Functional anomaly detection and robust estimation

Guillaume Staerman

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAT021

Thèse de doctorat

Functional Anomaly Detection and Robust Estimation

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 Ecole Doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Mathématiques et Informatique

Thèse présentée et soutenue à Palaiseau, le 12 Avril 2022, par

GUILLAUME STAERMAN

Composition du Jury :

Nicolas Vayatis Professor, École normale supérieure Paris-Saclay (CMLA)	Président
Zhi-Hua Zhou Professor, Nanjing University (LAMBDA)	Rapporteur
Zoltán Szabó Professor, London School of Economics	Rapporteur
Sara Lopez-Pintado Associate Professor, Northeastern University	Examinatrice
Rémi Flamary Assistant Professor (HDR), École Polytechnique (CMAP)	Examinateur
Florence d'Alché-Buc Professor, Télécom Paris (LTCI)	Directrice de thèse
Pavlo Mozharovskyi Associate Professor, Télécom Paris (LTCI)	Co-encadrant de thèse
Stephan Clémençon Professor, Télécom Paris (LTCI)	Invité (Co-directeur)

Acknowledgements

Je souhaite débuter cette thèse en remerciant les personnes qui m'ont accompagné et soutenu tout au long de ces années de doctorat.

Mes premiers remerciements vont à mes directeurs de thèse, Florence, Stéphan et Pavlo qui m'ont accordé leur confiance et m'ont ouvert les portes de la recherche. Je vous remercie pour votre disponibilité, votre gentillesse et vos précieux conseils. Florence et Stéphan, merci pour le savoir que vous m'avez transmis et la confiance que vous m'avez apporté en me prenant en thèse. Vos portes sont toujours ouvertes ce qui reflète votre bienveillance et votre gentillesse. Je m'estime chanceux et je suis fier d'avoir pu effectuer cette thèse sous votre supervision. Pavlo, merci pour ta bienveillance et ta disponibilité, malgré ton emploi du temps bien rempli tu as toujours réussi à trouver un moment pour me venir en aide. Tu as été le moteur de cette thèse et je suis convaincu que c'est l'aide précieuse que tu m'as apporté, en particulier au cours de ma première année, qui m'a conduit là où je suis aujourd'hui.

I sincerely thank Zhi-Hua Zhou and Zoltán Szabó for having reviewed this thesis. It was an honor for me. Your comments have truly improved the quality of this thesis. Many thanks to all of my jury Sara Lopez-Pintado, Rémi Flamary and Nicolas Vayatis to have accepted to be present at my PhD defense. I am very grateful.

Un grand merci à Florence Besnard, Laurence Zelmar et Delphine Laude pour leur gentillesse, leur compréhension et leur dévouement.

Je remercie aussi profondément Thomas Moreau and Alexandre Gramfort, pour m'avoir donné l'opportunité de poursuivre en postdoc avec vous. J'ai vraiment hâte de travailler avec vous et l'équipe sur de nouveaux projets !

Je remercie aussi Pablo Piantanida et son équipe, Pierre, Eduardo, Marine, Frederica, Marco et Francesco pour les nombreuses discussions intéressantes qui m'ont fait découvrir le monde du Deep Learning. Vous formez une belle équipe et j'espère que ce n'est que le début d'une longue et fructueuse collaboration !

Un point essentiel dans ma formation est mon passage dans le M2 stat/ML d'Orsay où la qualité des enseignements ainsi que celle des étudiants m'a fait énormément progresser. Je tiens à remercier les responsables (à l'époque), Cristophe Giraud et Pierre Alquier, pour leur honnêteté et leur bienveillance, me permettant d'aborder la monde de la recherche dans les meilleurs conditions. Amaury et Ryad, on a commencé la thèse ensemble en M2 et on fini ensemble ! Il me semble que la team "Kolmogorovisation du fonctionnel" a bien fait son travail.

Je remercie ensuite mes premiers partenaires de lab. Pierre pour ta sympathie, ta fidélité et ta détermination à toute épreuve, tu es inspirant ! Si il y a bien une personne qui mérite d'être là (bien ce qu'en dira le paysan breton), c'est toi ! Anas, pour nos pauses interminables à essayer de répondre à une question simple "c'est quoi la recherche ?" et qui m'ont donné une motivation inébranlable pour la thèse ! Les deux Anas.s qui ont rendus sympathiques les week-ends passé à bosser à Télécom, shukran akhwan. J'espère qu'on ne se perdra jamais de vue. Emile, j'espère que tu pourras finalement monter ton AMAP en Bretagne ! Emilia, merci pour ta gentillesse et ta sympathie ainsi que

les innombrables relecture de ce manuscrit, si il y a aussi peu de typos c'est en grande partie grâce à toi. Arturo, merci pour ta sympathie et ta présence au sein du labo ainsi que les relectures de mon manuscrit. Junjie, merci pour ta bonne humeur, j'espère que je ne t'ai pas trop traumatisé ! J'ai aussi une pensée pour Yannick et Nathan N. qui nous ont partagé leur vision de ML et leur expérience de scientifique avec grand plaisir, ce fut très enrichissant.

Merci à la génération d'après, Lucien, Dimitri, Luc, Jayneel, Rémi, bonne chance (et courage) pour votre fin de thèse ! Bon courage aussi à la nouvelle génération: Cyril, Nathan H., Sarah, Jérémy, Romain, Cédric. Merci aussi aux anciens de Paris qui m'ont beaucoup inspiré: Myrto, Hamid, Alex, Mastane, Robin, Pierre L., Anna, Eugène, Pierre A., Adil, Mathurin, Alexandre et tout ceux que j'oublie.

Je remercie évidemment ma famille et ma (belle)-famille pour le soutien perpétuel que vous m'avez apporté. Vous m'avez donné toutes les conditions pour réussir.

Enfin, merci à Marie, ma femme, qui partage ma vie depuis de nombreuses années et sans qui je ne serai pas là. Merci de me soutenir depuis le début malgré les innombrables contraintes qu'une (bonne) thèse engendre !

Abstract

Enthusiasm for Machine Learning is spreading to nearly all fields such as transportation, energy, medicine, banking or insurance as the ubiquity of sensors through IoT makes more and more data at disposal with an ever finer granularity. The abundance of new applications for monitoring complex infrastructures (e.g. aircrafts, energy networks) together with the availability of massive data samples has put pressure on the scientific community to develop new reliable Machine-Learning methods and algorithms. The work presented in this thesis focuses around two axes: *unsupervised functional anomaly detection* and *robust learning*, both from practical and theoretical perspectives.

The first part of this dissertation is dedicated to the development of efficient functional anomaly detection approaches. More precisely, we introduce Functional Isolation Forest (FIF), an algorithm based on randomly splitting the functional space in a flexible manner in order to progressively isolate specific function types. Also, we propose the novel notion of functional depth based on the area of the convex hull of sampled curves, capturing gradual departures from centrality, even beyond the envelope of the data, in a natural fashion. Estimation and computational issues are addressed and various numerical experiments provide empirical evidence of the relevance of the approaches proposed. In order to provide recommendation guidance for practitioners, the performance of recent functional anomaly detection techniques is evaluated using two real-world data sets related to the monitoring of helicopters in flight and to the spectrometry of construction materials.

The second part describes the design and analysis of several robust statistical approaches relying on robust mean estimation and statistical data depth. The Wasserstein distance is a popular metric between probability distributions based on optimal transport. Although the latter has shown promising results in many Machine Learning applications, it suffers from a high sensitivity to outliers. To that end, we investigate how to leverage Medians-of-Means (MoM) estimators to robustify the estimation of Wasserstein distance with provable guarantees. Thereafter, a new statistical depth function, the Affine-Invariant Integrated Rank-Weighted (AI-IRW) depth is introduced. Beyond the theoretical analysis carried out, numerical results are presented, providing strong empirical confirmation of the relevance of the depth function proposed. The upper-level sets of statistical depths—the depth-trimmed regions—give rise to a definition of multivariate quantiles. We propose a new discrepancy measure between probability distributions that relies on the average of the Hausdorff distance between the depth-based quantile regions w.r.t. each distribution and demonstrate that it benefits from attractive properties of data depths such as robustness or interpretability.

All algorithms developed in this thesis are open-sourced and available online.

Résumé

L'engouement pour l'apprentissage automatique s'étend à presque tous les domaines comme l'énergie, la médecine ou la finance. L'omniprésence des capteurs met à disposition de plus en plus de données avec une granularité toujours plus fine. Une abondance de nouvelles applications telles que la surveillance d'infrastructures complexes comme les avions ou les réseaux d'énergie, ainsi que la disponibilité d'échantillons de données massives, potentiellement corrompues, ont mis la pression sur la communauté scientifique pour développer de nouvelles méthodes et algorithmes d'apprentissage automatique fiables. Le travail présenté dans cette thèse s'inscrit dans cette ligne de recherche et se concentre autour de deux axes : *la détection non-supervisée d'anomalies fonctionnelles* et *l'apprentissage robuste*, tant du point de vue pratique que théorique.

La première partie de cette thèse est consacrée au développement d'algorithmes efficaces de détection d'anomalies dans le cadre fonctionnel. Plus précisément, nous introduisons Functional Isolation Forest (FIF), un algorithme basé sur le partitionnement aléatoire de l'espace fonctionnel de manière flexible afin d'isoler progressivement les fonctions les unes des autres. Nous proposons également une nouvelle notion de profondeur fonctionnelle basée sur l'aire de l'enveloppe convexe des courbes échantillonnées, capturant de manière naturelle les écarts graduels de centralité. Les problèmes d'estimation et de calcul sont abordés et diverses expériences numériques fournissent des preuves empiriques de la pertinence des approches proposées. Enfin, afin de fournir des recommandations pratiques, la performance des récentes techniques de détection d'anomalies fonctionnelles est évaluée sur deux ensembles de données réelles liés à la surveillance des hélicoptères en vol et à la spectrométrie des matériaux de construction.

La deuxième partie est consacrée à la conception et à l'analyse de plusieurs approches statistiques, potentiellement robustes, mêlant la profondeur de données et les estimateurs robustes de la moyenne. La distance de Wasserstein est une métrique populaire résultant d'un coût de transport entre deux distributions de probabilité et permettant de mesurer la similitude de ces dernières. Bien que cette dernière ait montré des résultats prometteurs dans de nombreuses applications d'apprentissage automatique, elle souffre d'une grande sensibilité aux valeurs aberrantes. Nous étudions donc comment tirer partie des estimateurs de la médiane des moyennes (MoM) pour renforcer l'estimation de la distance de Wasserstein avec des garanties théoriques. Par la suite, nous introduisons une nouvelle fonction de profondeur statistique dénommée Affine-Invariante Integrated Rank-Weighted (AI-IRW). Au-delà de l'analyse théorique effectuée, des résultats numériques sont présentés, confirmant la pertinence de cette profondeur. Les sur-ensembles de niveau des profondeurs statistiques donnent lieu à une extension possible des fonctions quantiles aux espaces multivariés. Nous proposons une nouvelle mesure de similarité entre deux distributions de probabilité. Elle repose sur la moyenne

de la distance de Hausdorff entre les régions quantiles, induites par les profondeur de données, de chaque distribution. Nous montrons qu'elle hérite des propriétés intéressantes des profondeurs de données telles que la robustesse ou l'interprétabilité.

Tous les algorithmes développés dans cette thèse sont accessible en ligne.

Contents

Notation

$:=$	Equal by definition
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^d$	Set of d-dimensional real-valued vectors
$\mathbb{S}^{d-1}$	Unit hypersphere of $\mathbb{R}^d$
$\mathbb{N}_*$	Positive integers: $\{1, 2, \dots\}$
λ_d	The d-dimensional Lebesgue measure
$\|\cdot\|_p$	ℓ_p-norm on vectors or functions for $p \in [1, +\infty]$
$\|\cdot\|$	ℓ_2-norm on vectors
$\mathbb{R}^{n\times m}$	Set of real matrices of size $n \times m$
I_n	Identity matrix of size $n \times n$
$B^\top$	Transpose of matrix B
$\|\cdot\|_{\mathrm{op}}$	Operator norm of a matrix
$\mathcal{N}(a, B)$	Gaussian distribution with mean a and covariance matrix B
$\mathrm{conv}(\cdot)$	Convex hull of a set
$\mathbb{I}\{\cdot\}$	Characteristic function of a set
$\mathbb{P}(\cdot)$	Probability of an event
$\mathbb{E}[\cdot]$	Expectation of a random variable
$\mathcal{P}(\cdot)$	The space of probability measure on a set
$\mathcal{F}(\mathcal{X}, \mathcal{Y})$	Space of functions from $\mathcal{X}$ to $\mathcal{Y}$
$\mathcal{F}(\mathcal{X})$	Space of real-valued functions defined on $\mathcal{X}$
$\mathcal{C}(\mathcal{X}, \mathcal{Y})$	Space of continuous functions from $\mathcal{X}$ to $\mathcal{Y}$
$L_2(\mathcal{X})$	Lebesgue square-integrable real-valued functions space defined on $\mathcal{X}$
$\mathcal{H}$	An arbitrary Hilbert space
$\langle\cdot,\cdot\rangle_{\mathcal{H}},\ \|\cdot\|_{\mathcal{H}}$	Inner product and norm in $\mathcal{H}$

1

Motivations and Contributions

Contents

The recent technological advances in data acquisition and management through IoT and distributed platforms offer new perspectives in many areas of human activity such as transportation, energy, health, commerce, insurance and confront these domains with major scientific challenges for exploiting these observations. The ever growing availability of massive data, often collected in quasi-real time, engendered high expectations, in particular the need of increased automation and computational efficiency, with the goal to design more and more "intelligent" systems. Despite the ubiquity of sensors collecting massive data which enhances inference accuracy of algorithms, it also foreshadows contaminated information either by flaws in measuring devices or by malicious attacks of the system, see e.g. Shafique et al. (2020). The presence of corrupted data may jeopardize the smooth operation of the system of interest leading to disastrous consequences in sensitive applications such as autonomous driving, aviation safety management or health monitoring systems.

Contaminated data arise in a variety of real-life processes while most of the classical methods from Machine Learning and Statistics assume that observed data correspond to the underlying distribution. Thus, there are possibly two ways to deal with such situations: (i) by designing a robust procedure that will be able to retrieve information present in uncorrupted data without being deteriorated by the corrupted ones and (ii) by finding outliers to remove them from the data set in order to apply a Machine Learning algorithm. In this thesis, our work revolve around two main aspects resulting from this problematic: *unsupervised anomaly detection* and *robust learning*. We introduce many efficient statistical procedures, both in terms of statistical accuracy and computational time, that contribute to the aforementioned areas. With the ubiquity of sensors in

the IoT era, statistical observations are becoming increasingly available in the form of massive time-series or functions. The case of functional data is thus of crucial interest in practice. Although unsupervised anomaly detection has been widely documented in the literature for multivariate data, the case of functional data remains understudied. Filling this gap is the angle embraced in the first part of this thesis. The second part introduces general statistical procedures, relying on the concept of data depth and robust mean estimation, that are able to handle corrupted data during inference.

1.1 Introduction

From a statistical perspective, the analysis of data that come out from industrial processes raises many challenging methodological issues such as identifying corruptions and recovering useful information from a contaminated data set. Several classical statistical models include, generally additive, noise corruptions (see e.g. Rousseeuw and Leroy, 1987): rather than observing an element x lying in a given space, we assume that $x + \zeta$ is observed, where ζ is a zero mean random variable. With recent advances in data collection, this heuristic appears limited to model real applications where the noise can be predominant. Instead, it is often more convenient to consider that a small fraction of the data set is thoroughly corrupted. The elements belonging to this small fraction will thus be denominated as *outliers* or *anomalies* and correspond to elements or patterns that deviate from the expected behavior, see e.g. Figure 1.1.

Introduced in the seminal work of Huber (1964), the Huber contamination model is probably the most popular corruption model. Given two probability distributions P_N and P_A corresponding to normal and abnormal distributions respectively, this model supposes that the observed sample $X_1, \ldots, X_n$ is independently and identically distributed (i.i.d.) from the mixture distribution $P = (1 - \epsilon)P_N + \epsilon P_A$ where $\epsilon \in (0, 1)$ is the fraction of corruption. Following this, a robust procedure needs to be designed to retrieve information present in uncorrupted data generated from P_N without being deteriorated by data generated from P_A. In general, the design of efficient robust approaches is challenging and involves a trade-off between: (i) the statistical accuracy: the quality of recovering information from the uncorrupted data, (ii) the robustness: the ability to recover this information up to a certain proportion of anomalies, and (iii) the computational efficiency. Alternatively, one can find data generated from P_A in order to remove them from the data set and then apply any Machine Learning algorithm using data from P_N. The first approach is known as *robust learning* whereas the second is known as *anomaly detection*.

It is worth noticing that many real-world problems can be formulated as an anomaly detection problem extending the scope of anomaly detection beyond a pre-processing step. For example, in a manufacturing line, the main concern is to avoid, or at least to reduce as most as possible, false negatives, i.e. releasing defective products on the market. The aim of ML approaches is then to help engineers to automatically monitor the production line in order to provide the following benefits: (i) to reduce the number of products with flaws that are considered as sane when leaving the production line and (ii) to detect as fast as possible flaws in a product to avoid useless and expensive operations and thus increase the number of intact items produced per day[1].

[1]The work presented in this dissertation has been funded by BPI France in the context of the PSPC Expresso project, with the objective of developing algorithms to monitor a production line in a Valeo's factory.

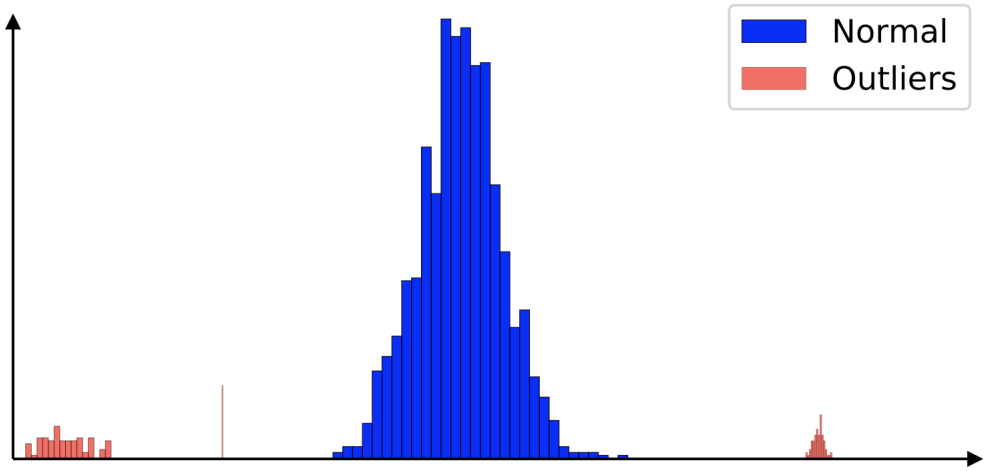


Figure 1.1 – Histogram of a data set with outliers.

In the following, we devote an introduction to our contributions by recalling the state-of-the-art and challenges arising in functional anomaly detection and robust metrics. In Section 1.2, we describe standard approaches dedicated to unsupervised anomaly detection for multivariate data and explain challenges arising when data take the form of functions. Focusing on discrepancy measures between multivariate probability distributions, with emphasis on the Wasserstein distance, robust variants are developed through *robust mean estimation* and *data depth*. In Section 1.3, after briefly recalling popular metrics between probability distributions with stress on the Wasserstein distance, we present Median-of-Means estimators and highlight our contributions. Section 1.4 presents the outline of this manuscript. Section 1.5 finally details the publications resulting from this work.

Notations. Throughout this manuscript, we denote by $(\Omega, \mathcal{A}, \mathbb{P})$ the probability space on which all the random quantities are defined. The set Ω is associated to the σ-algebra $\mathcal{A}$ of its subsets and $\mathbb{P}$ is a probability measure defined on the measurable space $(\Omega, \mathcal{A})$. Any measurable space $\mathcal{M}$ will be considered with its collection of Borel measurable subsets. Given a measurable space $\mathcal{M}$, $\mathcal{P}(\mathcal{M})$ stands for the space of probability measures defined on $\mathcal{M}$. Let $P \in \mathcal{P}(\mathcal{M})$, by $X \sim P$ is meant a random variable X taking its values in $\mathcal{M}$ following the distribution P. In this dissertation, we focus on two main types of data belonging to: (i) the finite-dimensional Euclidean space $\mathcal{M} = \mathbb{R}^d$, where $d \in \mathbb{N}_*$ (Chapter 2, Chapter 4 and Part III) and (ii) real-valued functional spaces $\mathcal{M} = \mathcal{F}(\mathcal{T})$, where $\mathcal{T}$ is any compact set of $\mathbb{R}$ (Chapter 3 and Part II). To avoid confusions, random variables taking values in $\mathcal{F}(\mathcal{T})$, probability measures defined on $\mathcal{F}(\mathcal{T})$ and any element of $\mathcal{F}(\mathcal{T})$ are bolded. The space $\mathcal{F}(\mathcal{T})$ will often be referred to as $\mathcal{F}$ when it is not necessary to specify $\mathcal{T}$. The measurable space $\mathcal{M}$ will be used in this chapter when assertions are true for any measurable spaces.

1.2 Functional Anomaly Detection

The word *anomaly* refers to anything that is not normal and has meaning in many domains such as in biology, sociology or industry (e.g. predictive maintenance, cybersecurity). When some quantitative information related to a process is available, the

objective is to leverage this knowledge to improve the underlying process. It is often performed by a domain expert that has first to identify which part of the data is worthwhile. When dealing with massive and complex data, relying on automatic procedures such as Machine Learning algorithms is then helpful: this is the aim of anomaly detection to identify the valuable part of the data. We refer the reader to Chandola et al. (2009) for a large review on anomaly detection.

In real-world applications, anomalies often appear unexpectedly in a process. Labels are therefore not induced by the process itself but rather require human resources such as domain experts. Assigning labels is then time-consuming and sub-optimal for companies. Anomaly detection is often performed in an unsupervised manner in practice and will be the setting used in this dissertation. Unsupervised anomaly detection corresponds to the case where no *a priori* knowledge on the normal nor on the abnormal are known. From this perspective, the common assumption is to consider that the normal behavior corresponds to what is usual/similar. In contrast, anomalies are considered rare. This implies that data have a hidden structure representing information of uncorrupted data while anomalies are elements that deviate from this structure. The goal of unsupervised anomaly detection is then to learn a model characterizing the normal behavior from data composed of normal and abnormal data. In an unsupervised setting, we assume that the observation sample $X_1, \ldots, X_n$ comes from n i.i.d. realizations of a probability distribution $P = (1-\epsilon)P_N + \epsilon P_A$ with a small $\epsilon \in (0,1)$ and $P_N, P_A \in \mathcal{P}(\mathcal{M})$ having no information about P_N or P_A. The key question is: *how to learn a function from the data, named often decision function or prediction function, that will be able to predict if a new observation $x \in \mathcal{M}$ has been generated from P_N or P_A?* Precisely, the goal is then to build a function $\eta : \mathcal{M} \to \{+1, -1\}$ such that given an observation $x \in \mathcal{M}$, $\eta(x) = 1$ if x is normal and $\eta(x) = -1$ if x is abnormal.

Unsupervised anomaly detection in the multivariate setting (when $\mathcal{M} \subset \mathbb{R}^d$ with $d \geq 1$) is well-documented in the literature and a large variety of dedicated techniques have been proposed and investigated. The vast majority of the heuristics considered consist in turning supervised learning methods into unsupervised approaches, where labels are replaced by rarity, anomalies being supposed to be rare by definition (e.g. SVM becoming one-class SVM, classification trees becoming isolation trees). In contrast, there has not been as much attention paid to the functional situation until now. The aim of Section 1.2.1 is to present general techniques used to perform multivariate unsupervised anomaly detection. In Section 1.2.2 we discuss the challenges raised by the functional framework. Finally, contributions related to this part are gathered in Section 1.2.3.

1.2.1 Anomaly Detection for Multivariate Data

From a statistical perspective, the general problem of anomaly detection can be formalized as a statistical hypothesis test. Assume for this subsection that distributions $P_N \in \mathcal{P}(\mathbb{R}^d)$ and $P_A \in \mathcal{P}(\mathbb{R}^d)$ of normal and abnormal observations are known. Given an observation $x \in \mathbb{R}^d$, the aim is to find if x has been generated by P_N or P_A leading to the statistical hypothesis test:

$$\begin{cases} \mathrm{H}_0 : x \sim P_N \\ \mathrm{H}_1 : x \sim P_A. \end{cases}$$

The optimal way to solve this test is the approach of Neyman-Pearson (Neyman and Pearson, 1933). For $\alpha \in (0,1)$, it consists in finding a critical region G^*_α of $\mathbb{R}^d$ that maximizes the power of the test, i.e. the probability $P_A(G^*_\alpha)$, such that the type-I error $P_N(G^*_\alpha)$ is lower than a confidence level $1-\alpha$. More precisely, it is formalized by the following optimization problem:

$$\max_G \; P_A(G) \;\; s.t. \;\; P_N(G) \leq 1-\alpha,$$

where the max is taken over all Borelian sets of $\mathbb{R}^d$. Equivalently, considering complement sets, it can be formulated as:

$$\min_S \; P_A(S) \;\; s.t. \;\; P_N(S) \geq \alpha. \tag{1.1}$$

Notice that the optimal region G^*_α can be taken as $G^*_\alpha = \mathbb{R}^d \backslash S^*_\alpha$ where S^*_α is a solution of the optimization problem (1.1). If furthermore P_N is absolutely continuous w.r.t. P_A, the set S^*_α can be characterized by the likelihood ratio between P_N and P_A: $S^*_\alpha = \left\{x \in \mathbb{R}^d \,:\, (\mathrm{d}P_N/\mathrm{d}P_A)(x) > t_\alpha\right\}$ with some $t_\alpha > 0$. It turns out that anomaly detection boils down to find upper-level sets of the underlying density of P_N w.r.t. P_A. This statistical framework is meaningful in the case of supervised anomaly detection where labels are available and is related to Neyman-Pearson classification (Cannon et al., 2002; Scott and Nowak, 2005; Rigollet and Tong, 2011). However, as anomalies are supposed to be rare, the structure of the probability distribution P_A cannot be properly observed in data in general even when labels are available. It turns out that the aforementioned likelihood ratio is therefore difficult to compute in practice.

In an unsupervised setting, we no longer have access to P_N nor P_A but only to $P = (1-\epsilon)P_N + \epsilon P_A$ and a different approach should be investigated. A common assumption made by statisticians in this situation is to consider anomalies to be rare, located in the tail of the distribution. If P has a density w.r.t. the Lebesgue measure, the problem boils down to find high density regions corresponding to normal observations while anomalies are located in the low density regions.

Density level sets estimation. Focusing on the finite-dimensional space $\mathbb{R}^d$, a widely used approach for unsupervised anomaly detection, assuming that anomalies are not concentrated, consists in finding a certain density level set $\{x \,:\, f(x) > t\}$ with a threshold $t > 0$ close to zero (see e.g. Figure 1.2), where f is assumed to be the density of the mixture $P = (1-\epsilon)P_N + \epsilon P_A$ w.r.t. the Lebesgue measure with $P, P_N, P_A \in \mathcal{P}(\mathbb{R}^d)$ (see e.g. Steinwart et al., 2005). Fixing the threshold $t = t^*$ then provides the normal region $\{x \,:\, f(x) > t^*\}$ leading to the prediction function $\eta : \mathbb{R}^d \to \{+1, -1\}$ defined by $\eta(x) = 2\mathbb{I}\{f(x) > t^*\} - 1$.

The probability distribution $P \in \mathcal{P}(\mathbb{R}^d)$ is unknown in practice and density level sets have to be estimated thanks to the observations. Let $X_1, \ldots, X_n$ an i.i.d. sample following P and denote by P_n the empirical measure defined as $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$, where δ_x means the Dirac mass at point x. There are essentially two approaches for estimating density level sets.

- **Plug-in approach.** A natural idea is to resort to *plug-in* methods where the density is replaced by an estimate $\hat{f}$ in the level sets (Tsybakov, 1997). The density is generally estimated using a non-parametric kernel estimator (Baillo et al., 2001;

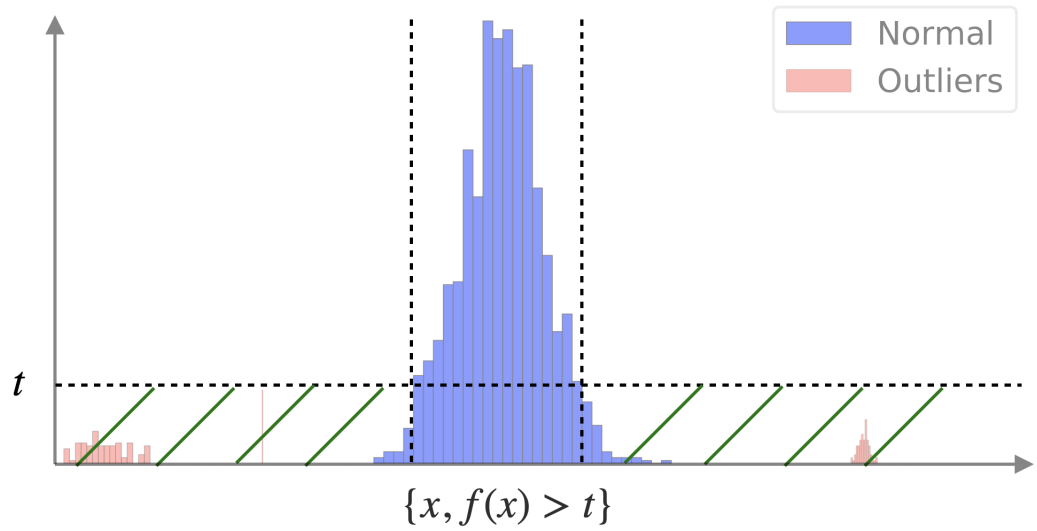


Figure 1.2 – Density level set with a fixed threshold t.

Baillo, 2003; Cadre, 2006). This approach has several drawbacks. Statistical rates are known to suffer from the *curse of dimensionality* and have been derived in (Tsybakov, 1997; Rigollet and Vert, 2009; Chen et al., 2017; Qiao and Polonik, 2019) among others (see also Mason and Polonik (2009) for asymptotic normality). In addition, this approach captures more information than needed such as local properties of f that are unnecessary for estimating density level sets.

- **Direct approach.** The second approach relies on *direct* methods based on minimum volume set estimation (Einmahl and Mason, 1992; Polonik, 1997; Scott and Nowak, 2006) or empirical excess-mass maximization (Hartigan, 1987; Polonik, 1995) that are closely related. Let $\alpha \in (0,1)$, a minimum volume set MV_α^* is a solution solving the constrained minimization problem:

$$\min_S \ \lambda_d(S) \ \ s.t. \ \ P(S) \geq \alpha,$$

where the minimum is taken over measurable subsets S of $\mathbb{R}^d$. It can be shown that MV_α^* is unique and there exists a unique t_α such that $MV_\alpha^* = \{x : f(x) > t_\alpha\}$ under mild assumptions on f (Einmahl and Mason, 1992; Polonik, 1995; Nunez-Garcia et al., 2003). However, these methods suffer both from the curse of dimensionaly as studied in Singh et al. (2009); Mammen and Polonik (2013) and from a computational burden that grows exponentially with the dimension d (Scott and Nowak, 2006).

Remark 1.1. *Density level sets are optimal regions of the Neyman-Pearson test when normal and abnormal distributions are known (supervised setting) or when the normal distribution is known and the abnormal distribution is assumed to be uniform on a compact set (semi-supervised setting). Nonetheless, these information are unavailable in the unsupervised setting.*

Remark 1.2 (Limitations)**.** *So far, computing density level sets of the mixture* $P = (1-\epsilon)P_N + \epsilon P_A$ *to solve the unsupervised anomaly detection task supposes that abnormal observations do not belong to a group of concentrated observations. However, it might*

fail in practice to detect clusters of anomalies and assume somehow that anomalies are uniformly distributed. In addition, the density is a local measure assigning the score of an element as the probability mass in an infinitesimal neighborhood that might fail to provide meaningful information on outliers since it cannot rank properly data that are isolated, i.e. without probability mass around them. Indeed, it will assign zero score to every $x \in \mathbb{R}^d$ that are far from concentrated regions regardless how far it is. Thus, the score returned by the density fails to reflect the degree of abnormality of data.

Learning a scoring function. An alternative approach to density level sets consists in learning a *scoring function* that provides a score reflecting the degree of abnormality of observations. The aim is to design a scoring function s defined as any measurable function $s : \mathbb{R}^d \longrightarrow \mathbb{R}_+$, where $\mathbb{R}_+$ is the set of non-negative reals, such that for any $x \in \mathbb{R}^d$, the smaller $s(x)$, the more abnormal is x. The function s thus provides a pre-order in $\mathbb{R}^d$ w.r.t. the probability distribution P. The prediction function $\eta : \mathbb{R}^d \to \{+1, -1\}$ is then defined by $\eta(x) = 2\, \mathbb{I}\{s(x) > t^*\} - 1$, where t^* is a fixed threshold chosen by the statistician. The choice of t^* is critical and must be addressed carefully. However, there is no optimal technique to address this choice in an unsupervised setting and it is often performed using domain expertise in applications.

Clearly, this formalism encompasses the density level sets approach and opens the way for multiple methods. Numerous anomaly detection algorithms, that return an anomaly score, have been introduced the last two decades ranging from classical Machine Learning algorithms to the most recent deep learning models. It includes among others: the One-Class Support Vector Machine (OCSVM; Schölkopf et al., 2001), Support Vector Data Description (SVDD; Tax and Duin, 2004), One-Class Neighbor Machine (OCNM; Moguerza and Muñoz, 2006), k-NN based approaches (Zhao and Saligrama, 2009; Sricharan et al., 2011), Local Outlier Factor (LOF; Breunig et al., 2000), Histogram-Based Outlier Score (HBOS; Goldstein and Dengel, 2012), Copula-Based Outlier Detection (CBOD; Li et al., 2020), Auto-Encoder (AE) where the reconstruction score is used as anomaly score (Zhou and Paffenroth, 2017) and Variational Auto-Encoder (VAE; An and Cho, 2015). We refer the reader to the survey of Chandola et al. (2009) for several additional techniques.

One of the most popular techniques is the Isolation Forest (IF) algorithm (Liu et al., 2008, 2012) (see also Hariri et al., 2019). This unsupervised algorithm can be viewed as an Ensemble Learning method insofar as it builds a collection of binary trees and an anomaly scoring function based on the aggregation of the latter. An isolation tree is a binary tree, representing a nested collection of partitions of the finite dimensional feature space, grown iteratively in a top-down fashion, where the cuts are axis perpendicular and random (uniformly, w.r.t. the splitting direction and the splitting value both at the same time). A tree is thus built by iterating this procedure until all training data points are isolated (or the depth limit set by the user is attained). An isolation tree constructed accordingly to a training subsample allows to assign to each training sample a path length, namely the depth at which it is isolated from the others. More generally, it can be used to define an anomaly score for any point $x \in \mathbb{R}^d$.

Remark 1.3. *In particular cases, the upper-level sets of the returned scoring function can be seen as approximations of density levels sets such as for OCSVM (Vert and Vert, 2006) or OCNM (Moguerza and Muñoz, 2006). Indeed, for any increasing map $g : \mathbb{R} \to \mathbb{R}$, the collection of level sets of $g \circ f$ is identical to that of the density f. Thus, it is sufficient to estimate any representative of the class of all increasing transformations*

of f to obtain its level sets. Density level sets can be approximated by level sets of any scoring function that provides an ordering close to that of the density. The quality of the approximation can be assessed by means of the Mass-Volume curve (Clémençon and Jakubowicz, 2013; Clémençon and Thomas, 2018) or Excess-Mass curve (Goix, 2016) criteria.

Multivariate Data depth. Widely developed in the statistical literature, multivariate data depths aim to provide a score that reflects the centrality of any element $x \in \mathbb{R}^d$ w.r.t a probability distribution and can therefore be included in scoring function approaches.

The natural left-to-right ordering of elements in the real line $\mathbb{R}$, given by the operator "$\leq$", is the cornerstone of many fundamental univariate statistics. This is the case of quantile and distribution functions, as well as their empirical counterparts such as ranks and signs, all strongly related to the notion of order. The problem of identifying points that deviate from $P_1 \in \mathcal{P}(\mathbb{R})$ is then straightforward when working in a univariate space. Given a univariate sample $X_1, \ldots, X_n$ drawing from a distribution $P_1 \in \mathcal{P}(\mathbb{R})$, empirical quantiles are then order statistics obtained by sorting the sample into ascending order $X_{(1)} \leq \ldots \leq X_{(n)}$. If an element $x \in \mathbb{R}$ is smaller (or larger) than the majority of the sample, i.e. x is close to $X_{(1)}$ (or to $X_{(n)}$) then x can be assigned as an outlier or a non-central point w.r.t. the distribution P_1. However, there is no canonical ordering in $\mathbb{R}^d$ when $d \geq 2$. A key question is then: *how to construct functions that provide statistical ordering for multivariate data?*

Noticing that the univariate median can be described as the element minimizing the maximum number of data points on one of its sides (Hotelling, 1929), John Tukey (Tukey, 1975) suggested the notion of depth within a data cloud for picturing bivariate data. Afterwards, so as to measure the depth of an arbitrary point $x \in \mathbb{R}^d$, Donoho and Gasko (1992) considered hyperplanes through x and determined its depth by the smallest portion of data that is separated by such hyperplanes. Since then, this idea has proved to be very fruitful and has lead to a rich statistical methodology, still in progress, in particular with the design of more general nonparametric depth statistics. In a nutshell, the depth of a data point describes its centrality, or outlyingness inversely, relatively to the data cloud, and thus defines a pre-order on the feature space: the smaller its depth, the likelier a point can be considered as an outlier. For a distribution $P \in \mathcal{P}(\mathbb{R}^d)$ with $d > 1$, by transporting the natural order on the real line to $\mathbb{R}^d$, a depth function $D(\cdot, P) : \mathbb{R}^d \to \mathbb{R}_+$ provides a center-outward ordering of points in the support of P and can be straightforwardly used to extend the notions of (signed) rank or order statistics to multivariate data, which find numerous applications in Statistics and Machine Learning (e.g. robust inference, hypothesis testing, novelty/anomaly detection), see e.g. Mosler and Mozharovskyi (2020). The earliest proposal is the halfspace depth developed in Tukey (1975), whose popularity arises in particular from its strong connection with the notion of distribution function in the univariate context. Indeed, for any probability measure $P_1 \in \mathcal{P}(\mathbb{R})$, constructed as a median-oriented distribution function, the univariate halfspace depth is given by: $\forall t \in \mathbb{R}$,

$$D_{\mathrm{H},1}(t, P_1) = \min\{F(t), 1 - F(t^-)\},$$

where F is the cdf associated to P_1 and $F(t^-)$ denotes the left-sided limit of F at t.

Considering a multivariate r.v. X with probability distribution $P \in \mathcal{P}(\mathbb{R}^d)$ with $d > 1$, its halfspace depth at $x \in \mathbb{R}^d$ is then defined as the infimum of the probability mass taken over all possible closed halfspaces passing through x:

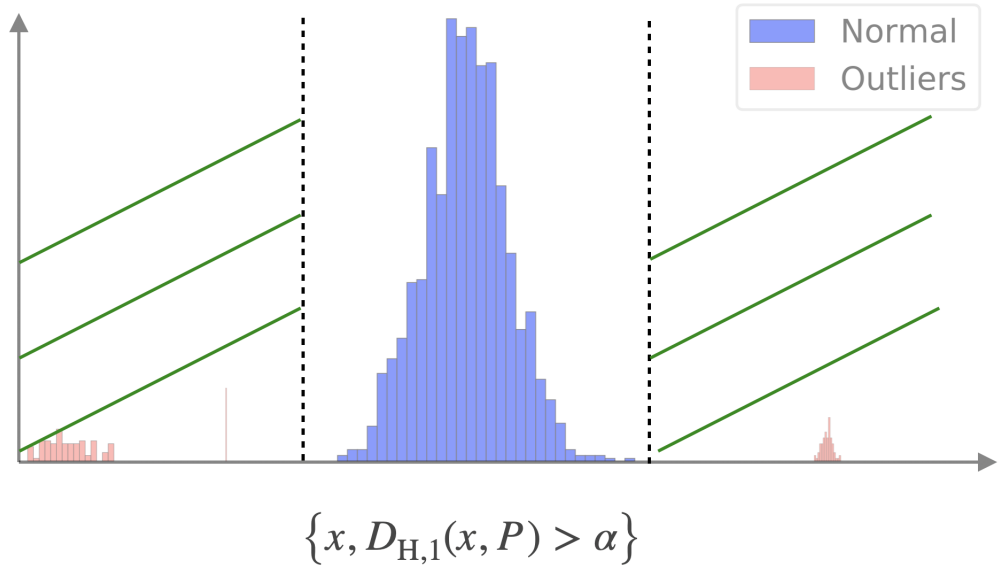


Figure 1.3 – Halfspace upper-level set with a fixed threshold α.

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

denoting by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the usual Euclidean inner product and norm on $\mathbb{R}^d$, and by $\mathbb{S}^{d-1} = \{z \in \mathbb{R}^d \ : \ ||z|| = +1\}$ the unit sphere of $\mathbb{R}^d$ w.r.t. the Euclidean norm. The halfspace depth, probably because of its appealing properties (see Chapter 2), is undeniably the most documented notion of depth function in the statistical literature. Numerous definitions have been proposed, as alternatives to the halfspace depth: among many others, the simplicial depth (Liu, 1990), the projection depth (Liu, 1992), the majority depth (Liu and Singh, 1993), the Oja depth (Oja, 1983), the zonoid depth (Koshevoy and Mosler, 1997), the spatial depth (see Chaudhuri, 1996 or Vardi and Zhang, 2000) or the Monge-Kantorovich depth (Chernozhukov et al., 2017).

In order to compare systematically merits and drawbacks of depth proposals, Zuo and Serfling (2000b) have devised an axiomatic nomenclature of statistical depths, listing key properties that should be ideally satisfied by a depth function. Roughly, as depth functions serve to define center-outward orderings, if a distribution P on $\mathbb{R}^d$ has a unique center $\theta \in \mathbb{R}^d$ (i.e. a symmetry center in a certain sense), the latter should be the deepest point and the depth should decrease along any fixed ray through it. One also expects that a depth function vanishes at infinity and does not depend on the coordinate system chosen. This latter property is usually formulated as *affine-invariance*. Today, in its variety of notions and applications, data depth constitutes a versatile methodology that has been successfully employed in a variety of Machine Learning tasks such as regression (Rousseeuw and Hubert, 1999a; Hallin et al., 2010), classification (Li et al., 2012; Lange et al., 2014), anomaly detection (Serfling, 2006a; Hubert et al., 2015) and clustering (Jörnsten, 2004). We refer the reader to Chapter 2 for a large review on multivariate data depth.

1.2.2 Functional Data Analysis

With the deployment of sensors monitoring their operation nearly continuously, Machine Learning is expected to provide solutions for predictive maintenance of sophisticated systems, such as electricity grids or aircrafts, facilitating the early detection of "weak" signals that announce breakdowns, and serving to plan the replacement of components before their probable failure. Beyond the control of the false alarm rate, the challenge now essentially consists in fully exploiting the information collected, taking often the form of measurements of a physical variable sampled at a very high frequency. In this case, the observations cannot be treated as multivariate data and a functional approach is required. Functional data are available in the form of functions, images and shapes or more general objects (Wang et al., 2016), their statistical analysis, referred to as functional data analysis (FDA in abbreviated form) has received much interest in the last two decades, see e.g. Ramsay and Silverman (2005), Ferraty and Vieu (2006) or Wang et al. (2016).

A functional random variable $\mathbf{X}$ is a random variable (r.v.) that takes its values in a space of functions. To be more specific, let $\mathcal{T} \subset \mathbb{R}$ be a compact interval and consider a r.v. taking its values in the space $\mathcal{F}(\mathcal{T})$ of real valued functions $\mathbf{x} : \mathcal{T} \to \mathbb{R}$:

$$\begin{aligned} \mathbf{X} \quad : \quad & \Omega \longrightarrow \mathcal{F}(\mathcal{T}) \\ & \omega \longmapsto \mathbf{X}(\omega) = \{\mathbf{X}(\omega, t),\ t \in \mathcal{T}\}. \end{aligned}$$

Common choices for the space $\mathcal{F}(\mathcal{T})$ are the space of real-valued continuous functions $\mathcal{C}(\mathcal{T})$ and the space of square integrable functions w.r.t. the Lebesgue measure denoted by $L_2(\mathcal{T})$. The space $\mathcal{T}$ is often considered as time but curves can be parametrized by any variable according to the applications. In practice, only a finite dimensional marginal $(\mathbf{X}(t_1), \ldots, \mathbf{X}(t_p))$, $t_1 < \ldots < t_p$, $p \geq 1$ and $(t_1, \ldots, t_p) \in \mathcal{T}^p$ can be observed. However, considering $(\mathbf{X}(t_1), \ldots, \mathbf{X}(t_p))$ as a discretized curve rather than a simple random vector of dimension p permits to take into account the dependence structure between the measurements over time, especially when the time points t_j are not equispaced. To come back to a function from discrete values, interpolation procedures or approximation schemes based on appropriate dictionaries can be used, combined with a preliminary smoothing step when the observations are noisy, see Ramsay and Silverman (2002). From a statistical perspective, the analysis is based on a functional sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ composed of $n \geq 1$ independent realizations of finite-dimensional marginals of the stochastic process $\mathbf{X}$, that may be very heterogeneous in the sense that these marginals may correspond to different time points and be of different dimensionality.

In this particular context, *functional* anomaly detection aims at detecting the curves that significantly differ from the others among the data set available. Given the richness of spaces of functions, the major difficulty lies in the huge diversity in the nature of the observed differences, which may not only depend on the locations of the curves. Following in the footsteps of Hubert et al. (2015), one may distinguish between three types of anomalies: *shift* (the observed curve has the same shape as the majority of the sample except that it is shifted away), *amplitude* or *shape* anomalies. All these three types of anomalies can be *isolated/transient* or *persistent*, depending on their duration with respect to that of the observations. In many applications, anomalies can also be an aggregation of all the aforementioned types, see e.g. Figure 1.4.

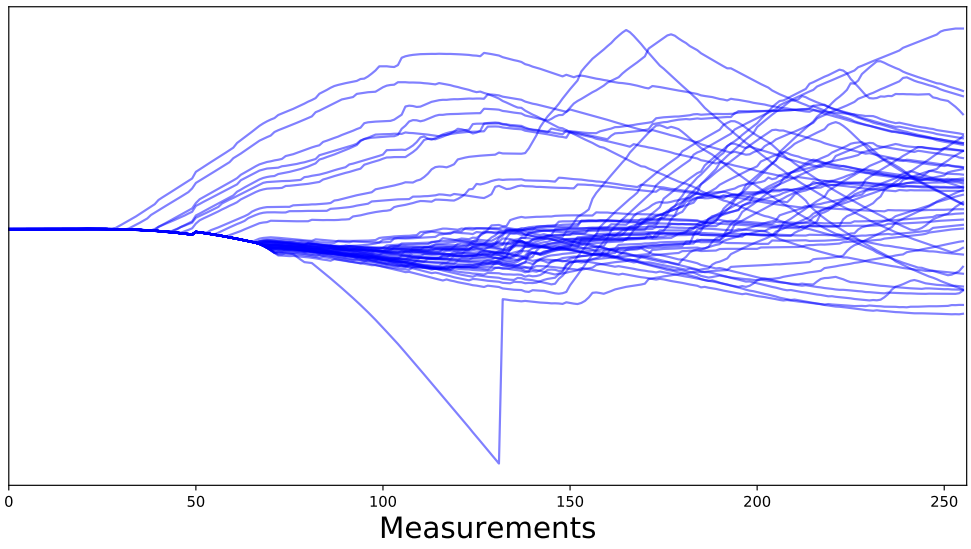


Figure 1.4 – A functional data set with unknown anomalies.

Extending multivariate approaches to the functional case is very difficult in general. There is no analogue of the Lebesgue measure in an infinite-dimensional Banach space. Considering a law $\boldsymbol{\mu}$ of reference (e.g. the Wiener or a Poisson measure) on the function space $\mathcal{F}$ of interest, the quantity $\boldsymbol{\mu}(\mathcal{F})$ of a measurable subset $\mathcal{F} \subset \mathcal{F}$ can be hardly computed in general. Thus, it is far from straightforward to extend multivariate approaches to the functional setup. Of course, anomaly detection for functional data can be reduced to the implementation of its counterpart in the multivariate case by means of straightforward dimensionality reduction techniques. Indeed, one may first project the observed curves onto a subspace of finite dimension by keeping the most informative components of a Kahrunen-Loève decomposition (i.e. those with largest empirical variance) referred to as *Functional Principal Components Analysis* (FPCA; see e.g. Ramsay and Silverman, 2005) or by truncating their expansion in a Hilbertian basis of reference, using a flexible dictionary of functions (e.g. Fourier basis, wavelets). This step is often referred to as the *filtering* stage. Next, any anomaly detection method designed for multivariate observations can be used for analyzing the (parsimonious) filtered data thus obtained.Though easy to implement, such methods have obvious drawbacks. In FPCA, estimation of the Kahrunen-Loève basis can be very challenging and lead to loose approximations, jeopardizing next the anomaly detection stage. The *a priori* representation offered by the "atoms" of a predefined basis or frame may unsuccessfully capture the patterns carrying the relevant information to distinguish abnormal curves from the others. Hence, a certain finite-dimensional representation tailored to the detection of a specific type of anomaly may completely fail in detecting the other types, see e.g. Figure 1.5.

To overcome these issues, one possible approach is to rely on designing scoring functions $s : \mathcal{F} \to \mathbb{R}_+$ working directly on functions. Though a very rich information can be carried by functional data, the downsides of FDA are challenges for designing feasible numerical procedures and for establishing a sound validity theoretical framework alike. Although FDA has been the subject of much attention in recent years, very few generic

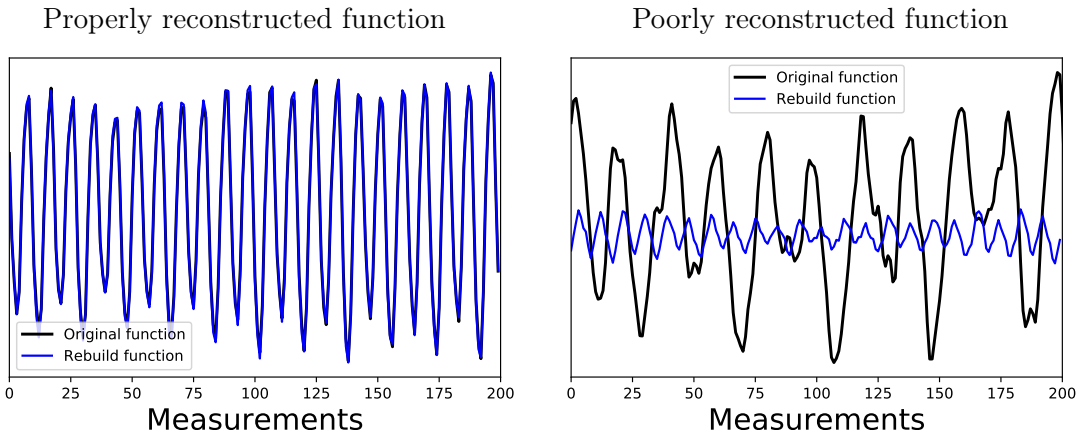


Figure 1.5 – Two functional observations (black curves) from an arbitrary data set and their reconstruction (blue curves) using 10 first principal components of FPCA: a normal observation, well reconstructed (left) and an abnormal one, poorly reconstructed (right).

and flexible methods tailored to functional anomaly detection are documented in the Machine Learning literature to the best of our knowledge, except for specific types of anomalies, see e.g. Rousseeuw and Hubert (2018b) and the references therein.

Functional Data Depth. The statistical literature on functional depths provides a rich methodology for designing scoring functions. Depths in a functional framework have been first considered in Fraiman and Muniz (2001), where authors proposed to define functional depths as simple integrals over time $\mathcal{T}$ of a univariate depth function $D : \mathbb{R} \times \mathcal{P}(\mathbb{R}) \to [0, 1]$. Due to the averaging effect, local changes for the curve $\mathbf{x} \in \mathcal{F}$ only induce slight modifications of the depth value which makes anomaly detection approaches based on such "poorly sensitive" functional depths ill-suited in general, especially for *isolated* anomalies. Alternatives have been introduced, among others, see López-Pintado and Romo (2009, 2011) for depths based on the geometry of the set of curves, Chakraborty and Chaudhuri (2014b) for a notion of depth based on the L_2 distance or Dutta et al. (2011) for a functional version of the halfspace depth. Direct extension of multivariate data depth methods to the functional setting turns to be impractical because the resulting depth functions then vanish everywhere during the optimization, as pointed out in e.g. Kuelbs and Zinn (2015), due to the richness of the feature space.

Let $\mathbf{X}$ be a functional random variable having a distribution $\mathbf{P} \in \mathcal{P}(\mathcal{F})$. Formally, a statistical functional depth is defined as follows:

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

This statistical depth function provides an ordering in the space of curves, see e.g. Figure 1.6. Depending on the depth functions considered, the two common choices for $\mathcal{F}$ in the literature are (i) $\mathcal{C}(\mathcal{T})$, the space of continuous functions defined on $\mathcal{T}$ equipped with the infinity norm $||\mathbf{x}||_\infty = \sup_{t \in \mathcal{T}} |\mathbf{x}(t)|$ and (ii) $L_2(\mathcal{T}, \lambda)$, the space of square integrable functions w.r.t. the Lebesgue measure λ equipped with the norm

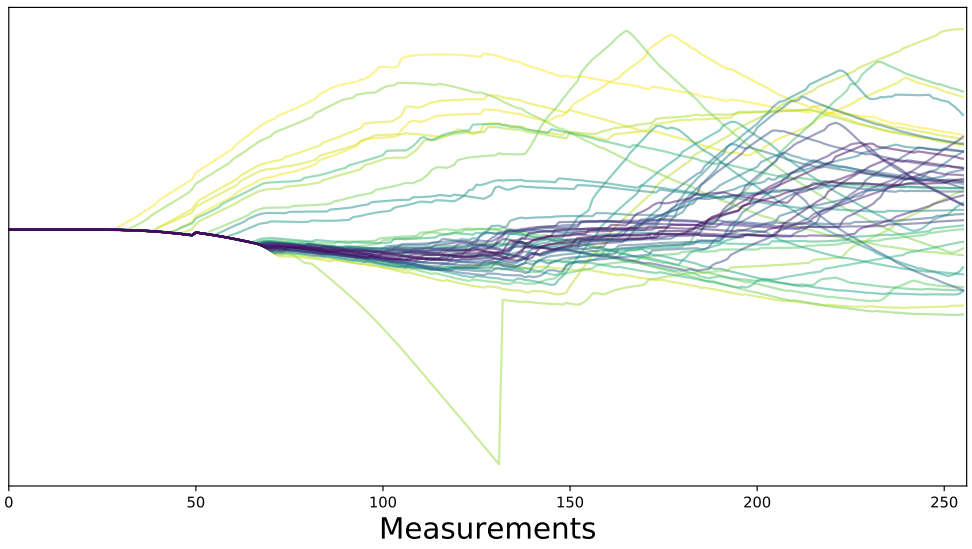


Figure 1.6 – Ordering returned by the functional halfspace depth. The darker the color, the deeper it is in the data set.

$||\mathbf{x}||_{L_2} = \sqrt{\int_{\mathcal{T}} \mathbf{x}(t)^2 \ \lambda(dt)}$.

As discussed in Nieto-Reyes and Battey (2016) and Gijbels and Nagy (2017), the axiomatic framework introduced in Zuo and Serfling (2000b) for multivariate depth is no longer adapted to the richness of the topological structure of functional spaces. Indeed, the vast majority of the functional depths documented in the literature do not fulfill versions of the most natural and elementary properties required for a depth function in a multivariate setup, cf. Gijbels and Nagy (2017). However, there is still no consensus about the set of desirable properties that a functional depth should satisfy, beyond the form of sensitivity mentioned above. We refer the reader to Chapter 3 for a large review on functional data depth.

1.2.3 Contributions

In this thesis, we propose two new tools for functional anomaly detection.

We extend the popular Isolation Forest (IF) approach to anomaly detection, originally dedicated to finite dimensional observations, to functional data. The major difficulty lies in the wide variety of topological structures that may equip a space of functions and the great variety of patterns that may characterize abnormal curves. We address the issue of (randomly) splitting the functional space in a flexible manner in order to isolate progressively any trajectory from the others, a key ingredient to the efficiency of the algorithm. Beyond a detailed description of the proposed algorithm, named Functional Isolation Forest (FIF), computational complexity, stability issues and performances are investigated at length. From the scoring function measuring the degree of abnormality of an observation provided by the proposed variant of the IF algorithm, a multivariate functional extension is defined and discussed.

We introduce a novel robust functional statistical depth measure dedicated to the analysis of functional data. Based on the area of the convex hull of collections of sampled curves, it is easy to compute and to interpret both at the same time. The general idea is to quantify the contribution of a curve $\mathbf{x} \in \mathcal{C}(\mathcal{T})$, on average, to the area of the convex hull (ACH in short) of random curves in $\mathcal{C}(\mathcal{T})$ with the same probability law. Precisely, this function, referred to as the ACH depth throughout the manuscript, is defined by the ratio of the ACH of the sample to that of the sample augmented by the curve $\mathbf{x}$. We prove here that it fulfills various properties desirable for functional depths. In particular, given its form, it exhibits sensitivity (i.e. the depth score of new/test curves that are further and further away from the training set of curves decreases smoothly), which property, quite desirable intuitively, is actually not satisfied by most statistical (functional) depth documented in the literature. In addition, the statistical depth we promote here is robust to outliers: adding outliers to the training set of curves has little or no effect on the returned score and ordering on a test set. For this reason, this functional depth is very well suited for unsupervised anomaly detection, especially to identify *isolated anomalies*.

Eventually, we investigate the performance of recent techniques for functional anomaly detection and compare their accuracy with that of simpler approaches, based on a preliminary dimensionality reduction, standing as natural competitors. In particular, specific attention is paid to those that are based on functional depth statistics or that extend multivariate methods by avoiding the filtering step. A benchmark study comparing the merits of the methods considered here regarding various metrics of reference is thus presented on aeronautics data gathered by Airbus and spectrometry measurements of sedimentary material collected by the Geological Survey of Austria for quality assessment on mining sites of Austria.

