, pages 513–520, 2007.

References vii


Markus Goldstein and Andreas Dengel.
Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm.
KI-2012: Poster and Demo Track, pages 59–63, 2012.


Corrado Gini and Luigi Galvani.
Di talune estensioni, dei concetti di media ai caratteri qualitativi.
Metron, 8:3–209, 1929.


Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.
Generative adversarial nets.
In *Advances in Neural Information Processing Systems (NeurIPS 2014)*, 2014.

References viii


John B. S. Haldane.
Note on the median of a multivariate distribution.
Biometrika, 35(3-4):414–417, 1948.


John F. Hayford.
What is the center of an area, or the center of a population?
Publications of the American Statistical Association, 8(58):47–58, 1902.


Joseph L. Hodges.
A Bivariate Sign Test.
The Annals of Mathematical Statistics, 26(3):523–527, 1955.


Mia Hubert, Peter J. Rousseeuw, and Pieter Segaert.
Multivariate functional outlier detection.
Statistical Methods & Applications, 24(2):177–202, 2015.

References ix


Trevor Harris, J. Derek Tucker, Bo Li, and Lyndsay Shand.
Elastic depths for detecting shape anomalies in functional data.
Technometrics, 63(4):466–476, 2021.


Emilien Joly and Gábor Lugosi.
Robust estimation of u-statistics.
Stochastic Processes and their Applications, 126(12):3760–3773, 2016.


Rebecka Jörnsten.
Clustering and classification based on the l1 data depth.
Journal of Multivariate Analysis, 90(1):67–89, 2004.


L. Kantorovitch.
On the translocation of masses (in russian).
In *Proceedings of the USSR Academy of Sciences.*, 1942.

References x


Sonja Kuhnt and André Rehage.
An angle-based multivariate functional pseudo-depth for shape outlier detection.
Journal of Multivariate Analysis, 146:325–340, 2016.


Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger.
From word embeddings to document distances.
In *International conference on machine learning*, pages 957–966. PMLR, 2015.


Jun Li, Juan A. Cuesta-Albertos, and Regina Y. Liu.
Dd-classifier: Nonparametric classification procedure based on dd-plot.
Journal of the American Statistical Association, 107(498):737–753, 2012.

References xi


Pierre Laforgue, Stephan Clémençon, and Patrice Bertail.
On medians of (randomized) pairwise means.
In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.


Tianyi Lin, Nhat Ho, and Michael I. Jordan.
On the acceleration of the sinkhorn and greenkhorn algorithms for optimal transport, 2019.


Regina Y. Liu.
On a notion of data depth based on random simplices.
The Annals of Statistics, 18(1):405–414, 1990.


Guillaume Lecué and Matthieu Lerasle.
Robust machine learning by median-of-means: Theory and practice.
The Annals of Statistics, 48(2):906–931, 2020.


Guillaume Lecué, Matthieu Lerasle, and Timlothée Mathieu.
Robust classification via mom minimization.
Machine Learning, 109(8):1635–1665, 2020.


Gabor Lugosi and Shahar Mendelson.
Risk minimization by median-of-means tournaments.
Journal of the European Mathematical Society, 2019.


Tatjana Lange, Karl Mosler, and Pavlo Mozharovskyi.
Fast nonparametric classification based on data depth.
Statistical Papers, 55(1):49–69, 2014.


Sara López-Pintado and Juan Romo.
Depth-based classification for functional data.
DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 72:103, 2006.


Sara López-Pintado and Juan Romo.
On the concept of depth for functional data.
Journal of the American Statistical Association, 104(486):718–734, 2009.


Sara López-Pintado and Juan Romo.
A half-region depth for functional data.
Computational Statistics & Data Analysis, 55(4):1679–1695, 2011.


Pierre Laforgue, Guillaume Staerman, and Stephan Clémençon.
Generalization bounds in the presence of outliers: a median-of-means study.
In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5937–5947, 2021.


Matthieu Lerasle, Zoltan Szabo, Timothée Mathieu, and Guillaume Lecué.
Monk–outlier-robust mean embedding estimation by median-of-means.
In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.


Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou.
Isolation forest.
In *In Proceedings 8th IEEE International Conference on Data Mining*, pages 413–422, 2008.


Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou.
Isolation-based anomaly detection.
ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1):1–39, 2012.


Debarghya Mukherjee, Aritra Guha, Justin M Solomon, Yuekai Sun, and Mikhail Yurochkin.
Outlier-robust optimal transport.
In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 7850–7860, 2021.


Pavlo Mozharovskyi, Julie Josse, and François Husson.
Nonparametric imputation by data depth.
Journal of the American Statistical Association, 115(529):241–253, 2020.


Sloan Nietert, Rachel Cummings, and Ziv Goldfeld.
Outlier-robust optimal transport: Duality, structure, and statistical analysis.
arXiv preprint arXiv:2111.01361, 2021.


Stanislav Nagy, Rainer Dyckerhoff, and Pavlo Mozharovskyi.
Uniform convergence rates for the approximated halfspace and projection depth.
Electronic Journal of Statistics, 14(2):3939–3975, 2020.


Stanislav Nagy, Irène Gijbels, and Daniel Hlubinka.
Depth-based recognition of shape outlying functions.
Journal of Computational and Graphical Statistics, 26(4):883–893, 2017.


Alicia Nieto-Reyes and Heather Battey.
A Topologically Valid Definition of Depth for Functional Data.
Statistical Science, 31(1):61–79, 2016.


Peter J. Rousseeuw, Jakob Raymaekers, and Mia Hubert.
A measure of directional outlyingness with applications to image data and video.
Journal of Computational and Graphical Statistics, 27(2):345–359, 2018.


Douglas Scates.
Locating the median of the population in the united states.
Metron, 11:49–66, 1933.


Robert Serfling.
Depth functions in nonparametric multivariate inference.
DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 72:1, 2006.

References xviii


Clayton Scott and Robert Nowak.
Learning minimum volume sets.
Journal of Machine Learning Resarch, 7:665–704, 2006.


Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex Smola, and Robert C. Williamson.
Estimating the support of a high-dimensional distribution.
Neural Computation, 13(7):1443–1471, 2001.


Kumar Sricharan, Raviv Raich, and Alfred O. Hero.
k-nearest neighbor estimation of entropies with confidence.
In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 1205–1209, 2011.


Galen R. Shorack and Jon A. Wellner.
Empirical Processes with Applications to Statistics.
John Wiley & Sons, 1986.


John W. Tukey.
Mathematics and the picturing of data.
In *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531, 1975.


Chong Zhou and Randy C Paffenroth.
Anomaly detection with robust deep autoencoders.
In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674, 2017.


Manqi Zhao and Venkatesh Saligrama.
Anomaly detection with score functions based on nearest neighbor graphs.
In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, page 2250–2258, 2009.