1.3 Probability Metrics, Data Depth and Robustness

Comparing probability distributions has attracted a long-standing interest in Information Theory (Kullback, 1959; Rényi, 1961; Csiszár, 1963), Probability Theory and Statistics (Rachev, 1991; Billingsley, 1999; Müller, 1997). Given two probability distributions P, Q defined on an arbitrary space, the goal is then to design metrics that are able to assess how close P and Q are, that differ in how the comparison is addressed. While they serve many purposes in Machine Learning (Cha and Srihari, 2002; MacKay and Mac Kay, 2003), they are of crucial importance as loss functions in automatic evaluation of natural language generation (see e.g. Kusner et al., 2015; Zhang et al., 2019), especially when leveraging deep contextualized embeddings such as the popular BERT (Devlin et al., 2019), graphical probabilistic modeling (Jordan, 1998) including generative adversarial modeling (see e.g. Goodfellow et al., 2014) as well as variational inference (see e.g. Blei et al., 2017). In these applications, choosing the right loss to be minimized between the two distributions is one of the key issues of the problem as its properties strongly influence the behavior of the associated algorithm. Thus, designing a measure to compare two probability distributions is considered as a challenging research field. This is certainly due to the inherent difficulty in capturing in a single measure typical desired properties such as: (i) metric or pseudo-metric properties, (ii) invariance under specific geometric transformations, (iii) efficient computation, (iv) efficient estimation from samples and (v) robustness to contamination. The latter has received little attention in the literature.

In Section 1.3.1, we present an overview of existing metrics between probability distributions with a focus on the Wasserstein distance. In Section 1.3.2, the problem of robust mean estimation is described and the popular Median-of-Means estimator is introduced. Eventually, contributions related to this part are presented in Sections 1.3.3 and 1.3.4.

1.3.1 Metrics between Probability Distributions

One can find in the literature a wide collection of discrepancies between probability distributions that rely on different principles.

The *φ-divergences* introduced in Rényi (1961) and Csiszár (1963) are defined as the weighted average by well-chosen functions φ of the odds ratio between the two distributions. The best known φ-divergence is the *Kullback-Leibler divergence* which is widely used in Machine Learning applications such as in Generative Adversarial Networks (GAN; Goodfellow et al., 2014) or Variational Auto-Encoders (VAE; Kingma and Welling, 2013) (see also Kingma and Welling, 2019). This family of divergences also includes, among others, the Jensen-Shannon and alpha divergences, the total variation and the squared Hellinger distances. However, φ-divergences do not metrize weak convergence which is a major issue. The metrization of weak-convergence is essential, as it ensures that the metrics remain stable under small perturbations of the support of the measures. It is illustrated by the degeneracy to $+\infty$ of φ-divergences when the supports of both distributions do not overlap, which appears to be a crucial limitation in many applications.

The family of Integral Probability Metric (IPM) introduced in Müller (1997) is based on a variational definition of the metric, i.e. the maximum difference in expectation for both distributions calculated over a class of measurable functions and give rise to various metrics (Maximum Mean Discrepancy (MMD), Dudley's metric, L_1-Wassertein distance) depending on the choice of this class (see e.g. Sriperumbudur et al., 2012). First of all, they metrize the weak convergence of measures under mild assumptions on the space of functions on which the maximum is computed. However, except in the case of MMD which appears to enjoy a closed-form solution, the variational definition raises issues in computation.

From the side of Optimal transport (OT) (see Villani, 2003; Peyré and Cuturi, 2019), the L_p-Wasserstein distance leverages a ground metric to take into account the geometry of the space on which the distributions are defined. Given two probability distributions, the latter is defined in terms of the solution to the Monge-Kantorovich optimal mass transportation problem. Its ability to handle non-overlapping support and to take into account the underlying geometry of the space makes OT a powerful tool. For these reasons, the Wasserstein distance stands out from the divergences usually exploited in generative modeling, like the φ-divergences, and has been successfully exploited in Wasserstein Generative Adversarial Networks (WGANs; Arjovsky et al., 2017; Gulrajani et al., 2017) as well as in Wasserstein Auto-Encoders (WAE; Tolstikhin et al., 2018), where the Wasserstein distance can advantageously replace a φ-divergence as the loss function. Given $p \in [1, \infty)$, the Wasserstein distance of order p between $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$, where $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, is defined through the solution of the Monge-Kantorovitch mass transportation problem (see e.g. Peyré and Cuturi, 2019):

$$\mathcal{W}_p(P, Q) = \min_{\pi \in \Pi(P,Q)} \left(\int_{\mathcal{X} \times \mathcal{Y}} \|x - y\|^p \, \mathrm{d}\pi(x \times y) \right)^{1/p}, \tag{1.2}$$

where $\Pi(P,Q) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \int \pi(x,y)\mathrm{d}y = P(x), \int \pi(x,y)\mathrm{d}x = Q(y)\}$ is the set of joint probability distributions with marginals P and Q.

A nice feature of the Wasserstein distance appears when computed between univariate distributions. Assuming $P_1, Q_1 \in \mathcal{P}(\mathbb{R})$, it holds (Rachev and Rüschendorf, 1998):

$$\mathcal{W}_p(P_1,Q_1) = \left(\int_0^1 |F_{P_1}^{-1}(t) - F_{Q_1}^{-1}(t)|^p \, \mathrm{d}t \right)^{1/p}, \tag{1.3}$$

where $F_{P_1}^{-1}, F_{Q_1}^{-1}$ are quantile functions of P_1 and Q_1, respectively. In the remainder of this manuscript, when dealing with the Wasserstein distance of order 1, $\mathcal{W}_1$, we omit the subscript 1 for notation simplicity. By the dual Kantorovich-Rubinstein formulation (Kantorovich and Rubinstein, 1958), with $\mathcal{F}_{\mathrm{Lip}}$ the unit ball of the Lipschitz functions space, a useful rewriting of the 1-Wasserstein distance is:

$$\mathcal{W}(P,Q) = \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \mathbb{E}_P\Big[\Psi(X)\Big] - \mathbb{E}_Q\Big[\Psi(Y)\Big],$$

where X and Y are random variables having distributions P and Q, respectively. In practice, the unit ball of Lipschitz functions can be replaced with a parameterized family of Lipschitz functions, more amenable for learning, see e.g. Wasserstein GANs (Arjovsky et al., 2017).

Of particular interest is the problem of estimating the Wasserstein distance between P and Q given a finite number of observations. The usual assumption is to rely upon two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, composed of i.i.d. realizations drawn respectively from P and Q. The corresponding empirical distributions denoted by $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$, and $Q_m = (1/m)\sum_{j=1}^m \delta_{Y_j}$. The natural questions are then: *how to compute the estimator $\mathcal{W}(P_n, Q_m)$, and does it converge towards $\mathcal{W}(P,Q)$?*

While this problem has long been theoretically studied under the i.i.d. assumption (Dudley, 1969; Bassetti et al., 2006; Weed and Bach, 2019), it has never been tackled through the lens of robustness to outliers, a crucial issue in Reliable Machine Learning. Despite its appealing properties, the Wasserstein distance suffers from sensitivity to outliers due to the marginal constraints in (1.2). Indeed, a small outlier mass can contribute significantly to the cost.

1.3.2 Robust Mean Estimation and the Median-of-Means Estimator

Univariate mean estimation plays a critical role in many statistical learning problems, ranging from classification and regression to ranking or generative modeling. Although the empirical mean appears as a natural candidate, it has been unfortunately shown to fail under contamination model. Consider a sample $X_1, \ldots, X_{n-1}$ composed of $n-1$ i.i.d. realizations of the real-valued random variable X, with a "concentrated" distribution $P_1 \in \mathcal{P}(\mathbb{R})$ (e.g. sub-Gaussian) and an outlier x independent from X that can be arbitrarly far from $\mathbb{E}[X]$. The sample mean, given by $\frac{1}{n}\left(\sum_{i=1}^{n-1} X_i + x\right)$, can be highly deteriorated by the presence of one single element such as x.

In contrast with the sample mean, the Median-of-Means (MoM) estimator, independently introduced during the 1980s (Nemirovsky and Yudin, 1983; Jerrum et al., 1986; Alon et al., 1999), exhibits attractive robustness properties. Let $X_1, \ldots, X_n$ be a sample of n i.i.d. realizations of the real-valued random variable X, with an arbitrary distribution $P_1 \in \mathcal{P}(\mathbb{R})$. Let $K_X \in \mathbb{N}_*$, $K_X \leq n$, and $\mathcal{B}_1^X, \ldots, \mathcal{B}_{K_X}^X$ be a partition of $\{1, \ldots, n\}$

into disjoint block of same size denoted by $B_X = |\mathcal{B}_k^X|$, $1 \leq k \leq K_X$. If n cannot be divided by K_X, some observations may be removed. The Median-of-Means estimator is defined as:

$$\mathrm{MoM}_X = \mathrm{med}\left\{\frac{1}{B_X}\sum_{i\in\mathcal{B}_k^X} X_i,\ 1 \leq k \leq K_X\right\},$$

where med stands for the median operator. The MoM estimator can be seen as an interpolation between the empirical mean and the empirical median involving the tuning parameter K_X. Thus, this estimator is not affected by contamination as long as there are at least $\lceil K_X/2 \rceil$ sane blocks which corresponds to having less than $\frac{1}{n}\lceil K_X/2 \rceil$ outliers.

Following the seminal deviation study by Catoni (2012), MoM has lately witnessed a surge of interest, mainly due to its attractive sub-Gaussian behavior, under the sole assumption that the underlying distribution has finite variance (Devroye et al., 2016). Originally devoted to scalar random variables, MoM has notably been extended to random vectors (Minsker, 2015; Hsu and Sabato, 2016; Lugosi and Mendelson, 2019a) and U-statistics (Joly and Lugosi, 2016; Laforgue et al., 2019). As a natural alternative to the empirical mean, MoM has become the cornerstone of several robust learning procedures in heavy-tailed situations, including bandits (Bubeck et al., 2013) and MoM-tournaments (Lugosi and Mendelson, 2019b). A more recent line of work now focuses on MoM's ability to deal with outliers. Aside from concentration results in a contaminated context (Depersin and Lecué, 2019), it has yielded promising applications in robust mean embedding (Lerasle et al., 2019), and the more general MoM-minimization framework (Lecué et al., 2020).

1.3.3 Contributions on Probability Metrics with Robustness

We focus on the Kantorovich-Rubinstein dual formulation of the Wasserstein distance and present three novel MoM-based estimators, leveraging in particular Medians of U-statistics (MoU). In the realistic setting of contaminated data, we show their strong consistency, and provide non-asymptotic bounds as well. We propose a dedicated algorithm to compute these three estimators in practice. Applied on a parametric family of Lipschitz functions, e.g. neural networks with clipped weights, it performs a MoM/MoU gradient descent algorithm. A sensitivity analysis of the unique parameter of these estimators is also provided through numerical experiments on toy data sets. We robustify Wassersein GANs (w.r.t. outliers) using a MoM-based estimator as loss function. We show the benefits of this approach through convincing numerical results on two contaminated well known benchmarks: CIFAR10 and Fashion MNIST.

1.3.4 Contributions on Data Depth

We introduce an extension of the *integrated rank-weighted* statistical depth (IRW depth in abbreviated form) originally introduced in Ramsay et al. (2019), modified in order to satisfy the property of *affine-invariance*, fulfilling thus all the four key axioms listed in the nomenclature elaborated by Zuo and Serfling (2000b). The variant we propose, referred to as the Affine-Invariant IRW depth (AI-IRW in short), involves the precision matrix of the (supposedly square integrable) d-dimensional random vector X under study, in order to take into account the directions along which X is most variable to assign a depth value to any point $x \in \mathbb{R}^d$. The accuracy of the sampling version of the

AI-IRW depth is investigated from a non-asymptotic perspective. Namely, a concentration result for the statistical counterpart of the AI-IRW depth is proved. Beyond the theoretical analysis carried out, applications to anomaly detection are considered and numerical results are displayed, providing strong empirical evidence of the relevance of the depth function we propose here.

Further, we introduce a novel discrepancy measure between probability distributions by leveraging the extension of univariate quantiles to multivariate spaces. This new pseudo-metric relies on the average of the Hausdorff distance between the depth-based quantile regions w.r.t. each distribution. Its good behavior w.r.t. major transformation groups as well as its ability to factor out translations are depicted. Robustness, an appealing feature of this pseudo-metric, is studied through the finite sample breakdown point. Moreover, we propose an efficient approximation method with linear time complexity w.r.t. the size of the data set and its dimension. The quality of this approximation as well as the performance of the proposed approach are illustrated at length in numerical experiments.

1.4 Outline of the Thesis

Part I provides a large review of the concept of data depth dedicated to multivariate and functional data as well as reminders on the Wasserstein distance. We now present the outline of this manuscript, summarized in the flowchart displayed in Figure 1.7.

Part II focuses on designing new efficient approaches to perform *functional anomaly detection*.

- Chapter 5 introduces the Functional Isolation Forest (FIF) algorithm.
- Chapter 6 describes a novel robust functional statistical depth measure dedicated to the analysis of functional data: the ACH depth.
- Chapter 7 investigates the performance of recent techniques for functional anomaly detection and compare their accuracy.

Part III is devoted to robust alternatives to standard metrics, defined between multivariate probability distribution belonging to $\mathcal{P}(\mathbb{R}^d)$, through the lens of the MoM estimator and data depths.

- Chapter 8 presents three novel MoM-based estimators, leveraging in particular Medians of U-statistics (MoU).
- Chapter 9 introduces the Affine-Invariant Rank-Weighted depth (AI-IRW depth in abbreviated form).
- Chapter 10 provides a pseudo-metric between probability distributions by leveraging the extension of univariate quantiles to multivariate spaces.

1.5 Publications

The contributions introduced in this dissertation have resulted in the following publications and preprints, presented in chronological order:

- **G. Staerman**, P. Mozharovskyi, S. Clémençon, F. d'Alché-Buc. Functional Isolation Forest. In *Proceedings of The Eleventh Asian Conference on Machine Learning (ACML)*, pages 332-347, 2019.
- **G. Staerman**, P. Mozharovskyi, S. Clémençon. The Area of the Convex Hull of Sampled Curves: a Robust Functional Statistical Depth measure. In *Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 570-579, 2020.
- **G. Staerman**, P. Laforgue, P. Mozharovskyi, F. d'Alché-Buc. When OT meets MoM: Robust estimation of Wasserstein distance. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136-144, 2021.
- P. Laforgue, **G. Staerman**, S. Clémençon. Generalization Bounds in the Presence of Outliers: a Median-of-Means Study. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 5937-5947, 2021.
- **G. Staerman**, P. Mozharovskyi, S. Clémençon. Affine-Invariant Integrated Rank-Weighted Depth: Definition, Properties and Finite Sample Analysis. *arXiv preprint arXiv:2106.11068*, 2021.
- **G. Staerman**, P. Mozharovskyi, P. Colombo, S. Clémençon, F. d'Alché-Buc. A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions. *arXiv preprint arXiv:2103.12711*, 2021.
- **G. Staerman**, P. Mozharovskyi, S. Clémençon. Functional Anomaly Detection: a Benchmark Study. *arXiv preprint arXiv:2201.05115*, 2022.

An additional publication that has been made during this thesis is not presented here:

- P. Colombo, **G. Staerman**, C. Clavel, P. Piantanida. Automatic Text Evaluation through the Lens of Wasserstein Barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10450-10466, 2021.

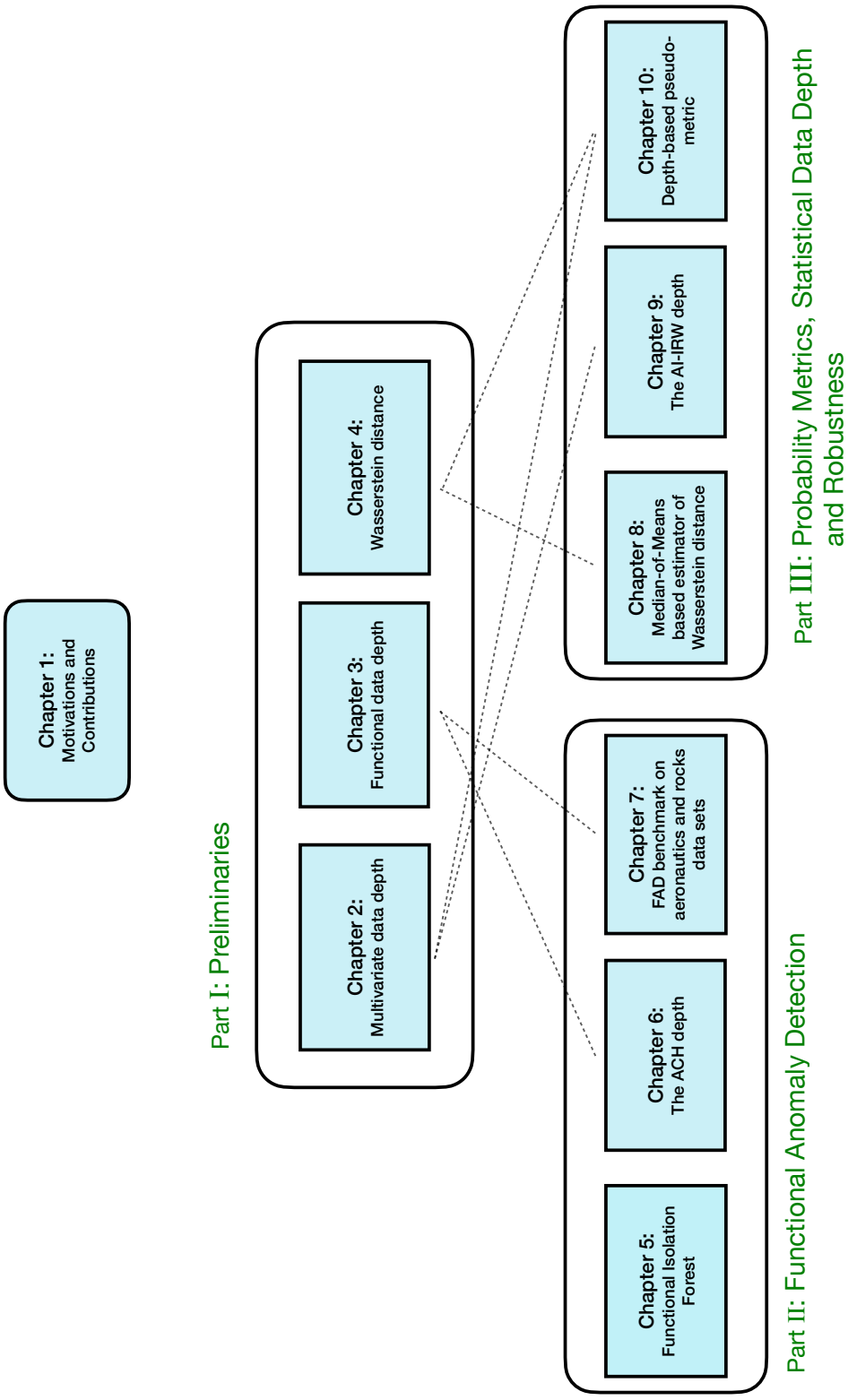


Figure 1.7 – Flowchart of this dissertation.

Part I

Preliminaries

2

Multivariate Data Depth

Contents

Since its introduction by John Tukey in 1975 in order to extend the notion of median to the multivariate setting, the concept of statistical depth has become increasingly popular in multivariate data analysis. A data depth function measures the centrality of any element $x \in \mathbb{R}^d$ w.r.t. a probability distribution $P \in \mathcal{P}(\mathbb{R}^d)$ (respectively, a data set). It provides a center-outward ordering of points in the support of P and can be straightforwardly used to extend the notions of (signed) rank or order statistics to multivariate data, which find numerous applications in statistics and machine learning beyond anomaly detection (e.g. robust inference, hypothesis testing, classification, clustering), see e.g. Mosler (2013). Numerous definitions have been proposed, as alternatives to the earliest proposal, the halfspace depth introduced in Tukey (1975). Among many others these include: the simplicial depth (Liu, 1990), the projection depth (Liu, 1992), the majority depth (Liu and Singh, 1993), the Oja depth (Oja, 1983), the zonoid depth (Koshevoy and Mosler, 1997), the regression depth (Rousseeuw and Hubert, 1999a), the location-scale depth (Mizera and Müller, 2004), the spatial depth (Chaudhuri, 1996; Vardi and Zhang, 2000) or the Monge-Kantorovich depth (Chernozhukov et al., 2017; Hallin et al., 2021) differing in their properties and applications. We refer the reader to Serfling (2006a); Cascos (2010); Mosler (2013); Mosler and Mozharovskyi (2020) or Van Bever (2013), Chapter 1 for excellent surveys on multivariate data depth.

After having precisely defined the concept and standard properties of multivariate data depth in Section 2.1 as well as its depth-trimmed regions in Section 2.2, we exhibit classical examples of data depth functions and highlight their particularities in Section 2.3. Beyond the verification of standard properties, the pros and cons of any data depth should be considered regarding statistical properties, robustness to outliers

and the possible existence of algorithms for exact computation in the case of empirical distributions, as discussed in Section 2.4.

2.1 Definition

Formally, a data depth function is defined as follows:

$$\begin{array}{rccl} D: & \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) & \longrightarrow & [0,1]\,, \\ & (x, P) & \longmapsto & D(x, P). \end{array} \tag{2.1}$$

The higher $D(x, P)$, the deeper x is in P. The depth-induced median of P is then defined by the set attaining $\sup_{x\in\mathbb{R}^d} D(x, P)$ in the case where it exists. Here and in the rest of the manuscript, the median of a depth function will refer, somewhat imprecisely, to the elements that reach this supremum. Since data depth naturally and in a nonparametric way defines a pre-order on $\mathbb{R}^d$ w.r.t. a probability distribution, it can be seen as a centrality-based alternative to the cumulative distribution function (cdf) for multivariate data. Clearly, (2.1) opens the door to a variety of existing definitions. While these differ in theoretical and practically related properties such as robustness or computational complexity, several postulates have been developed throughout the recent decades the "good" depth function should satisfy. Such properties have been thoroughly investigated in Liu (1990); Zuo and Serfling (2000b) and Dyckerhoff (2004) with slightly different sets of axioms (or postulates) to be satisfied by a depth function. They are recalled below.

($\mathbf{D}_1$) (Affine invariance) Denoting by P_X the distribution of any r.v. X taking its values in $\mathbb{R}^d$, we have:

$$\forall x \in \mathbb{R}^d, \;\; D(Ax + b, P_{AX+b}) = D(x, P_X),$$

for any $d \times d$ nonsingular matrix A with real entries and any vector b in $\mathbb{R}^d$.

($\mathbf{D}_2$) (Maximality at center) For any probability distribution P on $\mathbb{R}^d$ that possesses a symmetry center θ (in a sense to be specified), the depth function $D(\cdot, P)$ takes its maximum value at it:

$$D(\theta, P) = \sup_{x\in\mathbb{R}^d} D(x, P).$$

($\mathbf{D}_3$) (Monotonicity relative to deepest point) For any probability distribution P on $\mathbb{R}^d$ with deepest point x_P, the depth at any point x in $\mathbb{R}^d$ decreases as one moves away from x_P along any ray passing through it:

$$\forall \alpha \in [0,1], \;\; D(x_P, P) \geq D(x_P + \alpha(x - x_P), P).$$

($\mathbf{D}_4$) (Vanishing at infinity) For any probability distribution P on $\mathbb{R}^d$, the depth function D vanishes at infinity:

$$D(x, P) \to 0, \text{ as } ||x|| \to \infty.$$

($\mathbf{D}_5$) (Upper semi-continuity) For any probability distribution P on $\mathbb{R}^d$, the set $\left\{x \in \mathbb{R}^d \;\; D(x, P) < \alpha\right\}$ is an open set for every $\alpha \in (0, 1]$.

Here and throughout, we rely on the following definition of statistical data depth function.

Definition 2.1 (Zuo and Serfling, 2000b)**.** *A function* $D : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \longrightarrow [0,1]$ *is a statistical depth function if it satisfies* $(\mathbf{D}_1 - \mathbf{D}_4)$.

A stronger notion of $(\mathbf{D}_3)$, that gives more information on characteristics of the underlying distribution such as shape or symmetry (Serfling, 2004), is defined as:

$(\mathbf{D}_{3'})$ (QUASI-CONCAVITY) For any probability distribution P on $\mathbb{R}^d$, for every $\lambda \in [0,1]$ and $x, y \in \mathbb{R}^d$, $D(\lambda x + (1-\lambda)y, P) \geq \min\{D(x,P), D(y,P)\}$.

This property as well as the technical property $(\mathbf{D}_5)$ lead to a variant of the aforementioned definition.

Definition 2.2 (Dyckerhoff, 2004)**.** *A function* $D : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \longrightarrow [0,1]$ *is a convex statistical depth function if it satisfies* $(\mathbf{D}_1, \mathbf{D}_{3'}, \mathbf{D}_4, \mathbf{D}_5)$.

It is worth mentioning that properties $(\mathbf{D}_1, \mathbf{D}_3, \mathbf{D}_5)$ imply $(\mathbf{D}_2)$ (Dyckerhoff, 2004). The affine invariance property includes common transformations such as orthogonal, translation or scaling, and is useful in applications providing independence w.r.t. measurement units and coordinate system. Some depths that are based on distances such as Euclidean or spatial depths (Serfling, 2002) are generally invariant to orthogonal transformations and translations rather than affine transformations. Serfling (2010) has studied a general way, named *sphering* or *whitening*, making these depths affine invariant using a scatter matrix transform depending on the distribution. Precisely, the random variable X is first transformed by the following map:

$$x \mapsto S_X^{-1/2}(x - \theta_X) \quad \text{such that} \quad S_{AX+b} = \gamma A S_X A^\top$$

where $\gamma > 0$ is a scaling parameter, S_X is a $d \times d$ scatter matrix, and θ_X a location parameter of X. The aforementioned transformation leads to several choices of S_X and θ_X where widely used choices are the covariance matrix Σ_X and the expectation $\mathbb{E}[X]$ respectively.

For distributions having a uniquely defined center (e.g. symmetry center θ), data depths should be maximized at this center, as stated by $(\mathbf{D}_2)$. Several notions of symmetry have been widely used for multivariate distributions. A random variable X is said to be *centrally symmetric* (CS) about θ if $X - \theta \overset{\mathcal{L}}{=} \theta - X$. Further, X is said to be *angularly symmetric* (AS) if $(X-\theta)/||X-\theta||$ is centrally symmetric around the origin (Liu, 1990). Inspired from the halfspace depth, Zuo and Serfling (1999) have defined the *halfspace symmetric* (HS) notion such that $\mathbb{P}\left(X \in H_\theta\right) \geq 1/2$ for every closed halfspace containing θ denoted by H_θ. It is easy to see that CS $\Rightarrow$ AS $\Rightarrow$ HS making the halfspace symmetry the weaker notion, and then the most general, that $D(\cdot, P)$ should satisfy. Some classes of probability distributions also define symmetry notions by construction such as spherical or elliptical distributions. Indeed, X is spherically symmetric (SS) if $OX \overset{\mathcal{L}}{=} X$ for every othogonal matrix O while a r.v. Y has an elliptical distribution (ES) if $Y \overset{\mathcal{L}}{=} a + AX, a \in \mathbb{R}^d$ yielding SS $\Rightarrow$ ES $\Rightarrow$ CS. We refer the reader to Beran and Millar (1997) and Serfling (2006b) for further details on symmetry notions for multivariate

distributions. For example, skewness measures can be built from a depth function $D(\cdot,\cdot)$ and the symmetry notion it satisfies providing information on the asymmetry of the underlying probability distribution (Liu et al., 1999).

The property $(\mathbf{D}_3)$ is a consequence of the center-outward ordering construction of data depth. When a point $x \in \mathbb{R}^d$ moves away from the set of elements that reach the maximum value of the depth function (potentially reduced to a single element, e.g. for symmetric distributions defined above), $D(x,P)$ should decrease monotonically. Statistical functions satisfying $(\mathbf{D}_3)$ should tend to fail capturing and describing the multimodal behavior of probability distributions. Unfortunately, this crucial property limits the relevance of data depths to unimodal distributions.

Remark 2.3 (Multimodal distributions)**.** *As pointed out in Zuo and Serfling (2000b), one has to choose between center-outward ordering and the sensitivity to multimodal distributions. To design depth functions adapted to multimodal distributions, which can be required in some applications such as anomaly detection or classification, several data depths have been introduced. These depths, listed below, are conceptually close to the notion of density. The likelihood depth has been investigated by Fraiman and Meloche (1999) where the set of maximizers can be interpreted as modes of the probability distribution. Chen et al. (2009) and Hu et al. (2011) have proposed kernelized versions of the spatial depth and Mahalanobis depth respectively. A depth function based on quantiles of interpoint distances have also been proposed in Lok and Lee (2011) allowing a multitude of maximum depth and leading to disconnected depth-trimmed regions. Local versions of the halfspace and simplicial depths have been introduced and studied in Agostinelli and Romanazzi (2011) providing an interpolation between (global) depths and the density of the underlying distribution. A general notion of local depth based on the neighborhood notion of Paindaveine and Bever (2015), potentially applicable to any depth function, has been described in Paindaveine and bever (2013). For the purpose of classification and to obtain depth contours close to those of the density function of P, the localized spatial depth have been introduced by Dutta et al. (2016).*

$(\mathbf{D}_4)$ and $(\mathbf{D}_5)$ appear as natural properties since data depth is a (center-outward) generalization of cdf. Limit values vanish due to median-oriented construction. $(\mathbf{D}_{3'})$ allows to preserve the original center-outward ordering goal of data depth and induces convexity of the upper-level sets of depth functions. The property $(\mathbf{D}_{3'})$ is not mandatory but several introduced depth functions satisfy this property such as the halfspace depth or the projection depth (see Section 2.3)

Originally motivated to extend the notion of median to multivariate space, the set attaining $\sup_{x\in\mathbb{R}^d} D(x,P)$, if it exists, is considered as the median associated to the depth function D. In order to characterize the depth induced median, Dyckerhoff (2004) has provided conditions under which this set is non-empty.

Proposition 2.4 (Dyckerhoff, 2004, Proposition 3)**.** *If D is a convex depth function or a depth function satisfying $(\mathbf{D}_5)$ in the sense of Definition 2.1 and Definition 2.2, then there exists $z \in \mathbb{R}^d$ such that:*

$$D(z,P) = \max_{x\in\mathbb{R}^d} D(x,P)$$

Depending on the specific depth functions introduced, the maximum depth may be unique (see Section 2.3 for several depth examples and their properties).

2.2 Depth-Trimmed Regions

To describe global properties of a probability distribution such as location, dispersion or shape, one may be interested to the upper-level sets of depth functions named *depth-trimmed regions* or *central regions*. Precisely, for any $\alpha \in [0,1]$, the associated α-depth regions of a statistical depth function are defined as its upper-level sets:

$$D^\alpha(P) = \Big\{ x \in \mathbb{R}^d \,:\, D(x,P) \geq \alpha \Big\}.$$

The collection $(D^\alpha(P))_{\alpha\in[0,1]}$ reveals the geometry structure of the underlying probability distribution. It follows that these set-valued statistics are nested, i.e. $D^{\alpha'}(P) \subseteq D^\alpha(P)$ for any $\alpha < \alpha'$. Depth-trimmed regions correspond to the counterpart of quantile regions in $\mathbb{R}$ while their boundaries are analogous to quantiles for univariate distributions. It then allows computation of L-estimators of location parameters such as trimmed means in a multivariate space (see e.g. Serfling, 1984). Data depth properties $(\mathbf{D}_1, \mathbf{D}_3, \mathbf{D}_{3'}, \mathbf{D}_4, \mathbf{D}_5)$ mentioned in Section 2.1 can also be reformulated by equivalent properties attached to these nested regions as listed below.

($\mathbf{R}_1$) (Affine equivariance) Denoting by P_X the distribution of any r.v. X taking its values in $\mathbb{R}^d$, we have

$$D^\alpha(P_{AX+b}) = AD^\alpha(P_X) + b$$

for any $d \times d$ nonsingular matrix A with real entries and any vector b in $\mathbb{R}^d$.

($\mathbf{R}_3$) (Starshapedness) If $z \in \mathbb{R}^d$ is contained in all non-empty regions $D^\alpha(P)$, then the non-empty regions $D^\alpha(P)$, $\alpha \in (0,1]$ are starshaped w.r.t. z.

($\mathbf{R}_{3'}$) (Convexity) $D^\alpha(P)$ is convex for any $\alpha \in (0,1]$.

($\mathbf{R}_4$) (Boundedness) $D^\alpha(P)$ is bounded for any $\alpha \in (0,1]$.

($\mathbf{R}_5$) (Closedness) $D^\alpha(P)$ is closed for any $\alpha \in (0,1]$.

Proposition 2.5 (Serfling and Zuo, 2000; Dyckerhoff, 2004)**.** *Let $P \in \mathcal{P}(\mathbb{R}^d)$ and $D(\cdot, P) : \mathbb{R}^d \longrightarrow [0,1]$. Then it holds: $(\mathbf{D}_\ell) \Leftrightarrow (\mathbf{R}_\ell)$ for any $\ell \in \{1, 3, 3', 4, 5\}$.*

Let Y follow an elliptical distribution P_Y such that $Y \overset{\mathcal{L}}{=} a + AX$ for any $A \in \mathbb{R}^{d\times d}$ and $a \in \mathbb{R}^d$ where X is spherically distributed. Any depth function $D(\cdot, P_Y)$ that satisfies $(\mathbf{R}_1)$ has elliptical depth regions, i.e. upper-level sets of the quadratic form with matrix $AA^\top$ (see e.g. Mosler, 2013). Furthermore, an appealing behavior of satisfying $(\mathbf{R}_1)$ appears when P_Y has a unimodal Lebesgue density f_Y. Indeed, it can be shown that there exists a function h such that $f_Y(x) : h \circ D(x, P_Y)$ for every $x \in \mathbb{R}^d$ (Liu and Singh, 1993).

Remark 2.6. *Stochastic orders between random variables can be constructed from depth-trimmed regions. Massé and Theodorescu (1994) have derived an ordering notion between random variables based on the volume of depth regions. Let X, Y two random variables following $P_X, P_Y \in \mathcal{P}(\mathbb{R}^d)$ respectively. Authors have proposed to state $X \prec Y$ if*

$$\lambda_d(D^\alpha(P_X)) - \lambda_d(D^{\alpha'}(P_X)) \leq \lambda_d(D^\alpha(P_Y)) - \lambda_d(D^{\alpha'}(P_Y)),$$

for any $\alpha < \alpha'$. The quantities on either side of the inequality can be seen as analogous to the interquantile ranges of a univariate distribution. In addition, Zuo and Serfling (2000a) have proposed to define $X \prec Y$ if $\lambda_d(D^\alpha(P_X)) \leq \lambda_d(D^\alpha(P_Y))$ for every $\alpha \in (0, 1]$.

Remark 2.7. *Data depths can also be defined by means of regions in $\mathbb{R}^d$. Barnett (1976) and Eddy (1982) have introduced a way to order multivariate data by peeling convex hulls of a given set of sample points $\mathcal{S}_n = \{X_1, \ldots, X_n\}$ leading to the so-called convex hull peeling depth (= onion depth) even if it does not properly define a statistical depth function. Precisely, it consists in building a collection of subsets of $\mathbb{R}^d$ defined as $C_\ell = \text{conv}(\mathcal{S}_n \cap \text{int } C_{\ell-1})$ for $\ell = 2, \ldots$ with $C_1 = \text{conv}(\mathcal{S}_n)$. The "depth" of an element $x \in \mathbb{R}^d$ is then assigned as $\sum_{\ell \geq 2} \mathbb{I}\{x \in C_\ell\}$. Further, several data depths have been defined from a collection of subsets of $\mathbb{R}^d$ such as the zonoid depth (Koshevoy and Mosler, 1997) or the weighted-mean depths (Dyckerhoff and Mosler, 2011, 2012) following the construction:*

$$D(x, P) = \sup \{\alpha \in (0, 1] : x \in R^\alpha(P)\},$$

where $(R^\alpha(P))_{\alpha \in (0,1]}$ is a family of subsets of $\mathbb{R}^d$ that are affine equivariant, closed, bounded and starshaped.

2.3 Important Notions of Depth Functions

The *halfspace depth*, the first and probably the most studied depth in the literature, dates back to Tukey (1975). Many data depth functions based on various statistical tools ranging from U-statistics to outlyingness measures have been introduced in the continuation of the halfspace depth. In this section, we present precise definitions of some of them and discuss about their properties. With slight modifications to the nomenclature described in Zuo and Serfling (2000b), we group depth functions that share similar constructs in the following subsections. Structural properties of data depths presented in this section can be summarized in Table 2.1.

2.3.1 Depths based on Distribution Tails

Here we introduce the halfspace and the integrated rank-weighted depths.

- **Halfspace depth (= location depth, = Tukey depth).** Initially introduced as a visualization tool to reveal non-expected information present in data, the halfspace depth has become the most studied depth function in the literature probably due to its appealing properties as well as its connections with univariate quantiles. Indeed, when $d = 1$, for any $t \in \mathbb{R}$ and $P_1 \in \mathcal{P}(\mathbb{R})$, the halfspace depth, is defined by:

  $$D_{\text{H},1}(t, P_1) = \min\{F(t), 1 - F(t^-)\},$$

  where F is the cdf associated to P_1 and $F(t^-)$ denotes the left-sided limit of F at t. Therefore, this depth function, that corresponds to a center-oriented transformation of the cdf, is maximized by the median of the distribution P_1.

Generalizing the above expression by means of linear projections leads to the halfspace depth definition stated below. We refer the reader to Nagy et al. (2019) for an excellent account of halfspace depth characteristics.

Definition 2.8. *Let $x \in \mathbb{R}^d$ and X be a r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$. The halfspace depth of x w.r.t. P is defined as:*

$$\begin{aligned} D_{\mathrm{H}}(x, P) = & \inf \ \{P(H_x) \ : \ H_x \text{ is a closed halfspace containing } x\} \\ = & \inf_{u \in \mathbb{S}^{d-1}} \mathbb{P}\left(\langle u, X - x\rangle \leq 0\right). \end{aligned} \tag{2.2}$$

In order to enumerate halfspace depth properties, we need two regularity conditions on P. The first, trivially satisfied for absolutely continuous probability distributions, is a smoothness condition (see e.g. Dümbgen, 1992; Rousseeuw and Ruts, 1999) that says:

$$P(\partial H) = 0, \quad \forall \ H \text{ closed halfspace of } \mathbb{R}^d. \tag{2.3}$$

The second requires the support of P to be contiguous (see e.g. Kong and Zuo, 2010), i.e. there is no intersection of any two halfspaces with parallel boundaries that has nonempty interior but zero probability and divides the support of P into two parts. This condition is related to the fact that data depths are tools dedicated to unimodal distributions.

Convex depth. Zuo and Serfling (2000b) have shown that the halfspace depth satisfies conditions of the Definition 2.1. It also satisfies $(\mathbf{D}_5)$ (Donoho and Gasko, 1992, Lemma 6.1) and is further continuous in its first argument if (2.3) holds (see Mizera and Volauf, 2002, Proposition 1). Thus, combining $(\mathbf{D}_5)$ and quasi-concavity (see Rousseeuw and Ruts, 1999, Proposition 1), the halfspace depth belongs to the family of convex depth functions w.r.t. the Definition 2.2. It is worth noticing that the halfspace depth satisfies $(\mathbf{D}_2)$ for halfspace symmetric distributions.

Halfspace median. As the halfspace depth has the particularity to be maximized by the true median in the case of univariate data and possesses robustness properties (Donoho and Gasko, 1992), the set of halfspace medians has been extensively studied. In general, the set of all halfspace medians of P can be shown to be nonempty, compact and convex (Zuo and Serfling, 2000b; Dyckerhoff, 2004). If P has contiguous support and satisfies (2.3), then by the Proposition 7 in Mizera and Volauf (2002), the halfspace median of P is unique. After being conjectured by Rousseeuw and Hubert (1999b), a general lower bound on the set of halfspace medians has been provided by the Theorem 3.3 in Mizera (2002). Surprisingly, this result dates back to Neumann (1945) for $d = 2$ and Grünbaum (1960) for $d \geq 2$, much before the introduction of the halfspace depth.

We summarize the aforementioned properties of the halfspace depth in the following proposition.

Proposition 2.9. *For any $P \in \mathcal{P}(\mathbb{R}^d)$, the depth function $D_{\mathrm{H}}(\cdot, P)$ satisfies axioms of Definition 2.1 and Definition 2.2. For any $x \in \mathbb{R}^d$, the set $\sup_{x \in \mathbb{R}^d} D_{\mathrm{H}}(x, P)$ is non-empty, compact and convex and we have:*

$$\frac{1}{d+1} \leq \sup_{x \in \mathbb{R}^d} D_{\mathrm{H}}(x, P) \leq \left(1 + \sup_{x \in \mathbb{R}^d} P(\{x\})\right) \Big/ 2.$$

Furthermore, if P has contiguous support and satisfies (2.3), then the set of halfspace medians boils down to a unique element.

Bounds in Proposition 2.9 are known to be sharp (Grünbaum, 1960). It is possible to obtain a lower bound that does not depend on the dimension if P is distributed on a convex body or if P belongs to the family of s-concave probability distributions (see Theorem 3 in Nagy et al. (2019) for further details). There exist also bounds that depend on the sample size when P is replaced by the empirical distribution, see Donoho and Gasko (1992) and Liu et al. (2020).

In the univariate case, we know that the halfspace depth has 0.5 as maximum value. This property does not hold in general in multivariate spaces and authors of Dutta et al. (2011) have investigated conditions under which the maximum of the halfspace depth function is 0.5.

Characterization of P. In dimension one, cumulative distribution function as well as univariate quantiles uniquely characterize the underlying probability distribution. Many authors have investigated under which conditions this property may be satisfied by the halfspace depth starting by the work of Struyf and Rousseeuw (1999) that gives a positive answer for empirical distributions. Koshevoy (2002) showed that if $D(x, P) = D(x, Q)$, $\forall\, x \in \mathbb{R}^d$, where P, Q are two atomic measures with finite support then the measures are also identical. This assertion has been generalized by Cuesta-Albertos and Nieto-Reyes (2008b) when P, Q are general discrete distributions. Under some regularity conditions, the halfspace depth characterizes absolutely continuous probability distributions with compact support in $\mathbb{R}^d$ (Koshevoy, 2003); absolutely continuous distributions with connected supports (Hassairi and Regaieg, 2008); and absolutely continuous distributions with smooth depth contours including elliptical distributions (Kong and Zuo, 2010). Unfortunately, characterization is not true for general probability distributions (Nagy, 2021).

Continuity of $D^\alpha(P)$. We assume that the Tukey median x^*_{H} is unique (see Proposition 2.9 for mild sufficient conditions guaranteeing this property) and consider the map transporting $\alpha \in (0, x^*_{\mathrm{H}}]$ to the halfspace depth region $D^\alpha(P)$. From the properties $(\mathbf{D}_{3'}, \mathbf{D}_4, \mathbf{D}_5)$ and Section 2.2, we know that $D^\alpha(P)$ are convex bodies of $\mathbb{R}^d$ for every $\alpha \in (0, x^*_{\mathrm{H}}]$. By equipping the space of convex bodies with the Hausdorff distance $d_{\mathcal{H}}$, the continuity of $\alpha \in (0, x^*_{\mathrm{H}}] \mapsto D^\alpha(P)$ has been investigated in Massé and Theodorescu (1994); Mizera and Volauf (2002) and Dyckerhoff (2017) leading to the following proposition.

Proposition 2.10. *Let $P \in \mathcal{P}(\mathbb{R}^d)$ that has a contiguous support and satisfies (2.3). Then $\alpha \in (0, x^*_{\mathrm{H}}] \mapsto D^\alpha(P)$ is a continuous function w.r.t. the Hausdorff distance.*

This property will be crucial in order to derive statistical properties of the depth-trimmed regions that are further detailed in Section 2.4.1.

- **Integrated rank-weighted depth.** Recently, an integrated extension of the univariate halfspace depth in the same spirit of the integral dual depth (see Section 2.3.2) has been proposed in Ramsay et al. (2019). It consists in replacing the infimum by an integral in the halfspace depth definition (2.2).

  Definition 2.11. *Let $x \in \mathbb{R}^d$ and X be a r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$. The integrated rank-weighted depth (IRW) of x w.r.t. P is defined as:*

$$\begin{aligned} D_{\text{IRW}}(x, P) &= \int_{\mathbb{S}^{d-1}} D_{\text{H},1}(\langle u, x\rangle, P_u)\, \omega_{d-1}(\mathrm{d}u), \\ &= \mathbb{E}\left[D_{\text{H},1}(\langle U,\ x\rangle, P_U)\right], \end{aligned} \tag{2.4}$$

  where P_u is the pushforward distribution of P defined by the projection $x \in \mathbb{R}^d \mapsto \langle u, x\rangle$, i.e. $P_{\langle u,X\rangle}$, ω_{d-1} is the spherical probability measure on $\mathbb{S}^{d-1}$ and U is a r.v. uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$.

  As explained at length in Ramsay et al. (2019), the name of the data depth (2.4) originates from the fact that it can be represented as a weighted average of a finite set of normalized center-outward ranks. It has many advantages over the original halfspace depth (2.2). First, by construction it is *robust* to noisy directions and *sensitive* to new data point outside of the convex hull of the training data set both at the same time. Moreover, concerning numerical feasibility, the computation of the IRW depth does not require to implement any manifold optimization algorithm and can be approximately made by means of basic Monte-Carlo techniques, providing in addition confidence intervals as a by-product, see Section 2.4.2 below. Its contours $\{D_{\text{IRW}}(x, P) = \alpha\}$, $\alpha \in [0, 1]$, also exhibits a higher degree of smoothness in general. The depth function (2.4) is continuous at any point $x \in \mathbb{R}^d$ that is not an atom for P, cf. Proposition 1 in Ramsay et al. (2019), and properties $(\mathbf{D}_2, \mathbf{D}_3)$ for halfspace symmetric distributions, and $(\mathbf{D}_4)$ have been proved to be satisfied by IRW, see Theorem 2 in Ramsay et al. (2019). However, IRW fails to validate $(\mathbf{D}_1)$. This is the aim of Chapter 9 to build an affine invariant formulation of the IRW, satisfying the data depth definition, in order to exploit both its ease of computation and its advantageous properties.

2.3.2 Depths based on U-Statistics

Some data depth functions are based on U-statistics (see Lee, 1990 for a review) such as the simplicial depth and its integrated counterpart that are described below.

- **Simplicial depth.** The simplicial depth was first introduced and studied in Liu (1990). As the halfspace depth, the maximum depth is attained by the median in the univariate case. Indeed, the simplicial depth in dimension one is defined as $D_{\text{S},1}(t, P_1) = 2F(t)(1 - F(t^-))$ for any $t \in \mathbb{R}$ and $P_1 \in \mathcal{P}(\mathbb{R})$ where F is the cdf of P_1. It is equivalently defined as the probability that $t \in \mathbb{R}$ belongs to the segment between two r.v. $X_1, X_2 \sim P_1$ which is maximized by the median. The segment can be also considered as convex hull of the set $\{X_1, X_2\}$ leading to the following definition in the multivariate case.

Definition 2.12. *Let $x \in \mathbb{R}^d$ and X be a r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$. The simplicial depth of x w.r.t. P is defined as:*

$$D_{\mathrm{S}}(x, P) = \mathbb{P}\left(x \in \mathrm{conv}(X_1, \ldots, X_{d+1})\right),$$

where $X_1, \ldots, X_{d+1}$ are i.i.d. random variables following P.

The simplicial depth satisfies $(\mathbf{D}_1)$ in general, satisfies $(\mathbf{D}_2)$ for halfspace symmetric distributions, is continuous in x, and has a unique maximum depth for absolutely continuous distributions only. The property $(\mathbf{D}_3)$ holds for absolutely continuous distributions that are angularly symmetric. See Liu (1990) for further details about these properties.

- **Integrated dual depth.** It is the purpose of the integrated dual (ID) depth (Cuevas and Fraiman, 2009) to propose an extension of the univariate formulation of the simplicial depth to multivariate, and more general Banach spaces with separable dual. It relies on one-dimensional linear continuous projections and is defined as follows in the multivariate space.

  Definition 2.13. *Let $x \in \mathbb{R}^d$ and X be a r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$. The integrated dual depth of x w.r.t. P is defined as:*

$$D_{\mathrm{ID}}(x, P) = \int_{\mathbb{S}^{d-1}} D_{\mathrm{S},1}(\langle u, x\rangle, P_u)\, \omega_{d-1}(\mathrm{d}u),$$

  where P_u is the pushforward distribution of P defined by the projection $x \in \mathbb{R}^d \mapsto \langle u, x\rangle$, i.e. $P_{\langle u,X\rangle}$, and ω_{d-1} is the spherical probability measure on $\mathbb{S}^{d-1}$.

  ID depth properties have been extensively studied in Cuevas and Fraiman (2009). It is invariant to orthogonal transformations and translations but fails to be affine invariant. It is also continuous in $x \in \mathbb{R}^d$ and vanishes at infinity. In addition, it satisfies $(\mathbf{D}_2, \mathbf{D}_3)$ for halfspace symmetric distributions. It is worth noting that its maximum depth is equal to that of the IRW depth and might not be unique (Ramsay et al., 2019).

2.3.3 Depths based on Dispersion Measures

Some data depth functions involving dispersion measures are described in this section.

- **Oja depth.** The Oja depth is based on average volumes of the convex hull of the data (Zuo and Serfling, 2000b) generalizing the Oja multivariate median (Oja, 1983) into a depth function.

  Definition 2.14. *Let $x \in \mathbb{R}^d$ and X be a square integrable r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$. Let Σ be the covariance matrix of X. The Oja depth of x w.r.t. P is defined as:*

$$D_{\mathrm{Oja}}(x, P) = \left(1 + \frac{\mathbb{E}\left[\lambda_d(\mathrm{conv}(x, X_1, \ldots, X_d))\right]}{\sqrt{\det(\Sigma)}}\right)^{-1}$$

  where $X_1, \ldots, X_d$ are i.i.d. random variables following P.

The Oja depth satisfies properties $(\mathbf{D}_1, \mathbf{D}_{3'}, \mathbf{D}_4, \mathbf{D}_5)$ and $(\mathbf{D}_2)$ for centrally symmetric distributions, and is further continuous in $x \in \mathbb{R}^d$ (see Zuo and Serfling, 2000b). Thus, this depth belongs to the family of convex depths. However, its maximum is not unique in general (Oja, 1983). It is worth mentioning that the Oja depth characterizes distributions with compact supports of full dimension (Koshevoy, 2003).

- **Spatial depth.** The spatial depth relies on ideas of spatial quantiles (Chaudhuri, 1996; Koltchinskii and Dudley, 1996; Koltchinskii, 1997) and has been proposed in Vardi and Zhang (2000) (see also Serfling, 2002).

  Definition 2.15. *Let $x \in \mathbb{R}^d$ and X be a r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$. The spatial depth of x w.r.t. P is defined as:*

$$D_{\mathrm{Sp}}(x, P) = 1 - \left|\left| \mathbb{E}\left[\frac{x - X}{||x - X||} \right] \right|\right|,$$

  where 0/0=0 by convention.

  The spatial depth satisfies $(\mathbf{D}_2)$ for angularly symmetric distributions, $(\mathbf{D}_4)$, is invariant regarding translation and orthogonal transformations, and is continuous in x (Serfling, 2002). It is also known to have a unique median (Milasevic and Ducharme, 1987). Unfortunately, the spatial depth fails to satisfy $(\mathbf{D}_3)$ leading to non starshaped regions (Nagy, 2017). There exists a kernelized version replacing inner products (derived from the Euclidean norm) by kernel functions (Chen et al., 2009).

- **L_p depth.** This depth function has been introduced in Zuo and Serfling (2000b) and is based on the expected L_p norm between $x \in \mathbb{R}^d$ and the random variable X following $P \in \mathcal{P}(\mathbb{R}^d)$.

  Definition 2.16. *Let $x \in \mathbb{R}^d$ and X be a r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$. The L_p depth of x w.r.t. P is defined as:*

$$D_{\mathrm{L}_p}(x, P) = \left(1 - \mathbb{E}\left[||x - X||_p \right]\right)^{-1},$$

  with $1 \leq p < \infty$.

  The L_p depth satisfies $(\mathbf{D}_2)$ for centrally symmetric distributions, $(\mathbf{D}_{3'}, \mathbf{D}_4, \mathbf{D}_5)$ but is not affine invariant in general (Zuo and Serfling, 2000b). When $p = 2$, the deepest point of the L_2 depth is the spatial median that is unique (Milasevic and Ducharme, 1987). Furthermore, due to the presence of Euclidean distance, the L_2 depth is invariant to isometric transformations as the spatial depth.

2.3.4 Depths based on Outlyingness Measures

Constructed as rescaling of outlyingness measures, the Mahalanobis and the projection depths are introduced and discussed below.

- **Mahalanobis depth.** The Mahalanobis distance (Mahalanobis, 1936) is a distance between an element in $\mathbb{R}^d$ and a probability distribution having finite expectation and invertible covariance matrix differing from the Euclidean by taking account of correlations. Interpreting the Mahalanobis distance as an outlyingness measure, Liu (1992) suggested to rescale it to define a depth function as follows.

  Definition 2.17. *Let $x \in \mathbb{R}^d$ and X be a square integrable r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$ with invertible covariance matrix Σ. The Mahalanobis depth of x w.r.t. P is defined as:*

$$D_{\mathrm{M}}(x, P) = \Big(1 + (x - \mathbb{E}[X])^\top \Sigma^{-1} (x - \mathbb{E}[X])\Big)^{-1},$$

  where Σ^{-1} is the precision matrix of the r.v. X.

  This data depth satisfies $(\mathbf{D}_1, \mathbf{D}_{3'}, \mathbf{D}_4, \mathbf{D}_5)$, $(\mathbf{D}_2)$ for centrally symmetric distributions, and then belongs to the family of convex data depths. It also has a unique maximum at $\mathbb{E}[X]$. In practice, the Mahalanobis depth is very easy to compute as it only requires to compute the sample mean and the sample covariance matrix but lacks of robustness (to outliers) as pointed out in Liu and Singh (1993). However, these estimators may be replaced by robust ones, such as the Median-of-Means (see e.g. Devroye et al., 2016; Lecué and Lerasle, 2020) for the sample mean and the Minimum Volume Ellipsoid (Rousseeuw, 1985; Van Aelst and Rousseeuw, 2009) or the Minimum Covariance Determinant (Rousseeuw, 1984; Rousseeuw and Leroy, 1987) for the covariance matrix.

- **Projection depth.** The Stahel-Donoho outlyingness has been introduced independantly in Stahel (1981) and Donoho (1982). It is based on univariate measures of outlyingness extended to the multivariate case by means of projection pursuit (Friedman and Tukey, 1974). This outlyingness measure is then rescaled as data depth leading to the definition provided below (Liu, 1992).

  Definition 2.18. *Let $x \in \mathbb{R}^d$ and X be a r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$. The projection depth of x w.r.t. P is defined as:*

$$D_{\mathrm{P}}(x, P) = \left(1 + \sup_{u \in \mathbb{S}^{d-1}} \frac{|\langle u, x\rangle - \mathrm{med}(\langle u, X\rangle)|}{\mathrm{MAD}(\langle u, X\rangle)}\right)^{-1},$$

  where med *and* MAD *stand for the univariate median and median absolute deviation from the median, respectively.*

  The depth score of x is then based on the direction $u^* \in \mathbb{S}^{d-1}$ where the inner product $\langle u^*, x\rangle$ deviates the most from the univariate median of the projected r.v. $\langle u^*, X\rangle$. This data depth is then robust to outliers by construction (see Section 2.4.3 for further details). The projection depth has good properties since it satisfies $(\mathbf{D}_1, \mathbf{D}_{3'}, \mathbf{D}_4, \mathbf{D}_5)$ and $(\mathbf{D}_2)$ for halfspace symmetric distributions, is continuous in x and has a unique median (Zuo, 2003, 2013). Therefore, the projection depth belongs to the family of convex depths. It shares with the halfspace depth the nice feature to have continuous depth-trimmed regions in α if the mild conditions $\mathrm{med}(\langle u, X\rangle) < \infty$ and $\mathrm{MAD}(\langle u, X\rangle) < \infty$ for every $u \in \mathbb{S}^{d-1}$ hold. The projection depth is widely used in practice regarding its ordering properties, robustness (see Section 2.4.3) and computational feasibility (see Section 2.4.2).

2.3.5 Uncategorized Depths

Some data depths can not be integrated in the nomenclature given in Zuo and Serfling (2000b). It is the aim of this section to present some of them based on interesting construction.

- **Zonoid depth.** The zonoid depth has been introduced in Dyckerhoff et al. (1996) (see also Koshevoy and Mosler, 1997) and has the singularity to be constructed by means of subsets of $\mathbb{R}^d$ named zonoid regions. Let X be an integrable r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$, zonoid α-trimmed regions for any $\alpha \in (0, 1]$ are defined as:

$$R_{\mathrm{Z}}^{\alpha}(P) = \left\{ \mathbb{E}\left[Xg(X)\right] \middle| \, g : \mathbb{R}^d \longrightarrow [0, 1/\alpha] \text{ measurable and } \mathbb{E}\left[g(X)\right] = 1 \right\}$$

  Definition 2.19. *Let $x \in \mathbb{R}^d$ and X be an integrable r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$. The zonoid depth of x w.r.t. P is defined as:*

$$D_{\mathrm{Z}}(x, P) = \sup \{\alpha \, : \, x \in R_{\mathrm{Z}}^{\alpha}(P)\},$$

  with $D_{\mathrm{Z}}(x, P) = 0$ if $x \notin R_{\mathrm{Z}}^{\alpha}$ for every $\alpha \in (0, 1]$.

  The zonoid depth has desirable properties since it satisfies $(\mathbf{D}_1, \mathbf{D}_{3'}, \mathbf{D}_4, \mathbf{D}_5)$ and $(\mathbf{D}_2)$ for centrally symmetric distributions. It also possesses the nice feature to characterize any probability distribution with finite first moment (Mosler, 2002). The zonoid median is unique and is equal to $\mathbb{E}[X]$ that makes it sensitive to outliers. Furthermore, as the halfspace depth, the zonoid depth vanishes beyond the convex hull of the support of P. It also suffers from computational burden. The aforementioned drawbacks make it unsuitable for unsupervised anomaly detection.

- **Lens depth.** Following the idea behind the spherical depth (Elmore et al., 2006), the Lens depth has been introduced in Liu and Modarres (2011). It relies on the probability that an element $x \in \mathbb{R}^d$ belongs to a random hyper-lens defined by the intersection of two closed balls centered at two i.i.d. random variables following P.

  Definition 2.20. *Let $x \in \mathbb{R}^d$ and X be a r.v. following $P \in \mathcal{P}(\mathbb{R}^d)$. The Lens depth of x w.r.t. P is defined as:*

$$D_{\mathrm{Lens}}(x, P) = \mathbb{P}\left(x \in L(X, Y)\right),$$

  where X and Y are i.i.d. random variables from P and $L(X, Y)$ is the intersection of two closed balls of $\mathbb{R}^d$ with radius $||X - Y||$ centered at X and Y, respectively.

  Properties of the Lens depth have been extensively studied in Liu and Modarres (2011). It is invariant to orthogonal transformations, satisfies $(\mathbf{D}_2)$ for centrally symmetric distributions and $(\mathbf{D}_4)$, and is continuous in x for absolutely continuous distributions. The monotonicity property is valid only for centrally symmetric distributions.

- **Monge-Kantorovich depth.** Motivated by the desire to design a data depth that takes account of non-convex features of the underlying distribution, authors of Chernozhukov et al. (2017) have introduced the Monge-Kantorovich (MK) depth. This depth is constructed on the idea of transporting depth contours of any data depth that well-behaves on a reference distribution to a distribution P and relies on the optimal transport theory (see e.g. Villani, 2003; Santambrogio, 2015). Let $P, Q_{\text{ref}} \in \mathcal{P}(\mathbb{R}^d)$ be both absolutely continuous w.r.t. the Lebesgue measure. There exists a Q_{ref} a.s. unique $\nabla\psi$, the gradient of a convex, lower semi-continuous function ψ defined on a convex set $\mathcal{C}_1$ containing the support of Q_{ref} that map Q_{ref} to P, i.e. $\nabla\psi_\sharp Q_{\text{ref}} = P$ (Brenier, 1991; McCann, 1995). Furthermore, there exists a unique real function ψ^* defined on a convex set $\mathcal{C}_2$ containing the support of P such that $\nabla\psi^*_\sharp P = Q_{\text{ref}}$.

  Remark 2.21. *It can be shown (see Brenier, 1991) that if P and Q_{ref} have finite second moments, $\nabla\psi$ solves the Monge optimal transport problem:*

$$\inf_T \int_{\mathcal{C}_1} ||x - T(x)||^2 \, Q_{\text{ref}}(\mathrm{d}x) \quad \text{s.t. } T_\sharp Q_{\text{ref}} = P.$$

  We are now ready to give the definition of the Monge-Kantorovich depth.

  Definition 2.22. *Let $Q_{\text{ref}} \in \mathcal{P}(\mathbb{R}^d)$ be a reference probability distribution whose support is included in the convex set $\mathcal{C}_1$. Let $P \in \mathcal{P}(\mathbb{R}^d)$ and assume that both P and Q_{ref} are absolutely continuous w.r.t. the Lebesgue measure. Let $\nabla\psi$ the unique gradient of the convex and semi-lower continuous function ψ such that $\nabla\psi_\sharp Q_{\text{ref}} = P$. The Monge-Kantorovich depth of $x \in \mathbb{R}^d$ w.r.t. P is defined as:*

$$D_{\text{MK}}(x, P) = D(\text{rank}_P(x), Q_{\text{ref}}), \tag{2.5}$$

  where $\text{rank}_P(x) \in \underset{y \in \mathcal{C}_1}{\text{argsup}}\{\langle y, x\rangle - \psi(y)\}$ *and $D(\cdot,\cdot)$ is any data depth function.*

  The Monge-Kantorovich depth is invariant to isometric transformations but fails to be affine invariant. The function $D_{\text{MK}}(., P)$, is an homeomorphism under mild conditions on P (Figalli, 2018; Hallin et al., 2021). Its formulation is very general and to extract depth properties is challenging.

2.4 General Properties

Data depths such as those described in Section 2.3 require the knowledge of the probability distribution P. However, we cannot access to P in practice but rather to a sample of it. Let $X_1, \ldots, X_n$ be an i.i.d. sample following the distribution $P \in \mathcal{P}(\mathbb{R}^d)$ and denote by $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ the associated empirical distribution. The natural questions are then: *Does $D(x, P_n)$ converge towards $D(x, P)$? If yes, with which statistical rates? Does it have an asymptotic Gaussian behavior?* It is the aim of Section 2.4.1 to give answers to the aforementioned questions for the statistical depth functions introduced previously.

Recent advances in data collection allow the automatic acquisition of data, which often leads to large and contaminated data sets. Thus, two central challenges have emerged and can be summarized by: *how to compute efficiently the estimator $D(x, P_n)$?* and

	$\mathbf{D}_1$	$\mathbf{D}_2$	$\mathbf{D}_3$	$\mathbf{D}_{3'}$	$\mathbf{D}_4$	Continuity	Unique median
Halfspace	✓	$\checkmark_{\text{HS}}$	✓	✓	✓	✓	$\checkmark_{\text{cont}}$
IRW	✗	$\checkmark_{\text{HS}}$	$\checkmark_{\text{HS}}$	✗	✓	✓	✗
Simplicial	✓	$\checkmark_{\text{HS}}$	$\checkmark_{\text{AS}}$	✗	✓	✓	✓
ID	✗	$\checkmark_{\text{HS}}$	$\checkmark_{\text{HS}}$	✗	✓	✓	✗
Oja	✓	$\checkmark_{\text{CS}}$	✓	✓	✓	✓	✗
Spatial	✗	$\checkmark_{\text{AS}}$	✗	✗	✓	✓	✓
L_p	✗	$\checkmark_{\text{CS}}$	✓	✓	✓	✓	$\checkmark_{p=2}$
Mahalanobis	✓	$\checkmark_{\text{CS}}$	✓	✓	✓	✓	✓
Projection	✓	$\checkmark_{\text{HS}}$	✓	✓	✓	✓	✓
Zonoid	✓	$\checkmark_{\text{CS}}$	✓	✓	✓	✓	✓
Lens	✗	$\checkmark_{\text{CS}}$	✗	✗	✓	✓	✗
MK	✗	-	-	✗	-	✓	-

Table 2.1 – Structural properties of the presented depth functions. For the sake of clarity and for the purpose of this manuscript, P is assumed to be absolutely continuous w.r.t. the Lebesgue measure with finite first and second moments. We distinguish under which notion of symmetry (defined in Section 2.1) some properties hold giving the stronger result. By cont is meant contiguous support for the distribution P.

Which data depths are robust to contaminated data? The computational aspects of data depths have been widely studied by the statistical community but many challenges remain. It is the aim of Section 2.4.2 to discuss about the computational complexity of data depths. Further, the robustness is discussed in Section 2.4.3.

2.4.1 Statistical Analysis

This section gathers statistical results of the estimator $D(x, P_n)$ for the depth functions introduced in Section 2.3. Consistency has been thoroughly studied for most data depths. However, knowing the limit distribution of the depth process $x \mapsto D(x, P_n) - D(x, P)$ has received less attention while finite-sample rates have been derived for the halfspace depth only.

Consistency. Statistical consistency of multivariate data depths is studied through point-wise and uniform convergences. More precisely, in decreasing order of generality, we give the three following definitions:

($\mathbf{C}_{\text{pw}}$) For any $P \in \mathcal{P}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$, as $n \to \infty$ we have:

$$|D(x, P_n) - D(x, P)| \xrightarrow{a.s.} 0.$$

($\mathbf{C}_{\text{uc}}$) For any $P \in \mathcal{P}(\mathbb{R}^d)$, as $n \to \infty$ we have for any compact set $C \subset \mathbb{R}^d$:

$$\sup_{x \in C} |D(x, P_n) - D(x, P)| \xrightarrow{a.s.} 0.$$

($\mathbf{C}_{\mathrm{u}}$) For any $P \in \mathcal{P}(\mathbb{R}^d)$, as $n \to \infty$ we have:

$$\sup_{x \in \mathbb{R}^d} |D(x, P_n) - D(x, P)| \xrightarrow{a.s.} 0.$$

To capture the geometrical shape of the underlying probability distribution and more generally to understand its global features, one needs to rely on the set $\{D(x, P),\ x \in \mathbb{R}^d\}$. Therefore, this set needs to be well-approximated by its sample version which is provided by ($\mathbf{C}_{\mathrm{u}}$). When ($\mathbf{C}_{\mathrm{u}}$) cannot be satisfied, two weaker convergence definitions ($\mathbf{C}_{\mathrm{uc}}$) and ($\mathbf{C}_{\mathrm{pw}}$) may be valid. The property ($\mathbf{C}_{\mathrm{u}}$) is satisfied for most multivariate data depths but rarely holds when data depths are defined on functional spaces (see Chapter 3).

Uniform consistency ($\mathbf{C}_{\mathrm{u}}$) is satisfied by the halfspace depth (Donoho and Gasko, 1992), the integrated rank-weighted depth (Ramsay et al., 2019), the simplicial depth (Arcones and Gine, 1993; Dümbgen, 1992), the integrated dual depth (Cuevas and Fraiman, 2009), and the Lens depth (Liu and Modarres, 2011) in general. The Mahalanobis depth satisfies ($\mathbf{C}_{\mathrm{u}}$) if P has invertible covariance matrix (Liu and Singh, 1993). The projection depth also satisfies ($\mathbf{C}_{\mathrm{u}}$) under mild assumptions ensuring that $\mathrm{med}(\langle u, X\rangle)$ and $\mathrm{MAD}(\langle u, X\rangle)$ are unique for every $u \in \mathbb{S}^{d-1}$ and that their sample versions converge almost surely to them uniformly in $u \in \mathbb{S}^{d-1}$ (Serfling and Zuo, 2000). These assumptions are similar to those for univariate quantiles consistency see (Serfling, 1980, Theorem 2.3.2). The sample zonoid depth converges uniformly to its population version under (2.3) if P has a finite first moment (Cascos and López-Díaz, 2016). The Monge-Kantorovich depth rather satisfies probability convergence uniformly on any compact subset included in $\mathcal{C}_2$, a convex set that contains the support of P (Chernozhukov et al., 2017). The Oja depth, the spatial depth and the L_p depths trivially satisfy ($\mathbf{C}_{\mathrm{u}}$) thanks to the law of large numbers and its equivalent for U-statistics.

The convergence of sample depth contours for the halfspace depth function has been investigated, among others, in Nolan (1992); Massé and Theodorescu (1994); Serfling and Zuo (2000) and Kuelbs and Zinn (2016). These results have been unified in Dyckerhoff (2017): if (2.3) holds and if P has a contiguous support then $\sup_{\alpha \in L} d_{\mathcal{H}}(D^\alpha(P_n), D^\alpha(P)) \xrightarrow{a.s.} 0$ as n goes to infinity where L is any compact interval in $(0, x^*_{\mathrm{H}}]$. Uniform consistency of depth contours has been proved for elliptical distributions under some conditions on a general depth function (He and Wang, 1997). The consistency of $D(\cdot, P_n)$ is closely related to that of its associated depth-trimmed regions $D^\alpha(P_n)$, $\alpha \in [0, 1]$. Indeed, it can be shown that the continuity of the depth-trimmed regions map $\alpha \mapsto D^\alpha(P)$ is crucial in order to link ($\mathbf{C}_{\mathrm{pw}}$) and ($\mathbf{C}_{\mathrm{uc}}$) with their depth-trimmed regions counterparts (w.r.t. the Hausdorff distance), see Dyckerhoff (2017), Theorems 3.2, 4.5, 4.7. We refer the reader to Dyckerhoff (2017) for further convergence equivalence results and detailed discussions about the relation between the convergence of data depth and its depth-trimmed regions.

Asymptotic normality. Knowing the limit distribution of an estimator, whether Gaussian or any well-known probability distribution, is crucial in statistics in order to provide confidence bounds or to construct hypothesis testing. Regarding data depth, this property has been extensively investigated for the most popular depths such as halfspace or simplicial. We start by recalling the two following definitions of Central limit theorem (CLT) for data depths.

($\mathbf{CLT}_{\text{pw}}$) For any $P \in \mathcal{P}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$, as $n \to \infty$ we have:

$$\sqrt{n}\,|D(x, P_n) - D(x, P)| \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 > 0$.

($\mathbf{CLT}_{\text{u}}$) For any $P \in \mathcal{P}(\mathbb{R}^d)$, as $n \to \infty$ we have:

$$\sup_{x \in \mathbb{R}^d} \sqrt{n}\,|D(x, P_n) - D(x, P)| \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 > 0$.

Asymptotic normality of the Tukey median process has been investigated in Bai and He (1999) and Massé (2002). However, the limit distribution of the halfspace depth is not necessary Gaussian but conditions under which the CLT is satisfied on a subset of $\mathbb{R}^d$ can be stated (Massé, 2004). A Gaussian limit distribution ($\mathbf{CLT}_{\text{u}}$) has been provided for the simplicial depth (Dümbgen, 1992). Both median processes of the simplicial and the Oja depths have been derived through the study of U-processes in Arcones et al. (1994). Point-wise CLTs have been derived for integrated version of the halfspace/simplicial depths in Ramsay et al. (2019) and Cuevas and Fraiman (2009) respectively and for Lens depth in Liu and Modarres (2011). Asymptotic properties of the projection depth have been investigated at length. Zuo (2003) has derived the existence of a limit distribution of the empirical process of the projection median.

Some limit distributions of projection depth-based procedures such as location estimators (Maronna and Yohai, 1995), weighted means (Zuo et al., 2004a), L-statistics (Massé, 2009), trimmed means (Kim, 1992) or scatter estimators (Zuo and Cui, 2005) have been highlighted.

Finite-sample analysis. Nonasymptotic results about the accuracy of sample versions of statistical depths, such as those stated above, are seldom in the literature. To the best of our knowledge, rate bounds have only been derived in the halfspace depth case. The first result, where uniform rates of the sample version are provided, uses the fact that the set of halfspaces in $\mathbb{R}^d$ is of finite Vapnik-Chervonenkis dimension (see e.g. Vapnik, 1999) and is recalled below.

Proposition 2.23 (Shorack and Wellner (1986), Chapter 26)**.** *Let $P \in \mathcal{P}(\mathbb{R}^d)$. Let $X_1, \ldots, X_n$ a sample from P with empirical measure $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$. Denote by F_u and $F_{u,n}$ the cdf of P_u and $P_{u,n}$ respectively. Then for any $t > 0$ it holds:*

$$\mathbb{P}\left(\sup_{\substack{x \in \mathbb{R}^d \\ u \in \mathbb{S}^{d-1}}} \left| F_{u,n}(u^\top x) - F_u(u^\top x) \right| > t \right) \leq \frac{6(2n)^{d+1}}{(d+1)!} \exp(-nt^2/8).$$

Recently, this result has been refined under further assumptions (Burr and Fabrizio, 2017). An asymptotic convergence rate for the Monte-Carlo approximation of the halfspace depth, i.e. when the minimum over the unit hypersphere is approximated from a finite number of directions, has been recently established in Nagy et al. (2020b). Unfortunately, approximating a minimum over the unit sphere $\mathbb{S}^{d-1}$ using a Monte-Carlo scheme is not optimal. Indeed, when the distribution is assumed to belong to

a bounded subset of $\mathbb{R}^d$ with bounded density, the authors obtain slow rates of order $\mathcal{O}((\log(n_{\text{proj}})/n_{\text{proj}})^{1/(d-1)})$ where n_{proj} is the number of sampled directions to approximate the minimum over the hypersphere, suffering from the curse of dimensionality. Futhermore, they show that obtaining uniform rates of the halfspace depth approximation is not possible in absence of the bounded density assumption (see the second example in Section 4.2 in Nagy et al. (2020b)). Recently, Brunel (2019) provided a finite-sample analysis of the quantity $d_{\mathcal{H}}(D^\alpha(P_n), D^\alpha(P))$ and obtained parametric rates with explicit constants completing the work of Kim (2000).

2.4.2 Computational Issues

Most of sample versions of data depths introduced in Section 2.3 can be exactly computed. Some depth functions can be computed very efficiently such as the spatial and L_p depths that have sample version calculable in $\mathcal{O}(dn)$ but are sensitive to outliers. The Mahalanobis depth requires to invert the covariance matrix leading to a complexity of $\mathcal{O}(d^3 n)$. Unfortunately, computational complexity often grows exponentially with the dimension d for most data depths. Relying on approximations can be essential, leading to a well-known trade off between the computational efficiency and the accuracy of the approximation. Three computational challenges can be exhibited regarding the presented data depths. The first comes from combinatorial depths based on U-statistics such as the Oja depth where $\binom{n}{d+1}$ combinations are necessary to compute the sample mean. The second relies on computing statistics that have a geometric aspect such as the computation of the convex hull of sample points in the simplicial depth or its volume in the Oja depth. These two challenges have led to complexity of order $\mathcal{O}(n^{d+1})$ and $\mathcal{O}(n^d)$ for the simplicial and the Oja depth respectively yielding unfeasibility for most data sets. However, they both can be approximated by choosing a smaller of combinations instead of $\binom{n}{d+1}$ and $\binom{n}{d}$ or by taking a smaller portion of data to construct simplices (see e.g. Chapter 6 where both approximation techniques are investigated). The Lens depth, being a U-statistics of order 2 can be computed exactly in $\mathcal{O}(n^2 d)$. The third challenge comes from computing a minimum/maximum over the unit sphere of non-differentiable function (indicator function) such as for the halfspace or projection depths. A general algorithm, based on cutting a convex polytope with hyperplanes, to compute the projection depth and its contours with complexity $\mathcal{O}(n^d)$ has been described in Liu and Zuo (2014a). Algorithms to compute the Oja median and the spatial median can be found in Fischer et al. (2020) and Kent et al. (2015) respectively. Computation of the halfspace depth has received much attention and some references are given below.

Computation of halfspace depth. Starting with bivariate distributions, an algorithm that computes the halfspace depth of a point $x \in \mathbb{R}^2$ in $\mathcal{O}(n \log(n))$ has been proposed by Rousseeuw and Ruts (1996) and appears to be time-optimal (Langerman and Steiger, 2003). This algorithm relies on the idea of a *circular sequence* (see e.g. Edelsbrunner, 1987). A single depth-trimmed contour can be computed in $\mathcal{O}(n^2 \log(n))$ (Ruts and Rousseeuw, 1996b,a). Focusing on a small batch of the data set as well as a small number of depth contours to be computed, a faster algorithm, named FDC, has been introduced in Johnson et al. (1998) to compute several depth contours at the same time. Authors of Miller et al. (2003) provided a way to compute all depth regions in $\mathcal{O}(n^2)$ by means of *duality* and *toplogical sweep* of an arrangement of lines. When $d > 2$, to compute the depth function of a point and depth contours are a much more challenging task. The first introduced algorithm to compute the halfspace depth of an element $x \in \mathbb{R}^3$ requires $\mathcal{O}(n^3 \log(n))$ operations (Rousseeuw and Struyf, 1998). Later,

Bremner et al. (2006) proposed a primal-dual algorithm by successively updating upper and lower bounds of the depth in dimension d by means of *reserve search* technique, futher improved by Bremner et al. (2008). A faster algorithm has been developed in Liu and Zuo (2014b) using a breadth-first search algorithm to cover $\mathbb{R}^d$ and the QHULL algorithm to define cones (see also Liu, 2017). Dyckerhoff and Mozharovskyi (2016) proposed a whole class of algorithms allowing the computation of the exact halfspace depth in $\mathcal{O}(n^d)$ without requiring assumptions on the data set.

Following the discussion in Kong and Mizera (2012) about connections between halfspace trimmed regions and the intersection of directional quantile halfspaces, algorithms to compute halfspace contours, based on the division of the unit sphere into cones, have been investigated in Paindaveine and Šiman (2012). The idea to segment $\mathbb{R}^d$ into direction cones dates back to Mosler et al. (2009) for the computation of the zonoid depth and in Hallin et al. (2010) for the halfspace depth. A faster algorithm, employing the breadthfirst ridge-by-ridge search strategy, has been further introduced in (Liu et al., 2019).

Approximation of halfspace depth. Data depths possessing the projection property, such as the halfspace, projection and zonoid depths, can be approximated with a finite number, say n_{proj}, of sampled directions uniformly in $\mathbb{S}^{d-1}$ leading to a computational complexity in $\mathcal{O}(n_{\text{proj}}\, nd)$ (Dyckerhoff, 2004), see Algorithms 2.1 and 2.2 for the halfspace and projection depths respectively. For the halfspace depth, this approximation is known as the random Tukey depth (Cuesta-Albertos and Nieto-Reyes, 2008a). Approximating the minimum over the unit sphere by means of Monte-Carlo generates poor statistical accuracy involving the so-called curse of the dimension (Nagy et al., 2020b). However, this naive approximation can be improved. Indeed, Rousseeuw and Struyf (1998) (see also Struyf and Rousseeuw, 2000) have sampled directions based on a random combination of sample points while Chen et al. (2013) have proposed to sample directions uniformly but after projecting data into orthogonal subspaces. Afshani and Chan (2009) have presented an algorithm based on a randomized data structure and halfspace range counting queries techniques. Mozharovskyi et al. (2015) have provided an accelerated algorithm when the halfspace depth is computed on the sample itself. Further, the projection depth has been approximated by means of the Nelder-Mead algorithm (Dutta and Ghosh, 2012). Recently, Zuo (2019) has developped an algorithm that avoids $\mathcal{O}(n^d)$ polyhedral cones, using instead essentially pointwise distances of the order of $\mathcal{O}(n^2)$. This algorithm allows to compute the sample depth of a point that has values p/n in $\mathcal{O}(n^2(d + p + \log(n)))$ being linear w.r.t. the dimension d. An extensive study of further approximations based on zero order optimization algorithms and dedicated to depth satisfying the projection property can be found in Dyckerhoff et al. (2021). Namely, authors have investigated (*i*) random search, (*ii*) grid search, (*iii*) refined random search, (*iv*) refined grid search, (*v*) random simplices, (*vi*) simulated annealing, (*vii*) coordinate descent, (*viii*) Nelder-Mead, where (*iii*) and (*viii*) appear to reach the best performances.

Approximation of IRW depth. The exact computation of IRW and ID depths has not been investigated probably due to the fact that these depths can be well approximated by means of Monte-Carlo techniques leading to a complexity in $\mathcal{O}(n_{\text{proj}}nd)$. As the IRW depth will be the foundation of Chapter 9, we give more details on how to approximate IRW. Recall that a r.v. uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$ can be generated from a d-dimensional centered Gaussian random vector W with the identity I_d as covariance matrix: if $W \sim \mathcal{N}(0, \mathrm{I}_d)$, then $W/||W|| \sim \omega_{d-1}$, see Krantz

and Parks (2008). Hence, a basic Monte-Carlo method to approximate the IRW depth (2.4) would consist in generating $n_{\text{proj}} \geq 1$ independent realizations $W_1, \ldots, W_{n_{\text{proj}}}$ of $\mathcal{N}(0, \mathrm{I}_d)$ and compute

$$\frac{1}{n_{\text{proj}}} \sum_{j=1}^{n_{\text{proj}}} D_{\text{H},1}(\langle W_j/||W_j||,\ x\rangle, P_{W_j/||W_j||}), \tag{2.6}$$

refer to e.g. Kalos and Whitlock (2008) for an account of Monte-Carlo integration methods.

Because data depth is a very popular notion in the statistical community but not well known in machine learning, most data depth implementations can be found in R package such as `ddalpha` (Pokotylo et al., 2019), `fda.usc` (Febrero-Bande and de la Fuente, 2012), `depth` (Genest et al., 2019), `mrfdepth` (Segaert et al., 2020), and `depthproc` (Kosiorowski and Zawadzki, 2019). Python implementations of several multivariate data depths can be found in https://github.com/GuillaumeStaermanML. We refer the reader to Rousseeuw and Hubert (2018a) and Mosler and Mozharovskyi (2020) for further details on data depth computation.

Algorithm 2.1 Approximation of the halfspace depth.

input: $\mathcal{S}_n$,n_{proj}.

Construct $\mathbf{U} \in \mathbb{R}^{d \times n_{\text{proj}}}$ by sampling uniformly n_{proj} vectors $U_1, \ldots, U_{n_{\text{proj}}}$ in $\mathbb{S}^{d-1}$

Compute $\mathbf{M} = \mathbf{XV}$, where $\mathbf{X} = [X_1, \ldots, X_n]^\top \in \mathbb{R}^{n \times d}$

Compute the rank value $\sigma(i,j)$, the rank of index i in $\mathbf{M}_{:,j}$ for every $i \leq n$ and $j \leq n_{\text{proj}}$

Set $\widetilde{D}_{\text{H}}^{\text{MC}}(X_i, \mathcal{S}_n) = \min_{1 \leq j \leq n_{\text{proj}}} \sigma(i,j)$ for every $1 \leq i \leq n$

return $\left\{\widetilde{D}_{\text{H}}^{\text{MC}}(X_i, \mathcal{S}_n),\ \ 1 \leq i \leq n\right\}$

Algorithm 2.2 Approximation of the projection depth.

input: $\mathcal{S}_n$,n_{proj}.

Construct $\mathbf{U} \in \mathbb{R}^{d \times n_{\text{proj}}}$ by sampling uniformly n_{proj} vectors $U_1, \ldots, U_{n_{\text{proj}}}$ in $\mathbb{S}^{d-1}$

Compute $\mathbf{M} = \mathbf{XV}$, where $\mathbf{X} = [X_1, \ldots, X_n]^\top \in \mathbb{R}^{n \times d}$

Find $\mathbf{M}_{\text{med},j}$ the median value of $\mathbf{M}_{:,j}$, $\forall\, j \leq n_{\text{proj}}$

Compute $\text{MAD}_j = \text{median}\{\left|\mathbf{M}_{i,j} - \mathbf{M}_{\text{med},j}\right|,\ i \leq n\}$ for $1 \leq j \leq n_{\text{proj}}$

Compute $\mathbf{V}$ s.t. $\mathbf{V}_{i,j} = \left|\mathbf{M}_{i,j} - \mathbf{M}_{\text{med},j}\right| / \text{MAD}_j$

Set $D_i = \min_{1 \leq j \leq n_{\text{proj}}} 1/(1 + \mathbf{V}_{i,j})$ for every $i \leq n$

return $\left\{\widetilde{D}_{\text{P}}^{\text{MC}}(X_i, \mathcal{S}_n),\ \ 1 \leq i \leq n\right\}$

2.4.3 Robustness

The notion of breakdown point has been first introduced in Hampel (1968) as a global robustness measure. A simple notion of breakdown point, named finite-sample breakdown point, has been introduced in Donoho (1982) and further developped in Donoho and Huber (1983). The finite sample breakdown point (referred to as BP in its abbreviated form) is a more intuitive notion and has been studied at length in the literature. It corresponds to the smallest contamination fraction necessary to break down an estimator. More precisely, for any estimator $T := T(\mathcal{S}_n)$ the (additive) finite sample breakdown point is defined as

$$BP(T, \mathcal{S}_n) = \min \left\{ \frac{o}{n+o} : \sup_{\widetilde{\mathcal{S}}_o} d\left(T(\mathcal{S}_n \cup \widetilde{\mathcal{S}}_o),\ T(\mathcal{S}_n)\right) = +\infty \right\},$$

where d is a metric defined on the space where T takes its values and $\widetilde{\mathcal{S}}_o$ a contaminated sample of size o. A slight modified version, replacing (instead of adding) the standard sample by the contaminated one, named replacement breakdown point can be seen in the literature. The asymptotic breakdown point is given by making n goes to infinity. Data depths are usually nonnegative and bounded. Thus, the BP of data depths have been investigated through the lens of two similar settings: (1) by fixing $T(\cdot) = \log D(x, \cdot)$ and $d(\cdot, \cdot) = |. - .|$; (2) by means of its depth-trimmed regions at level α replacing $T(\cdot)$ by $D^\alpha(\cdot)$ and $d(\cdot, \cdot)$ by the Hausdorff distance. Lopuhaa and Rousseeuw (1991) have shown that the maximum breakdown point of any affine equivariant statistical estimators is $1/2$, that is attained by the median in the univariate case, while the sample mean breaks at $1/n$. It is an important property since the affine invariance property $\mathbf{D}_1$ makes the depth induced median affine equivariant and then provide an upper-bound on the BP of data depths.

The deepest point of the projection depth, relying on univariate medians of projected distributions and being robust by construction, has a breakdown point equal to $1/2$ extending the univariate median robustness. The L_2 and spatial medians have both a breakdown point equal to $1/2$ (Lopuhaa and Rousseeuw, 1991). The asymptotic breakdown point of the halfspace median has been shown to be higher than $1/(d+1)$ (Donoho and Gasko, 1992) being more robust to the simplicial median that has BP asymptotically lower than $1/(d+2)$ (Chen, 1995). It stipulates that halfspace median remains until at least $n/(d+1)$ outliers are added to the data set, meaning that the Tukey median is a robust statistics when d is not too high w.r.t. the sample size n, that is the case for most machine learning data sets. The IRW median has BP higher than $\frac{\lceil n/d \rceil}{n+\lceil n/d \rceil}$ and is therefore at least as robust as the Tukey median. The lens depth has an asymptotic breakdown point equal to $(\sqrt{2}-1)/\sqrt{2} \approx 0.3$ that is independent of the dimension. Several data depths have robust medians but are poorly robust in general such as the halfspace depth, that has BP of its larger depth-trimmed regions $D^\alpha(P)$ equal to $\frac{\lceil \alpha/(1-\alpha) n \rceil}{n+\lceil \alpha/(1-\alpha) n \rceil}$ (see Donoho and Gasko, 1992 and Nagy and Dvořák, 2021 for further details); and the L_p/simplicial depths that have both global BP equal to $1/n$. In contrast, some data depths should not be used when data are contaminated such as the zonoid depth, being maximized by the expectation $\mathbb{E}[X]$, and the Oja depth having both their asymptotic breakdown points equal to zero (Niinimaa et al., 1990). The Mahalanobis breakdown point relies on the BP of the mean and covariance estimates and suitable choices to ensure robustness can be made such as the Minimum Volume

Ellipsoid (Lopuhaa and Rousseeuw, 1991).

The local robustness of an estimator can be explored by studying its influence function, a robustness measure, based on the Gateaux derivative in the direction of a Dirac distribution at a point $y \in \mathbb{R}^d$, that has been introduced in the seminal paper Hampel (1971). The influence functions of data depths have been investigated in Romanazzi (2001) and Chen and Tyler (2002) for the halfspace median, in Niinimaa and Oja (1995) for the spatial/Oja and L_2 medians, in Zuo (2004) for the L_p median, and in Zuo et al. (2004b) for the projection median.

3
Functional Data Depth

Contents

The idea of extending the concept of multivariate data depth to the case of functional data first appears in Liu and Singh (1997) in order to define limiting p-values for hypothesis testing on infinite-dimensional parameters. Thereafter, the first proposal in a functional framework have been considered in Fraiman and Muniz (2001), where it is proposed to define functional depths as simple integrals, over the parametrization of functions (e.g. time), of a univariate depth function. Further, to contrast with integrated depths, Mosler and Polyakova (2012) have introduced the class of infimal depths replacing the integral by an infimum over its parametrization in order to identify outlying functions in a small part of their domain. Recently, alternative functional depths have been introduced, see López-Pintado and Romo (2009, 2011) for depths based on the geometry of the set of curves, Chakraborty and Chaudhuri (2014b) for a notion of depth based on the L_2 distance, Helander et al. (2020) for a depth based on a multivariate Pareto distribution, Dutta et al. (2011) and Sguera et al. (2014) for a functional version of the Tukey depth or Nagy et al. (2017) and Harris et al. (2021) for depths dedicated to detect outlying shapes.

In this chapter, we review the concept of data depth devoted to functional data. After having recalled the concept and standard properties in Section 3.1, we exhibit classical examples of functional depth and highlight their particularities in Section 3.2. Eventually, we discuss about recent developments on functional depth focusing on the shape of functions in Section 3.3.

3.1 Definition and Properties

Before defining functional statistical depth we first introduce some notations. Denote by $\mathcal{F}(\mathcal{T})$ a subset of the space of real-valued functions defined on a compact set $\mathcal{T} \subset \mathbb{R}$. Assume that $\mathcal{F}(\mathcal{T})$ is a normed vector space with an arbitrary norm $||\cdot||$ with the corresponding metric $d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$, for $\mathbf{x}, \mathbf{y} \in \mathcal{F}(\mathcal{T})$. Throughout this dissertation, $\mathcal{F}(\mathcal{T})$ is often denoted by $\mathcal{F}$ for simplicity. We consider the situation where the r.v.

$\mathbf{X}$, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, takes its values in the infinite dimensional space $\mathcal{F}$. By $\mathcal{P}(\mathcal{F})$ is meant the space of Borel probability measures on $\mathcal{F}$. The random variable $\mathbf{X}$ following $\mathbf{P} \in \mathcal{P}(\mathcal{F})$ is then defined as follows:

$$\begin{array}{rccc} \mathbf{X} & : & (\Omega, \mathcal{A}, \mathbb{P}) & \longrightarrow & \mathcal{F}(\mathcal{T}) \\ & & \omega & \longmapsto & \mathbf{X}(\omega) = \{\mathbf{X}(\omega, t),\ t \in \mathcal{T}\}. \end{array}$$

The two common choices for $\mathcal{F}$, depending on depth functions considered, are (i) $\mathcal{C}(\mathcal{T})$, the space of continuous functions defined on $\mathcal{T}$ equipped with the infinity norm $||\mathbf{x}||_\infty = \sup_{t \in \mathcal{T}} |\mathbf{x}(t)|$ and (ii) $L_2(\mathcal{T}, \lambda)$, the space of square integrable functions w.r.t. the Lebesgue measure λ equipped with the norm $||\mathbf{x}||_{L_2} = \sqrt{\int_{\mathcal{T}} \mathbf{x}(t)^2 \ \lambda(\mathrm{d}t)}$.

Now, we are ready to give the general definition of functional depth. Formally, a statistical functional depth is defined as follows:

$$\begin{array}{rccc} FD: & \mathcal{F} \times \mathcal{P}(\mathcal{F}) & \longrightarrow & [0,1]\,, \\ & (\mathbf{x}, \mathbf{P}) & \longmapsto & FD(\mathbf{x}, \mathbf{P}). \end{array} \tag{3.1}$$

Properties a notion of functional depth should satisfy have been discussed, among others, in López-Pintado and Romo (2009); Dutta et al. (2011); Mosler and Polyakova (2012); Chakraborty and Chaudhuri (2014a) and have recently been unified in Nieto-Reyes and Battey (2016) and Gijbels and Nagy (2017). The axiomatic framework recalled in Chapter 2 is no longer adapted to the richness of the topological structure of functional spaces. Indeed, the vast majority of the functional depths documented in the literature do not fulfill versions of the most natural and elementary properties required for a depth function in a multivariate setup, cf. Gijbels and Nagy (2017). However, there is still no consensus about the set of desirable properties that a functional depth should satisfy, beyond the form of sensitivity mentioned above. Those that appear to be the most relevant in our opinion are listed below.

($\mathbf{FD}_0$) (Non-degeneracy) For all non-atomic distribution $\mathbf{P}$ in $\mathcal{P}(\mathcal{F})$, we have:

$$\inf_{\mathbf{x} \in \mathcal{F}} FD(\mathbf{x}, \mathbf{P}) < \sup_{\mathbf{x} \in \mathcal{F}} FD(\mathbf{x}, \mathbf{P}).$$

($\mathbf{FD}_{1\mathbf{F}}$) (Function-affine invariance) Let $\mathbf{X}$ be a r.v. following $\mathbf{P}_\mathbf{X} \in \mathcal{P}(\mathcal{F})$. The depth FD is said to be (function-) affine invariant if for any $\mathbf{x}$ in $\mathcal{F}$ and all $\mathbf{a}, \mathbf{b}$ in $\mathcal{F}$, we have:

$$FD(\mathbf{x}, \mathbf{P}_\mathbf{X}) = FD(\mathbf{a}\mathbf{x} + \mathbf{b}, \mathbf{P}_{\mathbf{a}\mathbf{X}+\mathbf{b}}).$$

where by $\mathbf{a}\mathbf{x}$ is meant the pointwise product.

. . . . . .

($\mathbf{FD}_{1\mathbf{S}}$) (Scalar-affine invariance) Let $\mathbf{X}$ be a r.v. following $\mathbf{P}_\mathbf{X} \in \mathcal{P}(\mathcal{F})$. The depth FD is said to be (scalar-) affine invariant if for any $\mathbf{x}$ in $\mathcal{F}$ and all a, b in $\mathbb{R}$, we have:

$$FD(\mathbf{x}, \mathbf{P}_\mathbf{X}) = FD(a\mathbf{x} + b, \mathbf{P}_{a\mathbf{X}+b}).$$

($\mathbf{FD}_2$) (Maximality at the center) For any symmetric and non-atomic distribution $\mathbf{P} \in \mathcal{P}(\mathcal{F})$ with $\boldsymbol{\theta} \in \mathcal{F}$ as center of symmetry, we have:

$$FD(\boldsymbol{\theta}, \mathbf{P}) = \sup_{\mathbf{x} \in \mathcal{F}} FD(\mathbf{x}, \mathbf{P}).$$

($\mathbf{FD}_{3\mathbf{SD}}$) (Strictly decreasing w.r.t. the deepest point) For any $\mathbf{P}$ in $\mathcal{P}(\mathcal{F})$ such that $FD(\mathbf{z}, \mathbf{P}) = \sup_{\mathbf{x} \in \mathcal{F}} FD(\mathbf{x}, \mathbf{P})$, $FD(\mathbf{x}, \mathbf{P}) < FD(\mathbf{y}, \mathbf{P}) < FD(\mathbf{z}, \mathbf{P})$ holds for any $\mathbf{x}, \mathbf{y} \in \mathcal{F}$ such that $\min\{d(\mathbf{y}, \mathbf{z}), d(\mathbf{y}, \mathbf{x})\} > 0$ and $\max\{d(\mathbf{y}, \mathbf{z}), d(\mathbf{y}, \mathbf{x})\} < d(\mathbf{x}, \mathbf{z})$.

. . . . . .

($\mathbf{FD}_{3\mathbf{D}}$) (Decreasing w.r.t. the deepest point) For any $\mathbf{P}$ in $\mathcal{P}(\mathcal{F})$ such that $FD(\mathbf{z}, \mathbf{P}) = \sup_{\mathbf{x} \in \mathcal{F}} FD(\mathbf{x}, \mathbf{P})$, we have $FD(\mathbf{z}, \mathbf{P}) > \inf_{\mathbf{x} \in \mathcal{F}} FD(\mathbf{x}, \mathbf{P})$ and $FD(\mathbf{y}, \mathbf{P}) \leq FD(\mathbf{z} + \gamma(\mathbf{y} - \mathbf{z}), \mathbf{P})$ holds for all $\mathbf{y} \in \mathcal{F}$ and $\gamma \in [0, 1]$.

($\mathbf{FD}_4$) (Vanishing at ∞) For any non-atomic distribution $\mathbf{P}$ in $\mathcal{P}(\mathcal{F})$,

$$FD(\mathbf{z}, \mathbf{P}) \underset{||\mathbf{z}||_{\mathcal{F}} \longrightarrow \infty}{\longrightarrow} \inf_{\mathbf{x} \in \mathcal{F}} FD(\mathbf{x}, \mathbf{P}),$$

where $||\cdot||_{\mathcal{F}}$ is the norm associated to $\mathcal{F}$.

($\mathbf{FD}_5$) (Upper semi-continuity in $\mathbf{x}$) For any non-atomic distribution $\mathbf{P} \in \mathcal{P}(\mathcal{F})$, the function $\mathbf{x} \mapsto FD(\mathbf{x}, \mathbf{P})$ is upper semi-continuous w.r.t. the norm associated to $\mathcal{F}$.

($\mathbf{FD}_6$) (Continuity in $\mathbf{P}$) For all $\mathbf{x}$ in $\mathcal{F}$, the mapping $\mathbf{P} \in \mathcal{P}(\mathcal{F}) \mapsto FD(\mathbf{x}, \mathbf{P})$ is continuous w.r.t. the Lévy-Prohorov metric.

. . . . . .

($\mathbf{FD}_{6\mathbf{UC}}$) (Uniform continuity in $\mathbf{P}$ over compact sets) For all $\varepsilon > 0$, there exists $\delta > 0$ such that for any $\mathbf{P}, \mathbf{Q} \in \mathcal{P}(\mathcal{F})$ s.t. $d_{\mathrm{LP}}(\mathbf{P}, \mathbf{Q}) < \delta$, it holds for any compact sets $\mathcal{K} \subset \mathcal{F}$, $\sup_{\mathbf{x} \in \mathcal{K}} |FD(\mathbf{x}, \mathbf{P}) - FD(\mathbf{x}, \mathbf{Q})| < \varepsilon$ where $d_{\mathrm{LP}}(\mathbf{P}, \mathbf{Q})$ is the Lévy-Prohorov metric.

. . . . . .

($\mathbf{FD}_{6\mathbf{U}}$) (Uniform continuity in $\mathbf{P}$) For all $\varepsilon > 0$, there exists $\delta > 0$ such that for any $\mathbf{P}, \mathbf{Q} \in \mathcal{P}(\mathcal{F})$ s.t. $d_{\mathrm{LP}}(\mathbf{P}, \mathbf{Q}) < \delta$, it holds $\sup_{\mathbf{x} \in \mathcal{F}} |FD(\mathbf{x}, \mathbf{P}) - FD(\mathbf{x}, \mathbf{Q})| < \varepsilon$ where $d_{\mathrm{LP}}(\mathbf{P}, \mathbf{Q})$ is the Lévy-Prohorov metric.

The property ($\mathbf{FD}_0$) is specific to functional spaces as it is trivially satisfied for all multivariate data depths. It has been discussed in Nieto-Reyes and Battey (2016) in the case of Gaussian processes and further developed in Gijbels and Nagy (2017) to general distributions. Though it obviously appears as mandatory to make the other properties meaningful, *non-degeneracy*, is actually not fulfilled by all the functional depths proposed, see e.g. Dutta et al. (2011) and Chakraborty and Chaudhuri (2014a).

Known examples of depth that fail to be relevant for every $\mathbf{P} \in \mathcal{P}(\mathcal{F})$ include the band depth (López-Pintado and Romo, 2009), the halfregion depth (López-Pintado and Romo, 2011) or the infimal depth (Mosler and Polyakova, 2012).

Nieto-Reyes and Battey (2016) have first introduced the property named *distance invariance* relying on invariance to isometric transformations, i.e. transformations $\mathbf{f}$ s.t. the distance $||\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})|| = ||\mathbf{x} - \mathbf{y}||_{\mathcal{F}}$, $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{F}$ where $||\cdot||_{\mathcal{F}}$ is the norm of the functional space $\mathcal{F}$. However, this property is too restrictive in general. Indeed, Gijbels and Nagy (2017) have demonstrated that most of known depth functions fail to satisfy this strong notion of invariance. Therefore, weaker properties ($\mathbf{FD_{1F}}$) and ($\mathbf{FD_{1S}}$) have been discussed by the authors in order to develop meaningful properties for existing functional depth notions. In contrast to multivariate data depth, it is still not clear if a functional depth should fulfill the more general invariance properties. The property ($\mathbf{FD_{1F}}$), which implies scalar-affine invariance that is satisfied by most of functional depths, exhibits many transformations that completely modify the function structure.

The '*maximality at center*' and '*(strictly) decreasing w.r.t. the deepest point*' properties permit to preserve the original center-outward ordering goal as well as the goodness of fit for unimodal data of *data depth* in the functional framework. As described in the previous chapter, there are several notions of symmetry in the multivariate space that are not easily adaptable to functional spaces which led the authors of Nieto-Reyes and Battey (2016) to consider symmetry around the absciss axis of zero mean Gaussian processes. Further, Gijbels and Nagy (2017) have provided extensions of *central symmetry* and *halfspace symmetry* to functional data based on a characterization of functional symmetries by means of projections through continuous linear mappings $\varphi : \mathcal{F} \to \mathbb{R}$ that belong to the dual space $\mathcal{F}^*$. Relying on the previous work of Nagy et al. (2016b), authors argue that any random variable $\mathbf{X}$ following $\mathbf{P} \in \mathcal{P}(\mathcal{F})$ is symmetric around $\boldsymbol{\theta}$ if and only if for any $\varphi \in \mathcal{F}^*$ the distribution of $\varphi(\mathbf{X})$ is symmetric around $\varphi(\boldsymbol{\theta})$. Simple examples can be considered in $\mathcal{C}(\mathcal{T})$ or $L_2(\mathcal{T})$ spaces. Recall that L_2 is a Hilbert space equipped with the inner product $\langle \mathbf{x}, \mathbf{y} \rangle_{L_2} = \int_{\mathcal{T}} \mathbf{x}(t)\mathbf{y}(t)\mathrm{d}t$. Thus, the r.v. $\mathbf{X}$ following $\mathbf{P} \in \mathcal{P}(L_2)$ is symmetric around $\boldsymbol{\theta}$ if and only if $(\langle \mathbf{u}_1, \mathbf{X} \rangle, \ldots, \langle \mathbf{u}_p, \mathbf{X} \rangle)$ is symmetric around $(\langle \mathbf{u}_1, \boldsymbol{\theta} \rangle, \ldots, \langle \mathbf{u}_p, \boldsymbol{\theta} \rangle)$ for all $\mathbf{u}_1, \ldots, \mathbf{u}_p \in L_2$ and $p \in \mathbb{N}_*$. In particular, when $\mathcal{F} = \mathcal{C}$, the functional symmetry implies that the distribution of finite evaluation of random functions, i.e. $(\mathbf{X}(t_1), \ldots, \mathbf{X}(t_p))$ is symmetric around $(\boldsymbol{\theta}(t_1), \ldots, \boldsymbol{\theta}(t_p))$ for all $t_1, \ldots, t_p \in \mathcal{T}$ and $p \in \mathbb{N}_*$. In the following, we denote by ($\mathbf{FD_{2C}}$) and ($\mathbf{FD_{2H}}$) central and halfspace symmetry respectively. It is worth mentioning that ($\mathbf{FD_{2H}}$) $\Rightarrow$ ($\mathbf{FD_{2C}}$).

The property ($\mathbf{FD_{3D}}$) is the direct extension of ($\mathbf{D}_3$) to the functional case. It imposes the same behavior on the functional depth as in the multivariate case, i.e. star-shaped and connected depth-trimmed regions. However, the alternative and stronger property ($\mathbf{FD_{3SD}}$) relies on the topological properties of the considered metric space $(\mathcal{F}, d)$. Choosing the infinity norm as metric d, ($\mathbf{FD_{3SD}}$) ensures that the functional depth assigns lower values to elements of $\mathcal{F}$ that successively belong to balls centered in $\mathbf{z} = \underset{\mathbf{x} \in \mathcal{F}}{\text{argsup}}\ FD(\mathbf{x}, \mathbf{P})$ with growing radius. It can be shown that $\mathbf{z} = \underset{\mathbf{x} \in \mathcal{F}}{\text{argmax}}\ FD(\mathbf{x}, \mathbf{P})$ and ($\mathbf{FD}_0$) are consequences of satisfying ($\mathbf{FD_{3SD}}$) (Nieto-Reyes and Battey, 2016). Imposing a decay on the depth function relying on the considered topological space regardless of the probability distribution $\mathbf{P}$ is the main drawback of ($\mathbf{FD_{3SD}}$). As pointed out in Gijbels and Nagy (2017), the equivalent of this property in the multivariate space $\mathbb{R}^d$ is also more restrictive and stronger than ($\mathbf{D}_3$).

The '*semi-continuity in* $\mathbf{x}$' and '*decreasing w.r.t. the deepest point*' properties extend properties fulfilled by the cumulative distribution functions of univariate continuous distributions in the same way as for multivariate data depth.

From a statistical perspective, the '*continuity in* $\mathbf{P}$' property is essential, insofar as $\mathbf{P}$ must be replaced in practice by an estimator, built from finite-dimensional observations, i.e. a finite number of sampled curves. As in the multivariate case, it is more convenient to have the whole depth set $\{FD(\mathbf{x}, \mathbf{P}),\ \mathbf{x} \in \mathcal{F}\}$ well approximated from an inference angle. However, uniform continuity in $\mathbf{P}$ is very demanding in functional spaces. In the multivariate case, absolutely continuity w.r.t. Lebesgue measure of distributions is often required in order to have continuity of $x \in \mathbb{R}^d \mapsto D(x, P), \quad P \in \mathcal{P}(\mathbb{R}^d)$ allowing the uniformity of the consistency of the multivariate depth function. This is, of course, not easily expandable to functional spaces. Although obtaining uniform convergence over $\mathcal{F}$ is generally difficult on functional spaces, the uniformity can be limited to compact sets of $\mathcal{F}$. Indeed, local uniform continuity is satisfied by many functional depths under continuity assumptions on the distribution (see Section 3.2).

Remark 3.1 (Partially observed functional data)**.** *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *be a sample of i.i.d. random functions from* $\mathbf{P} \in \mathcal{P}(\mathcal{F})$ *and* $\mathbf{P}_n = (1/n)\sum_{i=1}^n \delta_{\mathbf{X}_i}$ *its associated empirical measure. In practice, functional data are discretely observed on a vector* $(t_1, \ldots, t_p) \in \mathcal{T}$ *for* $p \in \mathbb{N}_*$ *leading to the following observation set* $\{(\mathbf{X}_1(t_j), \ldots, \mathbf{X}_n(t_j)) : 1 \leq j \leq p\}$*. These observations are often referred to as time-series when the set of parametrization* $\mathcal{T}$ *is the time. The first step in functional data analysis (see e.g. Ferraty and Vieu (2006) and Ramsay and Silverman (2002)) is to reconstruct functions from their partial observations either by interpolation of by means of projection in smooth bases such as splines. The statistician then has access to a reconstructed sample* $\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_n$ *rather than* $\mathbf{X}_1, \ldots, \mathbf{X}_n$*. Given the reconstructed empirical measure* $\widetilde{\mathbf{P}}_n = (1/n)\sum_{i=1}^n \delta_{\widetilde{\mathbf{X}}_i}$*, the main challenge to be adressed is to answer the following question: is* $\widetilde{\mathbf{P}}_n$ *is an accurate approximation of* $\mathbf{P}$*? The asymptotic behavior of* $\widetilde{\mathbf{P}}_n$ *has been investigated by Nagy et al. (2016a) when* $\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_n$ *are obtained through linear interpolation. Authors have shown the convergence of* $\widetilde{\mathbf{P}}_n$ *to* $\mathbf{P}$ *as* n *and* p *go to infinity. Thus, the functional depth computed in practice asymptotically recovers the true depth function.*

Remark 3.2 (Multivariate functional data)**.** *Functional data can also be found in the form of multivariate functional data. In this case, the r.v.* $\mathbf{X}$ *is a multivariate stochastic process such that for each* $t \in \mathcal{T}$*,* $\mathbf{X}(t)$ *takes its values in* $\mathbb{R}^d$*. Many approaches have been introduced recently to deal with multivariate functional data such as the simplicial band depth (López-Pintado et al., 2014), the multivariate functional halfspace depth (Claeskens et al., 2014), the multivariate functional projection depth (Hubert et al., 2015) or the directional outlyingness (Dai and Genton, 2019b).*

3.2 Existing Notions of Functional Data Depths

In contrast to multivariate data depth, there is no well-established nomenclature that brings together functional depth approaches. In this part, we introduce a variety of (univariate) functional depths based on differing concepts and discuss their properties. We propose our own nomenclature which differs slightly from that of Nagy (2016), Chapter 2. More precisely, functional depths can be categorized into four classes:

- **Integrated depths:** univariate depth functions computed on one-dimensional projections are averaged, see Section 3.2.1.
- **Infimal depths:** the depth is chosen as the infimum of univariate depth functions computed on one-dimensional projections, see Section 3.2.2.
- **Geometry-based depths:** the graph of $\mathbf{x}$ is compared to a geometric shape induced by the graph of the r.v. $\mathbf{X}$, see Section 3.2.3.
- **Distance-based depths:** the depth of $\mathbf{x}$ decreases as its distance from $\mathbf{X}$ decreases w.r.t. a functional norm, see Section 3.2.4.

For the sake of completeness, structural properties of functional depths presented in this section can be summarized in the Table 3.1 while a similar table can be found in Gijbels and Nagy (2017). It is worth mentioning that most of the presented functional depths in this section can be easily extended to the case of multivariate functional data. As an example, univariate depth functions can be replaced by multivariate ones for integrated/infimal depths (see Section 3.2.1 and Section 3.2.2). However, it is not the purpose of this manuscript to present multivariate functional depth and we limit our review to the univariate functional case.

3.2.1 The Family of Integrated Depth

Integrated depths are based on low-dimensional projections and is an important family of functional depths. They are constructed as the mean of all univariate depth values computed over one-dimensional projections. We present here the main approaches.

- **Integrated depths.** The class of integrated depths, based on a integral over the space $\mathcal{T}$ of univariate depths of a curve observed at $t \in \mathcal{T}$, has been first introduced in Fraiman and Muniz (2001). This depth captures the overall nature of a function $\mathbf{x}$ by its point-wise averaged behavior making it very easy to compute in practice. The most popular depths that belong to this family are probably the functional halfspace depth (Claeskens et al., 2014) and the functional projection depth (also called functional Stahel-Donoho outlyingness) (Hubert et al., 2015).

  Definition 3.3. *Let $\mathbf{x} \in \mathcal{F}$ and $\mathbf{X}$ be a r.v. following $\mathbf{P} \in \mathcal{P}(\mathcal{F})$. Let $D(.,.) : \mathbb{R} \times \mathcal{P}(\mathbb{R}) \to [0,1]$ be an arbitrary univariate depth function. The integrated functional depth of $\mathbf{x}$ w.r.t. $\mathbf{P}$ is defined as:*

$$IFD(\mathbf{x}, \mathbf{P}) = \int_{\mathcal{T}} D(\mathbf{x}(t), \mathbf{P}_t)\ \mu(\mathrm{d}t), \tag{3.2}$$

  where μ is an arbitrary measure on $\mathcal{T}$ and $\mathbf{P}_t \in \mathcal{P}(\mathbb{R})$ is the marginal probability distribution of $\mathbf{X}(t)$.

  The most popular depths that belong to this family are probably the functional halfspace depth IFD_{H} (Claeskens et al., 2014) and the functional projection depth (also called functional Stahel-Donoho outlyingness) IFD_{P} (Hubert et al., 2015).

  This approach appears as a natural extension to univariate depth into functional data. However, the main drawback of this approach is its smoothing effect induced by the integral that cannot detect any atypical behavior on a small portion of $\mathcal{T}$ (Mosler and Polyakova, 2012).

Properties of *IFD* stem from features shared by the considered univariate depth. This family of functional depth has been extensively studied in Nagy et al. (2016b) at the lens of measurability, structural properties and consistency when $\mathcal{F} = \mathcal{C}(\mathcal{T})$. Authors first investigated the measurability of *IFD* which requires the measurability of the function $t \in \mathcal{T} \mapsto D(\mathbf{x}(t), \mathbf{P}_t)$ that follows for most of classical depth functions. However, authors shed light on a measurability issue. Indeed, when studying the uniform convergence of the sample version of *IFD*, the first step often relies on the following inequality:

$$\sup_{\mathbf{x}\in\mathcal{C}} \int_{t\in\mathcal{T}} \Big| D(\mathbf{x}(t), \mathbf{P}_{n,t}) - D(\mathbf{x}(t), \mathbf{P}_t) \Big| \mathrm{d}t \leq \int_{t\in\mathcal{T}} \sup_{\mathbf{x}\in\mathcal{C}} \Big| D(\mathbf{x}(t), \mathbf{P}_{n,t}) - D(\mathbf{x}(t), \mathbf{P}_t) \Big| \mathrm{d}t,$$

where $\mathbf{P}_{n,t}$ the empirical distribution of the marginal $\mathbf{P}_t$. To get the measurability of the term inside the integral on the right side of the inequality, we need a stronger notion of measurability on $D(.,.)$ as pointed out in Nagy et al. (2016b). More precisely, the function

$$\begin{array}{rccc} D: & \mathbb{R} \times \mathcal{P}(\mathbb{R}) & \longrightarrow & [0,1]\,, \\ & (x, P_1) & \longmapsto & D(x, P_1), \end{array}$$

needs to be jointly Borel measurable and $D(., P_1) \not\equiv 0$ for all $P_1 \in \mathcal{P}(\mathbb{R})$. This condition is satisfied by most of univariate depth functions such as the halfspace or simplicial depth. Structural properties of *IFD* relying on properties of D have been established in Nagy et al. (2016b). *IFD* satisfies $(\mathbf{FD}_{1\mathbf{S}})$ if D is invariant to rescaling and translations; $(\mathbf{FD}_2)$ for halfspace symmetric distributions if D satisfies $(\mathbf{D}_2, \mathbf{D}_4, \mathbf{D}_5)$; and $(\mathbf{FD}_{6\mathbf{U}})$ if D satisfies uniform continuity for $\mathbf{P}_t$ for almost all $t \in \mathcal{T}$. Furthermore, $(\mathbf{FD}_{3\mathbf{D}})$ holds if D satisfies $(\mathbf{D}_3)$; $(\mathbf{FD}_4)$ holds if D satisfies $(\mathbf{D}_4)$ and additional mild assumptions; and $(\mathbf{FD}_5)$ holds if D satisfies $(\mathbf{D}_5)$.

- **Functional integral dual depth.** The functional integral dual depth (Cuevas and Fraiman, 2009) can be considered as a generalization of the integrated depth. Indeed, it is defined as the value of any univariate depth (originally defined with the simplicial depth) integrated over all dual projections $\varphi : \mathcal{F} \to \mathbb{R}$.

  Definition 3.4. *Let $\mathbf{x} \in \mathcal{F}$ and $\mathbf{X}$ be a r.v. following $\mathbf{P} \in \mathcal{P}(\mathcal{F})$. Let $D(\cdot,\cdot) : \mathbb{R} \times \mathcal{P}(\mathbb{R}) \to [0,1]$ be a univariate depth function and $\boldsymbol{\mu}$ be a probability measure on $\mathcal{F}^*$, the dual space of $\mathcal{F}$. The functional integral dual depth of $\mathbf{x}$ w.r.t. $\mathbf{P}$ is defined as:*

$$FIFD(\mathbf{x}, \mathbf{P}) = \int_{\varphi\in\mathcal{F}^*} D(\varphi(\mathbf{x}), \mathbf{P}_\varphi)\ \mathrm{d}\boldsymbol{\mu}(\varphi),$$

  where $\mathbf{P}_\varphi \in \mathcal{P}(\mathbb{R})$ is the marginal probability distribution of $\varphi(\mathbf{X})$.

  The integrated depth function can be considered as a particular case of *FIFD*. Indeed, when the measure $\boldsymbol{\mu}$ is restricted to the set of Dirac measures $\{\delta_t \in \mathcal{F}^*,\ t \in \mathcal{T}\}$, the functional integrated dual depth boils down to the depth function (3.2). Surprisingly, theoretical properties of *FIFD* have not been clearly investigated yet through the scope of functional depth axioms. However, necessary conditions

under which continuity in $\mathbf{x}$ and in $\mathbf{P}$ hold are known from Cuevas and Fraiman (2009). Authors also shows that *FIFD* vanishes at infinity. Note that this depth is dedicated to any Banach space $\mathcal{F}$ making the approach valid for both $\mathcal{C}$ and L_2.

3.2.2 The Family of Infimal Depth

Infimal depths are based on low-dimensional projections and is an important family of functional depths. They are constructed as the infimum of all univariate depth values computed over one-dimensional projections. We present here the main approaches.

- **Random Tukey depth.** Cuesta-Albertos and Nieto-Reyes (2008a) have introduced the random Tukey depth in order to generalize the halfspace depth (approximated with a finite number of unit sphere vectors) to Hilbert functional spaces. Let $\mathcal{F} = L_2$ equipped with the inner product $\langle \cdot, \cdot \rangle_{L_2}$, the random Tukey depth aims at computing the minimum value of the univariate halfspace depth over a finite number of projections.

  Definition 3.5. *Let $\mathbf{x} \in L_2$ and $\mathbf{X}$ be a r.v. following $\mathbf{P} \in \mathcal{P}(L_2)$. Let $\mathcal{U} = \{\mathbf{U}_1, \dots, \mathbf{U}_m\}$ the set of m i.i.d. realizations drawn from a probability measure $\boldsymbol{\mu} \in \mathcal{P}(L_2)$. The random Tukey depth of $\mathbf{x}$ w.r.t. $\mathbf{P}$ is defined as:*

$$FD_{\mathrm{RT}}(\mathbf{x}, \mathbf{P}) = \min_{\mathbf{u} \in \mathcal{U}} D_{\mathrm{H},1}\left(\langle \mathbf{u}, \mathbf{x} \rangle_{L_2},\ \mathbf{P}_{\mathbf{u}}\right),$$

  where the probability measure $\boldsymbol{\mu}$ is taken as a non-degenerate stationary Gaussian measure on L_2 and by $\mathbf{P}_{\mathbf{u}}$ is meant $\mathbf{P}_{\langle \mathbf{u}, \mathbf{X} \rangle_{L_2}}$.

  Its properties have been studied in both Nieto-Reyes and Battey (2016) and Gijbels and Nagy (2017). It satisfies $(\mathbf{FD}_0, \mathbf{FD}_{1\mathbf{S}}, \mathbf{FD}_{3\mathbf{D}}, \mathbf{FD}_5)$ but violates $(\mathbf{FD}_2)$ and $(\mathbf{FD}_4)$. When the joint distribution $(\mathbf{P}_{\langle \mathbf{u}_1, \mathbf{X} \rangle_{L_2}}, \dots, \mathbf{P}_{\langle \mathbf{u}_m, \mathbf{X} \rangle_{L_2}})$ is absolutely continuous w.r.t. the Lebesgue measure or for empirical measures, the random Tukey depth is uniformly continuous in $\mathbf{P}$ $(\mathbf{FD}_{6\mathbf{U}})$. We refer to Cuesta-Albertos and Nieto-Reyes (2008a) for further discussions on FD_{RT} including the choice of m.

  The extension of the random Tukey depth defined above, as "continuous" or "infinite" version of the latter replacing the minimum over a finite number of projections with an infimum over all linear projections $\mathbf{x} \mapsto \langle \mathbf{u}, \mathbf{x} \rangle$, has been studied in Dutta et al. (2011); Kuelbs and Zinn (2013); Chakraborty and Chaudhuri (2014a). However, it turns out that taking an infinite number of projections leads to a degenerate behaviour for some commonly used probability distributions in infinite dimensional spaces of functions (Dutta et al., 2011; Chakraborty and Chaudhuri, 2014a) and to the lack of consistency (Kuelbs and Zinn, 2013).

- **Φ-depth.** Sharing the idea of Cuevas and Fraiman (2009), Mosler and Polyakova (2012) have introduced the general notion of infimal depth (=Φ-depth) replacing the integral in (3.2) by an infimum over all linear $\mathbb{R}$-valued functions φ generalizing the two previously introduced functional depths. Each function $\varphi : \mathcal{F} \to \mathbb{R}$ might be interpreted as a particular view of $\mathbf{x}$. With the same worst case aspect as in the multivariate halfspace depth, if there is one "direction" or rather one "characteristic" where $\mathbf{x}$ is far from the center of $\mathbf{P} \in \mathcal{P}(\mathcal{F})$, then $\mathbf{x}$ is considered as an outlier. Precisely, this approach is defined as follows.

Definition 3.6. *Let* $\mathbf{x} \in \mathcal{F}$ *and* $\mathbf{X}$ *be a r.v. following* $\mathbf{P} \in \mathcal{P}(\mathcal{F})$*. Let* $D(.,.) : \mathbb{R} \times \mathcal{P}(\mathbb{R}) \to [0,1]$ *be a univariate depth function. The infimal depth of* $\mathbf{x}$ *w.r.t.* $\mathbf{P}$ *is defined as:*

$$IFFD(\mathbf{x}, \mathbf{P}) = \inf_{\varphi \in \mathcal{F}^*} D(\varphi(\mathbf{x}), \mathbf{P}_\varphi),$$

where $\mathcal{F}^*$ *is the dual space of* $\mathcal{F}$ *and* $\mathbf{P}_\varphi \in \mathcal{P}(\mathbb{R})$ *is the marginal probability distribution of* $\varphi(\mathbf{X})$*.*

Properties of the infimal depth have been investigated in Mosler and Polyakova (2012) by establishing that if D is scale/translation invariant and satisfies properties $(\mathbf{D}_2, \mathbf{D}_3, \mathbf{D}_5)$ then *IFFD* satisfies $(\mathbf{FD}_{1\mathbf{S}}, \mathbf{FD}_{3\mathbf{D}}, \mathbf{FD}_5)$. Furthermore, if for any sequence $(\mathbf{z}_n)_{n\geq 1} \in \mathcal{F}$ such that $||\mathbf{z}_n|| \underset{n\to\infty}{\to} \infty$ there exists a sequence $(\varphi_n)_{n\geq 1}$ such that $||\varphi_n(\mathbf{z}_n)|| \underset{n\to\infty}{\to} \infty$ then *IFFD* satisfies $\mathbf{FD}_4$. Note that this depth is dedicated to any Banach space $\mathcal{F}$ making the approach valid for both $\mathcal{C}$ and L_2.

Remark 3.7. *Despite its appealing properties, the general form of the infimal depth makes him hard to compute in practice. Thus, this depth is often used in practice limiting* $\mathcal{F}^*$ *to a particular subspace of* $\mathcal{F}^*$*. A simple example, named graph depth, is the projection of* $\mathbf{x}$ *to its one-dimensional marginal s.t.* $\varphi_t(\mathbf{x}) = \mathbf{x}_t$ *for any* $t \in \mathcal{T}$*. The graph depth is then defined as:*

$$GFD(\mathbf{x}, \mathbf{P}) = \inf_{t \in \mathcal{T}} D(\mathbf{x}(t), \mathbf{P}_t).$$

The infimum differentiates the graph depth from the integrated depth by its ability to identify functions that have local deviations from the normal behavior. However, any $\mathbf{x} \in \mathcal{F}$ *that has one of its marginal outside of the interval of data, i.e. if there exists* $t \in \mathcal{T}$ *such that*

$$\mathbf{x}(t) \notin \left[\min_{1\leq i\leq n} \mathbf{X}_i(t),\ \max_{1\leq i\leq n} \mathbf{X}_i(t) \right]$$

will have a depth assigned to zero. This feature makes the ordering induced by the graph depth poor since almost all functions in $\mathcal{F}$ *will have zero value and considerably reduces the potential applications of this depth in practice. The graph depth satisfies* $(\mathbf{FD}_{1\mathbf{F}}, \mathbf{FD}_4, \mathbf{FD}_5, \mathbf{FD}_{6\mathbf{U}})$ *but fails to valid* $(\mathbf{FD}_0, \mathbf{FD}_2, \mathbf{FD}_3)$*.*

3.2.3 Functional Depths based on a Geometric Approach

As geometry in function spaces is abstract, it is often more relevant to consider the graphical representation of the curves. Let $\mathbf{x}$ be a function in $\mathcal{C}(\mathcal{T})$, the graph of $\mathbf{x}$ is defined as:

$$\text{graph}(\mathbf{x}) = \left\{ (t, \mathbf{x}(t)) \in \mathbb{R}^2 :\ t \in \mathcal{T} \right\}.$$

We present here functional depths that rely on a graphical approach.

- **(Modified) band depth.** López-Pintado and Romo (2009) have introduced the band depth dedicated to functions that belong to the separable Banach space $(\mathcal{C}(\mathcal{T}), ||\cdot||_\infty)$. The band depth is the first functional depth based on geometrical notions rather than low-dimensional projections that has been introduced in the literature. It is a very popular depth notion probably due to its simplicity and its ease of interpretation. Given the set of n curves $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in $\mathcal{C}(\mathcal{T})$, the band of this set of functions is defined as:

$$\text{band}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \Big\{(t, \mathbf{x}(t)) \in \mathbb{R}^2 \,:\, t \in \mathcal{T},\ \min_{1\leq i\leq n} \mathbf{x}_i(t) \leq \mathbf{x}(t) \leq \max_{1\leq i\leq n} \mathbf{x}_i(t)\Big\}.$$

The band depth is closely related to the notion of simplex in $\mathbb{R}^d$ and can be seen as an extension of the simplicial depth to the functional case by means of the band of a batch of functions.

Definition 3.8. *Let $\mathbf{x} \in \mathcal{C}$ and $\mathbf{X}$ be a r.v. following $\mathbf{P} \in \mathcal{P}(\mathcal{C})$. The band depth of $\mathbf{x}$ w.r.t. $\mathbf{P}$ is defined as:*

$$FD_{\mathrm{B},J}(\mathbf{x}, \mathbf{P}) = \frac{1}{J-1} \sum_{j=2}^{J} \mathbb{P}\left(\text{graph}(\mathbf{x}) \in \text{band}(\mathbf{X}_1, \ldots, \mathbf{X}_j)\right).$$

The band depth satisfies $(\mathbf{FD}_{1\mathbf{F}}, \mathbf{FD}_4, \mathbf{FD}_5)$ but violates $(\mathbf{FD}_2, \mathbf{FD}_{3\mathbf{D}}, \mathbf{FD}_6)$. However, it tends to degenerate for many distributions (see e.g. Chakraborty and Chaudhuri, 2014a and Gijbels and Nagy, 2017) and then fails to valid $(\mathbf{FD}_0)$. In order to correct the latter drawback, López-Pintado and Romo (2009) have also introduced a modified version of the band depth measuring the expected length of $\mathcal{T}$ where the graph of $\mathbf{x}$ belongs to the band. More precisely, it is defined as follows.

Definition 3.9. *Let $\mathbf{x} \in \mathcal{C}$ and $\mathbf{X}$ be a r.v. following $\mathbf{P} \in \mathcal{P}(\mathcal{C})$. The modified band depth of $\mathbf{x}$ w.r.t. $\mathbf{P}$ is defined as:*

$$FD_{\mathrm{MB},J}(\mathbf{x}, \mathbf{P}) = \frac{1}{J-1} \sum_{j=2}^{J} \frac{1}{\lambda(\mathcal{T})} \mathbb{E}\left[\lambda(\{t \in \mathcal{T} :\ \text{graph}(\mathbf{x}) \in \text{band}(\mathbf{X}_1, \ldots, \mathbf{X}_j)\})\right].$$

The modified band depth with $J = 2$ can be rewritten as the integrated depth (3.2) associated to the simplicial depth D_{S}. Although the modified band depth does not vanish at infinity, it has many additional properties compared to the band depth such as $(\mathbf{FD}_0, \mathbf{FD}_{1\mathbf{F}}, \mathbf{FD}_{3\mathbf{D}})$ and $(\mathbf{FD}_{6\mathbf{U}})$ for distributions $\mathbf{P}$ such that $\mathbf{P}_t$ has no atoms for each $t \in \mathcal{T}$ or for empirical measures. In addition, $(\mathbf{FD}_2)$ is valid for both halfspace and central symmetric distributions and is semi-continuous in $\mathbf{X}$ $(\mathbf{FD}_5)$. Further, many extensions of the (modified) band depth have been introduced such as the *corrected band depth* (López-Pintado and Jornsten, 2007), the *sparse band depth* (López-Pintado and Wei, 2011), the *local band depth* (Agostinelli and Romanazzi, 2013). Alternatives for the case of multivariate functional data have been introduced, among others, in Ieva and Paganoni (2013); López-Pintado et al. (2014) and Mirzargar et al. (2014).

- **(Modified) halfregion depth.** Sharing similar construction with the band depth, the halfregion depth was later introduced by López-Pintado and Romo (2011). As the band depth, the halfregion depth is defined for functions that belong to $\mathcal{C}(\mathcal{T})$.

  Definition 3.10. *Let* $\mathbf{x} \in \mathcal{C}$ *and* $\mathbf{X}$ *be a r.v. following* $\mathbf{P} \in \mathcal{P}(\mathcal{C})$. *The halfregion depth of* $\mathbf{x}$ *w.r.t.* $\mathbf{P}$ *is defined as:*

$$FD_{\mathrm{HR}}(\mathbf{x}, \mathbf{P}) = \min\left\{\mathbb{P}\left(\mathbf{x} \leq \mathbf{X}\right), \mathbb{P}\left(\mathbf{x} \geq \mathbf{X}\right)\right\},$$

  where by $\mathbf{x} \leq \mathbf{X}$ *is meant* $\mathbf{x}(t) \leq \mathbf{X}(t)$, $\forall\ t \in \mathcal{T}$.

  Structural properties of the halfregion depth have been investigated in the original paper López-Pintado and Romo (2009) and more recently in Gijbels and Nagy (2017) showing that it shares properties identical to those of the band depth. In addition, Kuelbs and Zinn (2015) have described several examples where the halfregion depth assigns zero to all sample functions such as Brownian motion or symmetric stable processes. They also provide a way to smooth FD_{HR} in order to derive strong consistency of the halfregion process. Further extensions such as the modified halfregion depth (López-Pintado and Romo, 2011) or a local version of the halfregion depth (Agostinelli, 2018) have been introduced. The modified version has been derived in the same way that the band depth taking the expectation of the Lebesgue measure instead of the probability.

  Definition 3.11. *Let* $\mathbf{x} \in \mathcal{C}$ *and* $\mathbf{X}$ *be a r.v. following* $\mathbf{P} \in \mathcal{P}(\mathcal{C})$. *The modified halfregion depth of* $\mathbf{x}$ *w.r.t.* $\mathbf{P}$ *is defined as:*

$$FD_{\mathrm{MHR}}(\mathbf{x}, \mathbf{P}) = \frac{1}{\lambda(\mathcal{T})} \min\left\{\mathbb{E}\left[\lambda(\{t \in \mathcal{T} :\ \mathbf{x}(t) \leq \mathbf{X}(t)\})\right], \mathbb{E}\left[\lambda(\{t \in \mathcal{T} :\ \mathbf{x}(t) \geq \mathbf{X}(t)\})\right]\right\}.$$

  As the FD_{MB}, the modified halfregion depth does not vanish at infinity and does not valid ($\mathbf{FD_{3D}}$) but satisfies ($\mathbf{FD}_0, \mathbf{FD_{1F}}, \mathbf{FD}_5$) and ($\mathbf{FD_{6U}}$) for distributions $\mathbf{P}$ such that $\mathbf{P}_t$ has no atoms for each $t \in \mathcal{T}$ or for empirical measures. The slight difference with FD_{MB} is its lack of the ability to recover the center of symmetric distributions (central and halfspace) when maximized. The modified halfregion depth can be written as the minimum between two integrated depths (see e.g. Nagy, 2016, Chapter 2.2.2) and can therefore also be considered as an integrated depth.

3.2.4 Functional Depths based on Distances

Functional depths based on norm-induced distance can be directly derived from the multivariate case replacing Euclidean norm by functional norms. Although satisfying most of the functional axioms and being easy to compute, they often capture little information about the distribution leading to a poor ordering. We briefly describe some of them below.

- **h-depth.** Let $\mathcal{F}$ be the vector space L_2 with the norm $||\cdot||_{L_2}$. Cuevas et al. (2007) have introduced the h-depth that is closely related to the density in the finite-dimensional case. The h-depth can be seen as a local measure of concentrated curves around $\mathbf{x}$ by means of kernel estimation involving the L_2 norm.

  Definition 3.12. *Let $\mathbf{x} \in L_2$ and $\mathbf{X}$ be a r.v. following $\mathbf{P} \in \mathcal{P}(L_2)$. The h-depth of $\mathbf{x}$ w.r.t. $\mathbf{P}$ is defined as:*

$$FD_{\mathrm{h}}(\mathbf{x}, \mathbf{P}) = \mathbb{E}\left[K_h\left(||\mathbf{x} - \mathbf{X}||_{L_2}\right)\right],$$

  where $h > 0$ and K_h is a re-scaled kernel of type $K_h(\cdot) = (1/h)K(\cdot/h)$ with K being a smoothing kernel function (e.g. Gaussian kernel).

  Properties of the h-depth have been extensively investigated in Nieto-Reyes and Battey (2016) and Gijbels and Nagy (2017). It satisfies $(\mathbf{FD_0}, \mathbf{FD_4}, \mathbf{FD_5}, \mathbf{FD_{6U}})$ but violates $(\mathbf{FD_2}, \mathbf{FD_{3D}})$. Furthermore, it is scalar translation invariant, i.e. it satisfies $(\mathbf{FD_{1S}})$ for $a = 1$.

- **Functional spatial and L_∞ depth.** Functional version of the multivariate Spatial and L_p- depths can be staightforwardly derived. The functional spatial depth has been introduced and studied in Chakraborty and Chaudhuri (2014b).

  Definition 3.13. *Let $\mathbf{x} \in L_2$ and $\mathbf{X}$ be a r.v. following $\mathbf{P} \in \mathcal{P}(L_2)$. The functional spatial depth of $\mathbf{x}$ w.r.t. $\mathbf{P}$ is defined as:*

$$FD_{\mathrm{Sp}}(\mathbf{x}, \mathbf{P}) = 1 - \left|\left|\mathbb{E}\left[\frac{\mathbf{x} - \mathbf{X}}{||\mathbf{x} - \mathbf{X}||_{L_2}}\right]\right|\right|_{L_2},$$

  where 0/0=0 by convention. The expectation of the L_2-valued r.v. corresponds to Bochner integral.

  The L_∞ depth is defined for continuous functions that belong to $\mathcal{C}(\mathcal{T})$ (Long and Huang, 2016).

  Definition 3.14. *Let $\mathbf{x} \in \mathcal{C}$ and $\mathbf{X}$ be a r.v. following $\mathbf{P} \in \mathcal{P}(\mathcal{C})$. The L_∞ depth of $\mathbf{x}$ w.r.t. $\mathbf{P}$ is defined as:*

$$FD_{\infty}(\mathbf{x}, \mathbf{P}) = \left(1 - \mathbb{E}\left[||\mathbf{x} - \mathbf{X}||_{\infty}\right]\right)^{-1},$$

  where $1/\infty = 0$ by convention.

  Both spatial and L_∞ depths share common properties such as $(\mathbf{FD_4}, \mathbf{FD_5}, \mathbf{FD_{6UC}})$. In addition, if $\mathbb{E}||\mathbf{X}||_\infty < \infty$, the L_∞ depth satisfies $(\mathbf{FD_0}, \mathbf{FD_{3D}})$ and is scalar-translation invariant but the property $(\mathbf{FD_2})$ does not hold. The functional spatial depth satisfies $(\mathbf{FD_0}, \mathbf{FD_{1S}})$ and $(\mathbf{FD_2})$ for centrally symmetric distributions but the property $(\mathbf{FD_{3D}})$ does not hold. More details can be found in the original papers and in Gijbels and Nagy (2017).

3.3 Discussion

Due to the richness of the functional spaces, capturing properties of the underlying distribution $\mathbf{P}$ is a challenging task. While the previously presented depth functions tend to focus on the amplitude/magnitude of the functional data, mainly by building statistics around the location of the observations $\mathbf{X}(t)$, $t \in \mathcal{T}$, several recent studies deal with the shape of the curves especially with the aim of detecting anomalies.

Beyond the incorporation of the derivatives of curves (Lange et al., 2014; Nagy et al., 2017), many depth functions focusing on both magnitude and shape of the underlying curves have been proposed in the literature. Especially designed for anomaly detection, the outliergram (Arribas-Gil and Romo, 2014), based on the modified band depth (see Definition 3.9) and the modified epigraph index (see López-Pintado and Romo, 2011), has been designed to visualize shape and magnitude components of functions. The notion of directional outlyingness introduced independently in Dai and Genton (2019b) and Rousseeuw et al. (2018) serves as the basis of visualization tools such as the magnitude-shape plot and the functional outlier map, respectively. Kuhnt and Rehage (2016) have developed the functional tangential angle (FUNTA) pseudo-depth based on local geometric features of the curves involving the tangential angles of the intersections of the centered data. Relying on the registration methods of Srivastava et al. (2011), Xie et al. (2017) have shown that dealing with the amplitude and phase components of functional data separately improves the detection of shape outliers. Further, a particular case of the integrated depth has been introduced in Huang and Sun (2019) to decompose the shape and the magnitude of functions. See also Narisetty and Nair (2016) and Myllymäki et al. (2017) for the extreme rank length depth; Dai et al. (2020) for a general procedure involving data transformations that turn shape outliers into magnitude outliers; Helander et al. (2020) for the functional Pareto depth based on a new multivariate Pareto depth applied after mapping functions to a vector of statistics of interest; and Harris et al. (2021) for the elastic depth that uses the elastic shape distance used in Xie et al. (2017) to measure the centrality of functions in the amplitude and phase spaces.

	$\mathbf{FD}_0$	$\mathbf{FD}_{1F}$	$\mathbf{FD}_{1S}$	$\mathbf{FD}_{2H}$	$\mathbf{FD}_{2C}$	$\mathbf{FD}_{3D}$	$\mathbf{FD}_4$	$\mathbf{FD}_5$	$\mathbf{FD}_{6U}$	$\mathbf{FD}_{6UC}$	$\mathbf{FD}_6$
$\mathcal{F}$	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{C}$	$\mathcal{C}$
IFD	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘	↘
FD_{RT}	↘	✗	↘	✗	✗	↘	✗	↘	✗	✗	✗
FD_{FT}	✗	✗	↘	✗	✗	✗	↘	✗	✗	✗	✗
GFD	✗	↘	↘	✗	✗	✗	↘	✗	↘	✗	✗
FD_{B}	✗	↘	↘	✗	✗	✗	↘	✗	↘	✗	✗
FD_{MB}	↘	↘	↘	↘	↘	↘	✗	↘	↘	↘	↘
FD_{HR}	✗	↘	↘	✗	✗	✗	↘	✗	✗	✗	✗
FD_{MHR}	↘	↘	↘	✗	✗	✗	✗	↘	✗	✗	✗
FD_{h}	↘	✗	✗	✗	✗	✗	↘	✗	↘	↘	↘
FD_{Sp}	↘	✗	↘	✗	↘	✗	↘	↘	✗	↘	↘
FD_{∞}	✗	✗	✗	✗	✗	✗	↘	↘	✗	↘	↘

Table 3.1 – Structural properties of the presented functional depths. *IFD*: integrated depth, FD_{RT}: random Tukey depth, FD_{FT}: functional Tukey depth, *GFD*: graph depth, FD_{B}: band depth, FD_{MB}: modified band depth, FD_{HR}: halfregion depth, FD_{MHR}: modified halfregion depth, FD_{h}: h-depth, FD_{Sp}: functional spatial depth, FD_{∞}: L_{∞} depth. For the aim of considering properties ($\mathbf{FD}_6, \mathbf{FD}_{6\mathbf{CU}}, \mathbf{FD}_{6\mathbf{U}}$) and for the sake of clarity, we assume that the probability distribution $\mathbf{P} \in \mathcal{P}(\mathcal{F})$ is such that if $\mathbf{X} \sim \mathbf{P}$ then $\mathbf{X}(t)$ has no atoms for each $t \in \mathcal{T}$. In addition, *IFD* and *GFD* are considered using continuous (in their first argument) univariate data depths satisfying the Definition 2.1.

4 Wasserstein Distance

Contents

Comparing probability distributions has attracted a long-standing interest in Information Theory (Kullback, 1959; Rényi, 1961; Csiszár, 1963), Probability Theory and Statistics (Rachev, 1991; Billingsley, 1999; Müller, 1997). Given two probability distributions P, Q defined on an arbitrary space, the goal is then to design metrics that are able to assess how close P and Q are, that differ in how the comparison is adressed. While they serve many purposes in Machine Learning (Cha and Srihari, 2002; MacKay and Mac Kay, 2003), they are of crucial importance as loss functions in automatic evaluation of natural language generation (see e.g. Kusner et al. 2015; Zhang et al. 2019), especially when leveraging deep contextualized embeddings such as the popular BERT (Devlin et al., 2019), graphical probabilistic modeling (Jordan, 1998) including generative adversarial modeling (see e.g. Goodfellow et al. 2014) as well as variational inference (see e.g. Blei et al. 2017). In the latter, choosing the right loss to be minimized between the two distributions is one of the key issues of the problem as its properties strongly influence the behavior of the associated algorithm. Thus, designing a measure to compare two probability distributions is considered as a challenging research field. This is certainly due to the inherent difficulty in capturing in a single measure typical desired properties such as: (i) metric or pseudo-metric properties, (ii) invariance under specific geometric transformations, (iii) efficient computation, (iv) efficient estimation from samples and (v) robustness to contamination.

From the side of Optimal Transport (OT) (see Villani 2003; Peyré and Cuturi 2019), the L_p-Wasserstein distance leverages a ground metric to take into account the geometry of the space on which the distributions are defined. Given two probability distributions, the latter is defined in terms of the solution to the Monge-Kantorovich optimal mass transportation problem. Its ability to handle non-overlapping support and appealing theoretical properties make OT a powerful tool. For these reasons, the Wasserstein distance stands out from the divergences usually exploited in generative modeling, like the φ-divergences, by its ability to take into account the underlying geometry of the space, capturing the difference between probability distributions even when they have non-overlapping supports. This appealing property has been successfully exploited in

Wasserstein Generative Adversarial Networks (WGANs; Arjovsky et al., 2017; Gulrajani et al., 2017) as well as in Wasserstein Auto-Encoders (WAE; (Tolstikhin et al., 2018)), where the Wasserstein distance can advantageously replace a φ-divergence as the loss function.

After having formally defined the Wasserstein distance and its basic properties in Section 4.1, its limitations are described in Section 4.2. In Section 4.3, we present the Sliced-Wasserstein distance originally introduced in order to remedy the previous drawbacks. Eventually, we briefly recall two important families of probability metrics in Section 4.4.

4.1 Definition and Properties

Originally introduced by Gaspard Monge in 1781 (Monge, 1781), Optimal Transport (OT) aims at finding a way to move the probability mass from one measure to another with least effort. The transportation effort of moving the probability mass of $P \in \mathcal{P}(\mathcal{X})$ to $Q \in \mathcal{P}(\mathcal{Y})$ is highlighted by a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, where $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$. This function c evaluates the distance between two elements $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The Monge OT problem consists in finding the measurable map $T : \mathcal{X} \to \mathcal{Y}$ that transports the source distribution $P \in \mathcal{P}(\mathcal{X})$ to its target $Q \in \mathcal{P}(\mathcal{Y})$ minimizing a total cost. Formally, the corresponding optimization problem is defined as:

$$\min_T \int_{\mathcal{X}} c(x, y) \, \mathrm{d}P(x) \quad \text{s.t.} \quad T_\sharp P = Q, \tag{4.1}$$

where $T_\sharp : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$ is the push-forward operator given by $T_\sharp P(A) = P(T^{-1}(A))$ for any measurable set A in $\mathcal{Y}$. This optimization problem is non-convex and feasible solutions may not exist. When measures P, Q are discrete, it boils down to a combinatorial assignment problem (see e.g. Peyré and Cuturi, 2019, Section 2.2).

Kantorovitch (1942) introduced a relaxed version of (4.1) such that the probability mass of any source point can be split into smaller masses which are then assigned to different target points, whereas the Monge problem performs a one-to-one assignment. The aim of this optimization problem consists in finding an optimal couplings, i.e. joint probability distributions on $\mathcal{X} \times \mathcal{Y}$ with marginals P and Q, minimizing a total cost. Precisely, it corresponds to:

$$\min_{\pi \, \in \, \Pi(P,Q)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, \mathrm{d}\pi(x \times y) \ ,$$

where $\Pi(P, Q) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \ : \ \int \pi(x, y) \, \mathrm{d}y = P(x), \int \pi(x, y) \, \mathrm{d}x = Q(y)\}$ is the set of joint probability distributions with marginals P and Q. There always exists a solution to this optimization problem as long as the map c is lower semi-continuous (see Santambrogio, 2015, Theorem 1.7). When the cost function is chosen as the Euclidean distance, the Kantorovich formulation leads to the L_p-Wasserstein distance described in the definition below. It defines an actual distance between probability measures which metrizes the weak-convergence.

Definition 4.1. *Given $p \in [1, \infty)$, the Wasserstein distance of order p between $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$, where $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, is defined through the resolution of the Monge-Kantorovitch mass transportation problem:*

$$\mathcal{W}_p(P, Q) = \min_{\pi \in \Pi(P,Q)} \left(\int_{\mathcal{X} \times \mathcal{Y}} \|x - y\|^p \, \mathrm{d}\pi(x \times y) \right)^{1/p}, \tag{4.2}$$

where $\Pi(P, Q) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \int \pi(x, y) \, \mathrm{d}y = P(x), \int \pi(x, y) \, \mathrm{d}x = Q(y)\}$ is the set of joint probability distributions with marginals P and Q.

The Wasserstein distance is then a particular case of the Kantorovich's relaxation. It leverages the information captured by the cost function c on the geometry of the supports of P and Q in order move the probability mass from P to Q in an optimal way. The comparison made via this discrepancy measure is then conceptually more powerful than with traditional divergences, e.g. φ-divergences which perform pointwise comparisons of the probability mass (see Section 4.4.1). This is a powerful tool that captures the underlying geometry of the measures, by relying on the cost function c which encodes somehow the geometry of the spaces $\mathcal{X}, \mathcal{Y}$, leading to meaningful comparisons even when the supports of the measures do not overlap (which is not the case for φ-divergences). Besides, the transport plan π gives a mapping between probability distributions which can be used in many applications such as in domain adaptation (Courty et al., 2014). We denote by $\mathcal{P}_p(\mathcal{X})$ the set of probability measures with finite p'th moment defined as:

$$\mathcal{P}_p(\mathcal{X}) = \left\{ P \in \mathcal{P}(\mathcal{X}) \; : \; \int_{\mathcal{X}} ||x - x_0||^p \, \mathrm{d}P(x), \text{ for some } \; x_0 \in \mathcal{X} \right\}.$$

Thus, assuming that $\mathcal{X} = \mathcal{Y}$, the Wasserstein distance $\mathcal{W}_p$ defines a metric on $\mathcal{P}_p(\mathcal{X})$ that metrizes the weak-convergence (see Villani, 2008, Chapter 6). The Wasserstein distance is a constrained convex minimization problem, and as such, it can naturally be associated with a so-called dual problem, which is a constrained concave maximization problem. For the purpose of this dissertation, we limit oursleves to the dual formulation of the L_1-Wasserstein.

Duality. When $p = 1$, by the dual Kantorovich-Rubinstein formulation (Kantorovich and Rubinstein, 1958), the L_1-Wasserstein distance can be reformulated as an Integral Probability Metric (see Section 4.4.2). Denote by $\mathcal{F}_{\text{Lip}}$ the unit ball of the Lipschitz functions space with the semi-norm $||\Psi||_{\text{Lip}} = \sup \left\{ \dfrac{|\Psi(x) - \Psi(y)|}{||x - y||} \; : \; x \neq y \in \mathbb{R}^d \right\}$, defined for any bounded continuous function Ψ.

Definition 4.2. *The dual form of the Wasserstein distance of order 1 between $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$, where $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, is defined as:*

$$\mathcal{W}(P, Q) = \sup_{\Psi \in \mathcal{F}_{\text{Lip}}} \mathbb{E}_P \left[\Psi(X) \right] - \mathbb{E}_Q \left[\Psi(Y) \right],$$

where X and Y are random variables having distributions P and Q, respectively.

In practice, the unit ball of Lipschitz functions can be replaced with a parameterized family of Lipschitz functions, more amenable for learning, see e.g. Wasserstein GANs (Arjovsky et al., 2017).

4.2 Limitations

Here, we present some limitations of the Wasserstein distance.

Curse of dimensionality. Of particular interest is the problem of estimating the Wasserstein distance between $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$ given a finite number of observations. The usual assumption is to rely upon two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, composed of i.i.d. realizations drawn respectively from P and Q. The corresponding empirical distributions denoted by $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$, and $Q_m = (1/m)\sum_{j=1}^m \delta_{Y_j}$. The natural questions are then: *how to compute the estimator $\mathcal{W}_p(P_n, Q_m)$, and does it converge towards $\mathcal{W}_p(P, Q)$?*

This problem has long been theoretically studied dating back to the seminal work of Dudley (1969). Assume that $P \in \mathcal{P}(\mathcal{X})$ is absolutely continuous w.r.t. the Lebesgue measure, the following optimal rate holds: $\mathbb{E}[\mathcal{W}_p(P_n, P)] \approx O(n^{-1/d})$. This result has been derived by Dudley (1969) for $p = 1$ for compactly supported measures and has been extended, among others, in Bassetti et al. (2006); Boissard (2011); Dereich et al. (2013); Boissard and Gouic (2014); Fournier and Guillin (2015); Chizat et al. (2020). They show that in general, the convergence rate of $\mathcal{W}_p(P_n, P)$ to zero (in expectation and with high probability) decreases exponentially when the dimension grows linearly. It is worth mentioning that when two samples are involved, the same rate holds since $|\mathcal{W}_p(P_n, Q_m) - \mathcal{W}_p(P, Q)| \leq \mathcal{W}_p(P_n, P) + \mathcal{W}_p(Q_m, Q)$. Thus, the sample sizes n and m must be very large in high dimension to obtain an accurate appproximation of $\mathcal{W}_p(P, Q)$. The Wasserstein distance then suffers from the curse of dimensionality. Despite of this result, the Wasserstein distance has been succesfully applied in generative modeling applied to images, that are high dimensional data (Arjovsky et al., 2017). However, it is known that high dimensional data such as images often belong to a manifold of smaller dimension. Following this, Weed and Bach (2019) have derived faster rates when the supports of measures are inherently of lower dimension.

Computational aspects. Solving (4.2) when dealing with empirical distributions P_n, Q_m boils down to:

$$\mathcal{W}_p^p(P_n, Q_m) = \min_{M \in \Pi(P_n, Q_m)} \sum_{i=1}^n \sum_{j=1}^m M_{i,j} C_{i,j},$$

where $\Pi(P_n, Q_m) = \{M \in \mathbb{R}_+^{n \times m} : M\mathbf{1}_n = \mathbf{1}_m/m \text{ and } M^\top = \mathbf{1}_n/n\}$ is the set of admissible joint probability matrices, and $C \in \mathbb{R}_+^{n \times m}$ storing the distances $||X_i - Y_j||^p$ for $i \leq n$ and $j \leq m$. The computation of the Wasserstein distance is therefore equivalent to solving a large-scale linear program (see Chapter 3 in Peyré and Cuturi, 2019 for further details). The computational complexity is then super-cubic $O(n^3 \log(n))$ assuming that $n = m$. It is worth mentioning that computing the cost matrix requires an additional cost of $O(n^2 d)$ often omitted in the literature. Due to its high computational cost it has long been disregarded by applied mathematicians.

Recently, many techniques, such as low-dimensional projections (Bonneel et al., 2015; Kolouri et al., 2019; Paty and Cuturi, 2019) or entropic regularization, have been developed to provide a computationally efficient approximation of the Wasserstein distance. Adding an entropic regularization makes the optimization problem strongly-convex and Cuturi et al. (2013) showed that it can be efficiently solved using Sinkhorn-Knopp matrix scaling algorithm (Sinkhorn, 1964), reducing the computational cost to $O(n^2)$, and can be accelerated with parallelization on CPU and GPU. The entropic

regularization makes the problem more robust to small changes in the distributions but tends to spread mass in the optimal transportation matrix which leads to difficult interpretation of the optimal transport problem. Many possible regularizations have been introduced to induce sparsity of the optimal transport plan and remedy to these problems (Schmitzer, 2016; Blondel et al., 2018).

Robustness. Despite its appealing properties, the Wasserstein distance suffers from a high sensitivity to outliers due to the marginal constraints in (4.2). Indeed, a small outlier mass can contribute highly to the cost. This is the aim of Chapter 8 to propose a robust estimation of the Wasserstein distance leveraging the Median-of-Means estimator, that is the first work, together with Balaji et al. (2020), dealing with the robustness of the Wasserstein distance. Thereafter, some works have been introduced (Nietert et al., 2021; Mukherjee et al., 2021) in the same line of Balaji et al. (2020) relying on slight modifications of the unbalanced optimal transport problem (Piccoli and Rossi, 2014; Chizat et al., 2018) relaxing the marginal constraints of (4.2).

4.3 Sliced-Wasserstein distance

To address these limitations, a line of research relies on the use of low-dimensional projections of probability distributions. Leveraging the computational benefits of the one-dimensional formula, the Sliced-Wasserstein (SW) discrepancy measure was introduced in Rabin et al. (2012) (see also Bonneel et al., 2015) and is presented in this section.

Univariate spaces. A nice feature of the Wasserstein distance appears when computed between univariate distributions. Assuming $P_1, Q_1 \in \mathcal{P}(\mathbb{R})$, it holds (Rachev and Rüschendorf, 1998):

$$\mathcal{W}_p(P_1, Q_1) = \left(\int_0^1 |F_{P_1}^{-1}(t) - F_{Q_1}^{-1}(t)|^p \, \mathrm{d}t \right)^{1/p}, \tag{4.3}$$

where $F_{P_1}^{-1}, F_{Q_1}^{-1}$ are quantile functions of P_1 and Q_1 respectively.

Considering two empirical measures $P_{1,n} = (1/n)\sum_{i=1}^n \delta_{X_i}$ and $Q_{1,n} = (1/n)\sum_{i=1}^n \delta_{Y_i}$, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ being two samples of same size having distributions P_1 and Q_1 respectively, (4.3) can be easily computed by performing a sorting of each sample (see Figure 4.1 for an illustration):

$$\mathcal{W}_p^p(P_{1,n}, Q_{1,n}) = \frac{1}{n}\sum_{i=1}^n |X_{(i)} - Y_{(i)}|^p, \tag{4.4}$$

where $X_{(1)} \geq \ldots \geq X_{(n)}$ and $Y_{(1)} \geq \ldots \geq Y_{(n)}$. The computation of (4.4) requires $O(n\log(n))$ operations induced by the sorting of data. Based upon this nice feature, Rabin et al. (2012) introduced a new quantity, named Sliced-Wasserstein, computing similarity between $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$ relying on an average of (4.3) over their one-dimensional projections.

Definition 4.3. *Given $p \in [1, +\infty)$ and ω_{d-1} the spherical probability measure on $\mathbb{S}^{d-1}$, the SW distance between $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$, is given by:*

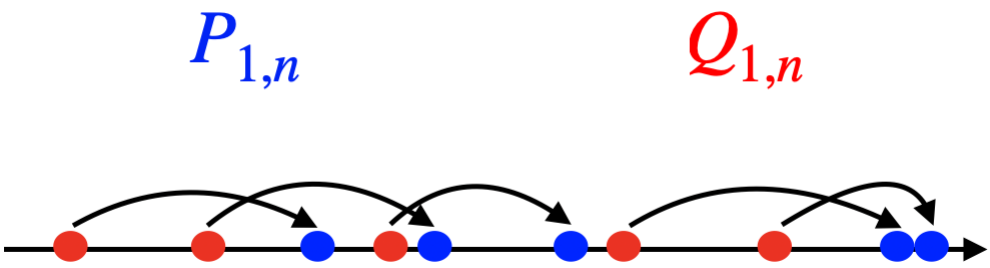


Figure 4.1 – Optimal transport on the real line.

$$SW_p(P, Q) = \left(\int_{\mathbb{S}^{d-1}} \mathcal{W}_p^p(P_u, Q_u)\, \omega_{d-1}(\mathrm{d}u) \right)^{1/p},$$

where P_u is the pushforward distribution of P defined by the projection $x \in \mathbb{R}^d \mapsto \langle u, x \rangle$.

With two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ having distributions $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$, the "projected" empirical measures are given by $P_{u,n} = (1/n) \sum_{i=1}^n \delta_{\langle u, X_i \rangle}$ and $Q_{u,n} = (1/n) \sum_{i=1}^n \delta_{\langle u, Y_i \rangle}$. Therefore, the Sliced-Wasserstein computed between $P_n = (1/n) \sum_{i=1}^n \delta_{X_i}$ and $Q_n = (1/n) \sum_{i=1}^n \delta_{Y_i}$ boils down to:

$$SW_p(P_n, Q_n) = \left(\int_{\mathbb{S}^{d-1}} \mathcal{W}_p^p(P_{u,n}, Q_{u,n})\, \omega_{d-1}(\mathrm{d}u) \right)^{1/p}. \tag{4.5}$$

Besides, the expectation that defines SW can easily be approximated with a standard Monte-Carlo scheme: one draws $n_{\text{proj}} \in \mathbb{N}_*$ samples i.i.d. from ω_{d-1}, denoted by $\{u_k\}_{k=1}^{n_{\text{proj}}}$, and approximates (4.5) with

$$SW_p^{\text{MC}}(P_n, Q_n) = \left(\frac{1}{n_{\text{proj}}} \sum_{k=1}^{n_{\text{proj}}} \mathcal{W}_p^p(P_{u_k,n}, Q_{u_k,n}) \right)^{1/p}.$$

This quantity then results in a computational complexity of $O(n_{\text{proj}}\, nd + n_{\text{proj}}\, n \log(n))$ due to the projecting and sorting operations. Several variants of the Sliced-Wasserstein exist such as the maximum Sliced-Wasserstein (Deshpande et al., 2019) where the integral over the unit sphere is replaced by a maximum or the generalized Sliced-Wasserstein (Kolouri et al., 2019) where linear projections are replaced by Radon transforms.

Following the seminal paper of Bonnotte (2013) showing that SW defines a distance on $\mathcal{P}_p(\mathbb{R}^d)$ and establishing its relation with the Wasserstein (see also Bayraktar and Guo, 2021), several recent works study statistical properties of the Sliced-Wasserstein distance and its variants (see, among others, Deshpande et al., 2019; Manole et al., 2019; Nadjahi et al., 2019, 2020).

Kullback-Leibler	$\varphi(x) = x\log(x)$
Reverse Kullback-Leibler	$\varphi(x) = -\log(x)$
Jensen-Shannon	$\varphi(x) = x\log(x) - (1+x)\log\Big((1+x)/2\Big)$
Squared Hellinger	$\varphi(x) = \Big(\sqrt{x}-1\Big)^2$

Table 4.1 – Examples of φ-divergences.

4.4 Alternative Metrics

In this section, we briefly recall two popular family of discrepancy measures: the φ-divergences and the Integral Probability Metrics namely.

4.4.1 φ-Divergences

The *φ-divergences* introduced in Rényi (1961) and Csiszár (1963) are defined as the weighted average by well-chosen functions φ of the odds ratio between the two distributions. The best known φ-divergence is the *Kullback-Leibler divergence* which is widely used in Machine Learning applications such as in Generative Adversarial Networks (GAN; Goodfellow et al., 2014) or Variational Auto-Encoders (VAE; Kingma and Welling, 2013) (see also Kingma and Welling, 2019). This family of divergences also includes, among others, the reverse Kullback-Leibler and the Jensen-Shannon divergences as well as the squared Hellinger distance, see Table 4.1.

Definition 4.4. *Let φ be a convex and lower semi-continuous function such that $\varphi(1) = 0$. Let $P, Q \in \mathcal{P}(\mathcal{X})$ be two probability distributions defined on the same set $\mathcal{X}$. The φ-divergences, denoted by d_φ, between P and Q are defined as:*

$$d_\varphi(P,Q) = \int_{\mathcal{X}} \varphi\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right)\mathrm{d}Q(x) + \varphi_\infty P^\perp(\mathcal{X}), \tag{4.6}$$

where $\varphi_\infty = \lim\limits_{x\to+\infty} \frac{\varphi(x)}{x}$ and $P^\perp(\mathcal{X})$ denotes the mass of the part of P that is not absolutely continuous w.r.t. Q in the Radon-Nikodym decomposition of P, i.e. $P = \frac{\mathrm{d}P}{\mathrm{d}Q}(x) + P^\perp$.

However, φ-divergences do not metrize weak convergence which is a major issue. The metrization of weak-convergence is essential, as it ensures that the metrics remain stable under small perturbations of the support of the measures. It is illustrated by the degeneracy to $+\infty$ of φ-divergences when the supports of both distributions do not overlap, which appears to be a crucial limitation in many applications.

4.4.2 Integral Probability Metrics

Integral Probability Metrics (IPMs) were introduced by Müller (1997) as the maximum difference in expectation for both distributions calculated over a class of measurable functions and regroup some well known distances.

Definition 4.5. *Let $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$ be two probability measures defined on $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$. Let $\mathcal{F}$ be an arbitrary functional set of measurable functions $\Psi : \mathbb{R}^d \to \mathbb{R}$. Integral Probability Metrics, denoted by $d_\mathcal{F}$, are defined as:*

Maximum Mean Discrepancy	$\mathcal{F} = \{\Psi : \|\|\Psi\|\|_{\mathcal{H}} \leq 1\}$, where $\mathcal{H}$ is an RKHS
Total Variation	$\mathcal{F} = \{\Psi : \|\|\Psi\|\|_{\infty} \leq 1\}$
L_1-Wasserstein	$\mathcal{F} = \{\Psi : \|\|\Psi\|\|_{\text{Lip}} \leq 1\}$
Dudley	$\mathcal{F} = \{\Psi : \|\|\Psi\|\|_{\text{Lip}} + \|\|\Psi\|\|_{\infty} \leq 1\}$

Table 4.2 – Examples of Integral Probability Metrics.

$$d_{\mathcal{F}}(P, Q) = \sup_{\Psi \in \mathcal{F}} \mathbb{E}_P \left[\Psi(X)\right] - \mathbb{E}_Q \left[\Psi(Y)\right],$$

where X and Y are r.v. having distributions P and Q respectively.

IPMs offer several theoretical guarantees supporting the relevance of this approach. First, they are pseudo-metrics (Sriperumbudur et al., 2009): they are non-negative, symmetric, verify the triangle inequality and for any $P \in \mathcal{P}(\mathbb{R}^d)$, $d_{\mathcal{F}}(P, P) = 0$. Besides, $d_{\mathcal{F}}$ metrizes weak convergence, provided that the span of $\mathcal{F}$ is dense in the space of continous and bounded functions on $\mathbb{R}^d$ endowed with the supremum norm (see Ambrosio et al., 2005, Section 5.1). Moreover, any IPM admits an empirical estimate which is consistent with explicit convergence rates (Sriperumbudur et al., 2012).

An important example of IPM which admits a consistent and statistically efficient empirical estimator is the Maximum Mean Discrepancy (MMD) (Gretton et al., 2007), see Table 4.2. Indeed, when the function space $\mathcal{F}$ is chosen as the unit ball of an Reproducing Kernel Hilbert Space (RKHS) (see e.g. Aronszajn, 1950), $d_{\mathcal{F}}$ boils down to the popular MMD metric. The reproducing property of RKHS allows to derive a much simpler expression for their associated IPMs. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel and $\mathcal{H}$ its corresponding RKHS with inner product $\langle ., . \rangle_{\mathcal{H}}$ and norm $\|.\|_{\mathcal{H}}$. Denote by $\mathcal{F}_{\mathcal{H}} = \{\Psi \mid \|\Psi\|_{\mathcal{H}} \leq 1\}$ the unit ball of $\mathcal{H}$. Assuming that $\sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} k(x, y) < \infty$, the Maximum Mean Discrepancy (MMD) between the two distributions $P \in cP(\mathcal{X})$ and $Q \in \mathcal{Q}(\mathcal{Y})$ associated with the kernel k is defined as:

$$\begin{aligned} \text{MMD}(P, Q) &= \sup_{\Psi \in \mathcal{F}_{\mathcal{H}}} \left| \mathbb{E}_P[\Psi(X)] - \mathbb{E}_Q[\Psi(Y)] \right| \\ &= \mathbb{E}_{P \otimes P}[k(X, X')] + \mathbb{E}_{Q \otimes Q}[k(Y, Y')] - 2\,\mathbb{E}_{P \otimes Q}[k(X, Y)]. \end{aligned}$$

The MMD can be estimated with a quadratic computational complexity $\mathcal{O}(n^2)$ where n is the sample size.

Part II

Functional Anomaly Detection

5
Functional Isolation Forest

Contents

In this chapter, we introduce a new generic algorithm, Functional Isolation Forest (FIF) that generalizes (Extended) Isolation Forest to the infinite dimensional context. This *ensemble learning* algorithm builds a collection of *Functional isolation trees* based on a recursive and randomized tree-structured partitioning procedure. Avoiding dimensionality reduction steps, this extension is shown to preserve the assets of the original algorithm concerning computational cost and interpretability. Its efficiency is supported by strong empirical evidence through a variety of numerical results.

The chapter is organized as follows. Section 5.1 recalls the principles of the Isolation Forest algorithm for anomaly detection in the multivariate case. In Section 5.2, the extension to the functional case is presented and its properties are discussed at length. In Section 5.3, we study the behavior of the new algorithm and compare its performance to alternative methods standing as natural competitors in the functional setup through experiments. In Section 5.4, extension to multivariate functional data is considered, as well as relation to the data depth function and an application to the supervised classification setting. Eventually, several concluding remarks are collected in Section 5.5. This chapter covers the contribution of:

- **G. Staerman**, P. Mozharovskyi, S. Clémençon, F. d'Alché-Buc. Functional Isolation Forest. In *Proceedings of The Eleventh Asian Conference on Machine Learning (ACML)*, pages 332-347, 2019.

5.1 Isolation Forest

As a first go, we describe the Isolation Forest algorithm for anomaly detection in the multivariate context in a formalized manner for clarity's sake, as well as the Extended Isolation Forest version, see Liu et al. (2008, 2012) and Hariri et al. (2019) respectively. These two unsupervised algorithms can be viewed as *ensemble learning* methods insofar as they build a collection of binary trees and an anomaly scoring function based on the aggregation of the latter. Let $\mathcal{S}_n = \{X_1, \ldots, X_n\}$ be a training sample composed of n independent realizations of a generic random variable, X, that takes its value in a finite dimensional Euclidean space, $\mathbb{R}^d$ say, $X = (X^{(1)}, \ldots, X^{(d)})$.

An *isolation tree* (*i*tree in abbreviated form) $\mathcal{T}$ of depth $J \geq 1$ is a proper binary tree that represents a nested sequence of partitions of the feature space $\mathbb{R}^d$. The root node corresponds to the whole space $\mathcal{C}_{0,0} = \mathbb{R}^d$, while any node of the tree, indexed by the pair (j, k) where j denotes the depth of the node with $0 \leq j < J$ and k, the node index with $0 \leq k \leq 2^j - 1$, is associated to a subset $\mathcal{C}_{j,k} \subset \mathbb{R}^d$. A non terminal node (j, k) has two children, corresponding to disjoint subsets $\mathcal{C}_{j+1,2k}$ and $\mathcal{C}_{j+1,2k+1}$ such that $\mathcal{C}_{j,k} = \mathcal{C}_{j+1,2k} \cup \mathcal{C}_{j+1,2k+1}$. A node (j, k) is said to be terminal if it has no children.

Each *i*tree is obtained by recursively filtering a subsample of training data of size n_s in a top-down fashion, by means of the following procedure. The data set composed of the training observations present at a node (j, k) is denoted by $\mathcal{S}_{j,k}$. At iteration $k + 2^j$ of the *i*tree growing stage, a direction p in $\{1, \ldots, d\}$, or equivalently a *split variable* $X^{(p)}$, is selected uniformly at random (and independently from the previous draws) as well as a *split value* κ in the interval $[\min_{x \in \mathcal{S}_{j,k}} x^{(p)}, \max_{x \in \mathcal{S}_{j,k}} x^{(p)}]$ corresponding to the range of the projections of the points in $\mathcal{S}_{j,k}$ onto the p-th axis. The children subsets are then defined by $\mathcal{C}_{j+1,2k} = \mathcal{C}_{j,k} \cap \{x \in \mathbb{R}^d : x^{(p)} \leq \kappa\}$ and $\mathcal{C}_{j+1,2k+1} = \mathcal{C}_{j,k} \cap \{x \in \mathbb{R}^d : x^{(p)} > \kappa\}$, the children training data sets being defined as $\mathcal{S}_{j+1,2k} = \mathcal{S}_{j,k} \cap \mathcal{C}_{j+1,2k}$ and $\mathcal{S}_{j+1,2k+1} = \mathcal{S}_{j,k} \cap \mathcal{C}_{j+1,2k+1}$.

An *i*tree $\mathcal{T}$ is thus built by iterating this procedure until all training data points are isolated (or the depth limit J set by the user is attained). A preliminary subsampling stage can be performed in order to avoid swamping and masking effects, when the size of the data set is too large. When it isolates any training data point, the *i*tree contains exactly $n_s - 1$ internal nodes and n_s terminal nodes. An *i*tree constructed accordingly to a training subsample allows to assign to each training datapoint x_i a path length $h_{\mathcal{T}}(x_i)$, namely the depth at which it is isolated from the others, i.e. the number of edges x_i traverses from the root node to the terminal node that contains the sole training data x_i. More generally, it can be used to define an anomaly score for any point $x \in \mathbb{R}^d$.

Anomaly Score prediction. As the terminal nodes of the *i*tree $\mathcal{T}$ form a partition of the feature space, one may then define the piecewise constant function $h_\tau : \mathbb{R}^d \to \mathbb{N}$ by: $\forall x \in \mathbb{R}^d$,

$$h_\tau(x) = j \text{ if and only if } x \in \mathcal{C}_{j,k} \text{ and } (j, k) \text{ is a terminal node.}$$

This random path length is viewed as an indication for its degree of abnormality in a natural manner: ideally, the more abnormal the point x, the higher the probability that the quantity $h_\tau(x)$ is small. Hence, the algorithm above can be repeated $N \geq 1$ times in order to produce a collection of *i*trees $\mathcal{T}_1, \ldots, \mathcal{T}_N$, referred to as an *i*forest, that defines the scoring function

$$s_n(x) = 2^{-\frac{1}{Nc(n_s)} \sum_{l=1}^{N} h_{\tau_l}(x)}, \tag{5.1}$$

where $c(n_s)$ is the average path length of unsuccessful searches in a binary search tree, see Liu et al. (2008) for further details.

Extended Isolation Forest. Observing that the geometry of the abnormal regions of the feature space is not necessarily well-described by perpendicular splits (i.e. by unions of hypercubes of the cartesian product $\mathbb{R}^d$), a more flexible variant of the procedure recalled above has been proposed in Hariri et al. (2019), in the purpose of bias reduction. Rather than selecting a direction in $\{1, \ldots, d\}$, one may choose a direction $u \in \mathbb{S}^{d-1}$, denoting by $\mathbb{S}^{d-1}$ the unit sphere of the euclidian space $\mathbb{R}^d$. A node is then cut by choosing randomly and uniformly a threshold value in the range of the projections onto this direction of the training data points lying in the corresponding region.

5.2 The FIF Algorithm

We consider the problem of learning a score function $s : \mathcal{H} \to \mathbb{R}$ that reflects the degree of anomaly of elements in an infinite dimensional space $\mathcal{H}$ w.r.t. $\mathbf{P} \in \mathcal{P}(\mathcal{H})$. By $\mathcal{H}$, we denote a functional Hilbert space equipped with a scalar product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ such that any $\mathbf{x} \in \mathcal{H}$ is a real-valued function defined on $[0, 1]$. In this chapter, we limit ourselves to $[0, 1]$ but our approach remains valid for any compact set $\mathcal{T} \subset \mathbb{R}$. A Functional Isolation Forest is a collection of *Functional isolation trees* (F-$i$trees) built from $\mathcal{S} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, a training sample composed of independent realizations of a functional random variable, $\mathbf{X}$, that takes its values in $\mathcal{H}$. Given a functional observation $\mathbf{x}$, the score returned by FIF is a monotone transformation of the empirical mean of the path lengths $h_{\mathcal{T}_l}(\mathbf{x})$ computed by the F-$i$trees $\mathcal{T}_l$, for $l = 1, \ldots, N$ as defined in (5.1) in the multivariate case. While the general construction principle depicted in Section 5.1 remains the same for a F-$i$tree, dealing with functional values raises the issue of finding an adequate feature space to represent various properties of a function. A function may be considered as abnormal according to various criteria of location and shape, and the features should permit to measure such properties. Therefore four ingredients have been introduced to handle functional data in a general and flexible way: (i) a set of candidate *Split variables* and (ii) a scalar product both devoted to function representation, (iii) a probability distribution to sample from this set and select a single *Split variable*, (iv) a probability distribution to select a *Split value*. The entire construction procedure of a F-$i$tree is described in Figure 5.1. An example of a F-$i$tree is also depicted in Figure 5.2.

Function representation. To define the set of candidate *Split variables*, a direct extension of the original IF algorithm (Liu et al., 2008) would be to randomly draw an argument value (e.g. time), and use functional evaluations at this point to split a node, but this boils down to only rely on instantaneous observations of functional data to capture anomalies, which in practice will be usually interpolated. Drawing a direction on a unit sphere as in Hariri et al. (2019) is no longer possible due to the potentially excessive richness of $\mathcal{H}$. To circumvent these difficulties, we propose to project the observations on elements of a dictionary $\mathcal{D} \subset \mathcal{H}$ that is chosen to be rich enough to explore different properties of data and well appropriate to be sampled in a representative manner. More explicitly, given a function $\mathbf{d} \in \mathcal{D}$, the projection of a function $\mathbf{x} \in \mathcal{H}$ on $\mathcal{D}$, $\langle \mathbf{x}, \mathbf{d} \rangle_{\mathcal{H}}$ defines a feature that partially describes $\mathbf{x}$. When considering all the functions of dictionary $\mathcal{D}$, one gets a set of candidate *Split variables* that provides a rich representation of function $\mathbf{X}$, depending on the nature of the dictionary. Dictionaries have been throughly studied in the signal processing community to achieve

Input: A subsample $\{\mathbf{X}_1, \ldots, \mathbf{X}_{n_s}\}$, a dictionary $\mathcal{D} \subset \mathcal{H}$, a probability measure $\boldsymbol{\nu} \in \mathcal{P}(\mathcal{D})$ and a scalar product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$.

(a) **Initialization:** The root node indexed by $(0, 0)$ is associated with the whole input space $\mathcal{C}_{0,0} = \mathcal{H}$. The construction starts with the training data set $\mathcal{S}_{0,0} = \{\mathbf{X}_1, \ldots, \mathbf{X}_{n_s}\}$ composed of n_s i.i.d.realizations of the random variable $\mathbf{X}$. Go to (b) with $(j = 0, k = 0)$.

(b) **Stopping criterion:** Test if the node (j, k) is terminal: a node (j, k) is declared as *terminal* if the intersection between the current set $\mathcal{C}_{j,k}$ and the current training set $\mathcal{S}_{j,k}$ is reduced to a single data point or to a set of predefined cardinal. If the node is terminal, then stop the construction for this node, otherwise go to (c).

(c) **Children node construction:** A non-terminal node (j, k) is split in three steps as follows:

1. Choose a *Split variable* $\mathbf{d}$ according to the probability distribution $\boldsymbol{\nu}$ on $\mathcal{D}$.
2. Choose randomly and uniformly a *Split value* κ 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

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

as well as the children training data sets

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

(d) **Recursion:** Apply the building procedure starting from (a) to nodes $(j+1, 2k)$ and $(j+1, 2k+1)$

Output: $(\mathcal{C}_{0,0}, \mathcal{C}_{1,1}, \ldots)$

Figure 5.1 – Construction procedure of a F-itree.

sparse coding of signals, see e.g. Mallat and Zhang (1993). They also provide a way to incorporate *a priori* information about the nature of the data, a property very useful in an industrial context in which functional data often come from the observation of a well known device and thus can benefit from expert knowledge.

Sampling a *Split variable*. Once a dictionary is chosen, a probability distribution $\boldsymbol{\nu}$ on $\mathcal{D}$ is defined to draw a *Split variable* $\mathbf{d}$. Note that the choice of the sampling distribution $\boldsymbol{\nu}$ gives an additional flexibility to orientate the algorithm towards the search for specific properties of the functions.

Sampling a *Split value*. Given a chosen *Split variable* $\mathbf{d}$ and a current training data set $\mathcal{S}_{j,k}$, a *Split value* is uniformly drawn in the real interval defined by the smallest and largest values of the projections on $\mathbf{d}$ when considering the observations present in the node.

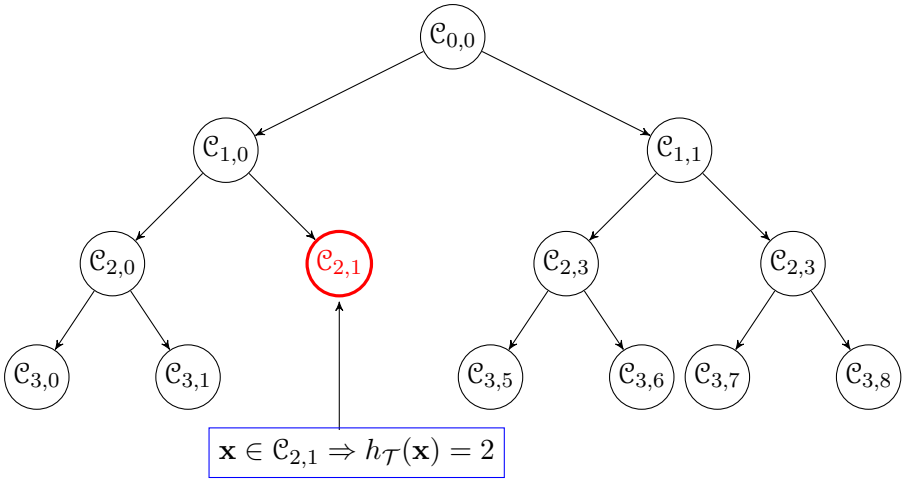


Figure 5.2 – An example of a functional isolation tree structure denoted by $\mathcal{T}$. Here, $\mathcal{C}_{2,1}$ is a cell associated with a terminal node.

5.2.1 Ability of FIF to Detect a Variety of Anomalies

As discussed in Section 1.2, most of state-of-the-art methods have a focus on a certain type of anomalies and are unable to detect various deviations from the normal behavior. The flexibility of the FIF algorithm allows for choosing the scope of the detection by selecting both the scalar product and the dictionary. Nevertheless, by choosing appropriate scalar product and dictionary, FIF is able to detect a great diversity of deviations from normal data. It is worth noticing than any discrepancy measure between functions may be choose instead of the scalar product. To account for both location and shape anomalies, we suggest the following discrepancy measure based on a normalized scalar product that provides a compromise between the both

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

where $\mathbf{f}, \mathbf{g}$ are both differentiable with $\mathbf{f}', \mathbf{g}'$ as derivatives and $\mathbf{f}, \mathbf{g}, \mathbf{f}', \mathbf{g}'$ belong to L_2. Thus, removing normalization terms, setting $\alpha = 1$ yields the classical L_2 scalar product, $\alpha = 0$ corresponds to the L_2 scalar product of derivative, and $\alpha = 0.5$ is the Sobolev $W_{1,2}$ scalar product. To illustrate the FIF's ability to detect a wide variety of anomalies at a time, we calculate the FIF anomaly scores with the Sobolev scalar product (normalized) and the *gaussian wavelets dictionary* for a sample consisting of 105 curves defined as follows (inspired by (Cuevas et al., 2007), see Fig. 5.3):

- 100 curves defined by $\mathbf{x}(t) = 30(1-t)^q t^q$ with q equispaced in $[1, 1.4]$,
- 5 *abnormal* curves composed by one isolated anomaly $\mathbf{x}_0(t) = 30(1-t)^{1.2}t^{1.2}$ with a jump in $t = 0.7$, one magnitude anomaly $\mathbf{x}_1(t) = 30(1-t)^{1.6}t^{1.6}$ and three kind of shape anomalies $\mathbf{x}_2(t) = 30(1-t)^{1.2}t^{1.2} + \sin(2\pi t)$, $\mathbf{x}_3(t) = 30(1-t)^{1.2}t^{1.2}$ noised by $\zeta \sim \mathcal{N}(0, 0.3^2)$ on the interval $[0.2, 0.8]$ and $\mathbf{x}_4(t) = 30(1-t)^{1.2}t^{1.2} + \frac{1}{2}\sin(10\pi t)$.

One can see that the five anomalies, although very different, are all detected by FIF with a significantly different score.

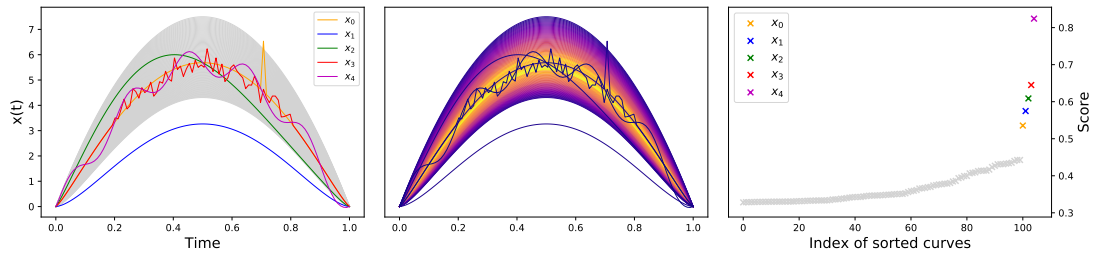


Figure 5.3 – The simulated data set with the five introduced anomalies (left). The scored data set (middle), the darker the color, the more the curves are considered anomalies. The sorted anomaly score of the data set (right).

5.2.2 Dictionary

The choice of a suited dictionary plays a key role in construction of the FIF anomaly score. The dictionary can consist of deterministic functions, incorporate stochastic elements, contain the observations from $\mathcal{S}$, or be a mixture of several mentioned options. In *Computational Harmonic Analysis*, a wide variety of bases or frames, such as wavelets, ridgelets, cosine packets, brushlets and so on, have been developed in the last decades in order to represent efficiently/parsimoniously functions, signals or images exhibiting specific form of singularities (e.g. located at isolated points, along hyperplanes) and may provide massive dictionaries. The following ones will be used throughout the chapter: *mexican hat wavelet dictionary* (MHW), *Brownian motion dictionary* (B), *Brownian bridge dictionary* (BB), *cosine dictionary* (Cos), *uniform indicator dictionary* (UI), *Uniform indicator derivative* (UId), *dyadic indicator dictionary* (DI), *Dyadic indicator derivative* (DId), and the *self-data dictionary* (Self) containing the data set itself. Precisely, they are defined as follows:

- *Mexican hat wavelet dictionary* (MHW) consists of the negative second derivatives of the normal density, shifted and scaled in a appropriate fashion:

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

  with $\theta \in [-0.8, 0.8]$ and $\sigma \in ([0.04, 0.2])$.

- *Brownian motion dictionary* (B) is a combination of the space of continuous function $\mathcal{D} = \mathcal{C}([0,1])$ and the Wiener measure $\boldsymbol{\gamma}$ on $\mathcal{D}$.

- *Brownian bridge dictionary* (BB) is a combination of the space of continuous function $\mathcal{D} = \mathcal{C}([0,1])$ and the Brownian bridge measure $\mathcal{G}$ on $\mathcal{D}$.

- *Cosine dictionary* (Cos) consists of curves with the following forms:

$$\mathbf{x}_{a,\omega}(t) = a\cos(2\pi\omega t)$$

  with $a \in [0,1]$ and $\omega \in [0,10]$.

- *Uniform indicator dictionary* (UI) consists of indicator function on $[a, b]$ where a and b are choosen uniformly on $[0, 1]$ such that $a < b$.

- *Uniform indicator derivative* (UId) consists of functions $t \mapsto t$ on $[a, b]$ where a and b are choosen uniformly on $[0, 1]$ such that $a < b$.

- *Dyadic indicator dictionary* (DI) consists of a set of indicator functions on the elements of binary partitioning, for a given J (chosen according to the granularity to be captured or from the discretisation considerations) having as elements $\left\{\left(\mathbf{x}_{k,j}\right)_{0\leq k<2^j}\right\}_{1\leq j\leq J}$:

$$\mathbf{x}_{k,j}(t) = \mathbb{1}\left(t \in \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right]\right).$$

- *Dyadic indicator derivative* (DId) consisting of a set of indicator functions on the elements of binary partitioning, for a given J (chosen according to the granularity to be captured or from the discretisation considerations) having as elements $\left\{\left(\mathbf{x}_{k,j}\right)_{0\leq k<2^j}\right\}_{1\leq j\leq J}$:

$$\mathbf{x}_{k,j}(t) = t\mathbb{1}\left(t \in \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right]\right).$$

- *Self-data dictionary* (Self) consists of the training data set itself.

To illustrate the incorporation of stochastic elements and external information, we bring an example of the use of the *Brownian motion dictionary*. Let $\boldsymbol{\gamma}$ be the Wiener measure defined on $\mathcal{C}([0, 1])$ the space of continuous function on [0,1] and $\mathcal{H}$ be the L_2 space. We define by *Brownian motion dictionary* (B) the *Split variables* space induced by $\boldsymbol{\nu} = \boldsymbol{\gamma}$ and $\mathcal{D} = \mathcal{C}([0, 1])$. Although seeming universal, this dictionary explores almost the entire argument space equivalently, and in practice can be unable to detect *isolated anomalies*. In Figure 5.4, we plot the following synthetic data set:

- 30 curves defined by $\mathbf{x}(t) = 30(1-t)^q t^q$ on $t \in [0, 0.2]$ and $\mathbf{x}(t) = 30(0.8)^q 0.2^q + \mathcal{N}(0, 0.3^2)$ on $t \in [0.2, 0.7]$ with q equispaced in $[0.5, 0.55]$.
- 1 *abnormal* curve with the same shape but that is shifted at the beginning and whose continuation is deep in the 30 preceding curves.

One can see that the anomaly is not detected, that indicated as anomaly curve (the one with highest anomaly score) is on the fringe of the data set though. Illustrative incorporation of the prior knowledge, in its simplified version, can consist, e.g. in adding to the measure $\boldsymbol{\gamma}$, a Dirac of the indicator function on the interval of interest $\tilde{\mathbf{x}}(t) = \mathbb{1}(t \leq 0.25)$ with weights: $\widetilde{\boldsymbol{\gamma}} := 0.2\ \boldsymbol{\gamma} + 0.8\ \delta_{\tilde{\mathbf{x}}}$; this assigns the highest anomaly score

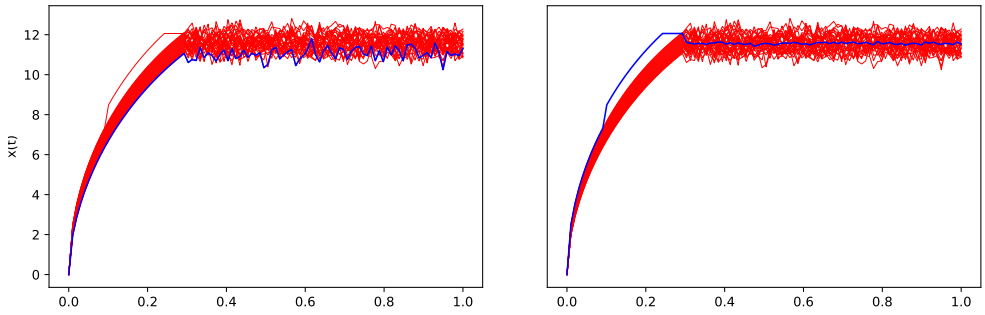


Figure 5.4 – Synthetic data containing an isolated anomaly and the observation having the highest anomaly score (blue), with dictionary being pure Brownian motion (left) and Brownian motion mixed with an indicator function in the area of interest in proportions 4 to 1 (right).

to the desired observation. In the sequel, $\boldsymbol{\nu}$ follows a uniform distribution if we do not explicitly mention its distribution.

When having not enough prior knowledge, e.g. just knowing to stick to local features of functional data but not the precise interval, one would like to use a dictionary exploring different localities. To illustrate possible advantage of this approach, we use *Mexican hat wavelet* and *Dyadic indicator* dictionaries on the "Chinatown" data set (Chen et al., 2015b), which represents pedestrian count in Chinatown-Swanston St North for 12 months during year 2017 with 14 functions (working days) representing normal observations and taking 4 functions (weekends) as anomalies (Figure 5.5). One observes that while the Mexican hat wavelet dictionary correctly detects part of the anomalies, due to its smooth nature it is distracted by two normal curves with high deviation on the second half of the domain. Having straight fronts and begin non-zero only in a small part of the domain, the dyadic indicator dictionary detects all four abnormal observations.

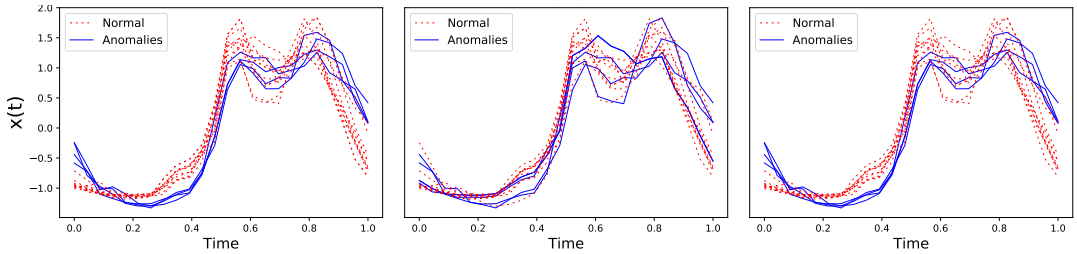


Figure 5.5 – The "Chinatown" data set, with normal observations in red and anomalies in blue: The data and the true anomalies (left), anomalies detected using the Mexican hat wavelet dictionary (middle), and anomalies detected using the dyadic indicator dictionary (right).

Before, we were considering dictionaries that are independent of data. Nevertheless one can use observations or their certain transform as a dictionary itself: projections on both normal and abnormal observations shall differ for normal ones and for anomalies; this suggests the *self-data dictionary* (Self). This can be extended to the *local self-data*

dictionary which consists of the product of the self-data dictionary with the uniform indicator dictionary. As an example, we apply this to the "ECG5000" data set plotted in Figure 5.6, where, different to the cosine dictionary, it allows to detect all abnormal observations.

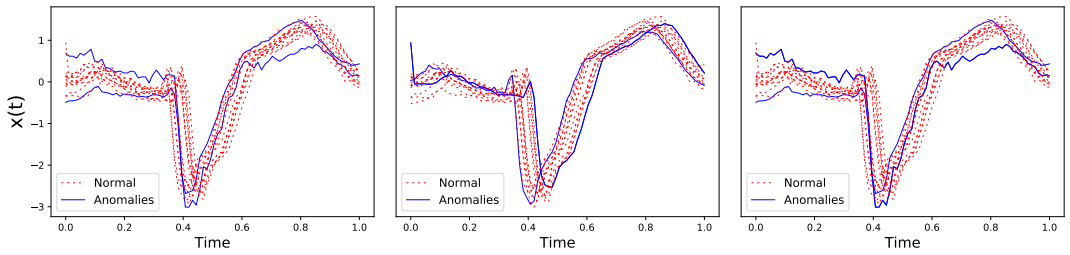


Figure 5.6 – The "ECG5000" data set, with normal observations in red and anomalies in blue: The data and the true anomalies (left), anomalies detected using the cosine dictionary (middle), and anomalies detected using the self data dictionary (right).

5.2.3 Scalar Product

Besides the dictionary, the *scalar product* defined on $\mathcal{H}$ brings some additional flexibility to measure different types of anomalies. While L_2 scalar product allows for detection of *location anomalies*, L_2 scalar product of derivatives (or slopes) would allow to detect anomalies regarding shape. This last type of anomalies can be challenging; e.g. Hubert et al. (2015) mention that *shape anomalies* are more difficult to detect, and Mosler and Mozharovskyi (2017) argue that one should consider both location and slope simultaneously for distinguishing complex curves. Beyond these two, a wide diversity of scalar products can be used, involving a variety of L_2-scalar products related to derivatives of certain orders, like in the definition of Banach spaces such as weighted Sobolev spaces, see Maz'ya (2011). To illustrate this, we provide an example (see Figure 5.7) where we highlight the impact of the scalar product choice. To illustrate the score change caused by different values of α, we calculate the FIF anomaly scores with the scalar product introduced in Section 5.2.1 with $\alpha = 1$ and $\alpha = 0$ for a sample consisting of 100 curves as follows (inspired by Cuevas et al. (2007), see Figure 5.7):

- 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 $\zeta \sim \mathcal{N}(0, 0.3^2)$ on the interval $[0.2, 0.8]$.

One can see that even though the 10 noisy curves are abnormal for the majority of the data, they are considered as normal ones when only location is taken into account. On the other hand, they are easily distinguished with the high anomaly score when derivatives are examined.

5.2.4 Direction Importance of Finite Size Dictionaries

Although feature importance has been tackled in supervised random trees (see e.g. Breiman, 2001 or Geurts et al., 2006), this has not been adressed in the *Isolation Forest* literature (see Liu et al., 2008, Liu et al., 2012 and Hariri et al., 2019). As a very randomized procedure, there is no incrementally way to define feature importance

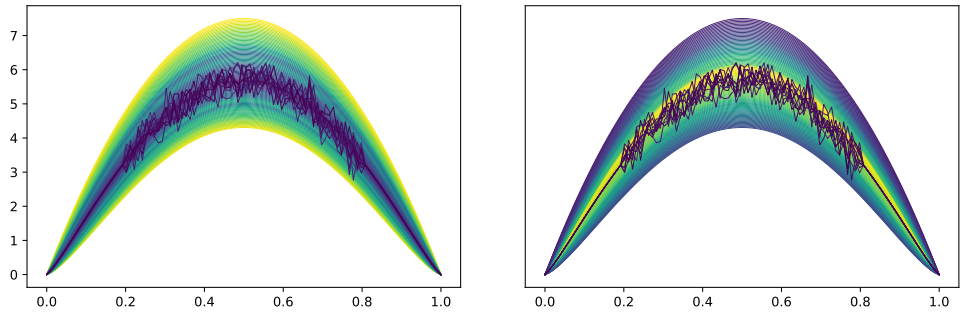


Figure 5.7 – FIF anomaly scores for a sample of 100 curves with $\alpha = 1$ (left) and $\alpha = 0$ (right). Anomaly score increases from purple to yellow in the left plot and decreases in the right plot.

from the supervised setting. Nevertheless, it is a matter of interest in many anomaly detection applications to get interpretability of models, especially when dealing with functional data where many information are contained in curves. Thus, it is rewarding to get an *a posteriori* sparse representation of the dictionary $\mathcal{D}$ which corresponds to the discriminating directions that have great importance in the construction of the model. Furthermore, it could bring some information on the distribution of normal data by studying the dispersion of the projection coefficients on a direction $\mathbf{d}$ (e.g. multi-modality). To extend this notion to the *Functional Isolation Forest* algorithm, we propose two ways to evaluate the importance of the elements of $\mathcal{D}$ to discriminate anomaly curves. The general idea is to give importance to elements of $\mathcal{D}$ which allows to discriminate between the sample. The naive idea is to add "+1" to the elements of $\mathcal{D}$ where an instance of the node sample is isolated (except for the cells with only two instances) such that good directions are those with a high score (after the forest construction). A clever one, more adaptive, would be to get weighted gain since curves isolated at nodes closer to the root should be more rewarding. To do this, we choose to give a reward depending on the size of the sample node where a curve is isolated. Precisely, the given reward is equal to the size of the node sample divided by the (sub)-sample used to build the tree. An example of the latter is given in Figure 5.8. The experiment is conducted on the real-world "CinECGTorso" data set. We use FIF with the *Dyadic indicator dictionary* and the L_2 scalar product. As we can see, the two most important elements of the dictionary are indicator functions which localize the peak around $t = 0.4$ where anomalies are really different from the normal ones. These leads to some interpretability from a "black-box" procedure.

5.3 Numerical Results

In this section, we provide an empirical study of the proposed algorithm. First, in Section 5.3.1 we explore the stability and consistency of the score function w.r.t. the probability distribution of a r.v. $\mathbf{X}$ and the sample size. Furthermore, we examine the influence of proposed dictionaries on the score function and bring performance comparisons with benchmark methods. Second, in Section 5.3.2, we benchmark the performance of FIF on several real labeled data sets by measuring its ability to recover an "abnormal" class on the test set. In all experiments, N the number of F-$i$trees is fixed to 100 and the height limit is fixed to $\lceil \log_2(n_s) \rceil$.

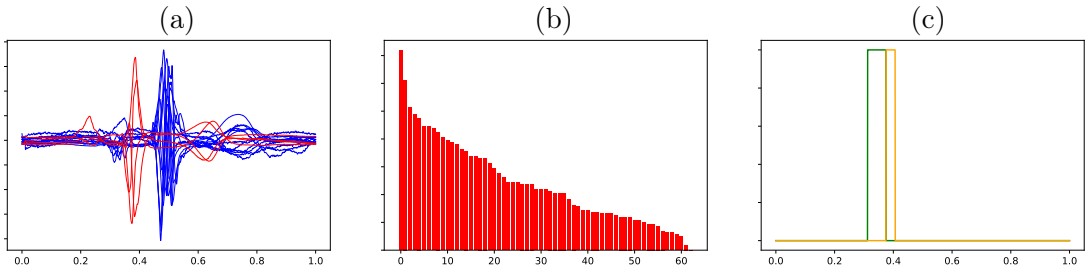


Figure 5.8 – "CinECGTorso" training data set (a), red curves correspond to anomalies while blue curves to normal data. The direction importances given by the "adaptative" way are represented by (b) and the two most important functions (from the dyadic dictionary) used by FIF to build the model are plotted in (c).

5.3.1 Impact of the Hyperparameters on Stability

Since functional data are more complex than multivariate data, and the dictionary constitutes an additional source of variance, the question of stability of the FIF anomaly score estimates is of high interest. This issue is even more important because of the absence of theoretical developments due to their challenging nature. The empirical study is conducted on two simulated functional data sets presented in Figure 5.9: Data set (a) is the standard Brownian motion being a classical stochastic process widely used in the literature. Data set (b) has been used by Claeskens et al. (2014) and has smooth paths. For each data set, we choose/add four observations for which the FIF anomaly score is computed after training: a normal observation $\mathbf{x}_0$, two anomalies $\mathbf{x}_1$ and $\mathbf{x}_2$, and a more extreme anomaly $\mathbf{x}_3$. We therefore expect the following ranking of the scores: $s_n(\mathbf{x}_0) < s_n(\mathbf{x}_1) \leq s_n(\mathbf{x}_2) < s_n(\mathbf{x}_3)$, for both data sets.

Further, we provide an illustration of the empirical convergence of the score. All other parameters being fixed, we increase the number of observations n when calculating the scores of the four selected observations; the empirical median and the boxplots of the scores computed over 100 random draws of the data set are shown in Figure 5.10. First, one observes score convergence and variance decrease in n. Further, let us take a closer look at the score tendencies on the example of $\mathbf{x}_0$ and $\mathbf{x}_3$. The score of $\mathbf{x}_3$ first increases (for data set (a)) and slightly decreases (for data set (b)) with growing n until n reaches $n_s = 64$, which happens because this abnormal observation is isolated quite fast (and

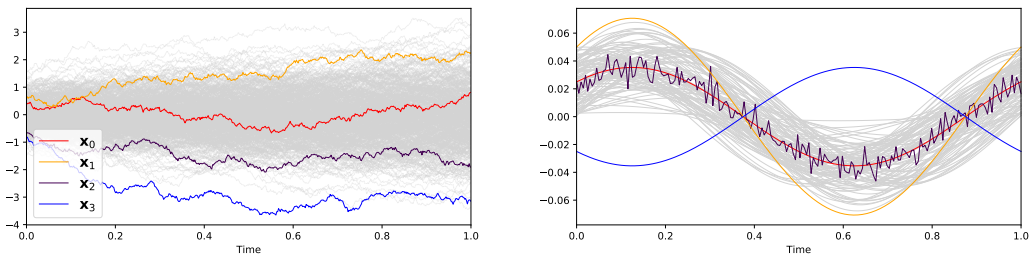


Figure 5.9 – Data sets (a) (left) and (b) (right) containing, respectively, 500 and 200 functional paths with 4 selected observations.

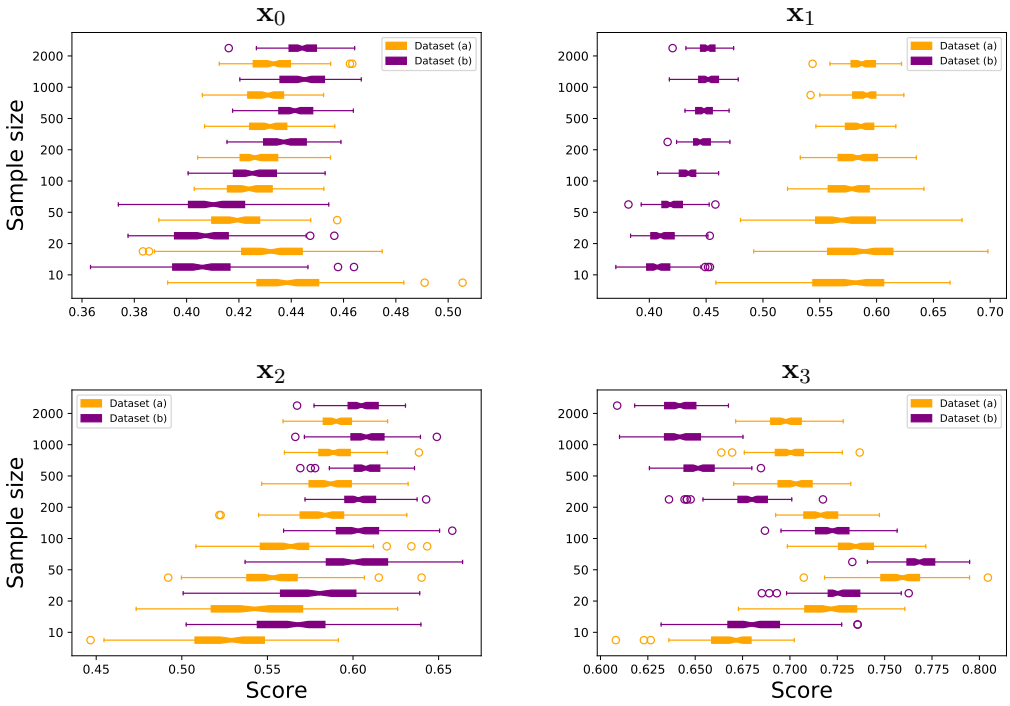


Figure 5.10 – Boxplot (over 100 repetitions) of the FIF score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for different sample sizes. The orange boxplots represent the data set (a) while the purple boxplots represent the data set (b).

thus has short path length) but the $c(n_s)$ in the denominator of the exponent of (5.1) increases in n_s. For $n > 64$, the score of $\mathbf{x}_3$ decreases in n since $h_i(\mathbf{x}_3)$ overestimates the real path length of $\mathbf{x}_3$ for subsamples in which it is absent; frequency of such subsamples grows in n and equals, e.g. 0.872 for $n = 500$. On the other hand, this phenomenon allows to unmask grouped anomalies as mentioned in Liu et al. (2008). The behavior is reciprocal for the typical observation $\mathbf{x}_0$. Its FIF anomaly score starts by decreasing in n since $\mathbf{x}_0$ tends to belong to the deepest branches of the trees and is always selected while $b < n$. For larger n, the path length of $\mathbf{x}_0$ is underestimated for subsamples where it is absent when growing the tree, which explains slight increase in the score before it stabilizes. A second experiment illustrated in Figure 5.11 is conducted to measure the impact of various dictionaries; L_2 scalar product is used. One observes that the variance of the score seems to be mostly stable across dictionaries, for both data sets. Thus, random dictionaries like *uniform indicator* (UI) or *Brownian motion* (B) do not introduce additional variance into the FIF score. Since we know the expected ranking of the scores, we can observe that FIF relying on the Self, UI, and DI dictionaries fail to make a strong difference between $\mathbf{x}_0$ and $\mathbf{x}_1$. Since $\mathbf{x}_1$ differs only slightly in the amplitude from the general pattern, these dictionaries seem insufficient to capture this fine dissimilarity: while Self and DI dictionaries simply do not contain enough elements, UI dictionary is too simple to capture this difference (it shares this feature with DI dictionary). For the scalar product L_2 on derivatives (see Figure A.9 in the Section A of Appendices), distinguishing anomalies for the Brownian motion becomes difficult since they differ mainly in location, while for a sine function the scores resemble those with the usual L_2 scalar product. Thus, even though—as seen in Section 5.2.1—capturing

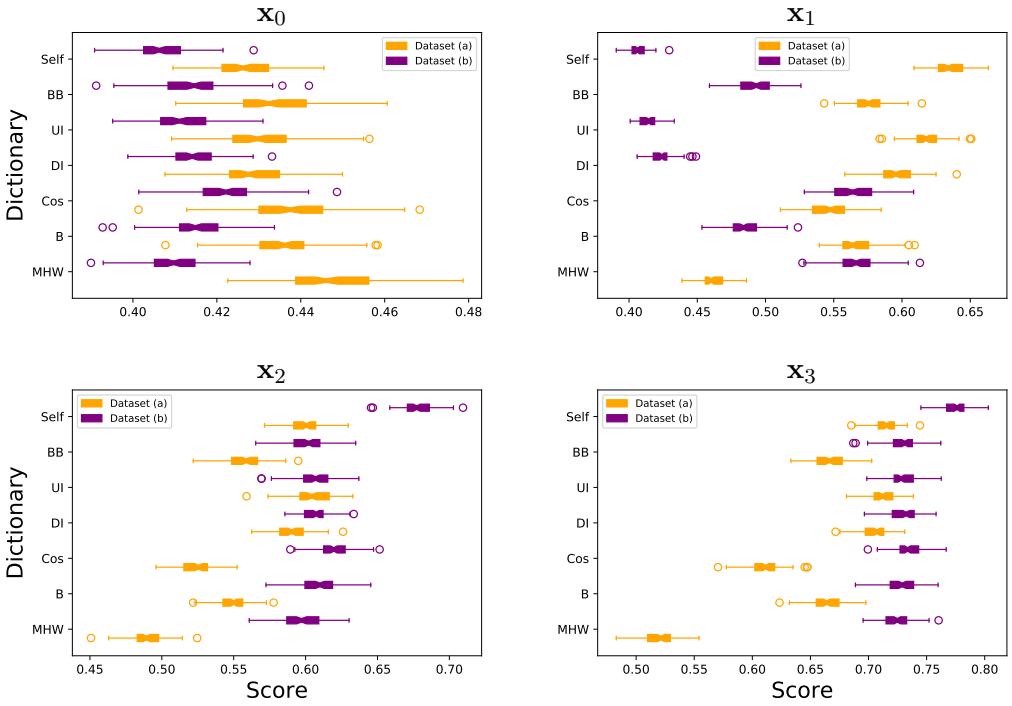


Figure 5.11 – Boxplot (over 100 repetitions) of the FIF score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for different dictionaries using the L_2 scalar product. The orange boxplots represent the data set (a) while the purple boxplots represent the data set (b).

different types of anomalies is one of the general strengths of the FIF algorithm, the dictionary may still have an impact on detection of functional anomalies in particular cases.

More experiments were run regarding the stability of the algorithm, but for sake of clarity, we describe them in the Section A of Appendices.

5.3.2 Real Data Benchmarking

To explore the performance of the proposed FIF algorithm, we conduct a comparative study using 13 classification data sets from the UCR repository (Chen et al., 2015b). We consider the larger class as normal and some of others as anomalies (see Table 5.1 for details). When classes are balanced, i.e. for 9 data sets out of 13, we keep only part of the anomaly class to reduce its size, always taking the same observations (at the beginning of the table) for a fair comparison. Since the data sets are already split into train/test sets, we use the train part (without labels) to build the FIF and compute the score on the test set. We assess the performance of the algorithm by measuring an Area Under the Receiver Operation Characteristic curve (AUC) on the test set. Both train and test sets are rarely used during learning in unsupervised setting since labels are unavailable when fitting the model. Thus, when fitting the models on unlabeled training data, good performances on the test set show a good generalization power.

Competitors. FIF is considered with two finite size dictionaries *dyadic indicator*, *the self-data* and the infinite size dictionary *cosines* (with $\alpha = 1$ and $\alpha = 0$); its parameters are set $N = 100$, $n_s = \min(256, n)$ and the height limit to $= \lceil \log_2(n_s) \rceil$). We contrast

the FIF method with three most used multivariate anomaly detection techniques and two functional depths, with default settings. The multivariate methods—isolation forest (IF; Liu et al., 2008), local outlier factor (LOF; Breunig et al., 2000), and one-class support vector machine (OCSVM; Schölkopf et al., 2001)— are employed after dimension reduction by Functional PCA keeping 20 principal components with largest eigenvalues after a preliminary step of filtering using Haar basis. The depths are the random projection halfspace depth (fT; Cuevas et al., 2007) and the functional Stahel-Donoho outlyingness (=functional projection depth) (fSDO; Hubert et al., 2015).

Analysis of the results. Taking into account the complexity of the functional data, as expected there is no method performing generally best. Nevertheless, FIF performs well in most of the cases, giving best results for 10 data sets and second best for 6 data sets. It is worth to mention that the dictionary plays an important role in identifying anomalies, while FIF seems to be rather robust w.r.t. other parameters: The "CinECGTorso" data set contains anomalies differing in location shift which are captured by the cosine dictionary. Dyadic indicator dictionary allows to detect local anomalies in "TwoLeadECG" and "Yoga" data sets. Self-data dictionary seems suited for Data sets "SonyRobotAI2" and "StarlightCurves" whose challenge is to cope with many different types of anomalies.

	p	training : $n_a/\ n$	testing : n_a/n	normal classes	abnormal classes
Chinatown	24	4 / 14 (29%)	95 / 345	2	1
Coffee	286	5 / 19 (26%)	6 / 19	1	0
ECGFiveDays	136	2 / 16 (12%)	53 / 481	1	2
ECG200	96	31 / 100 (31%)	36 / 100	1	-1
Handoutlines	2709	362 / 1000 (36 %)	133 / 370	1	0
SonyRobotAI1	70	6 / 20 (30 %)	343 / 601	2	1
SonyRobotAI2	65	4 / 20 (20 %)	365 / 953	2	1
StarLightCurves	1024	100 / 673 (15 %)	3482 / 8236	3	1 and 2
TwoLeadECG	82	2 / 14 (14 %)	570 / 1139	1	2
Yoga	426	10 / 173 (06 %)	1393 / 3000	2	1
EOGHorizontal	1250	10 / 40 (25 %)	30 / 61	5	6
CinECGTorso	1639	4 / 16 (25 %)	345 / 688	3	4
ECG5000	140	31 / 323 (10 %)	283 / 2910	1	3,4 and 5

Table 5.1 – Data sets considered in performance comparison: n is the number of instances, n_a is the number of anomalies. p is the number of discretization points.

5.4 Extensions of FIF

In this section, we present an extension of FIF to the case of multivariate functional data as well as highlight connections to data depth.

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 5.2 – AUC of different anomaly detection methods calculated on the test set. Bold numbers correspond to the best result while italics to the second best.

5.4.1 Extension to Multivariate Functions

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:

$$\begin{aligned}\mathbf{X} : \Omega &\longrightarrow (\mathcal{H}([0,1]))^{\otimes d} \\ \omega &\longmapsto \left((\mathbf{X}^{(1)}(\omega))_{t\in[0,1]}, \ldots, (\mathbf{X}^{(d)}(\omega))_{t\in[0,1]}\right)\end{aligned}$$

For this, the coordinate-wise sum of the d corresponding scalar products is used to project the data onto a chosen dictionary element:

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

Further, a dictionary should be defined in $(\mathcal{H}([0,1]))^{\otimes d}$. This can be done, e.g. by either component-wise application of one or several univariate dictionaries from Section 5.2.2, or by constructing of special d-variate ones. For illustration purposes, regard the following example constructed based on the MNIST (Lecun et al., 1998) data set. First, we extract the digits' contours (skeletons) using skimage Python library (van der Walt et al., 2014). Then each observation is transformed into a curve in $(L_2([0,1]) \times L_2([0,1]))$ (one vertical and one horizontal coordinates) using length parametrization on $[0,1]$. We construct the problem by taking 100 curves from class 7 and adding 10 observations from class 2. We apply FIF with two-dimensional *sinuscosine* dictionary and the following scalar product : $\langle \mathbf{f}, \mathbf{g} \rangle_{(L_2)^{\otimes d}}$. *sinuscosine* is constructed as a direct extension of *cosine* dictionary introduced for FIF by selecting randomly cosine or sinus function on each

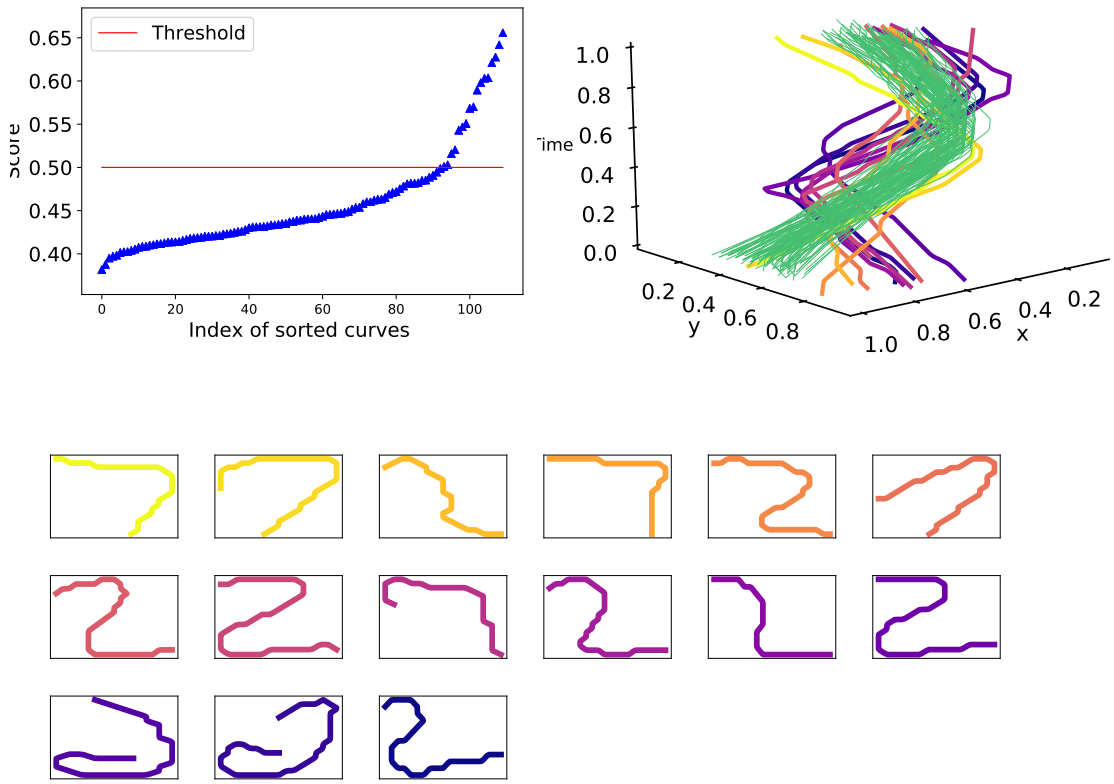


Figure 5.12 – FIF anomaly scores for a sample of 110 digits (100 seven and 10 two). Left plot corresponds to the sorted score of these curves. Right plot represents the digits in three dimensions, green ones correspond to normal data, anomalies score increases from orange to dark red. Bottom plot shows the fifteen detected anomalies.

coordinates. Figure 5.12 shows anomaly detection using the visual elbow rule to define the threshold. Among those detected, five digits are indeed 7s, but do not resemble them and thus are identified as anomalies.

5.4.2 Connection to Data Depth

Regarding FIF score as an anomaly ranking yields a connection to the notion of the *statistical depth function*, which has been successfully applied in outlier detection (see, e.g. Hubert et al., 2015). Statistical data depth has been introduced as a measure of centrality (or depth) of an arbitrary observation $\mathbf{x} \in (\mathcal{H}([0,1]))^{\otimes d}$ with respect to the data at hand $\mathcal{S}$. A data depth measure based on FIF score can be defined for (multivariate) functional data as: $FD_{FIF}(\mathbf{x};\mathcal{S}) = 1 - s_n(\mathbf{x};\mathcal{S})$. Data depth proves to be a useful tool for a low-dimensional data representation called *depth-based map*. Using this property, Li et al. (2012) defined a DD-plot classifier which consists in applying a multivariate classifier to the depth-based map. Low-dimensional representation is of particular interest for functional data and a DD-plot classifier can be defined using the FIF-based data depth. Let $\mathcal{S}^{trn} = \mathcal{S}^1 \cup ... \cup \mathcal{S}^q$ be a training set for supervised classification containing q classes, each subset $\mathcal{S}^j$ standing for class j. The depth map

is defined as follows:

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

As an illustration, we apply the depth map to 3 digits (1, 5 and 7, 100 observations per digit for training and 100 testing) of the MNIST data set after their transformation to two-variate functions using `skimage` Python library (see Figure 5.13). One observes appealing geometrical interpretation (observe, e.g. the location of the abnormally distant—from their corresponding classes—observations) and a clear separation of the classes. To illustrate separability, we apply linear multiclass (one-against-all) SVM in the depth space, which delivers the accuracy of 99% on the test data.

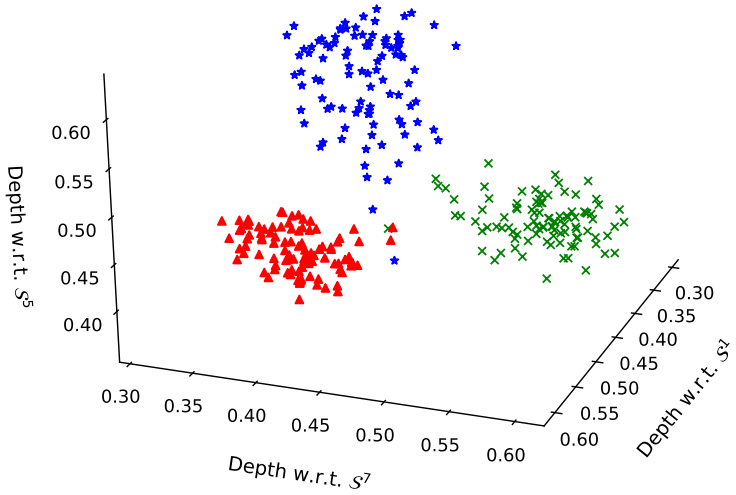


Figure 5.13 – Depth space embedding of the three digits (1, 5 and 7) of the MNIST data set.

5.5 Conclusion

The Functional Isolation Forest algorithm has been proposed, which is an extension of Isolation Forest to functional data. The combined choice of the dictionary itself, the probability distribution used to pick a *Split variable* and the scalar product used for the projection enables FIF to exhibit a great flexibility in detecting anomalies for a variety of tasks. FIF is extendable to multivariate functional data. When transformed in a data depth definition, FIF can be used for supervised classification via a low-dimensional representation—the depth space. The open-source Cython/C++ implementation of the method, along with all reproducing scripts, can be accessed at https://github.com/GuillaumeStaermanML/FIF.

6

The Area of the Convex Hull of Sampled Curves: a Robust Functional Statistical Depth Measure

Contents

Supported by sound theoretical and computational developments in the recent decades, data depth has proven to be extremely useful, in particular in functional spaces. However, most approaches documented in the literature consist in evaluating independently the centrality of each point forming the time series and consequently exhibit a certain insensitivity to possible shape changes. In this chapter, we propose a novel notion of functional depth based on the area of the convex hull of sampled curves, capturing gradual departures from centrality, even beyond the envelope of the data, in a natural fashion. We discuss practical relevance of commonly imposed axioms on functional depths and investigate which of them are satisfied by the notion of depth we promote here. Estimation and computational issues are also addressed and various numerical experiments provide empirical evidence of the relevance of the approach proposed.

The chapter is organized as follows. In Section 6.1 the functional statistical depth based on the area of the convex hull of a batch of curves is introduced at length and its theoretical properties are investigated, together with computational aspects. Section 6.2 presents numerical results in order to provide strong empirical evidence of the relevance of the novel depth function proposed, for the purpose of unsupervised functional anomaly detection especially. Eventually, concluding remarks are collected in Section 6.3. Technical proofs are deferred to Section 6.4. This chapter covers the contribution of:

- **G. Staerman**, P. Mozharovskyi, S. Clémençon. The Area of the Convex Hull of Sampled Curves: a Robust Functional Statistical Depth measure. In *Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 570-579, 2020.

6.1 The Area of the Convex Hull of (Sampled) Curves

In this section, we present the statistical depth function we propose for path-valued random variables. As shall be seen below, its definition is based on very simple geometrical ideas and various desirable properties can be easily checked from it. Statistical and computational issues are also discussed at length. By λ_2 is meant the Lebesgue measure on the plane $\mathbb{R}^2$. The graph of any function $\mathbf{x}$ in $\mathcal{C}([0,1])$ is denoted by

$$\mathrm{graph}(\mathbf{x}) = \{(t, y) \,:\, y = \mathbf{x}(t),\ t \in [0,1]\},$$

while we denote by $\mathrm{graph}(\{\mathbf{x}_1, \ldots, \mathbf{x}_n\})$ the set $\bigcup_{i=1}^n \mathrm{graph}(\{\mathbf{x}_i\})$ defined by a collection of $n \geq 1$ functions $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in $\mathcal{C}([0,1])$. We now give a precise definition of the statistical depth measure we propose for random variables valued in $\mathcal{C}([0,1])$.

Definition 6.1. *Let $J \geq 1$ be a fixed integer. The ACH depth of degree J is the function $FD^J_{\mathrm{ACH}} : \mathcal{C}([0,1]) \times \mathcal{P}(\mathcal{C}([0,1])) \longrightarrow [0,1]$ defined by: $\forall\, \mathbf{x} \in \mathcal{C}([0,1])$,*

$$FD^J_{\mathrm{ACH}}(\mathbf{x}, \mathbf{P}) = \mathbb{E}\left[\frac{\lambda_2(\mathrm{conv}(\mathrm{graph}(\{\mathbf{X}_1, \ldots, \mathbf{X}_J\})))}{\lambda_2(\mathrm{conv}(\mathrm{graph}(\{\mathbf{X}_1, \ldots, \mathbf{X}_J\} \cup \{\mathbf{x}\})))}\right],$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_J$ are i.i.d.ŕ.v.ś drawn from $\mathbf{P}$ with the convention $0/0 = 1$. Its average version $\overline{FD}^J_{\mathrm{ACH}}$ is defined by: $\forall\, \mathbf{x} \in \mathcal{C}([0,1])$,

$$\overline{FD}^J_{\mathrm{ACH}}(\mathbf{x}, \mathbf{P}) = \frac{1}{J}\sum_{j=1}^{J} FD^j_{\mathrm{ACH}}(\mathbf{x}, \mathbf{P}).$$

The choice of J leads to various views of distribution $\mathbf{P}$, the average variant permitting to combine all of them (up to degree J). Consider an i.i.d.sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ drawn from $\mathbf{P}$ in $\mathcal{P}(\mathcal{C}([0,1]))$ and $\mathbf{P}_n = (1/n)\sum_{i=1}^n \delta_{\mathbf{X}_i}$ its associated empirical measure. When $n \geq J$, an unbiased statistical estimation of $FD^J_{\mathrm{ACH}}(\mathbf{x}, \mathbf{P})$ can be obtained by computing the symmetric U-statistic of degree J, see Lee (1990): $\forall\, \mathbf{x} \in \mathcal{C}([0,1])$,

$$FD^J_{\mathrm{ACH}}(\mathbf{x}, \mathbf{P}_n) = \frac{1}{\binom{n}{J}} \sum_{1 \leq i_1 < \ldots < i_J \leq n} \frac{\lambda_2(\mathrm{conv}(\mathrm{graph}(\{\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_J}\})))}{\lambda_2(\mathrm{conv}(\mathrm{graph}(\{\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_J}, \mathbf{x}\})))}. \tag{6.1}$$

Considering the empirical average version given by

$$\forall\, \mathbf{x} \in \mathcal{C}([0,1]),\ \ \overline{FD}^J_{\mathrm{ACH}}(\mathbf{x}, \mathbf{P}_n) = \frac{1}{J}\sum_{j=1}^{J} FD^j_{\mathrm{ACH}}(\mathbf{x}, \mathbf{P}_n).$$

brings some "stability". However, the computational cost rapidly increasing with J, small values of J are preferred in practice. Moreover, as we illustrate in Section 6.2.1, J equals to two already yields satisfactory results.

Approximation from sampled curves. In general, one does not observe the batch of continuous curves $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ on the whole time interval $[0,1]$ but at discrete time points only, the number $p \geq 1$ of time points and the time points $0 \leq t_1 < t_2 < \ldots < t_p \leq 1$ themselves possibly varying depending on the curve considered. In such a case, the estimators above are computed from continuous curves reconstructed from the sampled curves available by means of interpolation procedures or approximation schemes based on appropriate basis. In practice, linear interpolation is used for this purpose with theoretical guarantees (refer to Theorem 6.4 below) facilitating significantly the computation of the empirical ACH depth, see Section 6.2.4.

6.1.1 Main Properties of the ACH Depth

Here, we study theoretical properties of the population version of the functional depths introduced above and next establish the consistency of their statistical versions. The following result reveals that, among the properties listed in Section 3.1, five are fulfilled by the (average) ACH depth function.

Proposition 6.2. *For all $J \geq 1$, the depth function FD^J_{ACH} (respectively, $\overline{FD}^J_{\text{ACH}}$) fulfills the following properties: (i) 'NON-DEGENERACY', (ii) 'SCALAR-AFFINE INVARIANCE', (iii) 'VANISHING AT INFINITY', (iv) 'CONTINUITY IN $\mathbf{x}$' and (v) 'CONTINUITY IN $\mathbf{P}$'. In addition, the following properties are not satisfied: (vi) 'MAXIMALITY AT CENTER' and (vii) 'DECREASING W.R.T. THE DEEPEST POINT'.*

Technical proofs are detailed in Section 6.4.2. In a functional space, not satisfying MAXIMALITY AT CENTER is not an issue. For instance, though the constant trajectory $\mathbf{y}(t) \equiv \mathbf{0}$ is a center of symmetry for the Brownian motion, it is clearly not representative of this distribution. However, the set of depth medians is not always meaningful and may be disjoint of the convex hull of the data. For example, assuming that $\mathbf{X}$ is a random variable defined by $\mathbf{X}(t) = t(1-t)Z$ for a real random variable $Z > 1$ a.s., then the set of depth medians is $\{(x, y) \in \mathbb{R}^2 : 0 < x < 1 \text{ and } 0 < y < x\ (1-x)\}$. In contrast, SCALAR-AFFINE INVARIANCE is relevant, insofar as it allows z-normalization of the functional data and CONTINUITY IN $\mathbf{P}$ is essential to derive the consistency of $FD^J_{\text{ACH}}(\cdot, \mathbf{P}_n)$ (respectively, of $\overline{FD}^J_{\text{ACH}}(\cdot, \mathbf{P}_n)$), as stated below.

Theorem 6.3. *Let $J \geq 1$ and $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be $n \geq J$ independent copies of a generic r.v. $\mathbf{X}$ with distribution $\mathbf{P} \in \mathcal{P}(\mathcal{C}([0,1]))$. As $n \to \infty$, we have, for any $\mathbf{x} \in \mathcal{C}([0,1])$, with probability one,*

$$\left| FD^J_{\text{ACH}}(\mathbf{x}, \mathbf{P}_n) - FD^J_{\text{ACH}}(\mathbf{x}, \mathbf{P}) \right| \longrightarrow 0,$$

and

$$\left| \overline{FD}^J_{\text{ACH}}(\mathbf{x}, \mathbf{P}_n) - \overline{FD}^J_{\text{ACH}}(\mathbf{x}, \mathbf{P}) \right| \longrightarrow 0.$$

Proof. For any $1 \leq j \leq J$, the term $|FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{P}_n) - FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{P})|$ goes to zero almost-surely by U-statistics consistency (see e.g. Hoeffding, 1961). Thus,

$$\mathbb{P}\left(\forall\, j : \left| FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{P}_n) - FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{P}) \right| \overset{n\to\infty}{\longrightarrow} 0 \right) = 1,$$

which is equivalent to

$$\mathbb{P}\left(\sum_{j=1}^{J} \left| FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{P}_n) - FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{P}) \right| \overset{n\to\infty}{\longrightarrow} 0 \right) = 1.$$

By triangle inequality, for any $\mathbf{x} \in \mathcal{C}([0,1])$,

$$\left| \sum_{j=1}^{J} FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{P}_n) - FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{P}) \right| \leq \sum_{j=1}^{J} \left| FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{P}_n) - FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{P}) \right|$$

which leads to the desired result. ■

Uniform consistency often relies on the equicontinuity of the set $\{FD^J_{\text{ACH}}(\mathbf{x},\mathbf{P}),\ \mathbf{x}\in\mathcal{F}\}$ where $\mathcal{F}$ is the functional space considered, which is not satisfied here. Indeed, let $J=1$, $\mathbf{x}_1=\mathbf{0}$, $\varepsilon=1/4$ and $\delta>0$. Let $\mathbf{y}_1\in\mathcal{C}([0,1])$ such that

$$\mathbf{y}_1(t)=2\delta t\ \mathbb{I}\{[0,0.5]\}+(2\delta-\delta t)\ \mathbb{I}\{[0,0.5]\}.$$

It follows that $||\mathbf{x}_1-\mathbf{y}_1||_\infty=\delta$. Now, taking $\mathbf{x}(t)=2\delta t$ we have $\Phi_{\mathbf{x}}(\mathbf{x}_1)=0$ and $\Phi_{\mathbf{x}}(\mathbf{y}_1)=1/2$ where the function $\Phi_{\mathbf{x}}$ is defined as:

$$\begin{aligned}\Phi_{\mathbf{x}}:\qquad (\mathcal{C}([0,1]))^{\otimes j}\ &\longrightarrow\ [0,1]\\ (\mathbf{x}_1,...,\mathbf{x}_j)\ &\longmapsto\ \frac{\lambda_2(\text{conv}(\text{graph}(\{\mathbf{x}_1,...,\mathbf{x}_j\})))}{\lambda_2(\text{conv}(\text{graph}(\{\mathbf{x}_1,...,\mathbf{x}_j,\mathbf{x}\})))}.\end{aligned}$$

Then for any $\delta>0$ there exists a function $\mathbf{y}_1$ close to $\mathbf{x}$ such that $||\Phi_{\mathbf{x}}(\mathbf{x}_1)-\Phi_{\mathbf{x}}(\mathbf{y}_1)||_\infty>0.25$. Thus, the set $\{\Phi_{\mathbf{x}},\ \mathbf{x}\in\mathcal{C}([0,1])\}$ is not equicontinuous at $\mathbf{x}_1$.

6.1.2 On Statistical/Computational Issues

As mentioned above, only sampled curves are available in practice. Each random curve $\mathbf{X}_i$ being observed at fixed time points $0=t_1^{(i)}<t_2^{(i)}<\ldots<t_{p_i}^{(i)}=1$ (potentially different for each $\mathbf{X}_i$) with $p_i\geq 1$, we denote by $\widetilde{\mathbf{X}}_1,\ldots,\widetilde{\mathbf{X}}_n$ the continuous curves reconstructed from the sampled curves $(\mathbf{X}_i(t_1^{(i)}),\ldots,\mathbf{X}_i(t_{p_i}^{(i)}))$, $1\leq i\leq n$, by linear interpolation. The measure $\widetilde{\mathbf{P}}_n$ is the empirical measure associated to the reconstructed sample. From a practical perspective, one considers the estimator $FD^J_{\text{ACH}}(\mathbf{x},\widetilde{\mathbf{P}}_n)$ of $FD^J_{\text{ACH}}(\mathbf{x},\mathbf{P})$ given by the approximation of $FD_{\text{ACH}}(\mathbf{x},\mathbf{P}_n)$ obtained when replacing the $\mathbf{X}_i$'s by the $\widetilde{\mathbf{X}}_i$'s in (6.1). The (computationally feasible) estimator $\overline{FD}^J_{\text{ACH}}(\mathbf{x},\widetilde{\mathbf{P}}_n)$ of $\overline{FD}^J_{\text{ACH}}(\mathbf{x},\mathbf{P})$ is constructed in a similar manner. The result stated below shows that this approximation stage preserves almost-sure consistency.

Theorem 6.4. *Let $J\leq n$. Suppose that, as $n\to\infty$,*

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

As $n\to\infty$, we have, for any $\mathbf{x}\in\mathcal{C}([0,1])$, with probability one,

$$\left|FD^J_{\text{ACH}}(\mathbf{x},\widetilde{\mathbf{P}}_n)-FD^J_{\text{ACH}}(\mathbf{x},\mathbf{P})\right|\longrightarrow 0$$

and

$$\left|\overline{FD}^J_{\text{ACH}}(\mathbf{x},\widetilde{\mathbf{P}}_n)-\overline{FD}^J_{\text{ACH}}(\mathbf{x},\mathbf{P})\right|\longrightarrow 0.$$

Proof. The result follows from the continuity in $\mathbf{P}$ and Theorem 3 in Nagy et al. (2016a). ■

Given the batch of continuous and piecewise linear curves $\widetilde{\mathbf{X}}_1,\ldots,\widetilde{\mathbf{X}}_n$, although the computation cost of the area of their convex hull is of order $O(p\log p)$ with $p=\max_i p_i$, that of the U-statistic $FD^J_{\text{ACH}}(\mathbf{x},\widetilde{\mathbf{P}}_n)$ (and *a fortiori* that of $\overline{FD}^J_{\text{ACH}}(\mathbf{x},\widetilde{\mathbf{P}}_n)$) becomes

very expensive as soon as $\binom{n}{J}$ is large. As pointed out in López-Pintado and Romo (2009), even if the choice $J = 2$ for statistics of this type, may lead to a computationally tractable procedure, while offering a reasonable representation of the distribution, varying J permits to capture much more information in general. For this reason, we propose to compute an *incomplete* version of the U-statistic $FD^J_{\mathrm{ACH}}(\mathbf{x}, \widetilde{\mathbf{P}}_n)$ using a basic Monte-Carlo approximation scheme with $n_{\mathrm{rep}} \geq 1$ replications: rather than averaging over all $\binom{n}{J}$ subsets of $\{1, \ldots, n\}$ with cardinality J to compute $FD^J_{\mathrm{ACH}}(\mathbf{x}, \widetilde{\mathbf{P}}_n)$, one averages over $n_{\mathrm{rep}} \geq 1$ subsets drawn with replacement, forming an *incomplete U-statistic*, see Enqvist (1978). The same approximation approach can be applied (in a randomized manner) to each of the U-statistics involved in the average $\overline{FD}^J_{\mathrm{ACH}}(\mathbf{x}, \widetilde{\mathbf{P}}_n)$, as described in the Figure 6.1.

Input: the sample $\tilde{\mathcal{S}} = \{\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_n\}$, the observed curve $\widetilde{\mathbf{x}}$, n_{rep}, $1 \leq J \leq n$ and the vector of weights $\mathbf{w} = (w_j)_{1 \leq j \leq J}$ such that $\mathbf{w}_j = \dfrac{\binom{j}{n}}{\sum_{m=1}^{J} \binom{m}{n}}$.

1. For $k = 1, \ldots, n_{\mathrm{rep}}$ do:
   (i) Select $l \in \{1, \ldots, J\}$ according to $\mathbf{w}$.
   (ii) Select randomly and uniformly $(i_1, \ldots, i_l) \in \{1, \ldots, n\}$.
   (iii) $s(\widetilde{\mathbf{x}}) \leftarrow s(\widetilde{\mathbf{x}}) + \dfrac{\lambda_2(\mathrm{conv}(\mathrm{graph}(\{\widetilde{\mathbf{X}}_{i_1}, \ldots, \widetilde{\mathbf{X}}_{i_l}\})))}{\lambda_2(\mathrm{conv}(\mathrm{graph}(\{\widetilde{\mathbf{X}}_{i_1}, \ldots, \widetilde{\mathbf{X}}_{i_l}, \widetilde{\mathbf{x}}\})))}$.
2. **Return:** $\overline{FD}^J_{\mathrm{ACH}}(\widetilde{\mathbf{x}}, \widetilde{\mathcal{S}}) = s(\widetilde{\mathbf{x}}) / \, n_{\mathrm{rep}}$.

Figure 6.1 – The approximation procedure

6.2 Numerical Experiments

From a practical perspective, this section explores certain properties of the functional depth proposed using simulated data. It also describes its performance compared with the state-of-the-art methods on (real) benchmark data sets. As a first go, we focus on the impact of the choice of the tuning parameter n_{rep}, which rules the trade-off between approximation accuracy and computational burden and parameter J. Precisely, it is investigated through the stability of the ranking induced by the corresponding depths. We next investigate the robustness of the ACH depth (ACHD in its abbreviated form), together with its ability to detect abnormal observations of various types. A simulation-based study of the variance of the ACH depth is also provided. Finally, the ACH depth is benchmarked against alternative depths standing as natural competitors in the functional setup using real data sets.

For the sake of simplicity, the two same simulated data sets, represented in Figure 6.2, are used throughout the section. The data set (a) corresponds to sample path segments of the *geometric Brownian motion* with mean 2 and variance 0.5, a stochastic process widely used in statistical modeling. The data set (b) consists of smooth curves given by $\mathbf{x}(t) = u\cos(2\pi t) + v\sin(2\pi t)$, $t \in [0, 1]$, where u and v are independently

and uniformly distributed on $[0, 0.05]$, as proposed by Claeskens et al. (2014). Four curves $\{\mathbf{x}_i : i \in \{0, 1, 2, 3\}\}$ have been incorporated to each data set: a deep curve and three atypical curves (*anomalies*), with expected depth-induced ranking $FD^J_{\text{ACH}}(\mathbf{x}_3) < FD^J_{\text{ACH}}(\mathbf{x}_2) \approx FD^J_{\text{ACH}}(\mathbf{x}_1) < FD^J_{\text{ACH}}(\mathbf{x}_0)$.

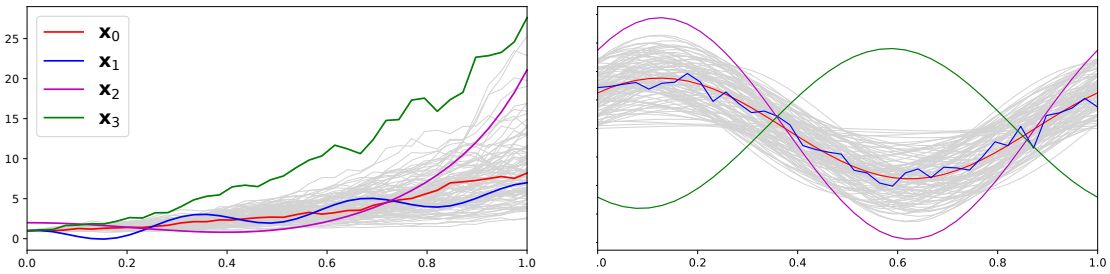


Figure 6.2 – Data sets (a) (left) and (b) (right) containing 100 paths with four selected observations. The colors are the same for the four selected observations of both data sets (a) and (b).

6.2.1 Choosing Tuning Parameters n_{rep} and J

Parameter n_{rep} reflects the trade-off between statistical performance and computational time. In order to investigate its impact on the stability of the method, we compute depths of the deepest and most atypical curves ($\mathbf{x}_0$ and $\mathbf{x}_3$) for data set (b), taking $J = 2, 3, 4$. Figure 6.3 presents boxplots of the approximated ACH depth (together with the exact values of ACH depth) over 100 repetitions. Note that, as expected, depth values grow with J. The variance of the depth decreases taking sufficiently small values for $n_{\text{rep}} = 5n$ and almost disappearing for $n_{\text{rep}} \geq 20n$, while decreasing pattern remains the same for different values of n_{rep}. For these reasons, we keep $n_{\text{rep}} = 5n$ in what follows.

The choice of J is less obvious, and clearly when describing an observation in a functional space a substantial part of information is lost anyway. Nevertheless, one observes that computational burden increases exponentially with J and thus smaller values are preferable. Figure 6.4 shows the rank-rank plots of data sets (a) and (b) for small values of $J = 2, 3, 4$ and indicates, that depth-induced ranking does not change much with J. Thus, for saving computational time, we use value $J = 2$ in all subsequent experiments.

6.2.2 Asymptotic Variance of the Exact and Approximate Versions

To obtain further insights about the stability of the proposed depth notion, we explore its asymptotic variance. For this, we compute (exact and approximate) ACH depth of $\mathbf{x}_i$, $i = 1, 2, 3, 4$ for different sample sizes. The boxplots over 100 repeated simulation for data sets (a) and (b) are indicated in Figure 6.5 and Figure 6.6. One observes not only stable decrease of the variance of ACH with the sample size, but also the similarity between exact and approximate versions, which hints on stability and precision of the exact algorithm even when exploring a small portion of combinations (e.g. when $n = 500$ only 2% of all pairs are explored for $n_{\text{rep}} = 5n$).

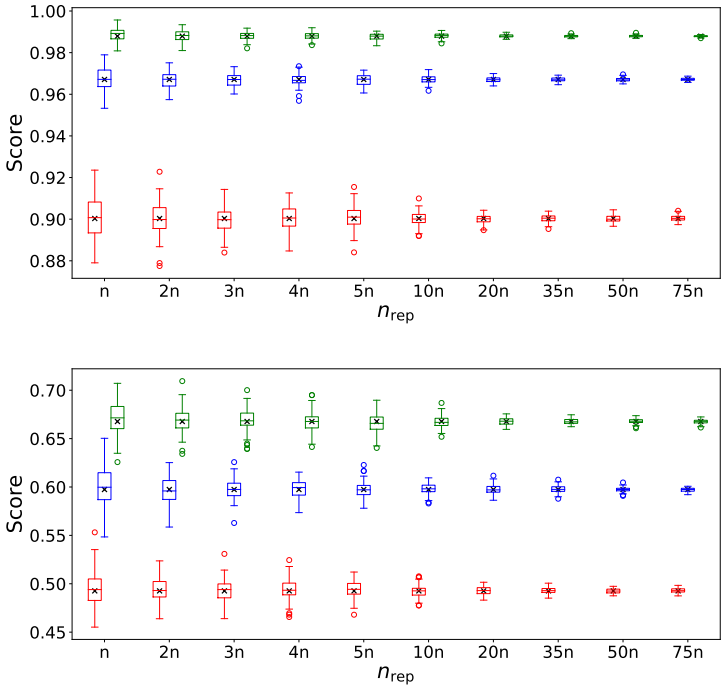


Figure 6.3 – Boxplots of the approximations of $FD^J_{\text{ACH}}(\mathbf{x}_0)$ (top) and $FD^J_{\text{ACH}}(\mathbf{x}_3)$ (bottom) over different size of n_{rep}. The black crosses correspond to the exact depth measure FD^J_{ACH} for each J respectively.

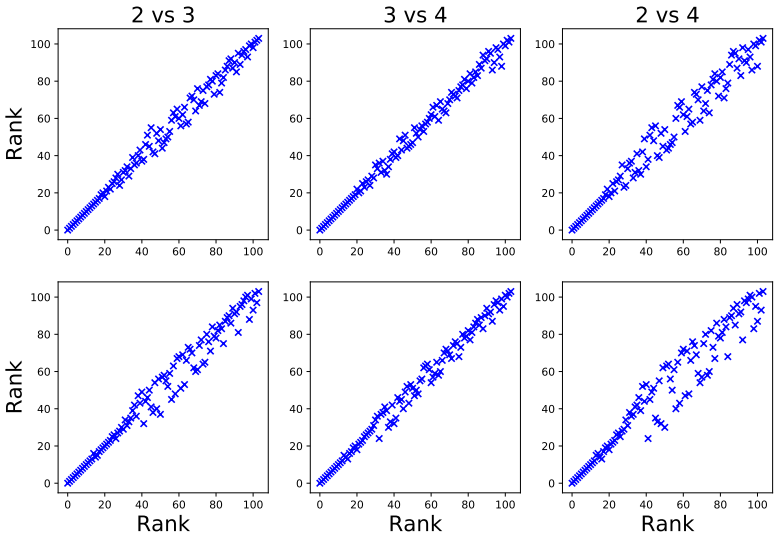


Figure 6.4 – Rank-Rank plot for different values of J (2, 3 and 4). The first line represents the rank over the data set (a) while the second line represents the data set (b).

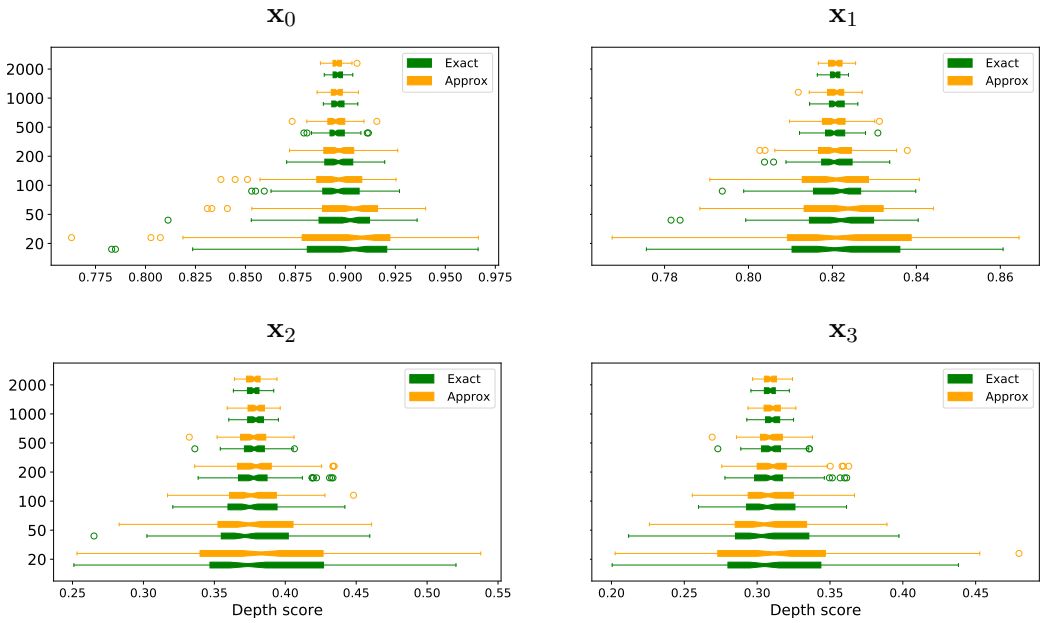


Figure 6.5 – Boxplot (over 100 repetitions) of the depth score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for the two following settings on the data set (a): the green boxplots represent the exact computation while the orange boxplots represent the approximation both with $J = 2$.

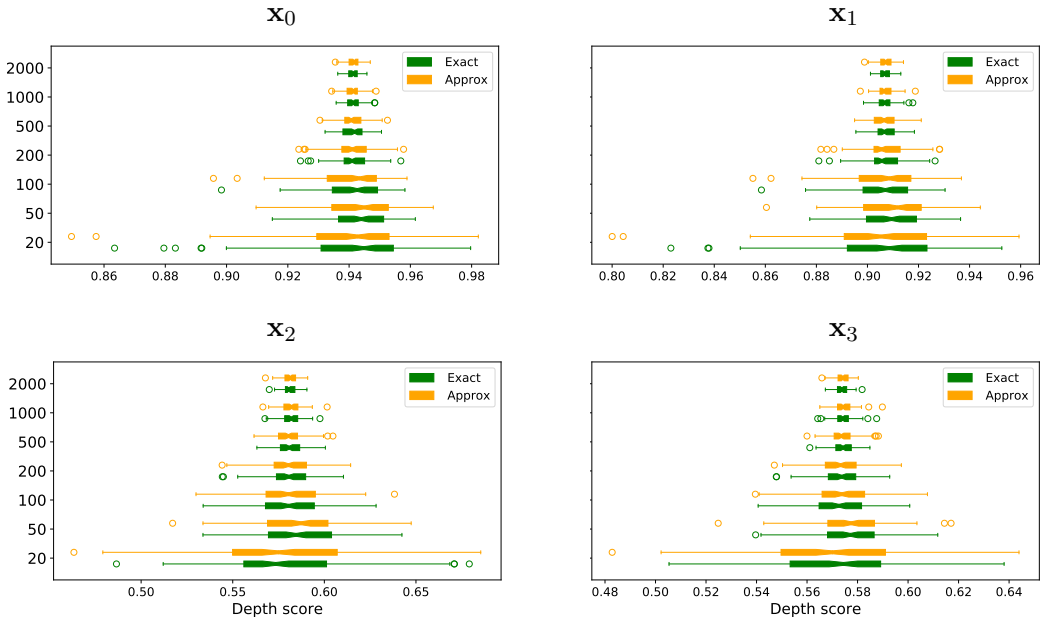


Figure 6.6 – Boxplot (over 100 repetitions) of the depth score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for the two following settings on the data set (b): the green boxplots represent the exact computation while the orange boxplots represent the approximation both with $J = 2$.

6.2.3 Robustness

Under robustness of a statistical estimator on understands its ability not to be "disturbed" by atypical observations. We explore robustness of the ACH depth in the following simulation study: between the original data set and the same data set contaminated with anomalies, we measure (averaged over 10 random repetitions) Kendall's τ distance of two depth-induced rankings σ and σ', respectively, of the original data:

$$d_\tau(\sigma, \sigma') = \frac{2}{n(n-1)} \sum_{i<j} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\}\,.$$

In their overview work, Hubert et al. (2015) introduce taxonomy for atypical observations, focusing on *location*, *isolated*, and *shape* anomalies. Here, we add *location* anomalies to data set (a) and *isolated* and *shape* anomalies to data set (b). The abnormal functions are constructed as follows. *Location* anomalies for data set (a) are $\breve{\mathbf{x}}(t) = \mathbf{x}(t) + a\mathbf{x}(t)$ with a drawn uniformly on $[0, 1]$. *Isolated* anomalies for data set (b) are constructed by adding a peak at t_0 (drawn uniformly on $[0, 1]$) of amplitude b (drawn uniformly on $[0.03, 0.06]$) such that $\breve{\mathbf{y}}(t_0) = \mathbf{y}(t_0) + b$ and $\breve{\mathbf{y}}(t) = \mathbf{y}(t)$ for any $t \neq t_0$. *Shape* anomalies for data set (b) are $\breve{\mathbf{z}}(t) = \mathbf{z}(t) + 0.01 \times \cos(2\pi t u) + 0.01 \times \sin(2\pi t u)$ with u drawn uniformly from $\{1, 2, \ldots, 10\}$. By varying the percentage of abnormal observations $\alpha_{\mathcal{O}}$, we compare ACHD to several of the most known in the literature depth approaches: the *functional Stahel-Donoho outlyingness* (fSDO) (= functional projection depth) (Hubert et al., 2015) and the *functional Tukey depth* (fT) (= functional halfspace depth) (Claeskens et al., 2014), and also to the *functional isolation forest* (FIF) algorithm (see Chapter 5) which proves satisfactory anomaly detection; see Table 6.1. One can observe that ACH consistently preserves depth-induced ranking despite inserted abnormal observation, even if their fraction $\alpha_{\mathcal{O}}$ reaches 30%. fSDO behaves competitively giving slightly better results than ACH for shape anomalies.

$d_\tau(\sigma_0, \sigma_{\alpha_{\mathcal{O}}})(\times 10^{-2})$							
	$\boldsymbol{\alpha_{\mathcal{O}}}$	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

Table 6.1 – Kendall's tau distances between the rank returned with normal data (σ_0) and contamined data ($\sigma_{\alpha_{\mathcal{O}}}$, over different portion of contamination $\alpha_{\mathcal{O}}$ with location, isolated and shape anomalies) for ACH and three state-of-the-art methods. Bold numbers indicate best stability of the rank over the contaminated data sets.

6.2.4 Applications to Anomaly Detection

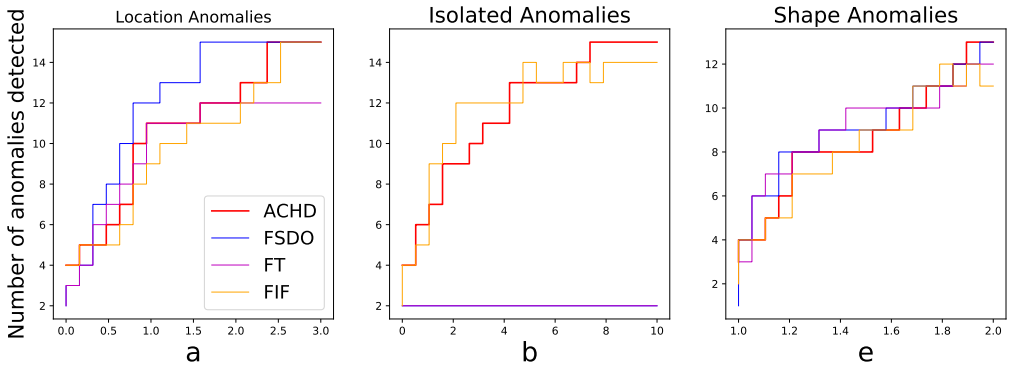


Figure 6.7 – Number of anomalies detected over a grid of parameters for three types of anomalies (location, isolated, and shape) for ACH and three further state-of-the-art methods.

Further, we explore the ability of ACH depth to detect atypical observations. For this, we conduct an experiment in settings similar to those in Section 6.2.3, while changing degree of abnormality gradually for 15 (out of 100 curves) in data set (a). Thus, we alter a in $[0, 3]$ for *location anomalies*, b in $[0, 10]$ for *isolated anomalies*, and e in $[1, 2]$ for *shape anomalies* to amplify the "spikes" of oscillations such that $\check{\mathbf{z}}(t) = e\mathbf{z}(t)$. Figure 6.7 illustrates number of anomalies detected by ACH, fSDO, fT, and FIF for different parameters of abnormality. While it is difficult to find the general winner, ACH behaves favorably in all the considered cases and clearly outperforms the two other depths when the data is contaminated with *isolated anomalies*.

We conclude this section with a *real-world data benchmark* based on three data sets: "Octane" (Esbensen, 2001), "Wine" (Larsen et al., 2006), and "EOG" (Chen et al., 2015b). The "Wine" data set consists of 397 measurements of proton nuclear magnetic resonance (NMR) spectra of 40 different wine sanamples, the "Octane" data set are 39 near infrared (NIR) spectra of gasoline samples with 226 measurements, while the "EOG" data set represents the electrical potential between electrodes placed at points close to the eyes with 1250 measurements. As pointed out by Hubert et al. (2015), it is difficult to detect anomalies in the first two data sets, while they are easily seen during the human eye inspection. For the "EOG" data set, we assign smaller of the two classes to be abnormal. To the existing state-of-the-art methods, we add here *Isolation Forest* (IF; Liu et al., 2008) and the *One-Class SVM* (OC; Schölkopf et al., 2001)—multivariate methods applied after a proper dimension reduction (to the dimension 10) using *Functional Principal Component Analysis* (FPCA; see Ramsay and Silverman, 2002). Portions of detected anomalies (by all the considered methods), indicated in Table 6.2, hint on very competitive performance of ACH depth in the addressed benchmark.

6.3 Conclusion

In this chapter, we have introduced a novel functional depth function on the space $\mathcal{C}([0, 1])$ of real valued continuous curves on $[0, 1]$ that presents various advantages. Regarding interpretability first, the depth computed at a query curve $\mathbf{x}$ in $\mathcal{C}([0, 1])$ takes

	ACH	fSDO	fT	FIF	IF	OC
Octane	**1**	0.5	0.33	**1**	0.5	0.5
Wine	**1**	0	0	**1**	0	**1**
EOG	**0.73**	0.55	0.48	0.43	0.63	0.6

Table 6.2 – Portion of detected anomalies of benchmark methods for the "Octane", "Wine", and "EOG" data sets.

the form of an expected ratio, quantifying the relative increase of the area of the convex hull of i.i.d. random curves when adding $\mathbf{x}$ to the batch. We have shown that this depth satisfies several desirable properties and have explained how to solve approximation issues, concerning the sampled character of observations in practice and scalability namely. Numerical experiments on both synthetic and real data have highlighted a number of crucial benefits: reduced variance of its statistical versions, robustness with respect to the choice of tuning parameters and to the presence of outliers in the training sample, capacity of detecting (possibly slight) anomalies of various types, surpassing competitors such as depths of integral type, for isolated anomalies in particular. The open-source Cython/C++ implementation of the method can be accessed at https://github.com/GuillaumeStaermanML/ACHD.

6.4 Technical Details

6.4.1 Auxiliary Lemma

Lemma 6.5. *(Continuity of the Band function) Let $\mathbf{x}_1, \dots, \mathbf{x}_j$, $j \in \{1, \dots, J\}$ be fixed curves in $\mathcal{C}([0,1])$. The function*

$$\begin{aligned} \mathcal{C}([0,1]) &\longrightarrow \mathcal{K}^2 \\ \mathbf{x} &\mapsto \mathrm{band}(\mathbf{x}_1, \dots, \mathbf{x}_j, \mathbf{x}) \end{aligned}$$

is continuous if $\mathcal{K}^2$ is equipped with the Hausdorff distance $d_{\mathcal{H}}$.

Proof. Let $\mathbf{x}_0 \in \mathcal{C}([0,1])$ and j be fixed in $\{1, \dots, J\}$. Let $\zeta > 0$, and write $\mathrm{band}_{\mathbf{x}} := \mathrm{band}(\mathbf{x}_1, \dots, \mathbf{x}_j, \mathbf{x})$ and $\mathrm{band}_{\mathbf{x}_0} := \mathrm{band}(\mathbf{x}_1, \dots, \mathbf{x}_j, \mathbf{x}_0)$ for the sake of clarity. We have:

$$d_{\mathcal{H}}\left(\mathrm{band}_{\mathbf{x}}, \mathrm{Band}_{\mathbf{x}_0}\right) = \max\left(\sup_{\mathbf{z}\in\mathrm{band}_{\mathbf{x}}} d(\mathbf{z}, \mathrm{band}_{\mathbf{x}_0}), \sup_{\mathbf{z}\in\mathrm{band}_{\mathbf{x}_0}} d(\mathbf{z}, \mathrm{band}_{\mathbf{x}})\right)$$

with $d(\cdot,\cdot)$ being the distance induced by $||.||_\infty$. It is easy to see that for any $\mathbf{z} \in \mathrm{band}_{\mathbf{x}}$, $\inf_{\mathbf{y}\in\mathrm{band}_{\mathbf{x}_0}} ||\mathbf{z} - \mathbf{y}||_\infty$ is minimized by the function:

$$\mathbf{y}^*(t) = \mathbf{z}(t)\, \mathbb{I}\{(\mathbf{z}(t) \in \mathrm{band}_{\mathbf{x}_0})\} + \max\left(\mathbf{x}_1(t), \dots, \mathbf{x}_j(t), \mathbf{x}(t)\right)\, \mathbb{I}\{(\mathbf{z}(t) \notin \mathrm{band}_{\mathbf{x}_0})\}.$$

Following this, $||\mathbf{z} - \mathbf{y}^*||_\infty$ is equal to the maximum between

$$\sup_{t:\ \mathbf{z}(t)>\mathrm{band}_{\mathbf{x}_0}} \left|\mathbf{z}(t) - \max\left(\mathbf{x}_1(t), \dots, \mathbf{x}_j(t), \mathbf{x}(t)\right)\right|,$$

and

$$\sup_{t:\ \mathbf{z}(t)<\text{band}_{\mathbf{x}_0}} \left|\mathbf{z}(t) - \min\left(\mathbf{x}_1(t), \ldots, \mathbf{x}_j(t), \mathbf{x}(t)\right)\right|.$$

Furthermore, as $\mathbf{z} \in \text{Band}_{\mathbf{x}}$, it follows:

$$\forall\ t, \quad \min\left(\mathbf{x}(t), \min_{i=1,\ldots,j} \mathbf{x}_i(t)\right) \leq \mathbf{z}(t) \leq \max\left(\mathbf{x}(t), \max_{i=1,\ldots,j} \mathbf{x}_i(t)\right).$$

If $\mathbf{z}(t) > \text{band}_{\mathbf{x}_0}$ we have $\max\left(\max_{i=1,\ldots,j} \mathbf{x}_i(t), \mathbf{x}_0(t)\right) < \mathbf{z}(t) \leq \mathbf{x}(t)$ and then

$$\sup_{t:\ \mathbf{z}(t)>\text{band}_{\mathbf{x}_0}} \left|\mathbf{z}(t) - \max\left(\mathbf{x}(t), \max_{i=1,\ldots,j} \mathbf{x}_i(t)\right)\right| = \sup_{t:\ \mathbf{z}(t)>\text{band}_{\mathbf{x}_0}} |\mathbf{z}(t) - \mathbf{x}(t)|.$$

With the same argument we have:

$$\sup_{t:\ \mathbf{z}(t)<\text{band}_{\mathbf{x}_0}} \left|\mathbf{z}(t) - \max\left(\mathbf{x}(t), \max_{i=1,\ldots,j} \mathbf{x}_i(t)\right)\right| = \sup_{t:\ \mathbf{z}(t)<\text{band}_{\mathbf{x}_0}} |\mathbf{z}(t) - \mathbf{x}(t)|.$$

It follows that for every $\mathbf{z} \in \text{band}_{\mathbf{x}}$, $d(\mathbf{z}, \text{band}_{\mathbf{x}_0}) \leq ||\mathbf{x} - \mathbf{x}_0||_\infty \leq \zeta$. We then have

$$\sup_{\mathbf{z}\in\text{band}_{\mathbf{x}}} d(\mathbf{z}, \text{band}_{\mathbf{x}_0}) \leq \zeta\,.$$

We can prove that

$$\sup_{\mathbf{z}\in\text{band}_{\mathbf{x}_0}} d(\mathbf{z}, \text{band}_{\mathbf{x}}) \leq \zeta$$

with the same argument which concludes the proof.

■

6.4.2 Proof of Proposition 3.2

(*ii*) Scalar-affine invariance. Let $a, b \in \mathbb{R}$, it is clear that

$$\text{conv}(\text{graph}(\{a\mathbf{X}_1 + b, \ldots, a\mathbf{X}_n + b\})) = a \times \text{conv}(\text{graph}(\{\mathbf{X}_1, \ldots, \mathbf{X}_n\})) + b$$

where $a \times \text{conv}(\text{graph}(\{\mathbf{X}_1, \ldots, \mathbf{X}_n\})) + b = \{(t, a\mathbf{x} + b) : (t, \mathbf{x}) \in \text{conv}(\text{graph}(\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}))\}$. The result follows from properties of the Lebesgue measure. However, this property is not satisfied when $\mathbf{a}, \mathbf{b} \in \mathcal{C}([0, 1])$. Indeed, let $J = 2$ and $\mathbf{X}$ be a random variable following a distribution $\mathbf{P}$ such that $\mathbb{P}(\mathbf{X} \equiv \mathbf{x}_1) = \frac{1}{2}$ and $\mathbb{P}(\mathbf{X} \equiv \mathbf{x}_2) = \frac{1}{2}$ with $\mathbf{x}_1 \equiv \mathbf{1}, \mathbf{x}_2 \equiv \mathbf{2}$. Let $\mathbf{X}_1, \mathbf{X}_2$ two i.i.d.random variables from $\mathbf{P}$ and $\mathbf{b}$ be continuous function $t \mapsto (10t - 4)\ \mathbb{I}\{[0.4, 0.5]\} + (-10t + 6)\ \mathbb{I}\{[0.5, 0.6]\}$. Let $\mathbf{x} \equiv \mathbf{0}$. It is easy to see that $FD^J_{\text{ACH}}(\mathbf{x}|P) = \frac{1}{8} \neq FD^J_{\text{ACH}}(\mathbf{x} + \mathbf{b}|P_{\mathbf{X}+\mathbf{b}})$ since

$$\begin{aligned} FD^J_{\text{ACH}}(\mathbf{x} + \mathbf{b}, P_{\mathbf{X}+\mathbf{b}}) &= \frac{1}{2} \times \left(\frac{1}{2} \times \frac{0.5}{1.5} + \frac{1}{2} \times \frac{0.5}{2.5}\right)^{j=1} + \frac{1}{2} \times \left(\frac{1}{4} \times \frac{0.5}{1.5} + \frac{1}{4} \times \frac{0.5}{2.5} + \frac{1}{2} \times \frac{1.5}{2.5}\right)^{j=2} \\ &= \frac{17}{60}. \end{aligned}$$

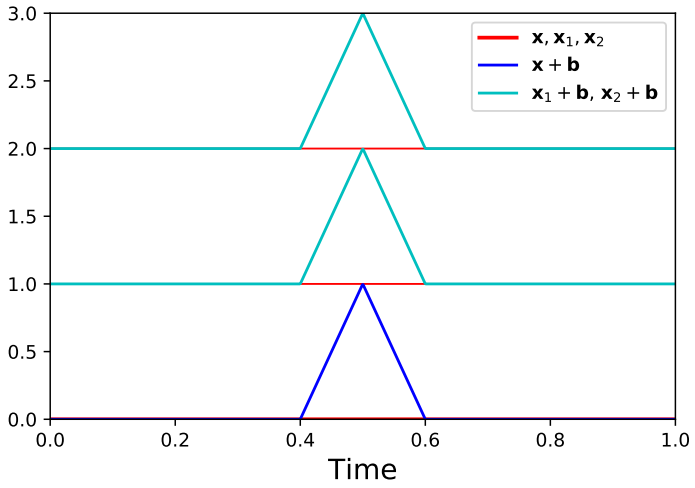


Figure 6.8 – Plots of the functions used in the case of $\mathbf{a}, \mathbf{b} \in \mathcal{C}([0,1])$. The three red lines in ascending order are $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2$. The cyan curves correspond to $\mathbf{x}_1 + \mathbf{b}$ and $\mathbf{x}_2 + \mathbf{b}$ and the blue curve to $\mathbf{x} + \mathbf{b}$.

Even if we restrict $j > 1$ to avoid the fact that the convex hull of constant function have zero Lebesgue measure, $FD^J_{\text{ACH}}(\mathbf{x}, P)$ and $FD^J_{\text{ACH}}(\mathbf{x} + \mathbf{b}, P_{\mathbf{X}+\mathbf{b}})$ remain different, see Figure 6.8.

(iii) Vanishing at infinity. Let J be fixed and $\mathbf{x}_n$ be a sequence of functions such that $||\mathbf{x}_n||_\infty$ tends to infinity when n grows, for every $j \in \{1, ..., J\}$ we define :

$$\begin{aligned} \Phi_{\mathbf{x}_n} : \quad (\mathcal{C}([0,1]))^{\otimes j} \quad &\longrightarrow [0,1] \\ (\mathbf{x}_1, \ldots, \mathbf{x}_j) &\longmapsto \frac{\lambda_2(\text{conv}(\text{graph}(\{\mathbf{x}_1, \ldots, \mathbf{x}_j\})))}{\lambda_2(\text{conv}(\text{graph}(\{\mathbf{x}_1, \ldots, \mathbf{x}_j, \mathbf{x}_n\})))}. \end{aligned}$$

As a continuous function on compact set, $\mathbf{x}_1, ..., \mathbf{x}_j$ are bounded, then $\Phi_{\mathbf{x}_n} \underset{||\mathbf{x}_n||_\infty \to \infty}{\longrightarrow} 0$. The result follows from dominated convergence theorem since $\Phi_{\mathbf{x}_n}$ is bounded by 1.

(iv) Continuity in $\mathbf{x}$. Let $\mathbf{x}_1, ..., \mathbf{x}_j$, $j \in \{1, ..., J\}$ be fixed curves in $\mathcal{C}([0,1])$ with at least two different curves, i.e there exists a $t \in [0,1]$ and $l, k \in \{1, \ldots, j\}$ such that $\mathbf{x}_k(t) \neq \mathbf{x}_l(t)$. If $j = 1$, $\mathbf{x}_1$ is assumed not to be a constant function. From Lemma 6.5, we know that the function

$$g : \mathbf{x} \longmapsto \text{band}(\mathbf{x}_1, ..., \mathbf{x}_j, \mathbf{x}),$$

is continuous w.r.t. the infinity norm. Let $\mathfrak{K}^2$ be the set of all compact sets in $\mathbb{R}^2$ and $\mathfrak{K}^2_C$ the set of all convex bodies (compact, convex set with non-empty interior). We equip both spaces with the topology induced by the Hausdorff distance. The two following

maps:

$$\begin{aligned}\text{conv} : \mathfrak{K}^2 &\longrightarrow \mathfrak{K}_C^2 \\ \mathcal{K}^2 &\longmapsto \text{conv}(\mathcal{K}^2),\end{aligned}$$

and

$$\begin{aligned}\text{vol} : \mathfrak{K}_C^2 &\longrightarrow \mathbb{R}_+ \\ \mathcal{K}_C^2 &\longmapsto \lambda_2(\mathcal{K}_C^2),\end{aligned}$$

are continuous with respect to the Hausdorff distance, see e.g. Theorem 12.3.5 in Schneider and Weil (2008) for conv and Theorem 1.8.16 in Schneider (1993) for vol. Then the function $\text{vol} \circ \text{conv} \circ g : \mathbf{x} \mapsto \lambda_2(\text{conv}(\text{graph}\{\mathbf{x}_1, ..., \mathbf{x}_j, \mathbf{x}\}))$ is continuous. Now, from the dominated convergence theorem it follows that:

$$\mathbf{x} \longmapsto \mathbb{E}\left[\frac{\lambda_2(\text{conv}(\text{graph}(\{\mathbf{x}_1, ..., \mathbf{x}_j\})))}{\lambda_2(\text{conv}(\text{graph}(\{\mathbf{x}_1, ..., \mathbf{x}_j, \mathbf{x}\})))}\right],$$

is continuous. The final result holds with the continuity of the sum of continuous functions.

(v) Continuity in $\mathbf{P}$. Since $\left(\mathcal{C}([0,1]), ||.||_\infty\right)$ is a Polish space, the set of all probability measures defined on it equipped with the Lévy-Prohorov metric d_{LP} is still Polish. Denoting by $\xrightarrow{\mathcal{L}}$ the convergence in law, by the Portmanteau theorem (see e.g. Theorem 11.3.3 in Dudley (2002)), it follows that $d_{LP}(\mathbf{Q}_n, \mathbf{Q}) \to 0$ is equivalent to $\mathbf{Q}_n \xrightarrow{\mathcal{L}} \mathbf{Q}$ for $\mathbf{Q}$ and $\mathbf{Q}_n$ respectively a measure and a sequence of measures on $\mathcal{C}([0,1])$. It implies, as n goes to ∞:

$$\int \mathbf{f} \, \mathrm{d}\mathbf{Q}_n \longrightarrow \int \mathbf{f} \, \mathrm{d}\mathbf{Q},$$

for any $\mathbf{f}$ bounded continuous real function on $\left(\mathcal{C}([0,1]), ||.||_\infty\right)$. Let $j \in \mathbb{N}_*$ and define the following function:

$$\begin{aligned}\Phi_{\mathbf{x}} : \quad (\mathcal{C}([0,1]))^{\otimes j} &\longrightarrow [0,1] \\ (\mathbf{x}_1, ..., \mathbf{x}_j) &\longmapsto \frac{\lambda_2(\text{conv}(\text{graph}(\{\mathbf{x}_1, ..., \mathbf{x}_j\})))}{\lambda_2(\text{conv}(\text{graph}(\{\mathbf{x}_1, ..., \mathbf{x}_j, \mathbf{x}\})))}.\end{aligned}$$

If $\mathcal{C}([0,1])^{\otimes j}$ is equipped with the infinity norm $|||.|||_{\infty,j}$ defined by:

$$|||\mathbf{f}|||_{\infty,j} = \max(||\mathbf{f}_1||_\infty, \ldots, ||\mathbf{f}_j||_\infty),$$

following the same arguments from the proof of the assertion (iv), $\Phi_{\mathbf{x}}$ is bounded and continuous. Now, let $J \leq n$ be fixed and $\mathbf{Q}_n$ be a sequence of measures on $\mathcal{C}([0,1])$ such that $d_{LP}(\mathbf{Q}_n, \mathbf{Q}) \to 0$. we have:

$$\begin{aligned}\lim_{n\to\infty} \sum_{j=1}^{J} FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{Q}_n) &= \sum_{j=1}^{J} \lim_{n\to\infty} \int_{\mathcal{C}([0,1])^{\otimes j}} \Phi_{\mathbf{x}} \mathrm{d}\mathbf{Q}_n^{\otimes j} \\ &= \sum_{j=1}^{J} \int_{\mathcal{C}([0,1])^{\otimes j}} \Phi_{\mathbf{x}} \mathrm{d}\mathbf{Q}^{\otimes j} \\ &= \sum_{j=1}^{J} FD^j_{\text{ACH}}(\mathbf{x}, \mathbf{Q}).\end{aligned}$$

Then the results holds for $\sum_{j=1}^{J} FD^{j}_{\text{ACH}}$ (and trivially for FD^{J}_{ACH}).

(*vi*) MAXIMALITY AT THE CENTER. We restrict ourselves for simplicity to $J = 2$. Let $\mathbf{X} \sim \mathbf{P}$ be a distribution such that $\mathbb{P}(\mathbf{X} \equiv \mathbf{y}_1) = \mathbb{P}(\mathbf{X} \equiv \mathbf{y}_2) = \frac{1}{2}$ with

$$\mathbf{y}_1 = (-2t+1)\, \mathbb{I}\{[0, 0.25]\} + (2t)\, \mathbb{I}\{[0.25, 0.5]\} + (-2t+2)\, \mathbb{I}\{[0.5, 0.75]\} + (2t-1)\, \mathbb{I}\{[0.75, 1]\}\,,$$
$$\mathbf{y}_2 = -\mathbf{y}_1.$$

The distribution is clearly centrally and halfspace symmetric around $\boldsymbol{\theta} \equiv \mathbf{0}$ but we have

$$FD^{J}_{\text{ACH}}(\boldsymbol{\theta}, \mathbf{P}) < FD^{J}_{\text{ACH}}(\mathbf{y}_1, \mathbf{P}) = FD^{J}_{\text{ACH}}(\mathbf{y}_2, \mathbf{P})\,.$$

Since

$$FD^{J}_{\text{ACH}}(\mathbf{0}, \mathbf{P}) = \frac{1}{2} \times \left(\frac{1}{2} \times \frac{3}{8} + \frac{1}{2} \times \frac{3}{8}\right)^{j=1} + \frac{1}{2} \times \left(\frac{1}{2} + \frac{1}{4} \times \frac{3}{8} + \frac{1}{4} \times \frac{3}{8}\right)^{j=2} = \frac{17}{32} \approx 0.53$$

and

$$FD^{J}_{\text{ACH}}(\mathbf{y}_1, \mathbf{P}) = \frac{1}{2} \times \left(\frac{1}{2} \times + \frac{1}{2} \times \frac{3}{16}\right)^{j=1} + \frac{1}{2} \times \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{4} \times \frac{3}{16}\right)^{j=2} = \frac{70}{128} \approx 0.546$$

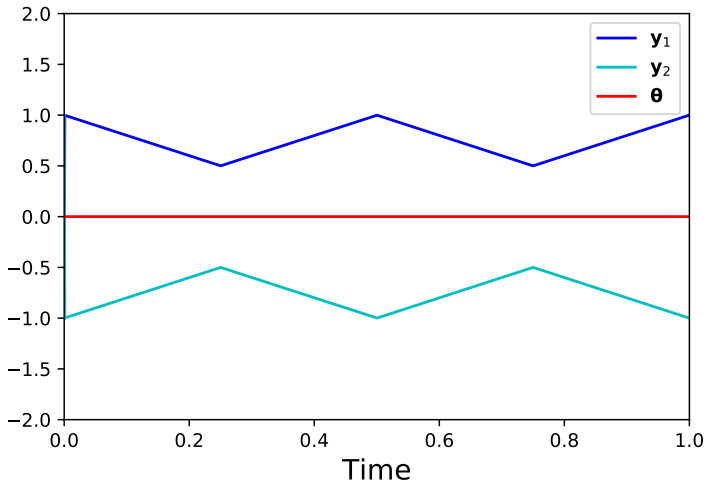


Figure 6.9 – Plot of the functions used in the counter example of the maximality at the center property. $\mathbf{y}_1$ (blue curve) and $\mathbf{y}_2$ (cyan curve) correspond to the distribution and $\boldsymbol{\theta} \equiv \mathbf{0}$ corresponds to the red curve.

(*vii*) DECREASING W.R.T. THE DEEPEST POINT. We restrict ourselves for simplicity to $J = 2$. Let $\mathbf{X}$ be a r.v. following $\mathbf{P}$ such that

$$\mathbb{P}(\mathbf{X} \equiv \mathbf{0}) = \mathbb{P}(\mathbf{X} \equiv \mathbf{1}) = \mathbb{P}(\mathbf{X} \equiv -\mathbf{1}) = \frac{1}{3}.$$

It is clear from this distribution that $\mathbf{z} := \mathbf{0} \in \sup_{\mathbf{x} \in \mathcal{C}([0,1])} FD^J_{\text{ACH}}(\mathbf{x}, \mathbf{P})$ with $FD^J_{\text{ACH}}(\mathbf{0}, \mathbf{P}) = \frac{1}{4}$. We define $\mathbf{y} \equiv \mathbf{1.5}$ and $\mathbf{x}(t) = 4t\ \mathbb{I}\{t \in [0, 0.5]\} + (-4t + 4)\ \mathbb{I}\{t \in [0.5, 1]\}$. We have $||\mathbf{x} - \mathbf{z}||_\infty = 2$, $||\mathbf{x} - \mathbf{y}||_\infty = 0.5$ and $||\mathbf{y} - \mathbf{z}||_\infty = 1.5$. Computing the depth of $\mathbf{x}$ and $\mathbf{y}$ we have:

$$FD^J_{\text{ACH}}(\mathbf{y}, \mathbf{P}) = \frac{1}{2} \times \frac{2}{9} \times \left(\frac{4}{5} + \frac{2}{5} + \frac{2}{3} \right) = \frac{23}{135}$$

and

$$FD^J_{\text{ACH}}(\mathbf{x}, \mathbf{P}) = \frac{1}{2} \times \frac{2}{9} \times \left(\frac{4}{5} + \frac{1}{2} + \frac{8}{9} \right) = \frac{197}{810}\,.$$

The result follows. It is worth mentioning that the result remains true if conv is replaced by the band function.

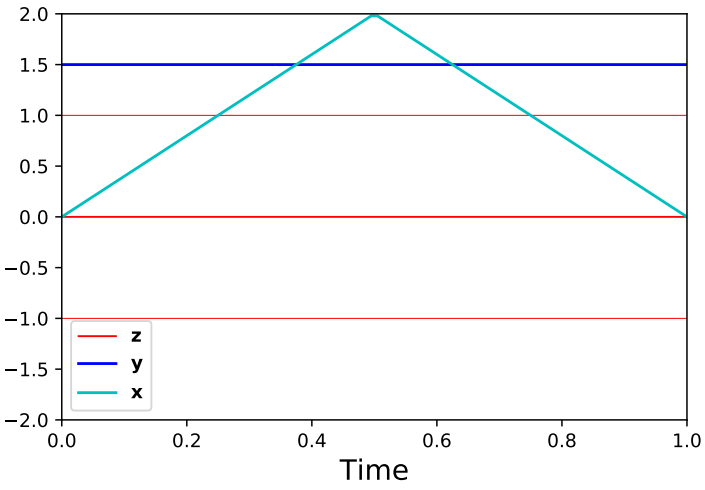


Figure 6.10 – Plots of the functions used in the counter example of the decreasing property. The three red lines come from the distribution, the thicker red curve corresponds to the maximal depth. The cyan curve corresponds to $\mathbf{x}$ and the blue curve to $\mathbf{y}$.

7

Functional Anomaly Detection: a Benchmark Study

Contents

The increasing automation in many areas of the industry expressly demands to design efficient Machine-Learning solutions for the detection of abnormal events. With the ubiquitous deployment of sensors monitoring nearly continuously the health of complex infrastructures, anomaly detection can now rely on measurements sampled at a very high frequency, providing a very rich representation of the phenomenon under surveillance. In order to exploit fully the information thus collected, the observations cannot be treated as multivariate data anymore and a functional analysis approach is required. It is the purpose of this chapter to investigate the performance of recent techniques for anomaly detection in the functional setup on real data sets. While taxonomies of abnormalities (e.g. shape, location) in the functional setting are documented in the literature, assigning a specific type to the identified anomalies appears to be a challenging task. Thus, strengths and weaknesses of the existing approaches are benchmarked in view of these highlighted types in a simulation study. Anomaly detection methods are next evaluated on two data sets, related to the monitoring of helicopters in flight and to the spectrometry of construction materials namely. The benchmark analysis is concluded by a recommendation guidance for practitioners.

The chapter is organized as follows. In Section 7.1, we introduce data sets and benchmarked methods. In Section 7.2, the performance metrics used for measuring the accuracy of anomaly detection rules learned in an unsupervised manner on (labeled) test data are described and the experiments on synthetic data are displayed. The experiments on real data and the results obtained are presented at length in Section 7.3. Finally, some concluding remarks are collected in Section 7.4. This chapter covers the contribution of:

- **G. Staerman**, P. Mozharovskyi, S. Clémençon. Functional Anomaly Detection: a Benchmark Study. *arXiv preprint arXiv:2201.05115*, 2022.

7.1 Introduction

It is the goal of this chapter to investigate the performance of recent techniques for functional anomaly detection and compare their accuracy with that of simpler approaches, based on a preliminary dimensionality reduction, standing as natural competitors. In particular, specific attention is paid to those that are based on functional depth statistics or that extend multivariate methods by avoiding the filtering step. A benchmark study comparing the merits of the methods considered here regarding various metrics of reference is thus presented on aeronautics data gathered by Airbus and spectrometry measurements of sedimentary material collected by the Geological Survey of Austria for quality assessment on mining sites of Austria. Specifically, the aeronautics data set consists of one-minute-sequences of accelerometer data measured at a 1024Hz frequency. Airbus data set is divided into two parts: the training set composed of 1677 curves with no available labels that may contains "abnormal" observations and the validation/test set composed of 2511 time-series with 1794 "normal" curves. In contrast, the test data are labeled, in order to evaluate the performance of the anomaly detection rules learned in the training stage. These measurements were made on test helicopters at various locations, in various angles, on different flights. Data sets are displayed in Figure 7.1. The learning framework is unsupervised: in the experiment, all accelerometer data series at disposal for training automatically a classifier to detect abnormal changes are considered as normal. The spectrometry of rocks data consists of one data set of 2096 curves with 600 measurements. It represents materials of two types whose labels are available, with limestone being the desirable (normal) rock type and the intrusive (abnormal) cellular dolomite. The task is thus, given the reflectance spectrum (with noise subtracted and normalized with respect to the reference spectrum) of multiple samples of mined examples, separate those abnormal.

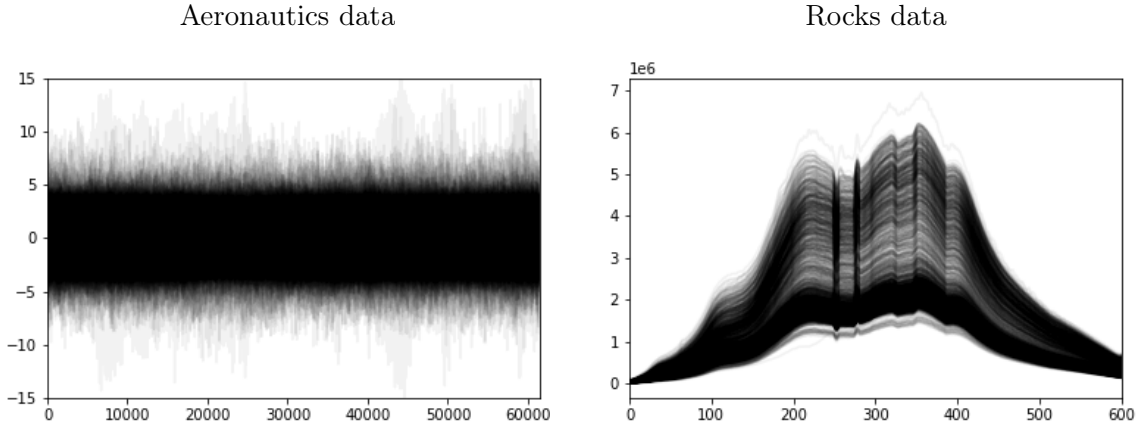


Figure 7.1 – The aeronautics and the rock data sets.

Functional anomaly detection benchmarked methods, all along this chapter, are the following: (*i*) Functional Isolation Forest (FIF; see Chapter 5), (*ii*) ACH depth (see Chapter 6), (*iii*) functional Adjusted Outlyingness (fAO; Hubert et al., 2015), (*iv*)

functional Stahel-Donoho Outlyingness (fSDO; Hubert et al., 2015, (*v*) functional bag-distance (fbd; Hubert et al., 2015, (*v*) functional Tukey depth (fT; Claeskens et al., 2014), (*vi*) Outliergram (Arribas-Gil and Romo, 2014), (*vii*) Magnitude-Shape (MS; Dai and Genton, 2019a) plot combined with Isolation Forest, (*viii*) Functional Outlier Map (FOM; Rousseeuw et al., 2018) combined with Isolation Forest, (*ix*) Isolation Forest (IF; Liu et al., 2008) combined with Functional Principal Component Analysis (FPCA) (see e.g. Ramsay and Silverman, 2002), (*x*) Local Outlier Factor (LOF; Breunig et al., 2000) combined with FPCA, (*xi*) One-Class Support Vector Machine (OC; Schölkopf et al., 2001) combined with FPCA.

7.2 A Preparatory Simulation Study

This section is devoted to empirical analysis of the performance of the functional anomaly detection techniques. Their accuracy is investigated from simulated data inspired from the real data set collected by Airbus, composed of one-minute sequences of accelerometer data measured on helicopters. As a first go, we recall the standard performance metrics commonly used to quantify it.

7.2.1 Performance Metrics in Anomaly Detection

Although the learning procedure does not rely on data of the same nature, the anomaly detection problem can be formulated in the same probabilistic framework as binary classification, the flagship problem in statistical learning. In the standard setup, the binary random variable Z indicates the occurrence of an anomaly: the label is positive, i.e. $Z = +1$, when an anomaly occurs, and negative, i.e. $Z = -1$, otherwise. The random variable $\mathbf{X}$, taking its values in the feature space $\mathcal{F}$, models the measurement at disposal to predict Z. The goal pursued is to build an anomaly scoring function $s : \mathcal{F} \to \mathbb{R} \cup \{+\infty\}$ in an unsupervised manner (without observing Z), so that, ideally, the larger $s(\mathbf{X})$, the likelier an anomaly occurs, i.e. the more probable is $Z = +1$. A decision to raise an alarm can be then built by thresholding the scoring function $s(\mathbf{x})$ at a critical level, ruling the trade-off between errors of type I and type II. Equipped with this notation, decreasing transforms of a depth function w.r.t. $\mathbf{X}$'s marginal distribution provide anomaly scoring functions in a natural fashion. Precisely, when using data depths in the following study, the transformation $1 - FD(\cdot, \cdot)$ is performed to rescale them as an anomaly score.

ROC analysis. The golden standard to quantify theoretically the accuracy of an anomaly scoring function s is the PP-plot of the false positive rate *vs* the true positive rate, namely $t \in \mathbb{R} \mapsto (\mathbb{P}\{s(\mathbf{X}) \geq t \mid Z = -1\}, \mathbb{P}\{s(\mathbf{X}) \geq t \mid Z = +1\})$, referred to as the ROC curve (standing for Receiver Operator Characteristic curve), see e.g. Fawcett (2006). The higher the curve, the more accurate the anomaly scoring function. A simple Neyman-Pearson argument shows that optimal scoring functions are increasing transforms of the likelihood ratio $\Psi(\mathbf{X}) := (\mathrm{d}\mathbf{F}_+/\mathrm{d}\mathbf{F}_-)(\mathbf{X})$, denoting by $\mathbf{F}_\sigma$ the conditional distribution of $\mathbf{X}$ given $Z = \sigma 1$, $\sigma \in \{-, +\}$: their ROC curve dominating everywhere the ROC curve of any other anomaly scoring function. For this reason, this functional performance measure is generally summarized by the Area Under the ROC curve (AUC in abbreviated form), a popular scalar criterion that can be classically interpreted as the rate of concordance of pairs: $\mathrm{AUC}(s) = \mathbb{P}\{s(\mathbf{X}) > s(\mathbf{X}') \mid Z = +1,\ Z' = -1\} + \mathbb{P}\{s(\mathbf{X}) = s(\mathbf{X}') \mid Z = +1,\ Z' = -1\}/2$, where $(\mathbf{X}', Z')$ denotes an independent copy of the pair $(\mathbf{X}, Z)$.

PR analysis. Alternatively, one may evalute the accuracy of any score by plotting the precision-recall (shortly PR) curve, namely $t \mapsto (\mathbb{P}\{s(\mathbf{X}) \geq t \mid Z = +1\}, \mathbb{P}\{Z = +1 \mid s(\mathbf{X}) \geq t\})$. The higher its PR curve, the more accurate an anomaly scoring function. Of course, as may be immediately shown by means of the Bayes formula, PR and ROC curves are in one to one correspondence, see e.g. Clémençon and Vayatis (2009). Like the ROC curve, the PR curve may be summarized by the area under it, referred to as the Average Precision (AP).

When labeled data are available, it is possible to compute the performance measures recalled above, replacing the probabilities involved by their statistical counterparts, in order to assess the accuracy of any anomaly scoring function candidate $s(\mathbf{x})$. However, in unsupervised anomaly detection, the scoring function $s(\mathbf{x})$ cannot be learned using labeled training data, in contrast to classification or bipartite ranking: only "negative" observations, i.e. an i.i.d. sample drawn from (a possibly noisy version of) distribution $\mathbf{F}_-$, are available in the training stage. Hence, the learning task cannot be achieved by optimizing empirical versions of the aforementioned criteria, which makes it extremely challenging.

7.2.2 Simulating Anomalies of Specific Types

Data sets containing various types of anomalies are usually more challenging to analyze. As a first go, we start by investigating to which extent the techniques recalled above may permit to detect simulated anomalies of well-identified types according to the usual taxonomy, see e.g. Hubert et al. (2015). Figure 7.2 illustrates the four types of anomalies addressed in detail in these experiments: isolated, magnitude (of two different kinds) and shape anomalies.

To reproduce a controlled version of each type of anomalies, four data sets have been built from a collection of 1794 "normal" functional observations from the validation data set collected by Airbus. Each functional observation corresponds to accelerometer data measured on helicopters at a 1024 Hz frequency over time windows of 1 minute: the curves $\mathbf{X} = (\mathbf{X}(t))_{t\in[0,1]}$ are built by means of an affine interpolation of the 61440 sampled points. One per anomaly type, four data sets have been constructed by adding a specific contamination to 5% of these "normal" observations, drawn uniformly at random. Cases when 1%, 2% 3% and 4% are added in the Section B.1 of Appendices for completeness and show similar behavior of methods than for 5%. The four contamination models defined below, are used to generate independent curves $\mathbf{Y}$ (independently from the original data set) that are next added to the selected above 1794 "normal" observations $\mathbf{X}$. By $\mathcal{U}([a, b])$ is meant the uniform distribution on the interval $[a, b]$, while δ_u denotes the Dirac mass at point u.

Model 1 (Isolated Anomalies)
$\mathbf{Y}(t) = \gamma u_1 \mathbb{I}\{t = \tau\}$, where $u_1 \sim \mathcal{U}([3, 4])$ and $\gamma \sim (1/2)(\delta_{-1} + \delta_1)$ are independent random variables, with τ being the time at which the isolated anomaly occurs that is chosen randomly in a uniform manner among the set of sampling points, independently of u_1 and γ.

Model 2 (Magnitude Anomalies I)

$\mathbf{Y}(t) \equiv u_2$, with $u_2 \sim \mathcal{U}([-12, -15])$.

Model 3 (Magnitude Anomalies II)

$\mathbf{Y}(t) = u_3 \mathbb{I}(t \in I)$, where $u_3 \sim \mathcal{U}([0, 15])$ and I is a subinterval of $[0, 1]$ of length $1/10$ which location is chosen uniformly at random, independently of u_3.

Model 4 (Shape Anomalies)

$\mathbf{Y}(t) = \sin(2\pi u_4 t)$, where $u_4 \sim \mathcal{U}([0.2, 2])$.

Simulated anomalies, together with a small subset of "normal" data, are illustrated in Figure 7.3.

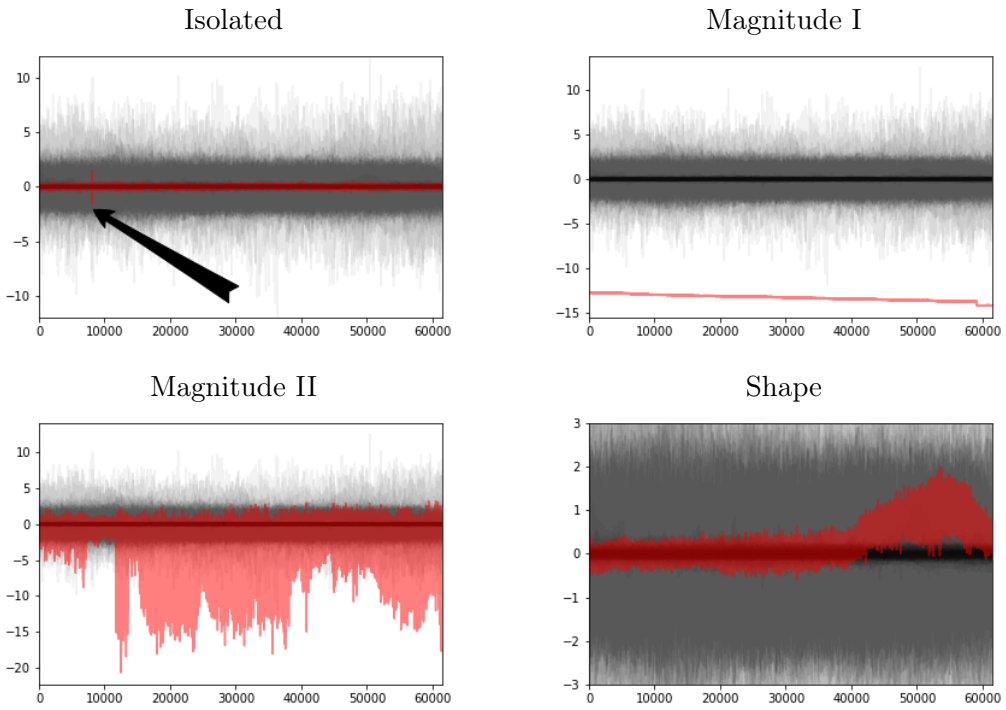


Figure 7.2 – Examples for each of the identified types of anomalies in the aeronautics data set. In order from top to bottom and from left to right: each time one isolated, magnitude, magnitude/shape and shape anomaly, with a subsample of normal data. Grey curves are normal data while red curves are anomalies.

7.2.3 Naive Approaches - Sampled Curves Viewed as Multivariate Data

Being the most straightforward idea, and still frequently employed in functional anomaly detection, direct consideration of (discretized) functional data in the multivariate space (i.e. $\mathbb{R}^p$) of their measurements can be seen as a first naïve approach. It is important to notice that the sampling design should not vary with the curve/signal and preferably correspond to regularly spaced points in the observation domain, so that no dimensions are disadvantaged. If not, one can resort to importance-weighting techniques though, e.g. giving more weights to coordinates with higher marginal variance as it is suggested, e.g. in Claeskens et al. (2014). When implementing such a naïve strategy to deal with functional observations, specific attention must be paid to the possibly very large dimension of the data (compared to the size of the population sample), due to the high frequency character of the measurements, as it is the case for the aeronautics data set

considered in this chapter where $61440 = p \gg n = 1677$. We have applied three of the most widely used methods for multivariate anomaly detection, namely Isolation Forest (IF), Local Outlier Factor (LOF), and One-Class Support Vector Machine (OC) to the four settings described in Section 7.2.2. Their True Positive Rate (TPR) and Area Under Receiver Operating Characteristic (AUC) are reported in Table 7.1.

Anomaly type		IF	LOF	OC
Isolated	TPR	0	0	0
	AUC	0.41	0.25	0.44
Magnitude I	TPR	1	0.48	1
	AUC	1	0.97	1
Magnitude II	TPR	0	0	0
	AUC	0.54	0.03	0.7
Shape	TPR	0	0.48	0
	AUC	0.67	0.97	0.67

Table 7.1 – Methods considered in performance comparison with the TPR and the Area Under the Receiver Operating Characteristic (AUC) for the four simulated models.

7.2.4 Results and Discussion

We now consider the set of methods which are in the center of attention of the current work. These treat the data directly in their original functional space, and—as we shall see right below—prove beneficial for anomaly detection.

The performance of these unsupervised methods is evaluated on the four simulated data sets described in Section 7.2.2 using the sensitivity (portion of correctly identified anomalies, TPR) and AUC as metrics. All parameters of the used algorithms are set to their default values (as it is pre-defined in the corresponding software packages). Thus, fSDO, fAO, fbd, FOM are available in the `mrfDepth` `R`-package (Segaert et al., 2020); the outliergram is available in the `roahd` `R`-package (Tarabelloni et al., 2018); IF, LOF and OC are available in `sklearn` `python` library (Pedregosa et al., 2011); Magnitude-Shape plot (MS) is available in the `scikit-fda` `python` library[1]; Functional Isolation Forest (FIF) and ACH open-source python codes are available under the following link [2]; fT can be easily coded from scratch.

Results of the simulation study are displayed in Table 7.2. For completeness, the ROC curves of the four best methods for each contamination setting are displayed in Figure 7.4. As expected, the score drastically varies across contamination models and anomaly detection methods. Isolated anomalies (especially short ones) of the Contamination Model 1 are difficult to detect with projections on most bases as well as by integrating depths, whilst ACH is sensitive to this kind of anomalies. Magnitude (especially type I) anomalies are known to be easier to detect and a number of methods (fAO, fbd, fSDO, fT and outliergram) perform well by managing to detect all of them. The difficulty that differentiates Magnitude II anomalies (from those in Magnitude I) is that the anomalies are expressed only for a subset of time points. This impedes many methods from detecting this kind of anomalies, while ACH seems to perform best among differing results, most probably due to slight resemblance of Magnitude II anomalies with the isolated ones. Shape anomalies is the least identifiable type, and FIF delivers

[1] https://github.com/GAA-UAM/scikit-fda
[2] https://github.com/GuillaumeStaermanML

better performance than other methods while taking into account both location and slope of the functional curves (due to the employed Sobolev-type scalar product).

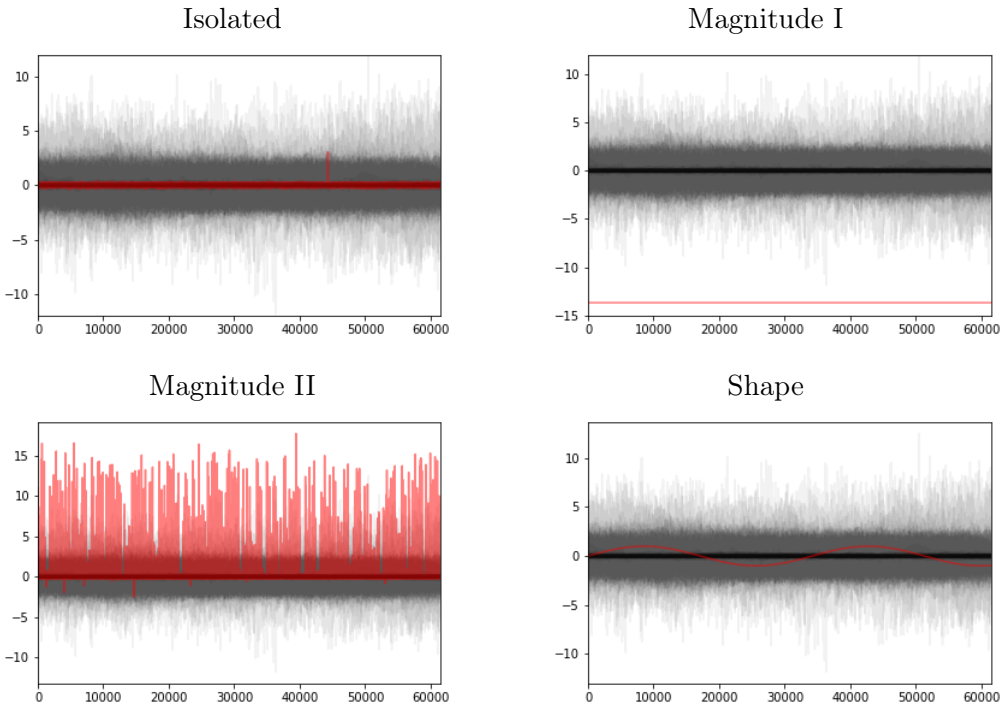


Figure 7.3 – Examples for each of four types of simulated anomalies. In order from top to bottom and from left to right: each time one isolated, magnitude, magnitude/shape and shape anomaly, with a subsample of normal curves (out of 1794). Grey curves are normal data while red curves are simulated anomalies.

7.3 Benchmarking Methods for Functional Anomaly Detection using Real Data

This section is devoted to the empirical analysis of the performance of the functional anomaly detection techniques. Their accuracy, previously investigated based on artificially contaminated data, shall be now benchmarked using real labeled data sets. We analyze the complete Airbus data set (i.e. including normal and abnormal accelerometer data) as well as data arising from the spectrophotometry of rocks with wavelengths of light source ranging from 382 to 930 nanometers corresponding to visible-IR spectrum.

The complete Airbus data set is split into an unlabeled training data set (1677 sampled curves observed at 61440 equally spaced time points) used for the learning purposes and a labeled test data set (2511 observations among which 717 are labeled as abnormal) to evaluate performance using the metrics described in Section 7.2.1. The second data set contains linearly interpolated spectroscopy measurements of construction materials (mined rocks) from different locations and mining sites. Precisely, it includes 2038 limestones and 58 cell limes of 600 wavelength measures. Since the measured spectra are very noisy, and with the task of the current work being rather comparative (than absolute) performance of the methods, we modified the original data set by removing

	Isolated		**Magnitude I**		**Magnitude II**		**Shape**	
Methods	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC
FIF	0	0.21	0.93	0.99	0	0.30	**0.63**	**0.98**
fAO	0	0.44	**1**	**1**	0	0.54	0	0.66
fbd	0	0.44	**1**	**1**	0	0.54	0	0.68
fSDO	0	0.42	**1**	**1**	0	0.43	0	0.77
fT	0	0.43	**1**	**1**	0	0.44	0	0.71
ACH	0	**0.63**	0.48	0.97	**0.80**	**0.99**	0	0.56
Outliergram	0	0.55	**1**	**1**	0	0.54	0	0.47
MS + IF	0	0.05	0.66	0.98	0	0.70	0.33	0.74
FOM (fSDO) + IF	0	0	0.85	0.99	0.64	0.99	0.06	0.80
FOM (fAO) + IF	0	0.14	0.81	0.99	0.55	0.98	0.02	0.87
FPCA + IF	0	0.11	0	0.91	0	0.71	0.45	0.97
FPCA + LOF	0	0.5	0	0.16	0	0.38	0	0.79
FPCA + OC	0	0.04	0	0.93	0	0.77	0.3	0.96

Table 7.2 – Methods considered in performance comparison with the TPR and the Area Under the Receiver Operating Characteristic (AUC) for the four simulated models with 5% of added anomalies.

	Airbus data				**Rocks data**			
Method	F1-Score	AP	AUC	TPR	F1-Score	AP	AUC	TPR
FIF	**0.81**	**0.88**	**0.76**	0.52	0.974	**0.992**	0.772	0.10
fAO	0.78	0.77	0.63	0.45	**0.989**	0.991	**0.833**	**0.58**
fbd	0.78	0.77	0.63	0.45	0.977	0.988	0.751	0.19
fSDO	0.78	0.77	0.63	0.45	0.972	0.980	0.555	0
fT	0.78	0.77	0.63	0.45	0.984	0.989	0.780	0.43
ACH	0.77	0.77	0.62	0.44	0.972	0.961	0.280	0
Outliergram	0.71	0.76	0.55	0.28	0.974	0.981	0.66	0.03
MS + IF	0.80	0.76	0.64	0.51	0.972	0.981	0.601	0
FOM (fSDO) + IF	**0.81**	0.76	0.66	0.53	0.972	0.978	0.530	0.02
FOM (fAO) + IF	0.80	0.77	0.65	0.51	0.984	0.991	0.804	0.448
FPCA + IF	**0.81**	0.80	0.70	0.53	0.972	0.971	0.446	0
FPCA + LOF	0.72	0.73	0.52	0.31	0.972	0.969	0.445	0.02
FPCA + OC	**0.81**	0.79	0.70	**0.54**	0.972	0.971	0.463	0

Table 7.3 – Methods considered in performance comparison with F1-Score, Average Precision (AP), AUC and TPR for Airbus and Rocks data. Bold numbers correspond to the best result.

most difficult to detect anomalies in a supervised manner (which only underlines the difficulty of the real-data anomaly detection task). More precisely, we removed 860

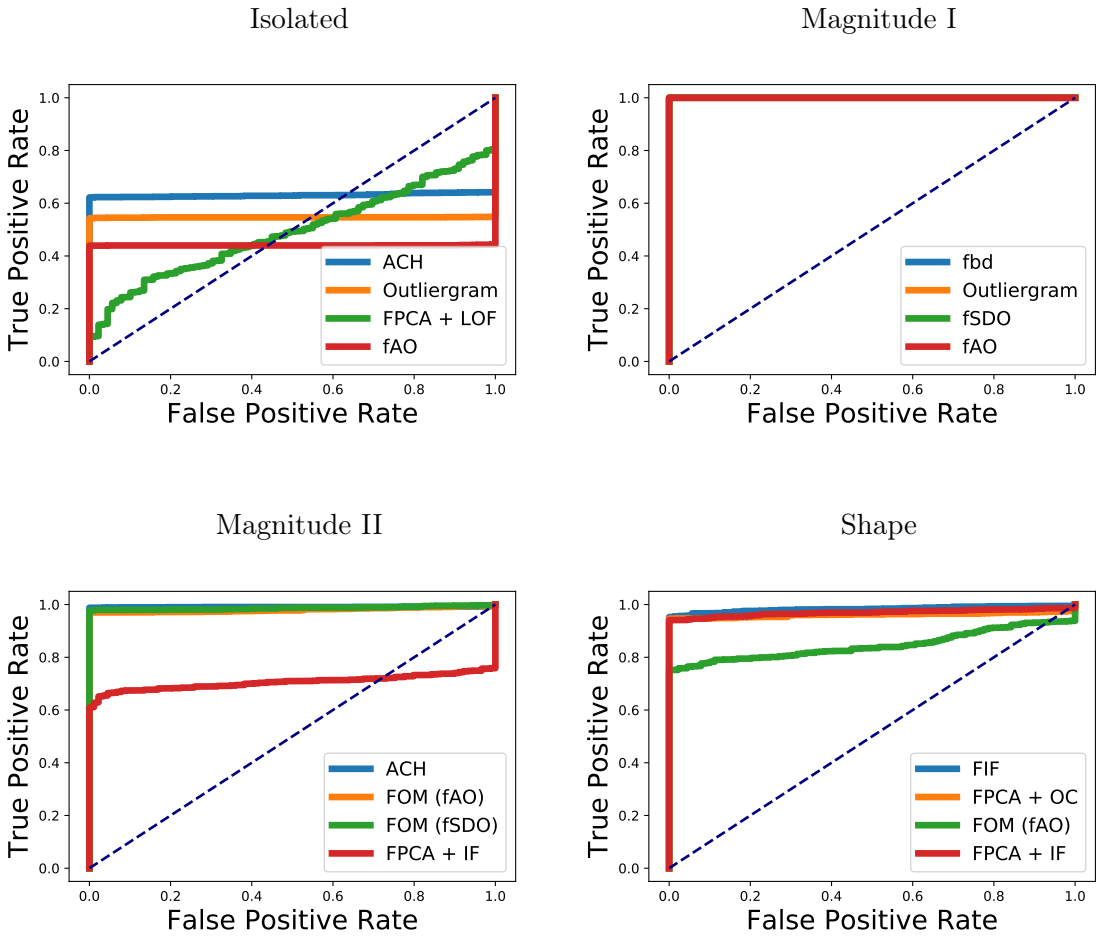


Figure 7.4 – ROC curves of the four best methods for the simulation study.

normal data that have (on an average) same values as the average value of anomalies and that can not be distinguished by any of the algorithms. Such removal shall not prioritize any of the methods. We further perform a linear interpolation in order to obtain data with equidistant measurement points.

7.3.1 Visualization

While a simple plot of the rocks data provides relevant information on the data shape (see Section 7.1), the richness of the aeronautics curves makes it impossible. To get a first insight into the structure of the aeronautics data set under study, we start with meaningful visualization. Recently, many graphical tools have been proposed in the literature, such as general-purpose functional highest density region plots (Hyndman and Shang, 2010), functional boxplots (Sun and Genton, 2011) or amplitude and phase boxplot displays (Xie et al., 2017), as well as those especially designed for anomaly detection such as the Outliergram (Arribas-Gil and Romo, 2014), the Functional Outlier Map (FOM; Hubert et al., 2015; Rousseeuw et al., 2018) or the Magnitude-Shape plot (MS; Dai and Genton, 2019a). This last group, together with the generic Functional Principal Component Analysis (FPCA) constitute our interest.

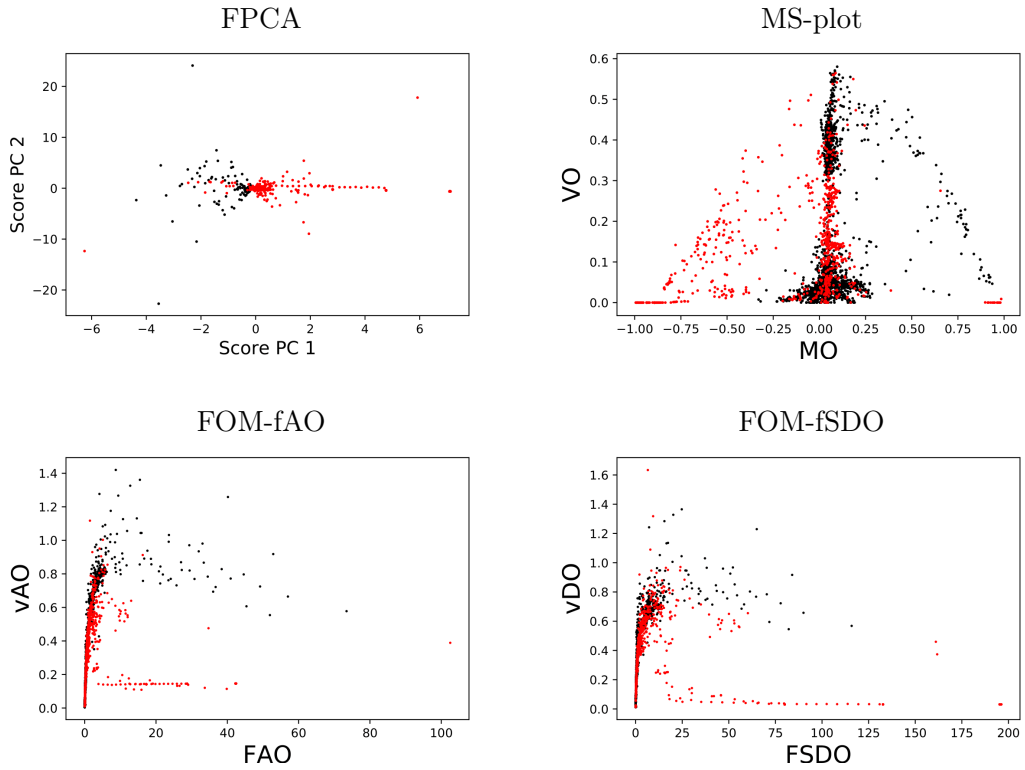


Figure 7.5 – Four visualization plots from Airbus data. In order from top to bottom and from left to right: FPCA, Magnitude-Shape plot with Integrated Tukey depth, Functional Outlier map with fAO and Functional Outlier map with fSDO. Black points corresponds to normal curves, red points to anomalies.

The Magnitude-Shape plot is two-dimensional outlyingness (or data depth)-based graphical tool, based on the outlyingness mean and variance—over the time domain—of the functional observation. For a sample of curves $\mathcal{S}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ observed on $\{t_1, \ldots, t_p\}$, the MS plot is then the scatter plot of the points $(\mathrm{MO}_i, \mathrm{VO}_i)_{1 \leq i \leq n}$ with the Mean Outlyingness (MO)

$$\mathrm{MO}_i = \frac{1}{p} \sum_{j=1}^{p} \mathrm{O}(\mathbf{X}_i(t_j) \mid \mathcal{S}_n(t_j)),$$

and the Variational Outlyingness (VO)

$$\mathrm{VO}_i = \frac{1}{p} \sum_{j=1}^{p} \Big(\mathrm{O}(\mathbf{X}(t_j) \mid \mathcal{S}_n(t_j)) - \mathrm{MO}_i \Big)^2,$$

where the outyingness itself is usually a monotone decreasing transform of data depth: $\mathrm{O}(\mathbf{X}_i(t_j), \mathcal{S}_n(t_j)) = \dfrac{1}{\mathrm{D}(\mathbf{X}_i(t_j) \mid \mathcal{S}_n(t_j)) - 1}$.

The Functional Outlier Map (FOM) is similar to the MS plot in the case of univariate functional data, with the only difference in measuring relative instead of absolute variability adapting thus to the actual variability of the function, see Rousseeuw et al. (2018). It is defined as a plot of points

$$\left(\mathrm{MO}_i, \frac{\sqrt{\mathrm{VO}_i}}{1 + \mathrm{MO}_i}\right)_{1 \leq i \leq n}.$$

For explainability purposes, vizualisation tools are computed on the test set where labels are available. Visualization plots, displayed in Figure 7.5, allow to identify certain anomalies (red points correspond to labeled anomalies), while others substantially overlap with the normal data (black points). They reveal the difficulty of marking the entity of abnormal observations for the auronautics data, and explain variations in the performance of different methods used in Section 7.3.2.

7.3.2 Benchmark Study on Aeronautics and Rocks Data

In Section 7.2.2 only four types of anomalies were identified and studied in more detail, while many others remain that are not easy to associate with any existing taxonomy, which is often the challenge of real-world data. While many methods are designed for aiming at certain types of anomalies, clearly no universality in detecting abnormal observations of different types can be generally expected.

To account for multiplicity of possible goals, we use several performance metrics: F1-Score, Average Precision (AP), AUC, and TPR; see Section 7.2.1 for more details.

Table 7.3 displays the results. Methods perform differently and there is no general winner. First observation is the evidence of complexity of the real-world aeronautics data set, which leads to non-perfect results across all the considered methods. This is also due to the fact that test data contains types of anomalies not present in the train data. Second, one should note that FIF appears to have very good performance, being best method in this benchmark in general. Even if its AUC of 76% is not high, it can be still seen as satisfactory provided existence of 29% of anomalies in the test data. Third, the depth-based methods indicate very stable results (mostly relatively satisfactory, except TPR), which comes from the fact that ordering of observations may be very similar for univarite functional curves due to (almost) coincidence of orderings for different univariate depths. While the rocks data set was artificially simplified, the comparison remains similar and thus strengthens the conclusions.

Additionally, we display the ROC curves of the four best methods for both data sets (see Figure 7.6). The true positive rate is maximized by the FIF algorithm when the false positives are very close to zero which is crucial in many applications such as predictive maintenance in the case of airbus. This additional analysis makes FIF preferable to fAO in practice.

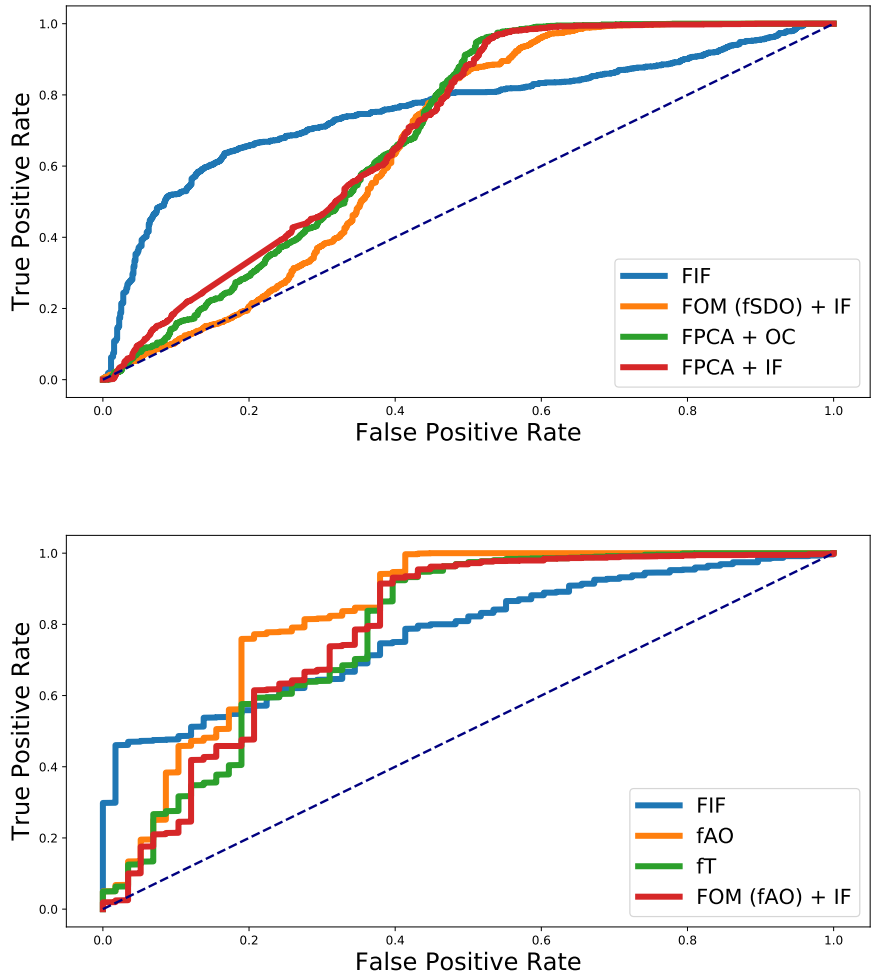


Figure 7.6 – ROC curves of the four best methods for the aeronautics (top) and rocks (bottom) data.

7.4 Conclusion and Perspectives

Due to its unsupervised nature, the anomaly detection problem is extremely challenging. Because the learning algorithms used to build a decision rule in this context rely on "normal" data solely, they mostly consist in identifying regions of the feature space where normal observations are less likely to fall in and cannot be based on empirical estimates of the performance metrics that shall be used afterwards to evaluate them when labels become available. This is even much more challenging in the case of observations valued in a vast functional space. In absence of prior knowledge about the types of the functional anomalies to be detected in the future, it is key to implement very flexible algorithms, exploiting the statistical information at disposal as far as possible. In this chapter, we investigated the performance of recent anomaly detection techniques tailored to the functional framework such as Functional Isolation Forest and ACH depth, by means of real data sets, composed of sequences of accelerometer data measured on helicopters and spectrometry of construction materials. As confirmed by

additional simulation studies, the benchmark revealed a clear advantage of these techniques compared to more traditional methods, relying on a preprocessing of the data (e.g. FPCA). These very encouraging results call for further applications and extensions, to anomaly detection based on multivariate time-series in particular, the behavior of complex infrastructure being often continuously monitored by several sensors and not just one.

Part III

Probability Metrics, Statistical Data Depth and Robustness

8

When OT meets MoM: Robust Estimation of Wasserstein Distance

Contents

Originated from Optimal Transport, the Wasserstein distance has gained importance in Machine Learning due to its appealing geometrical properties and the increasing availability of efficient approximations. It owes its recent ubiquity in generative modeling and variational inference to its ability to cope with distributions having non overlapping support. In this chapter, we consider the problem of estimating the Wasserstein distance between two probability distributions when observations are polluted by outliers. To that end, we investigate how to leverage a Median of Means (MoM) approach to provide robust estimates. Median of Means has been recognized as a versatile tool for robust mean estimation: it substitutes the median of means computed on a partition of the sample to the empirical mean. Exploiting the dual Kantorovitch formulation of the Wasserstein distance, we introduce and discuss novel MoM-based robust estimators whose consistency is studied under a data contamination model and for which convergence rates are provided. In order to assess robustness, we place ourselves in the realistic $\mathcal{O} \cup \mathcal{I}$ framework devoted to data contamination (see e.g. Huber and Ronchetti, 2009 and Lecué and Lerasle, 2020). Beyond computational issues, the choice of the partition size, i.e. the unique parameter of theses robust estimators, is investigated in numerical experiments. Furthermore, these MoM estimators make Wasserstein Generative Adversarial Network (WGAN) robust to outliers, as witnessed by an empirical study on

two benchmarks CIFAR10 and Fashion MNIST.

As a first go, we provide in Section 8.1 a unified and insightful study of the concentration properties of (univariate) MoM-based estimators under a contaminated framework. In Section 8.2 we introduce three new robust estimators of the Wasserstein distance based on median-of-means and median-of-U-statistics and investigate their theoretical properties based upon results of Section 8.1. In Section 8.3, we propose three algorithms in order to compute the previously introduced estimators and robustify the popular Wasserstein GANs. Some concludings remarks are collected in Section 8.4. Eventually, technical proofs are provided in Section 8.5. This chapter gathers contributions of:

- **G. Staerman**, P. Laforgue, P. Mozharovskyi, F. d'Alché-Buc. When OT meets MoM: Robust estimation of Wasserstein distance. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136-144, 2021.
- P. Laforgue, **G. Staerman**, S. Clémençon. Generalization Bounds in the Presence of Outliers: a Median-of-Means Study. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 5937-5947, 2021.

8.1 Median-of-Means

The Median-of-Means (MoM) is a robust mean estimator firstly introduced in complexity theory during the 80's (Nemirovsky and Yudin, 1983; Jerrum et al., 1986; Alon et al., 1999). Following the seminal deviation study by Catoni (2012), MoM has lately witnessed a surge of interest, mainly due to its attractive sub-Gaussian behavior, under the sole assumption that the underlying distribution has finite variance (Devroye et al., 2016). Originally devoted to scalar random variables, MoM has notably been extended to random vectors (Minsker, 2015; Hsu and Sabato, 2016; Lugosi and Mendelson, 2019a) and U-statistics (Joly and Lugosi, 2016; Laforgue et al., 2019). As a natural alternative to the empirical mean, MoM has become the cornerstone of several robust learning procedures in heavy-tailed situations, including bandits (Bubeck et al., 2013) and MoM-tournaments (Lugosi and Mendelson, 2019b). A more recent line of work now focuses on MoM's ability to deal with outliers. Aside from concentration results in a contaminated context (Depersin and Lecué, 2019), it has yielded promising applications in robust mean embedding (Lerasle et al., 2019), and the more general MoM-minimization framework (Lecué et al., 2020).

8.1.1 Definition

Let X be a r.v. following a probability distribution $P \in \mathcal{P}(\mathcal{X})$ where $\mathcal{X} \subset \mathbb{R}^d$. Given a sample $X_1, \ldots, X_n$ i.i.d. as X, the Median-of-Means (MoM) is an estimator of $\mathbb{E}_P[X]$ built as follows. First, choose $K_X \leq n$, and partition $\{1, \ldots, n\}$ into K_X disjoint blocks $\mathcal{B}_1^X, \ldots, \mathcal{B}_{K_X}^X$ of size $B_X = n/K_X$. If n cannot be divided by K_X, some observations may be removed. Then, empirical means are computed on each of the K_X blocks. The estimator returned is finally the median of the empirical means thus computed. For a function $\Psi \colon \mathcal{X} \to \mathbb{R}$, the MoM estimator of $\mathbb{E}_P[\Psi(X)]$ is then formally given by:

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

This estimator provides an attractive alternative to $\overline{\Psi}_X = (1/n)\sum_{i=1}^n \Psi(X_i)$ for robust learning. Indeed, it has been shown to (i) exhibit a sub-Gaussian behavior under only a finite variance assumption, making it particularly suited to heavy-tailed distributions, and (ii) be non-sensitive to outliers (see Section 1.3.2 and references therein). When the function Ψ is the identity function, the MoM estimator is denoted by MoM_X.

MoM also nicely adapts to multisample U-statistics of arbitrary degrees (Lee, 1990). Medians-of-U-statistics (MoU) naturally extend the MoM approach by considering the median of U-statistics built on disjoint blocks (see Joly and Lugosi (2016) for the case of degenerate U-statistics, or Laforgue et al. (2019) for a general study on randomized, possibly overlapping, blocks). Assume that one is interested in estimating $\mathbb{E}_{P\otimes Q}[h(X,Y)]$, for some kernel $h\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ and some r.v. Y drawn from $Q\in\mathcal{P}(\mathcal{Y})$ with $\mathcal{Y}\subset\mathbb{R}^d$. Given two samples $X_1,\dots,X_n$ and $Y_1,\dots,Y_m$ i.i.d. from P and Q respectively, a natural idea then consists in partitioning both $\{1,\dots,n\}$ and $\{1,\dots,m\}$ into $\mathcal{B}_1^X,\dots,\mathcal{B}_{K_X}^X$ and $\mathcal{B}_1^Y,\dots,\mathcal{B}_{K_Y}^Y$ respectively, with $K_Y\leq m$, and $B_Y = m/K_Y$. One may then compute U-statistics on each pair of blocks (k,l) for $k\leq K_X$ and $l\leq K_Y$, and return the median of the $K_X\times K_Y$ U-statistics. However, this construction introduces dependence between the base estimators, making the theoretical study more difficult. An alternative then consists in choosing $K_X = K_Y = K$, and considering only the diagonal blocks (see Figure 8.1c). These two estimators are referred to as (diagonal) Median-of-U-statistics (MoU), and using $\mathcal{B}_{k,l}^{XY}$ to denote the block of tuples (X_i,Y_j) such that $X_i\in\mathcal{B}_k^X$ and $Y_j\in\mathcal{B}_l^Y$, they are formally given by:

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

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

8.1.2 Concentration Bounds for MoM and MoU

The recent resurgence of interest for MoM in the statistical literature dates back to the seminal deviation studies by Audibert and Catoni (2011) and Catoni (2012), that propose to assess an estimator through its deviation probabilities, rather than by computing its quadratic risk. Extensively studied since then, MoM now benefits from a large corpus of concentration results. For instance, a proof of its behavior under finite variance assumption can be found in Devroye et al. (2016).

Proposition 8.1. *(Devroye et al., 2016) Suppose that an i.i.d. sample $X_1,\dots,X_n$ is drawn from $P\in\mathcal{P}(\mathbb{R})$ having finite variance σ^2. Then, for any $\delta\in[e^{1-n/2},1[$, choosing $K_X = \lceil\log(1/\delta)\rceil$, it holds with probability at least $1-\delta$:*

$$\left|\mathrm{MoM}_X - \mathbb{E}_P[X]\right| \;\leq\; 2\sqrt{2}e\;\sigma\sqrt{\frac{1+\log(1/\delta)}{n}}.$$

These concentration results have further been extended to random vectors, through different generalizations of the median in a multidimensional setting (Minsker, 2015; Hsu and Sabato, 2016; Lugosi and Mendelson, 2019a), and to U-statistics (Joly and Lugosi (2016) for the degenerate case, Laforgue et al. (2019) with randomized blocks) among other extensions. Such interesting properties in the presence of heavy-tailed

data has given birth to numerous applications in statistical learning. This includes an adaptation of the Upper Confidence Bound (UCB) bandit algorithm in Bubeck et al. (2013), of Empirical Risk Minimization (ERM) in Brownlees et al. (2015), or the more general framework of MoM-tournaments (Lugosi and Mendelson, 2019b) and Le Cam's approach (Lecué and Lerasle, 2019).

A recent line of work is now trying to change perspective, abandoning the heavy-tailed framework (i.e. finite variance assumption) to focus on MoM's behavior within contamination models. In this setting, the i.i.d. assumption is relaxed, and the following assumption is instead adopted.

Assumption 8.2. *Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ be two samples. The first sample is polluted with $n_{\mathcal{O}} < n/2$ (possibly adversarial) outliers. The remaining $n - n_{\mathcal{O}}$ points are informative data, or* inliers*, independently distributed according to $P \in \mathcal{P}(\mathcal{X})$, where $\mathcal{X} \subset \mathbb{R}^d$. A similar assumption is made on the second sample, which is supposed to contain $m_{\mathcal{O}} < m/2$ arbitrary outliers, and $m - m_{\mathcal{O}}$ inliers drawn from $Q \in \mathcal{P}(\mathcal{Y})$, where $\mathcal{Y} \subset \mathbb{R}^d$. In contrast, no assumption is made on the outliers. The proportions of outliers in both samples are denoted by $\tau_X = n_{\mathcal{O}}/n$ and $\tau_Y = m_{\mathcal{O}}/m$ respectively.*

Assumption 8.2 is thus addressed through the general angle of MoM-minimization in Lecué et al. (2020), while Lerasle et al. (2019) develop an application to Maximum Mean Discrepancy and outlier-robust mean embedding. Depersin and Lecué (2019) propose a sub-Gaussian MoM-inspired multidimensional estimator computable in almost linear time, and Depersin (2020) studies a multivariate estimator based on one-dimensional projections. However, all these works rely on *ad-hoc* assumptions that are quite difficult to interpret. For instance, Lecué et al. (2020) use unusual outlier-adapted Rademacher complexities, while the choice of K_X is based on unknown constants in Depersin (2020), or defined implicitly in Lerasle et al. (2019). In Depersin and Lecué (2019), the choice of K_X incidentally reduces the analysis to the case where $\tau_X \leq 0.33\%$.

In contrast, this section proposes a unified study of the concentration properties of (univariate) MoM-based estimators under the contamination regime of Assumption 8.2. In particular, we show that MoM is able to handle up to 50% of outliers, at the price of a degraded constant though. Indeed, our bounds allow to encapsulate the impact of the proportion of outliers τ_X into constant terms only achieved through an explicit value for the number of blocks K_X. Assuming the inliers to be sub-Gaussian, we show that MoM becomes efficient on a wide interval, allowing next to derive bounds in expectation (we are not aware of similar results for MoM) under the following assumption stipulating that the number of outliers $n_{\mathcal{O}}$ grows sub-linearly with n.

Roughly, we want $K_X > 2n_{\mathcal{O}}$ to ensure that blocks without outliers are in majority. However, if K_X is too large MoM tends to the median, which is a bad estimator of the mean in general. To correctly calibrate K_X, we propose to choose $\tau_X \mapsto \sqrt{2\tau_X}$ as an upper bound of $\tau_X \mapsto 2\tau_X$. This way, setting $K_X \approx \sqrt{2\tau_X}n > 2\tau_X n = 2n_{\mathcal{O}}$ satisfies the outlier constraint, while refraining from choosing too large values. It is worth mentioning that any function $\varrho : \tau_X \mapsto \varrho(\tau_X)$ that satisfies $2\tau_X < \varrho(\tau_X) < 1$ can be chosen instead of $\sqrt{2\tau_X}$ leading to slightly different bounds.

We now derive the concentration of MoM under the contamination regime of Assumption 8.2 in the next proposition and can be found in Laforgue et al. (2021).

Proposition 8.3. *Let $P \in \mathcal{P}(\mathcal{X})$ be ρ sub-Gaussian, suppose that the sample $X_1, \ldots, X_n$ satisfies Assumption 8.2, and define $\Gamma\colon \tau \mapsto \sqrt{1+\sqrt{2\tau}}/\sqrt{1-2\tau}$. Thus, for all $\delta \in]0, e^{-4n\sqrt{2\tau_X}}]$, with $K = \lceil \sqrt{2\tau_X} n \rceil$, it holds with probability at least $1-\delta$:*

$$\left|\mathrm{MoM}_X - \mathbb{E}_P[X]\right| \;\leq\; 4\rho\, \Gamma(\tau_X) \sqrt{\frac{\log(1/\delta)}{n}}. \tag{8.1}$$

The technical proof is given in Section 8.5.1. Its argument essentially consists in using that the MoM estimator has a similar behavior to that of a majority of block means. The condition $K > 2n_{\mathcal{O}}$ is strengthened into $K \geq \sqrt{2\tau_X} n$ ensuring that a fraction $(\sqrt{2\tau_X} - \tau_X)/\sqrt{2\tau_X} > 1/2$ of "sane" blocks (i.e. including none of the $n_{\mathcal{O}}$ outliers) actually constitutes a majority of blocks. One may then focus on the sane blocks deviations only, which is controlled by means of the concentration properties of a Binomial random variable.

A δ-limited sub-Gaussian tail bound. We first point out that the main price to pay for extending the sub-Gaussian tail behavior of MoM to the contaminated framework of Assumption 8.2 is the limited range of acceptable confidence levels $1-\delta$. This type of limitation is typical of MoM's concentration results. The upper limit value comes from the constraint $2n_{\mathcal{O}} < K_X$ (or $\sqrt{2\tau_X} n \leq K_X$), and is specific to the contaminated framework. It should be noticed that this restriction vanishes (i.e. the upper limit value is 1) when $\tau_X = 0$.

About the constants. An interesting property of the bound derived in Proposition 8.3 is that it fully encapsulates the impact of the proportion of outliers τ_X into the constant $\Gamma(\tau_X)$. Naturally, the latter increases with τ_X, and tends to infinity as τ_X goes to $1/2$.

Accuracy vs range of confidence levels. The choice of the mapping $\tau_X \mapsto \sqrt{2\tau_X}$ as upper-bound of the $\tau_X \mapsto 2\tau_X$ determines at the same time the range $(0, e^{-4n\sqrt{2\tau_X}}]$ for which the equation (8.1) holds true with probability at least $1-\delta$, and the constant $\Gamma(\tau_X)$. Therefore, there is a trade-off between the size of the range for the confidence levels and the order of magnitude of the constant $\Gamma(\tau_X)$, both decreasing with τ_X.

Rate bound. MoM trades the ability of discarding outliers for the degradation of its statistical guarantees to those of one single sane block, of order $1/\sqrt{B_X} = \sqrt{K_X/n} \approx \sqrt{n_{\mathcal{O}}/n}$, as K_X is roughly of the order of $n_{\mathcal{O}}$. Hence, if $n_{\mathcal{O}}$ grows linearly with n, then B_X stays bounded and guarantees do not improve with n. This also highlights the importance of not choosing a too rough upper bound of $2n_{\mathcal{O}}$. We finally highlight that this rate is optimal. Indeed, our bounds are obtained after conditioning upon the observations and, as can be seen by examining proofs, they cannot be refined, insofar as they simply rely on exact computations of the binomial distribution.

Unknown τ_X. In practice, the proportion of outliers τ_X is generally unknown, preventing from using it to calibrate K_X. We emphasize that the above stated bounds may still be used with an overestimation of τ_X, at the price of a deterioration of $\Gamma(\tau_X)$ though.

Related work. Although they are quite similar in spirit, six critical points distinguish Proposition 8.3 from Theorem 1 in Lerasle et al. (2019). (*i*) It is important to notice first that Proposition 8.3 focuses on the deviations of scalar MoMs, while Theorem 1 in Lerasle et al. (2019) addresses that of particular kernel mean embeddings, defined as MoM minimizers. (*ii*) This being said, our choice of K_X can be computed explicitly from the total proportion of outliers τ_X, and the targeted confidence δ. In contrast,

the number of blocks in Lerasle et al. (2019) depends on the proportion of outliers with respect to the number of blocks itself, resulting in a recursive definition, hard to disambiguate. This inherent difficulty is typically overcome here by reparameterizing using $(\sqrt{2\tau_X} - \tau_X)/\sqrt{2\tau_X}$. (*iii*) As a consequence, our bound features the true and fixed proportion of outliers τ_X within the sample, while Lerasle et al. (2019) use the proportion w.r.t. the number of blocks, that may change with it. (*iv*) Additionally, their range of admissible confidence levels $1-\delta$ is defined implicitly, whereas we provide an explicit interval, that depends only on τ_X and n. (*v*) Lerasle et al. (2019) require $2n_{\mathcal{O}} \leq K_X \leq n/2$, meaning they allow at most 25% of outliers, while we can handle up to 50%. (*vi*) They only prescribe a rough estimate of K_X, that might not be an integer.

The next proposition details the concentration guarantees of $\mathrm{MoU}^{\mathrm{diag}}_{XY}[h]$ with a symmetric kernel $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, that corresponds to independent U-statistics, when the samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are contaminated according to Assumption 8.2.

Proposition 8.4. *Suppose that samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ satisfy Assumption 8.2. Define $\Gamma : \tau \mapsto \sqrt{1+\sqrt{2\tau}}/\sqrt{1-2\tau}$ and assume that $\tau_X + \tau_Y < 1/2$. In addition, if $n = m$ and if the essential supremum $\|h(X,Y)\|_{ess,\infty}$ is finite and upper bounded by M, then for all $\delta \in]0, e^{-4n\sqrt{2(\tau_X+\tau_Y)}}]$, choosing $K_X = \lceil \sqrt{2(\tau_X+\tau_Y)}n \rceil$, it holds with probability at least $1-\delta$:*

$$\left|\mathrm{MoU}^{diag}_{XY}[h] - \mathbb{E}[h(X,Y)]\right| \leq \; 8M\; \Gamma(\tau_X+\tau_Y)\; \sqrt{\frac{\log(1/\delta)}{n}}.$$

The technical proof is detailed in 8.5.2. Notice that the constraint $n = m$ can be relaxed, as long as $2(n_{\mathcal{O}} + m_{\mathcal{O}}) \leq \min(n, m)$ still holds. However, the case $n = m$ is the only one documented in MoM's literature to our knowledge (Lerasle et al., 2019), while it nicely exhibits the critical point $\tau_X + \tau_Y = 1/2$. When estimating Integral Probability Metrics (Sriperumbudur et al., 2012), one typically relies on two-sample U-statistics, built upon kernels of the form $h_\Psi(X,Y) = \Psi(X) - \Psi(Y)$, for Ψ in the functional set considered. Hence, one might use a MoM-MoM estimate, instead of a MoU or a $\mathrm{MoU}^{\mathrm{diag}}$ estimate (see the next section). The corresponding proportions of outliers admitted would be $\tau_X < 1/2$, and $\tau_Y < 1/2$, representing a less stringent constraint, as shown in Figure 8.1d. In the next section, we introduce MoM/MoU based estimators of the Wasserstein distance described in Staerman et al. (2021a).

8.2 When Wasserstein meets MoM

In this section, we investigate how MoM/MoU estimators can be leveraged to define and analyze new estimators of $\mathcal{W}(P,Q)$ that exhibit strong theoretical guarantees in presence of outliers.

8.2.1 MoM and MoU-based Estimators

Starting from the expression of the dual of the Wasserstein distance given by:

$$\mathcal{W}(P,Q) = \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \mathbb{E}_P\left[\Psi(X)\right] - \mathbb{E}_Q\left[\Psi(Y)\right],$$

we observe that it can be considered with a two-fold perspective. The first one consists in considering the Wasserstein distance as the supremum of the difference between two expected values. The second one, obtained by linearity of the expectation, rather regards $\mathcal{W}(P,Q)$ as the supremum of single expected values, but taken with respect to the tuple (X,Y), and associated to the kernel: $h_\Psi\colon (X,Y)\mapsto \Psi(X)-\Psi(Y)$.

Although quite elementary at first sight, this two-fold perspective gains complexity when applied to the empirical distributions $P_n=(1/n)\sum_{i=1}^n \delta_{X_i}$ and $Q_m=(1/m)\sum_{j=1}^m \delta_{Y_j}$. Indeed, following the first perspective, the natural estimator obtained is the supremum of the differences between two empirical averages, while the second one leads to the supremum of 2-samples U-statistics of degrees $(1,1)$ and kernels h_Ψ. So far, both points of view are strictly equivalent by linearity of the expectation and the empirical mean. However, this equivalence breaks down as soon as non-linearities are introduced, through MoM-like estimators for instance. We therefore introduce three distinct estimators of $\mathcal{W}(P,Q)$, that differ upon which estimator of Section 8.1.1 is used.

Definition 8.5. *We define the Median-of-Means and the Median-of-U-statistics estimators of the* 1*-Wasserstein distance as follows:*

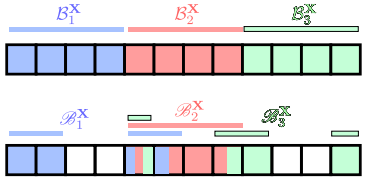
$$\mathcal{W}_{\mathrm{MoM}}(P_n,Q_m)=\sup_{\Psi\in\mathcal{F}_{\mathrm{Lip}}}\{\mathrm{MoM}_X[\Psi]-\mathrm{MoM}_Y[\Psi]\},$$

$$\mathcal{W}_{\mathrm{MoU}}(P_n,Q_m)=\sup_{\Psi\in\mathcal{F}_{\mathrm{Lip}}}\{\mathrm{MoU}_{XY}[h_\Psi]\},$$

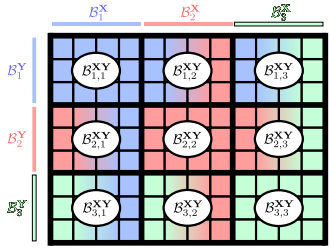
$$\mathcal{W}_{\mathrm{MoU-diag}}(P_n,Q_m)=\sup_{\Psi\in\mathcal{F}_{\mathrm{Lip}}}\{\mathrm{MoU}^{\mathrm{diag}}_{XY}[h_\Psi]\}.$$

While $\mathcal{W}_{\mathrm{MoM}}$ relies on the difference between individual median blocks, $\mathcal{W}_{\mathrm{MoU}}$ considers the median over all possible combinations of blocks between X and Y. As an intermediate step, $\mathcal{W}_{\mathrm{MoU-diag}}$ looks after diagonal blocks only. The latter formulation is used in Lerasle et al. (2019) to derive robust mean embedding and Maximum Mean Discrepancy estimators. The theoretical analysis is made simpler by the independence between the blocks, but the estimator suffers from an increased variance due to the important loss of information, see Figure 8.1c and Joly and Lugosi (2016). It should be noticed however that $\mathcal{W}_{\mathrm{MoU-diag}}$ enjoys a much lower computational cost in practice.

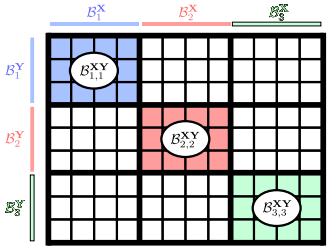
One elegant way to combine both benefits, i.e. small loss of information and low computational cost, is to consider randomized blocks (Laforgue et al., 2019). Instead of partitioning the data set, this method builds blocks by sampling them independently through simple Sampling Without Rejection (SWoR). One consequence is the possibility for the randomized blocks to overlap (see Figure 8.1a), making the estimator's concentration analysis more difficult. Nevertheless, guarantees similar to that of MoM have been established (up to constants), and the extension to 2-sample U-statistics built on randomized blocks allows for a better exploration of the grid than through $\mathrm{MoU}^{\mathrm{diag}}$, see Figure 8.1e. However, despite the possibility to reach every part of the grid, the exploration scheme illustrated in Figure 8.1e have a fixed structure (e.g. always 3 cells per column, 4 cells per row). The *totally free* alternative, as depicted in Figure 8.1f, consists in sampling directly from the pairs of observations, which generates incomplete U-statistics. If no theoretical guarantees have been established for this extension due to the complex replication setting between blocks, it still benefits from good empirical results (Laforgue et al., 2019), consistent with the grid covering it allows.



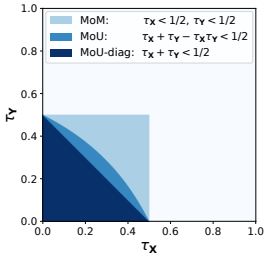
(a) 1D standard and randomized MoM



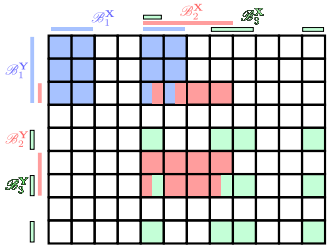
(b) MoU_{XY}



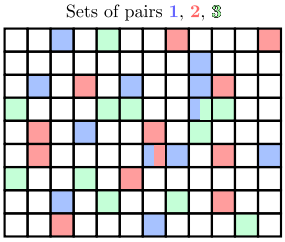
(c) $\mathrm{MoU}_{XY}^{\mathrm{diag}}$



(d) Admitted proportion of outliers



(e) MoU based on randomized blocks



(f) MoU based on randomized pairs

Figure 8.1 – Sampling strategies to build MoM and MoU, as well as admitted proportion of outliers.

Another important question to be addressed is: *how to handle the non-differentiability introduced by the median operator?* Indeed, the Wasserstein distance often acts as a loss function, e.g. in generative modeling (VAEs, GANs), and optimizing through a MoM/MoU-based criterion then becomes crucial. One answer is to uses a MoM-gradient descent algorithm (Lecué et al., 2020). It consists in performing a mini-batch gradient step based on the median block. In order to avoid local minima, authors propose to shuffle the partition at each step of the descent, leading to the minimization of an expected MoM loss (w.r.t. the shuffling) that is more stable. Notice that this method goes beyond random partitions, and easily adapts to the randomized extensions discussed above.

8.2.2 Theoretical Guarantees

We now establish the statistical guarantees satisfied by the estimators introduced in Definition 8.5 under Assumption 8.2. First notice that if P_n^{MoM} denotes with a language abuse the *measure* such that for all function $\Psi\colon \mathbb{R}^d \to \mathbb{R}$ it holds $\mathbb{E}_{P_n^{\mathrm{MoM}}}\left[\Psi\right] = \mathrm{MoM}_X[\Psi]$, it is direct to see that $\mathcal{W}_{\mathrm{MoM}}(P_n, Q_m) = \mathcal{W}(P_n^{\mathrm{MoM}}, Q_m^{\mathrm{MoM}})$. Then, it holds $\mathcal{W}_{\mathrm{MoM}}(P_n, Q_m) - \mathcal{W}(P, Q) \leq \mathcal{W}(P, P_n^{\mathrm{MoM}}) + \mathcal{W}(Q_m^{\mathrm{MoM}}, Q)$, and one may only focus on

the theoretical guarantees of the right-hand side terms. Before stating our main results, we need additional assumptions on the feature spaces and on the numbers of outliers $n_{\mathcal{O}}$ and $m_{\mathcal{O}}$, which are assumed to grow sub-linearly with respect to n and m.

Assumption 8.6. *The feature spaces $\mathcal{X}$ and $\mathcal{Y}$ of inliers defined in Assumption 8.2 both lie in a compact set $\mathcal{K}^d \subset \mathbb{R}^d$.*

Assumption 8.7. *There exist $C_{\mathcal{O}} \geq 1$ and $0 \leq \alpha_{\mathcal{O}} < 1$ such that $n_{\mathcal{O}} \leq C_{\mathcal{O}}^2\, n^{\alpha_{\mathcal{O}}}$ and $m_{\mathcal{O}} \leq C_{\mathcal{O}}^2\, m^{\alpha_{\mathcal{O}}}$.*

We start by an asymptotic result establishing the strong consistency of estimators in Definition 8.5. It highlights the different outlier configurations allowed through conditions on the proportions of outliers τ_X and τ_Y.

Proposition 8.8. *Suppose that samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ satisfy Assumptions 8.2, 8.6 and 8.7. Then, choosing $K_X = \lceil \sqrt{2\tau_X}\ n \rceil$, as $n \to \infty$, it holds:*

$$\mathcal{W}(P_n^{\mathrm{MoM}}, P) \xrightarrow{a.s} 0.$$

If moreover $\tilde{\tau} \coloneqq \tau_X + \tau_Y - \tau_X\tau_Y < 1/2$, then choosing $K_X = \lceil \sqrt{2\tilde{\tau}}\ n \rceil$ and $K_Y = \lceil \sqrt{2\tilde{\tau}}\ m \rceil$, as $n \to \infty$ and $m \to \infty$, it holds:

$$\left| \mathcal{W}_{\mathrm{MoU}}(P_n, Q_m) - \mathcal{W}(P, Q) \right| \xrightarrow{a.s} 0.$$

If finally $\tau_X + \tau_Y < 1/2$ and $n = m$, then choosing $K_X = K_Y = \lceil \sqrt{2(\tau_X + \tau_Y)}\ n \rceil$, as $n \to \infty$, it holds:

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

The key argument in the proof of Proposition 8.8 consists in converting the convergence of the different estimators into the convergences of blocks containing no outliers. The proof is postponed in Section 8.5.3. Numbers of blocks K_X and K_Y are chosen such that (i) such blocks are always in majority, and (ii) their sizes n/K_X and m/K_Y go to infinity as n and m go to infinity. Any other choice of K_X and K_Y that satisfies this two conditions also ensures convergence. If the outliers proportions are unknown, building K_X and K_Y from upper bounds of τ_X and τ_Y thus does not impact Proposition 8.8. The conditions on K_X and K_Y also constrain the proportions of outliers admitted, as illustrated in Figure 8.1d. The assumption $n = m$ for $\mathcal{W}_{\mathrm{MoU-diag}}$ is necessary to be able to build a majority of sane blocks.

Our next proposition now investigates the nonasymptotic behavior of the proposed estimators.

Proposition 8.9. *Suppose that samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ satisfy Assumptions 8.2 and 8.6, and define $\Gamma \colon \tau \mapsto \sqrt{1 + \sqrt{2\tau}}/\sqrt{1 - 2\tau}$. Then, for all $\delta \in]0, \exp(-4n\sqrt{2\tau_X})]$, choosing $K_X = \lceil \sqrt{2\tau_X}\ n \rceil$, it holds with probability at least $1 - \delta$:*

$$\mathcal{W}(P_n^{\mathrm{MoM}}, P) \leq \frac{C_1(\tau_X)}{n^{1/(d+2)}} + C_2(\tau_X)\sqrt{\frac{\log(1/\delta)}{n}},$$

with $C_1(\tau) = 2 + C_{\mathrm{Lip}}C_2(\tau)$, $C_2(\tau) = 4\ \mathrm{diam}(\mathcal{K}^d)\ \Gamma(\tau)$, and C_{Lip} a universal constant depending only on $\mathcal{F}_{\mathrm{Lip}}$.

If furthermore $\tau_X + \tau_Y < 1/2$ *and* $n = m$*, then for all* $\delta \in]0, \exp(-4n\sqrt{2(\tau_X+\tau_Y)}\,)]$*, choosing* $K_X = K_Y = \lceil \sqrt{2(\tau_X+\tau_Y)}\, n \rceil$*, it holds with probability at least* $1-\delta$*:*

$$\left| \mathcal{W}_{\mathrm{MoU-diag}}(P_n, Q_m) - \mathcal{W}(P,Q) \right| \leq \frac{2C_1(\tau_X+\tau_Y)}{n^{1/(d+2)}} + 2C_2(\tau_X+\tau_Y)\sqrt{\frac{\log(1/\delta)}{n}}.$$

The proof derives from concentration results established in Proposition 8.3, combined with a generic chaining argument. It should be noticed that constant $C_2(\tau_X)$ explodes as τ_X goes to $1/2$, which is expected: the more outliers, the more difficult it is to estimate $\mathcal{W}(P,Q)$. We also stress that the dependence in $1/\sqrt{1-2\tau_X}$ is better than the $1/(1-2\tau_X)^{3/2}$ term exhibited in Lerasle et al. (2019). Integrating the deviation probabilities of Proposition 8.9 and using Assumption 8.7, we finally obtain our main theorem, that provides a nonasymptotic control on the expected value of our estimators deviations from $\mathcal{W}(P,Q)$.

Theorem 8.10. *Suppose that samples* $X_1, \ldots, X_n$ *and* $Y_1, \ldots, Y_m$ *satisfy Assumptions 8.2, 8.6 and 8.7, and recall the notation used in Proposition 8.9. Let* $\beta \in [0,1]$*, then for all* n *such that* $n^{\frac{1}{d+2}+\frac{1-\beta}{2}} \geq C_1(\tau_X)/(2C_2(\tau_X)(2\tau_X)^{\frac{1}{4}})$*, it holds:*

$$\mathbb{E}\left[\mathcal{W}(P_n^{\mathrm{MoM}}, P)\right] \;\leq\; \frac{\kappa_1(\tau_X)}{n^{1/(d+2)}} + \frac{\kappa_2(\tau_X)}{n^{(\beta-\alpha_{\mathcal{O}})/2}} + \frac{\kappa_3(\tau_X)}{n^{\beta/2}},$$

with $\kappa_1(\tau) = C_1(\tau)$*,* $\kappa_2(\tau) = 2C_{\mathcal{O}}C_2(\tau)(2/\tau)^{1/4}$*, and* $\kappa_3(\tau) = \sqrt{\pi}C_2(\tau)/2$*.*

Of course, the above bound only makes sense if $\beta > \alpha_{\mathcal{O}}$*. In particular, if* $\alpha_{\mathcal{O}} \leq d/(d+2)$*, setting* $\beta = 1$ *gives that for all* n *such that* $n^{\frac{1}{d+2}} \geq C_1(\tau_X)/(2C_2(\tau_X)(2\tau_X)^{\frac{1}{4}})$*, with the notation* $\kappa = \kappa_1 + \kappa_2 + \kappa_3$*, it holds:*

$$\mathbb{E}\left[\mathcal{W}(P_n^{\mathrm{MoM}}, P)\right] \;\leq\; \kappa(\tau_X)\; n^{-1/(d+2)}.$$

If furthermore $\tau_X + \tau_Y < 1/2$ *and* $n = m$*, then for all* n *s.t.* $n^{\frac{1}{d+2}} \geq C_1(\tau_X + \tau_Y)/(2C_2(\tau_X+\tau_Y)(2(\tau_X+\tau_Y))^{\frac{1}{4}})$*, with the notation* $\kappa' = 2\kappa_1 + 2\sqrt{2}\kappa_2 + 2\kappa_3$*, it holds:*

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

The proof relies on Proposition 8.9 and is postponed to Section 8.5.5. Theorem 8.10 highlights that the estimators proposed in Definition 8.5 remarkably resist to the presence of outliers in the training data sets. The price to pay is a slightly slower rate of order $O(n^{-1/(d+2)})$, that becomes equivalent in high dimension – the usual setting of Optimal Transport – to the standard $O(n^{-1/d})$ rate. Interestingly, the dependence in the outliers growing rate $\alpha_{\mathcal{O}}$ is made explicit, and is in line with expectations (see below). Unfortunately, the dependency between the blocks makes the nonasymptotic analysis harder for $\mathcal{W}_{\mathrm{MoU}}$ and the computationally cheap randomized extensions discussed in Section 8.2.1. This theoretical challenge is left for future work. We stress that there is no *median-of-means miracle*. If the number of blocks allows to cancel the outliers impact, the statistical performance then scales with the block size, i.e. as $1/\sqrt{B_X} = \sqrt{K_X/n}$. Since K_X is roughly $2n_{\mathcal{O}}$, this means a $\sqrt{n_{\mathcal{O}}/n}$ rate. So if one allows $n_{\mathcal{O}}$ to grow proportionally to n, the bound becomes vacuous. To get guarantees improving with n, we thus need $n_{\mathcal{O}}$ to scale as $n^{\alpha_{\mathcal{O}}}$, for some $\alpha_{\mathcal{O}} < 1$, and the resulting

rate is $n^{(1-\alpha_{\mathcal{O}})/2}$, as found in Theorem 8.10. We finally point out that the condition on n ensures $\mathcal{W}(P_n^{\mathrm{MoM}}, P) \geq C_1(\tau_X)/n^{1/(d+2)} - C_2(\tau_X)\sqrt{\log(1/\delta)/n^{\beta}}$, as the right hand side is negative while $\mathcal{W}(P_n^{\mathrm{MoM}}, P)$ is positive by construction.

Remark 8.11. *The unique property of the Wasserstein distance we used, compared to other Integral Probability Metrics (IPMs; Sriperumbudur et al., 2012), is the way to bound the entropy of the unit ball of Lipschitz functions. The present analysis can thus be extended in a direct fashion to any other IPM that has finite entropy.*

8.3 MoM-based Estimators in Practice

In this section, we first propose a novel algorithm to approximate the MoM/MoU-based estimators using neural networks and provide an empirical study of its behaviour on two toy data sets. Then, we show how to robustify Wasserstein-GANs and present MoMWGAN, a MoM-based variant of GAN, which is evaluated on two well-known image benchmarks.

8.3.1 Approximation Algorithm

As show in Section 8.2, MoM/MoU-based estimation of the Wasserstein distance offers a robust alternative to the classical empirical estimator of $\mathcal{W}$. Indeed, the empirical estimator of $\mathcal{W}$ would not converge towards the target in the $\mathcal{O} \cup \mathcal{I}$ framework. The proposed estimators are consistent and have convergence rates of order $O(n^{-1/(d+2)})$ with the $\mathcal{O} \cup \mathcal{I}$ framework. These convergence rates are similar, when d is not too small, to those of the empirical estimator of $\mathcal{W}$ in a non-contaminated setting. Nevertheless, the question of computing these estimators raises two major difficulties: (i) the optimization over the unit ball of Lipschitz functions is intractable, which is a difficulty common to the approximation of the standard Wasserstein distance, and (ii) the non-differentiability of the median-based loss. The first issue is well known of the practioners of the Wasserstein distance who usually prefer to rely on its primal definition with an entropy-based regularization (Cuturi et al., 2013). However, learning algorithms devoted to Wasserstein GANs overcome this by weight clipping (Arjovsky et al., 2017) or gradient penalization (Gulrajani et al., 2017) to impose to the GAN a Lipchitz constraint. Similarly we propose here to limit Ψ to be a neural network with similar constraints on weights to ensure its ϑ-Lipschitzianity. This enables to approximate the Wasserstein distance up to a (unknown) multiplicative coefficient ϑ. To overpass (ii), one can adopt MoM/MoU gradient descent. Exploited in the context of robust classification (Lecué et al., 2020), using MoM/MoU gradient descent has been proved to be equivalent to minimize the expectation over the sampling strategy of blocks of $\mathcal{W}_{\mathrm{MoM}}, \mathcal{W}_{\mathrm{MoU-diag}}$ and $\mathcal{W}_{\mathrm{MoU}}$. Combining these techniques, we design novel algorithms to compute approximations of the proposed estimators: $\widetilde{\mathcal{W}}_{\mathrm{MoM}}$ (see Algorithm 8.1), $\widetilde{\mathcal{W}}_{\mathrm{MoU-diag}}$ (see Algorithm 8.2) and $\widetilde{\mathcal{W}}_{\mathrm{MoU}}$ (see Algorithm 8.3).

8.3.2 Empirical Study

We denote I_2 the identity matrix of dimension 2 and $\mathbf{v}$, the vector $(v, v)^\top$ with $v \in \mathbb{R}$.

Toy data sets. Two simulated data sets in 2D space with different kinds of anomalies are used in the experiments. The random vectors X_1 and X_2 are chosen to be distributed according a mixture of a standard Gaussian distribution and an "abnormal" distribution,

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 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 $\mathcal{B}_{1,1}^{XY}, \mathcal{B}_{2,2}^{XY}, \ldots, \mathcal{B}_{k,k}^{XY}, \ldots \mathcal{B}_{K,K}^{XY}$ from a sampling scheme and find the median block $\mathcal{B}_{med}^{XY}$

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

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

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

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

Algorithm 8.3 Computation of $\mathcal{W}_{\mathrm{MoU}}(P_n, Q_m)$.

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 $K_X \times K_Y$ disjoint blocks $\mathcal{B}_{1,1}^{XY}, \ldots, \mathcal{B}_{k,l}^{XY}, \ldots \mathcal{B}_{K_X,K_Y}^{XY}$ from a sampling scheme and find the median block $\mathcal{B}_{med}^{XY}$

  $G_w \longleftarrow \left\lfloor K_X/n \right\rfloor \times \left\lfloor K_Y/m \right\rfloor \sum_{(i,j) \in \mathcal{B}_{med}^{XY}} \nabla_w \left[\Psi_w(X_i) - \Psi_w(Y_j) \right]$

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

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

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

respectively $\mathcal{A}_1$ and $\mathcal{A}_2$ defined as follows. $\mathcal{A}_1$ is the uniform distribution in $[-\mathbf{50}, \mathbf{50}]$ that mimics *isolated outliers* while $\mathcal{A}_2$ is the standard Cauchy distribution shifted by 25, defined to mimic *aggregated outliers* (see e.g. Chandola et al., 2009). The random vector Y is Gaussian with $Y \sim \mathcal{N}(\mathbf{5}, \mathrm{I}_2)$, Data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ contain 500 independent and identical copies of (X_1, Y), (X_2, Y) respectively, with the same proportion of outliers τ_X.

Evaluation metrics. The Lipchitz constant ϑ being unknown and highly depending of the clipping parameter choice, it wouldn't be appropriate to compare the true L_1-Wasserstein value, equal to $\sqrt{50}$, with $\widetilde{\mathcal{W}}_{\text{MoM}}, \widetilde{\mathcal{W}}_{\text{MoU}-\text{diag}}$ and $\widetilde{\mathcal{W}}_{\text{MoU}}$. Therefore, we propose to compare $\widetilde{\mathcal{W}}_{\text{MoM}}$, $\widetilde{\mathcal{W}}_{\text{MoU}-\text{diag}}$ and $\widetilde{\mathcal{W}}_{\text{MoU}}$ to $\widetilde{\mathcal{W}}$, the L_1-Wasserstein distance approximated by Algorithm 8.1, when MoM is not used, e.g. $K_X = K_Y = 1$, by measuring the absolute error between them.

Influence of K_X, K_Y. The numbers of blocks, K_X and K_Y, are crucial parameters for computation. They define the trade-off between the robustness of the estimator and computational burden. However the theory does not give enough insights about their value: the necessary assumption for the consistency is only that they should be greater than $2\tau_X n$ (see Section 8.2.2). An empirical study of the influence of their values on the behavior of the approximations of $\mathcal{W}_{\text{MoM}}, \mathcal{W}_{\text{MoU}-\text{diag}}$ and $\mathcal{W}_{\text{MoU}}$ is therefore much useful. For sake of simplicity, we set $K_X = K_Y$ in the subsequent experiments.

In a first experiment, we explore the ability of Algorithm 8.1, Algorithm 8.2 and Algorithm 8.3 to override outliers according to the values of K_X and with different rates of outliers τ_X. The approximations $\widetilde{\mathcal{W}}_{\text{MoM}}$,$\widetilde{\mathcal{W}}_{\text{MoU}-\text{diag}}$ and $\widetilde{\mathcal{W}}_{\text{MoU}}$ are computed using a simple multilayer perceptron (MLP) with one hidden layer and MoM gradient descent over several τ_X and K_X on both data sets. The experiment is repeated 20 times with various seeds. Mean results are displayed. Figure 8.2 represents absolute deviations between the L_1-Wasserstein distance approximated with a MLP when $\tau_X = 0$ and $\widetilde{\mathcal{W}}_{\text{MoU}-\text{diag}}$ with various anomalies settings and different values of K_X. Shaded areas, in Figure 8.2 represent 25%-75% quantiles over the 20 repetitions. On both data sets and for the three estimators, we observe that the approximation algorithm succeeds to provide an estimation of $\mathcal{W}_{\text{MoU}-\text{diag}}$, able to override outliers with different $\tau_{\mathbf{X}}$ while K_X is high enough. From Section 8.2.2, we know that K_X needs to be higher than $2\tau_X n$ to have theoretical guarantees. Experiments show that in practice, this condition is not necessary in every situations. For example, when $\tau_X = 0.1$ (i.e. 50 anomalies) in Figure 8.2 (left), only 70 blocks are needed to override outliers. The reason is that hypothesis makes things work in the worst case, i.e. when each outlier is isolated in one block which lead to have $\tau_X n$ contaminated blocks. This is rarely the case in practice, several blocks can be contaminated by many outliers and this is why fewer blocks are needed. Results for $\widetilde{\mathcal{W}}_{\text{MoU}}$ and $\widetilde{\mathcal{W}}_{\text{MoM}}$ are displayed in Figure 8.3 and are quite similar due to the simplicity of the problem.

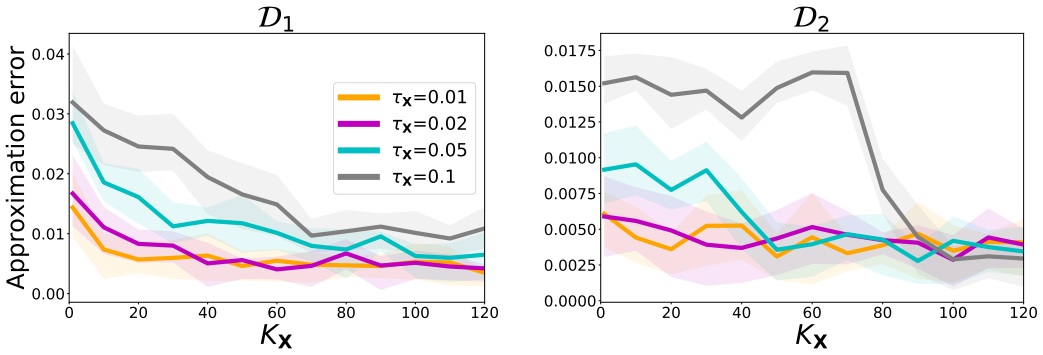


Figure 8.2 – $\widetilde{\mathcal{W}}_{\text{MoU}-\text{diag}}$ over K_X for different anomalies proportion τ_X on $\mathcal{D}_1$ (left) and $\mathcal{D}_2$ (right).

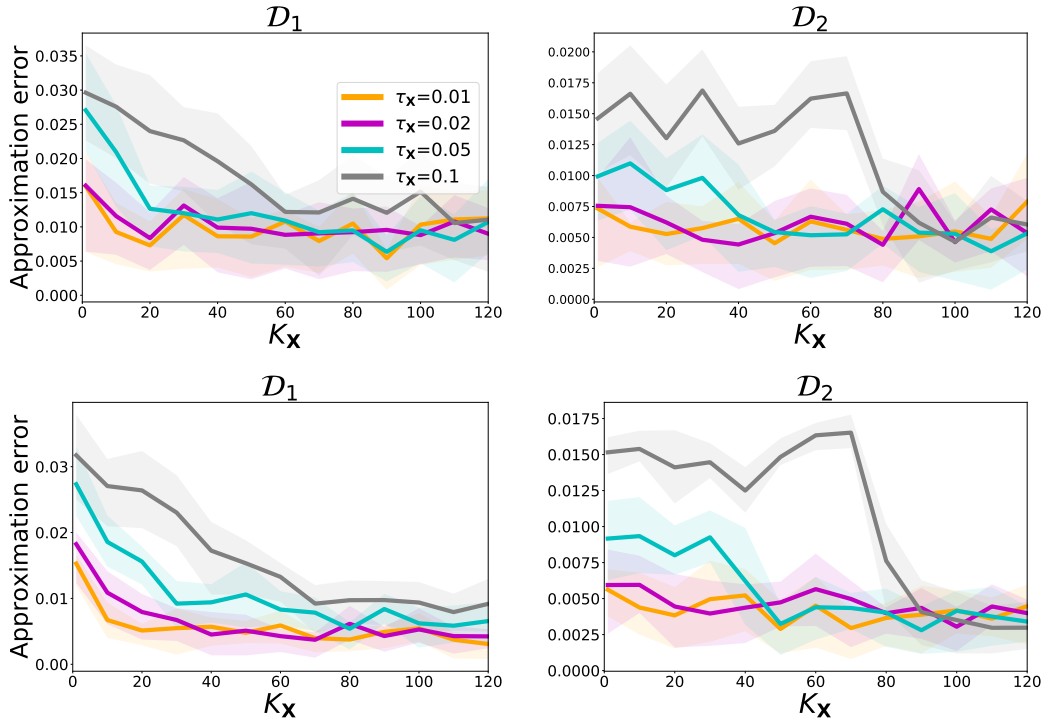


Figure 8.3 – $\widetilde{\mathcal{W}}_{\text{MoU}}$ (top) and $\widetilde{\mathcal{W}}_{\text{MoM}}$ (bottom) over K_X for different anomalies proportion τ_X on $\mathcal{D}_1$ (left) and $\mathcal{D}_2$ (right).

In a second experiment illustrated by Figure 8.4, we study the convergence of the approximation algorithm with and without anomalies for different values of K_X on $\mathcal{D}_1$. To get a fair comparison between the different settings of algorithms, we compare the predicted values across the "learning" epoch. Here during one epoch, the algorithm has made a gradient pass over the whole data set, which means that one epoch corresponds one iteration of the approximation algorithm if $K_X = 1$ (no MoM estimation), and to K_X iterations, in the other cases. In both cases (with or without anomalies), the higher K_X is, the faster the approximation algorithm converges. Without surprise, the MoM approach benefits from the same properties than a mini-batch approach. When there are no anomalies, the distance values reached after convergence are close to the "true" value (obtained with the plain estimator when $K_X = 1$), especially when K_X is lower. This means that the MoM-based algorithm can be used routinely instead of the plain estimator. With 5% of anomalies, one can see that distance values reached after convergence get closer to the target as K_X grows. Results for $\widetilde{\mathcal{W}}_{\text{MoU}}$ and $\widetilde{\mathcal{W}}_{\text{MoM}}$ are displayed in Figure 8.5 and are quite similar due to the simplicity of the problem.

8.3.3 Application to Robust Wasserstein GANs

In this part, we introduce a robust modification of WGANs, named MoMWGAN, using one of the three proposed estimators in Section 8.2.1. The behaviour of likelihood-free generative modeling such as Generative Adversarial Networks in the presence of outliers, i.e. with heavy-tails distributions or contaminated data, has been poorly investigated up to very recently. At our knowledge, the unique reference is Gao et al. (2018). In particular, Gao et al. (2018) have studied theoretically and empirically the robustness of φ-GAN in the special case of mean estimation for elliptical distributions. In contrast,

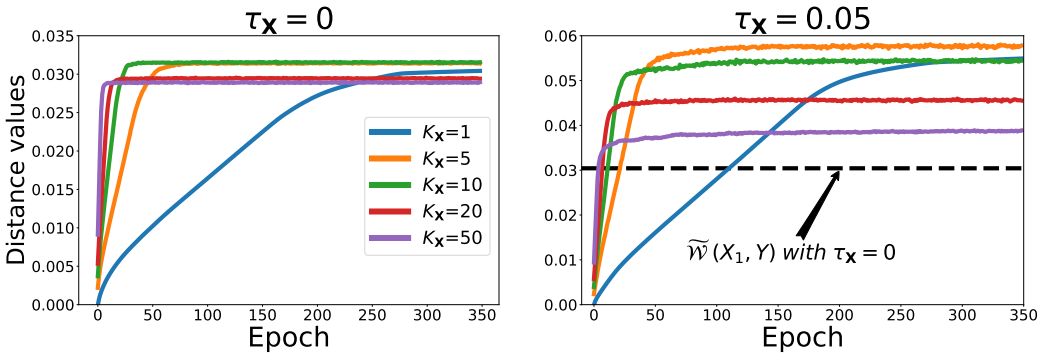


Figure 8.4 – Convergence of $\widetilde{\mathcal{W}}_{\mathrm{MoU-diag}}$ without anomalies (left) and with 5% anomalies (right) for different K_X.

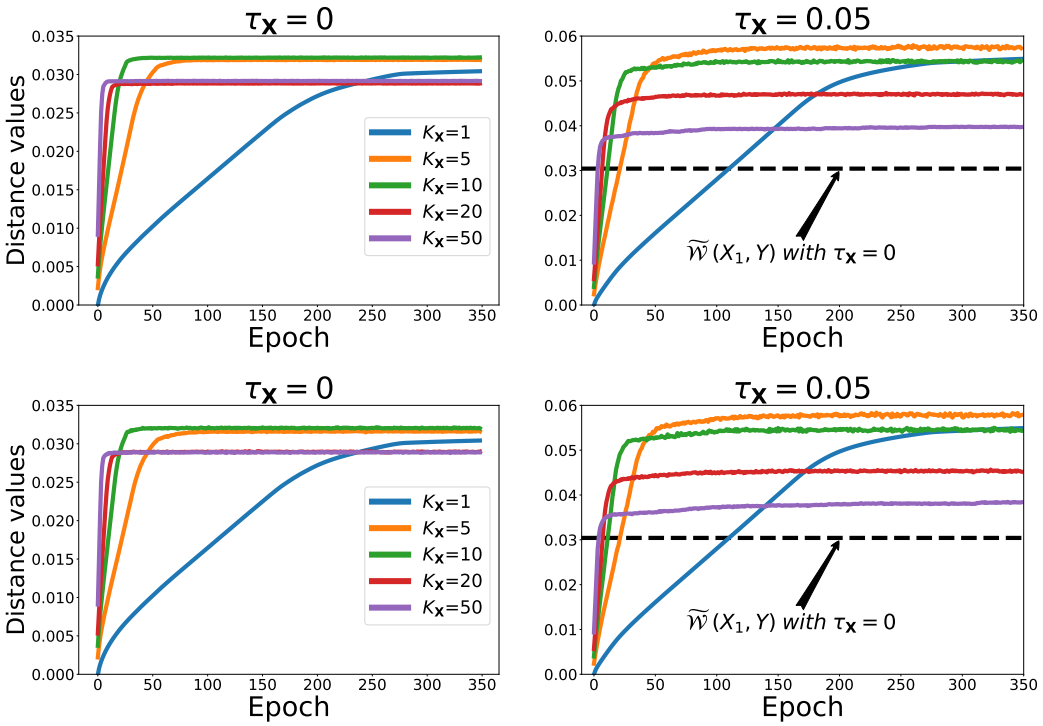


Figure 8.5 – Convergence of $\widetilde{\mathcal{W}}_{\mathrm{MoU}}$ (top) and $\widetilde{\mathcal{W}}_{\mathrm{MoM}}$ (bottom) without anomalies (left) and with 5% anomalies (right) for different K_X.

we illustrate here the theoretical results of Section 8.2 by applying a MoM approach to robustify WassersteinGAN and show on two real-world image benchmarks how this new variant of GAN behaves when learned with contaminated data.

Reminder on GAN: Let us briefly recall the GAN principle. A GAN learns a function $g_\theta : \Xi \to \mathcal{X}$ such that samples generate by $g_\theta(\xi) \sim Q_\theta$, taking as input a sample ξ (from some reference measure R, often Gaussian) in a latent space Ξ, are close to those of the true distribution $P \in \mathcal{P}(\mathcal{X})$ of data. Wasserstein GANs (Arjovsky et al., 2017; Gulrajani et al., 2017) use the L_1-Wasserstein Distance under its Kantorovich-Rubinstein dual formula as the loss function. Instead of maximizing over the unit ball of Lipschitz functions, one uses a parametric family of ϑ-Lipschitz functions under the form of neural

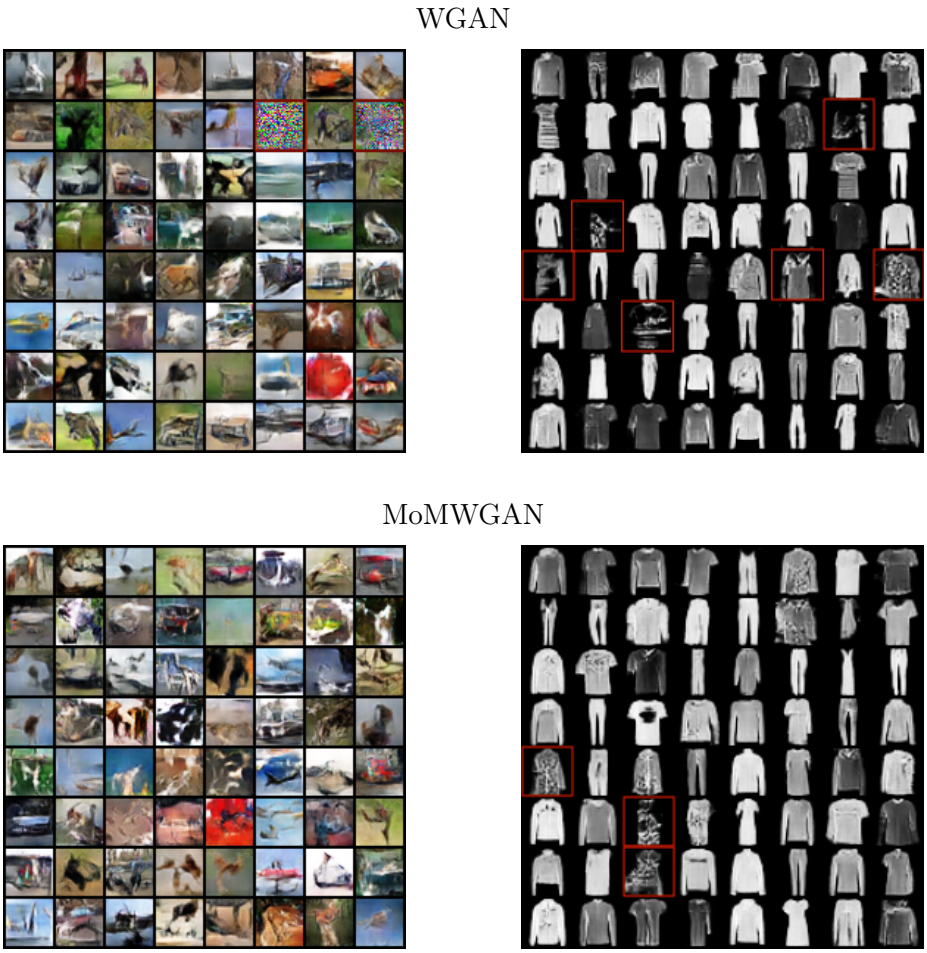


Figure 8.6 – Generated samples from trained WGAN and MoMWGAN on CIFAR10 and Fashion MNIST data sets.

net with clipped weights w (Arjovsky et al., 2017). Following up the theoretical analysis of Section 8.2, we introduce a MoM-based WGAN (MoMWGAN) model, combining the $\mathcal{W}_{\mathrm{MoM}}$ estimator studied in 8.2 and WGAN's framework. Following the weight clipping approach, MoMWGAN boils down to the problem:

$$\min_{\theta} \max_{w} \left\{ \mathrm{MoM}_X[\Psi_w] - \frac{1}{m}\sum_{j=1}^{m} \Psi_w(g_\theta(\xi_j)),\ k \leq K_X \right\}.$$

Note that the MoM procedure is chosen to be only applied on the observed contaminated sample. It is not clear in which way the sample drawn from the currently learned density is polluted and thus defining the number of blocks would be an issue. Optimization in WGAN is usually performed by taking mini-batches to reduce the computational load. In the same spirit, we apply MoM inside contaminated mini-batches as described in Algorithm 8.4. To get the outliers-robust property observed in the numerical experiments, we pay the price of finding the median block at each step by evaluating the loss which significantly increases the computational complexity.

Algorithm 8.4 MoMWGAN

input : ς, the learning rate. c, the clipping parameter. b, the batch size. n_c, the number of critic iterations per generator iteration, K_X the number of blocks. w_0, θ_0 the initial critic/generator's parameters.

while θ *has not converged* **do**
  for $t = 0, \dots, n_c$ **do**
    Choose a subsample $X_{i_1}, \dots, X_{i_b}$ to get $P_b = (1/b) \sum_{i \in \{i_1, \dots, i_b\}} \delta_{X_i}$ and sample $\xi_{j_1}, \dots, \xi_{j_b}$ from R to get $R_b = (1/b) \sum_{j \in \{j_1, \dots, j_b\}} \delta_{\xi_j}$
    Updating w with step 2-6 of Algorithm 8.1 with P_b and R_b
  Sample $\{\xi_j\}_{\ell=1}^b$ from R
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{b} \sum_{\ell=1}^b \Psi_w(g_\theta(\xi_j))$
  $\theta \leftarrow \theta - \varsigma \times \mathrm{RMSProp}(\theta, g_\theta)$
return θ, g_θ.

Numerical Experiments To test the robustness of MoMWGAN we contaminated two well-known image data sets, CIFAR10 and Fashion MNIST, with two anomalies settings. *Noise* based-anomalies are added to CIFAR10, i.e. images with random intensity pixels drawn from a uniform law. For Fashion MNIST, the five first classes are considered as "informative dat" while the sixth (Sandal) contains the anomalies. In both settings, WGAN and MoMWGAN are trained on the training samples contaminated in a uniform fashion with a proportion of 1.5% of outliers in both data sets. Both models use standard parameters of WGAN. $K_X = 4$ blocks have been used by MoMWGAN in both experiments. To assess performance of the resulting GANs, we generated 50000 generated images using each model (WGAN and MoMGAN) and measured the Fréchet Inception Distance (FID; Heusel et al., 2017) between the generated examples in both cases and the (real) test sample. Table 8.1 shows that MoMWGAN improves upon WGAN in terms of outliers-robustness. Furthermore, some generated images are represented in Figure 8.6. One can see that outliers do not affect MoMWGAN generated samples while WGAN reproduce noise on contaminated CIFAR10 data set. For Fashion MNIST, one may see that fewer images are degraded with MoMWGAN generator.

	WGAN	MoMWGAN
Polluted CIFAR10	57	55.9
Polluted Fashion MNIST	13.8	13.2

Table 8.1 – FID on polluted data sets.

8.4 Conclusion and Perspectives

In this chapter, we provided a view of the robustness properties of the MoM and MoU estimators with clear dependence on the proportion of outliers τ_X, τ_Y and the number of blocks K_X, K_Y. These bounds are incidentally shown to supply a sound theoretical basis for the reliability of MoM-based learning techniques when the training data are possibly contaminated by outliers with arbitrary distribution. Further, we have introduced three robust estimators of the Wasserstein distance based on MoM methodology. We have shown asymptotic and non-asymptotic results in the context of polluted data, i.e. the $\mathcal{O} \cup \mathcal{I}$ framework. Surpassing computational issues, we have designed an al-

gorithm to compute, in a efficient way, these estimators. Numerical experiments have highlighted the behavior of these estimators over their unique tuning parameter. Finally, we proposed to robustify WGANs using one of the introduced estimators and have shown its benefits on convincing numerical results. The open-source implementation of the method can be accessed at https://github.com/GuillaumeStaermanML/MoM-Wasserstein.

The theoretically well-founded MoM approaches to robustify the Wasserstein distance open the door to numerous applications beyond WGAN, including variational generative modeling. The promising MoMGAN deserves more attention and future work will concern the analysis of the estimator it provides.

8.5 Proofs

8.5.1 Proof of Proposition 8.3

Roughly speaking, the median has the same behavior as that of a majority of observations. Similarly, the MoM has the same behavior as that of a majority of blocks. In presence of outliers, the key point consists in focusing on *sane* blocks only, i.e. on blocks that do not contain a single outlier, since no prediction can be made about blocks *hit* by an outlier, in absence of any structural assumption concerning the contamination. One simple way to ensure the sane blocks to be in (almost) majority is to consider twice more blocks than outliers. Indeed, in the worst case scenario each outlier contaminates one block, but the sane ones remain more numerous. Let K_X denote the total number of blocks chosen, $K_X^{\mathcal{O}}$ the number of blocks containing at least one outlier, and $K_X^{\mathcal{I}}$ the number of sane blocks containing no outlier. The crux of our proofs then consists in determining some $\gamma > 1/2$ (that eventually depends on τ_X) such that $K_X^{\mathcal{I}} \geq \gamma K_X$. As discussed before, we thus need to consider at least twice more blocks than outliers. On the other hand, K_X is by design upper bounded by n. The global constraint can be written:

$$2n_{\mathcal{O}} = 2\tau_X n < K_X \leq n. \tag{8.2}$$

Choosing the geometric mean $\sqrt{2\tau_X}$, (8.2) is satisfied as soon as K_X verifies:

$$\sqrt{2\tau_X} n \leq K_X \leq n.$$

It directly follows that:

$$K_X^{\mathcal{I}} = K_X - K_X^{\mathcal{O}} \geq K_X - n_{\mathcal{O}} \geq K_X - \tau_X n \geq \left(1 - \frac{\tau_X}{\sqrt{2\tau_X}}\right) K_X = \frac{\sqrt{2\tau_X} - \tau_X}{\sqrt{2\tau_X}}\ K_X,$$

and one then may use:

$$\gamma = \gamma(\tau_X) = \frac{\sqrt{2\tau_X} - \tau_X}{\sqrt{2\tau_X}}.$$

Once $\gamma(\tau_X)$ is determined, a standard MoM deviation study can be carried out. If at least $K_X/2$ sane blocks have an empirical estimate that is t close to the expectation, then so is the MoM. Reversing the implication gives:

$$\mathbb{P}\left\{\left|\mathrm{MoM}_X - \mathbb{E}[X]\right| > t\right\} \leq \mathbb{P}\left\{\sum_{\text{blocks without outlier}} \mathbb{1}\left\{\left|\mathrm{MoM}_{\text{block}} - \mathbb{E}[X]\right| > t\right\} \geq K_X^{\mathcal{I}} - \frac{K_X}{2}\right\}$$
$$\leq \mathbb{P}\left\{\sum_{\text{blocks without outlier}} \mathbb{1}\left\{\left|\mathrm{MoM}_{\text{block}} - \mathbb{E}[X]\right| > t\right\} \geq \frac{2\gamma(\tau_X)-1}{2\gamma(\tau_X)} K_{\mathcal{I}}\right\}, \quad (8.3)$$

with $\mathrm{MoM}_{\text{block}} = (1/B_X)\sum_{i\in\text{block}} X_i$ the block empirical mean. Now observe that equation (8.3) describes the deviation of a binomial random variable, with $K_X^{\mathcal{I}}$ trials and parameter $p_t = \mathbb{P}\{|\mathrm{MoM}_{\text{block}} - \mathbb{E}[X]| > t\}$. It can thus be upper bounded by:

$$\sum_{k=\left\lceil \frac{2\gamma(\tau_X)-1}{2\gamma(\tau_X)} K_X^{\mathcal{I}} \right\rceil}^{K_X^{\mathcal{I}}} \binom{K_X^{\mathcal{I}}}{k} p_t^k \, (1-p_t)^{K_X^{\mathcal{I}}-k} \leq p_t^{\frac{2\gamma(\tau_X)-1}{2\gamma(\tau_X)} K_X^{\mathcal{I}}} \sum_{k=1}^{K_X^{\mathcal{I}}} \binom{K_X^{\mathcal{I}}}{k}$$
$$\leq p_t^{\frac{2\gamma(\tau_X)-1}{2\gamma(\tau_X)} K_X^{\mathcal{I}}} \, 2^{K_X^{\mathcal{I}}}$$
$$\leq p_t^{\frac{2\gamma(\tau_X)-1}{2} K} \, 2^{\gamma(\tau_X) K_X}.$$

Assume now that X is ρ sub-Gaussian. Chernoff's bound gives that $p_t \leq 2e^{-B_X t^2/2\rho^2}$. Plugging this bound into MoM's deviation yields:

$$\mathbb{P}\left\{\left|\mathrm{MoM}_X - \mathbb{E}[X]\right| > t\right\} \leq \exp\left(\frac{2\gamma(\tau_X)-1}{2} K_X \cdot \log\left[2^{\frac{4\gamma(\tau_X)-1}{2\gamma(\tau_X)-1}} e^{-B_X t^2/2\rho^2}\right]\right)$$
$$\leq \exp\left(-\frac{2\gamma(\tau_X)-1}{16\rho^2} n t^2\right),$$

for all t such that:

$$t^2 \geq \frac{4\rho^2}{B_X} \frac{4\gamma(\tau_X)-1}{2\gamma(\tau_X)-1} \log 2.$$

Reverting in δ gives that it holds with probability at least $1-\delta$:

$$\left|\mathrm{MoM}_X - \mathbb{E}[X]\right| \leq \frac{4\rho}{\sqrt{2\gamma(\tau_X)-1}} \sqrt{\frac{\log(1/\delta)}{n}},$$

for all δ that satisfies:

$$\delta \leq e^{-\frac{\log 2}{4}(4\gamma(\tau_X)-1)\frac{n}{B_X}}, \quad \text{and in particular} \quad \delta \leq e^{-4n\sqrt{2\tau_X}}. \quad (8.4)$$

Indeed it holds $B_X = \lfloor n/K_X \rfloor \geq n/(2K_X)$, so that $n/B_X \leq 2K_X = 2\lceil \sqrt{2\tau_X} n \rceil \leq 2(\sqrt{2\tau_X} n + 1) \leq \sqrt{2\tau_X} n$, since $1 \leq 2n_{\mathcal{O}} = 2\tau_X n \leq \sqrt{2\tau_X} n$. When $n_{\mathcal{O}} = \tau_X = 0$, one may choose $K_X = 1$, $B_X = n$, and $\delta \leq 1/e$.

The final writing is obtained by setting:

$$\Gamma(\tau_X) = \frac{1}{\sqrt{2\gamma(\tau_X)-1}} = \sqrt{\frac{\sqrt{2\tau_X}}{\sqrt{2\tau_X} - 2\tau_X}}.$$

8.5.2 Proof of Proposition 8.4

The proof can be directly adapted from that of Section 8.5.1. Assume that $K_X = K_Y = K$ and that $B_X = B_Y = B$. The first difference lies in the constraint K needs to satisfy. It now writes: $2(n_{\mathcal{O}} + m_{\mathcal{O}}) = 2(\tau_X + \tau_Y)n < K \leq n$, and the reasoning can then be reused in totality with $\tau_X + \tau_Y$ instead of τ_X. The second difference lies in the use of Chebyshev's inequality instead of Chernoff's bound. Indeed, the variance of the U-statistics of one block (e.g. the k-th block $(1/B^2)\sum_{(i,j)\in\mathcal{B}_{k,k}^{XY}} h(X_i, Y_j)$), denoted by $\sigma^2_{B,B}(h)$, is given by (see e.g. van der Vaart (2000)):

$$\begin{aligned}\sigma^2_{B,B}(h) &= \frac{1}{B^2}\sigma^2(h) + \frac{B-1}{B^2}\sigma_1^2(h) + \frac{B-1}{B^2}\sigma_2^2(h),\\ &\leq \frac{\sigma^2(h) + \sigma_1^2(h) + \sigma_2^2(h)}{B},\end{aligned}$$

where $\sigma^2(h) = \mathrm{Var}(h(X,Y))$, $\sigma_1^2(h) = \mathrm{Var}(h_1(X))$ and $\sigma_2^2(h) = \mathrm{Var}(h_2(Y))$, with $h_1(x) = \mathbb{E}_Q\Big[h(x,Y)\Big]$ and $h_2(y) = \mathbb{E}_P\Big[h(X,y)\Big]$. Finally, when $\|h\|_{ess,\infty}$ is finite, using the notation $\mathcal{S}_n = (X_1, \ldots, X_n)$, one may bound p_t as follows:

$$\begin{aligned}p_t &= \mathbb{P}\left\{\left|\mathrm{MoU}_{\mathrm{block}}[h] - \mathbb{E}[h(X,Y)]\right| > t\right\},\\ &= \mathbb{P}\left\{\left|\frac{1}{B^2}\sum_{i\in\mathcal{B}_1^X}\sum_{j\in\mathcal{B}_1^Y} h(X_i,Y_j) - \mathbb{E}[h(X,Y)]\right| > t\right\},\\ &\leq \mathbb{P}\left\{\left|\frac{1}{B}\sum_{j\in\mathcal{B}_1^Y}\left(\sum_{i\in\mathcal{B}_1^X}\frac{h(X_i,Y_j)}{B} - \mathbb{E}\left[\sum_{i\in\mathcal{B}_1^X}\frac{h(X_i,Y_j)}{B}\;\middle|\;\mathcal{S}_n\right]\right)\right| > \frac{t}{2}\;\middle|\;\mathcal{S}_n\right\}\\ &+ \mathbb{P}\left\{\left|\frac{1}{B}\sum_{i\in\mathcal{B}_1^X}\mathbb{E}_Q\Big[h(X_i,Y)\Big] - \mathbb{E}[h(X,Y)]\right| > \frac{t}{2}\right\},\\ &\leq 2e^{-Bt^2/8\|h\|^2_{ess,\infty}} + 2e^{-Bt^2/8\|h\|^2_{ess,\infty}},\end{aligned}$$

where we have used Hoeffding's inequality twice: on the $\sum_{i\in\mathcal{B}_1^X}\frac{h(X_i,Y_j)}{B}$ for $j \in \mathcal{B}_1^Y$, conditionally to the X_i's, and a second time to the $\mathbb{E}_Y\Big[h(X_i,Y)\Big]$ for $i \in \mathcal{B}_1^X$, both random variables being bounded by $\|h\|_\infty$. The rest of the proof is similar to that of Section 8.5.1.

8.5.3 Proof of Proposition 8.8

We first show the strong consistency of $\mathcal{W}_{\mathrm{MoU}}(P_n, Q_m)$, that of $\mathcal{W}(P_n^{\mathrm{MoM}}, P)$ and $\mathcal{W}_{\mathrm{MoU-diag}}(P_n, Q_m)$ being then straightforward adaptations. Assume that $\tilde{\tau} = \tau_X + \tau_Y - \tau_X\tau_Y < 1/2$, and $K_X, K_Y > 0$ such that $2\tilde{\tau} < K_X K_Y/(nm)$. The latter condition implies that the blocks containing no outlier are in majority. Indeed, the number of contaminated blocks is upper bounded by:

$$n_{\mathcal{O}}K_Y + m_{\mathcal{O}}K_X - n_{\mathcal{O}}m_{\mathcal{O}} \leq \tilde{\tau}nm < K_X K_Y/2.$$

One may choose K_X and K_Y the lower as possible such that the above condition is respected. Following this, it is a natural choice to set $K_X = \lceil \sqrt{2\tilde{\tau}}\ n \rceil$ and $K_Y = \lceil \sqrt{2\tilde{\tau}}\ m \rceil$.

Let $\mathcal{I}_X$ (respectively $\mathcal{I}_Y$) denote the set of indices of blocks of $X_1, \dots, X_n$ (respectively the blocks of $Y_1, \dots, Y_m$) containing no outlier. Let $\mathcal{K}^d$ be a bounded subspace of $\mathbb{R}^d$, and assume that X, Y are valued in $\mathcal{X}, \mathcal{Y} \subset \mathcal{K}^d$. Finally, we denote by $\overline{\Psi}_{X,k}$ and $\overline{\Psi}_{Y,l}$ the quantities:

$$\overline{\Psi}_{X,k} = \frac{1}{B_X} \sum_{i \in \mathcal{B}_k^X} \Psi(X_i), \qquad \text{and} \qquad \overline{\Psi}_{Y,l} = \frac{1}{B_Y} \sum_{j \in \mathcal{B}_l^Y} \Psi(Y_j).$$

Using the shortcut notation $\mathbb{E}_P\left[\Psi\right] = \mathbb{E}_{X\sim P}\left[\Psi(X)\right]$ and $\mathbb{E}_Q\left[\Psi\right] = \mathbb{E}_{Y\sim Q}\left[\Psi(Y)\right]$, first notice that:

$$\begin{aligned}
\mathcal{W}_{\mathrm{MoU}}(P_n, Q_m) &= \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \mathrm{MoU}_{XY}[h_\Psi] \\
&= \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \operatorname*{med}_{\substack{1 \le k \le K_X \\ 1 \le l \le K_Y}} \left\{ \overline{\Psi}_{X,k} - \overline{\Psi}_{Y,l} \right\} \\
&= \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \operatorname*{med}_{\substack{1 \le k \le K_X \\ 1 \le l \le K_Y}} \left\{ \overline{\Psi}_{X,k} - \mathbb{E}_P[\Psi] + \mathbb{E}_P[\Psi] - \mathbb{E}_Q[\Psi] + \mathbb{E}_Q[\Psi] - \overline{\Psi}_{Y,l} \right\} \\
&\le \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \operatorname*{med}_{\substack{1 \le k \le K_X \\ 1 \le l \le K_Y}} \left\{ \overline{\Psi}_{X,k} - \mathbb{E}_P[\Psi] + \mathbb{E}_Q[\Psi] - \overline{\Psi}_{Y,l} \right\} + \mathcal{W}(P, Q). \qquad (8.5)
\end{aligned}$$

Conversely, it holds:

$$\begin{aligned}
\mathcal{W}(P, Q) &= \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \left\{ \mathbb{E}_P\left[\Psi\right] - \mathbb{E}_Q\left[\Psi\right] \right\} \\
&\le \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \left\{ \mathbb{E}_P[\Psi] - \overline{\Psi}_{\mathcal{B}^X_{\mathrm{med}}} + \overline{\Psi}_{\mathcal{B}^Y_{\mathrm{med}}} - \mathbb{E}_Q[\Psi] + \overline{\Psi}_{\mathcal{B}^X_{\mathrm{med}}} - \overline{\Psi}_{\mathcal{B}^Y_{\mathrm{med}}} \right\} \\
&\le \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \operatorname*{med}_{\substack{1 \le k \le K_X \\ 1 \le l \le K_Y}} \left\{ \mathbb{E}_P[\Psi] - \overline{\Psi}_{X,k} + \overline{\Psi}_{Z,l} - \mathbb{E}_Q[\Psi] \right\} + \mathcal{W}_{\mathrm{MoU}}(P_n, Q_m), \qquad (8.6)
\end{aligned}$$

where $\mathcal{B}^X_{\mathrm{med}}$ and $\mathcal{B}^Y_{\mathrm{med}}$ are the median blocks of $\overline{\Psi}_{X,k} - \overline{\Psi}_{Y,l}$ for $1 \le k \le K_X$ and $1 \le l \le K_Y$. From (8.5) and (8.6), we deduce that:

$$\begin{aligned}
\left| \mathcal{W}_{\mathrm{MoU}}(P_n, Q_m) - \mathcal{W}(P, Q) \right| &\le \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \operatorname*{med}_{\substack{1 \le k \le K_X \\ 1 \le l \le K_Y}} \left\{ \left| \overline{\Psi}_{X,k} - \mathbb{E}_P[\Psi] + \mathbb{E}_Q[\Psi] - \overline{\Psi}_{Y,l} \right| \right\} \qquad (8.7) \\
&\le \sup_{k \in \mathcal{I}_X,\ l \in \mathcal{I}_Y} \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \left| \overline{\Psi}_{X,k} - \mathbb{E}_P[\Psi] + \mathbb{E}_Q[\Psi] - \overline{\Psi}_{Y,l} \right| \\
&\le \sup_{k \in \mathcal{I}_X} \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \left| \overline{\Psi}_{X,k} - \mathbb{E}_P[\Psi] \right| + \sup_{l \in \mathcal{I}_Y} \sup_{\Psi \in \mathcal{F}_{\mathrm{Lip}}} \left| \mathbb{E}_Q[\Psi] - \overline{\Psi}_{Y,l} \right|,
\end{aligned}$$

where we have used the subadditivity of the supremum and the fact that $\mathcal{I}_X \times \mathcal{I}_Y$ represents a majority of blocks. By independence between X and Y, it holds:

$$\mathbb{P}\left\{\left|\mathcal{W}_{\mathrm{MoU}}(P_n, Q_m) - \mathcal{W}(P, Q)\right| \underset{\substack{n\to+\infty \\ m\to+\infty}}{\longrightarrow} 0\right\}$$
$$\geq \prod_{k\in\mathcal{I}_X} \mathbb{P}\left\{\sup_{\Psi\in\mathcal{F}_{\mathrm{Lip}}} \left|\overline{\Psi}_{X,k} - \mathbb{E}_P[\Psi]\right| \underset{n\to+\infty}{\longrightarrow} 0\right\} \cdot \prod_{l\in\mathcal{I}_Y} \mathbb{P}\left\{\sup_{\Psi\in\mathcal{F}_{\mathrm{Lip}}} \left|\overline{\Psi}_{Y,l} - \mathbb{E}[\Psi]\right| \underset{m\to+\infty}{\longrightarrow} 0\right\}.$$

Now, the arguments to get the right-hand side equal to 1 are similar to those used in Lemma 3.1 and Proposition 3.2 in Sriperumbudur et al. (2012). We expose them explicitly for the sake of clarity.

Let $N(r, \mathcal{F}_{\mathrm{Lip}}, L_1(P))$ be the *covering number* of $\mathcal{F}_{\mathrm{Lip}}$ which is the minimal number of $L_1(\mathbb{P})$ balls of radius $r > 0$ needed to cover $\mathcal{F}_{\mathrm{Lip}}$. Let $\mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, L_1(P))$ be the *entropy* of $\mathcal{F}_{\mathrm{Lip}}$, defined as $\mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, L_1(P)) = \log N(r, \mathcal{F}_{\mathrm{Lip}}, L_1(P))$. Let E be the minimal envelope function such that $E(x) = \sup_{\phi\in\mathcal{B}_L} |\phi(x)|$. We need to check that $\int E\mathrm{d}P$ and $\int E\mathrm{d}Q$ are finite and that $(1/n)\mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, L_1(P_n))$ and $(1/m)\mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, L_1(Q_m))$ go to zero when n and m go to infinity. Thus, we can apply Theorem 3.7 in van de Geer (2000) which ensures the uniform (a.s.) convergence of empirical processes. Noticing that we can include in the definition of $\mathcal{F}_{\mathrm{Lip}}$ the property that $\Psi(0) = 0$ without changing the supremum, one has:

$$\Psi(x) \leq \sup_{x\in\mathcal{K}^d} |\Psi(x)| \leq \sup_{x,y\in\mathcal{K}^d} |\Psi(x) - \Psi(y)| \leq \sup_{x,y\in\mathcal{K}^d} \|x - y\| = \mathrm{diam}(\mathcal{K}^d) < +\infty. \quad (8.8)$$

Therefore $E(x)$ is finite, and following Lemma 3.1 in Kolmogorov and Tikhomirov (1961) we have:

$$\mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, \|\cdot\|_\infty) \leq N(r/4, \mathcal{K}^d, \|\cdot\|_2) \log\left(2\left\lceil\frac{2\mathrm{diam}(\mathcal{K}^d)}{r}\right\rceil + 1\right).$$

Since $\mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, L^1(P_n)) \leq \mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, \|\cdot\|_\infty)$ and $\mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, L^1(Q_m)) \leq \mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, \|\cdot\|_\infty)$ then, as n and m go to infinity, we have:

$$\frac{1}{n}\mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, L^1(P_n)) \xrightarrow{P} 0, \qquad \text{and} \qquad \frac{1}{m}\mathcal{H}(r, \mathcal{F}_{\mathrm{Lip}}, L^1(Q_m)) \xrightarrow{Q} 0,$$

which lead to the desired result.

Adaptation to other estimators. The above proof can be adapted in a straightforward fashion to $\mathcal{W}(P_n^{\mathrm{MoM}}, P)$ and $\mathcal{W}_{\mathrm{MoU-diag}}(P_n, Q_m)$. Indeed, it holds:

$$\mathcal{W}(P_{\mathrm{MoM}}, P) = \sup_{\Psi\in\mathcal{F}_{\mathrm{Lip}}} \underset{1\leq k\leq K_X}{\mathrm{med}} \left|\overline{\Psi}_{X,k} - \mathbb{E}_P\left[\Psi\right]\right|,$$

and

$$\left|\mathcal{W}_{\mathrm{MoU-diag}}(P_n, Q_m) - \mathcal{W}(P, Q)\right| \leq \sup_{\Psi\in\mathcal{F}_{\mathrm{Lip}}} \underset{\substack{1\leq k\leq K_X \\ 1\leq l\leq K_Y}}{\mathrm{med}} \left|\overline{\Psi}_{X,k} - \mathbb{E}_P\left[\Psi\right] + \mathbb{E}_Q\left[\Psi\right] - \overline{\Psi}_{Y,k}\right|.$$

It is then direct to adapt the reasoning starting from (8.7).

8.5.4 Proof of Proposition 8.9

In order to prove the Proposition 8.9, we first recall a simple lemma on the difference between two median vectors.

Lemma 8.12. *Let x and y be two vectors of $\mathbb{R}^d$. Then it holds*

$$\Big|\text{median}(x) - \text{median}(y)\Big| \leq \|x - y\|_\infty.$$

Proof. It is direct to see that:

$$x \preceq y \preceq z \quad \Rightarrow \quad \text{median}(x) \leq \text{median}(y) \leq \text{median}(z).$$

Thus, for all y within the infinite ball of center x and radius r it holds:

$$\text{median}(x) - r = \text{median}(x - r\mathbf{1}_d) \leq \text{median}(y) \leq \text{median}(x + r\mathbf{1}_d) = \text{median}(x) + r,$$

where $\mathbf{1}_d$ is a d-dimensional vector of ones. ∎

Let $\Phi \in \mathcal{F}_{\text{Lip}}$. From (8.8), we know that $-\text{diam}(\mathcal{K}^d) \leq \Phi(X) \leq \text{diam}(\mathcal{K}^d)$, so that $\Phi(X)$ is in particular sub-Gaussian with parameter $\rho = \text{diam}(\mathcal{K}^d)$. A direct application of Proposition 8.3 then gives that for all $\delta \in]0, e^{-4n\sqrt{2\tau_X}}]$ and $K_X = \lceil \sqrt{2\tau_X} n \rceil$, it holds with probability at least $1 - \delta$:

$$\left|\text{MoM}_X[\Phi] - \mathbb{E}_P\left[\Phi\right]\right| \leq 4\ \text{diam}(\mathcal{K}^d)\ \Gamma(\tau_X)\ \sqrt{\frac{\log(1/\delta)}{n}}, \tag{8.9}$$

with $\Gamma \colon \tau_X \mapsto \sqrt{1 + \sqrt{2\tau_X}}/\sqrt{1 - 2\tau_X}$. Using Lemma 8.12, observe also that $\forall (\Psi, \Phi) \in (\mathcal{F}_{\text{Lip}})^2$, it holds:

$$\begin{aligned}\left|\text{MoM}_X[\Psi] - \mathbb{E}_P\left[\Psi\right]\right| &\leq \left|\text{MoM}_X[\Psi] - \text{MoM}_X[\Phi]\right| + \left|\mathbb{E}_P\left[\Psi\right] - \mathbb{E}_P\left[\Phi\right]\right|, \\ &\quad + \left|\text{MoM}_X[\Phi] - \mathbb{E}_P\left[\Phi\right]\right|, \\ &\leq 2\|\Psi - \Phi\|_\infty + \left|\text{MoM}_X[\Phi] - \mathbb{E}_P\left[\Phi\right]\right|. \end{aligned} \tag{8.10}$$

Now, let $\zeta > 0$, and $\Phi_1, \ldots, \Phi_{N(\zeta, \mathcal{F}_{\text{Lip}}, \|\cdot\|_\infty)}$ be a ζ-coverage of $\mathcal{F}_{\text{Lip}}$ with respect to $\|\cdot\|_\infty$. We know from Sriperumbudur et al. (2012) that there exists $C_{\text{Lip}} > 0$ such that for all $\zeta > 0$, it holds:

$$\log(N(\zeta, \mathcal{F}_{\text{Lip}}, \|\cdot\|_\infty)) \leq C_{\text{Lip}}^2 (1/\zeta)^d. \tag{8.11}$$

From now on, we use $N = N(\zeta, \mathcal{F}_{\text{Lip}}, \|\cdot\|_\infty)$ for notation simplicity. Let Ψ be an arbitrary element of $\mathcal{F}_{\text{Lip}}$. By definition, there exists $i \leq N$ such that $\|\Psi - \Phi_i\|_\infty \leq \zeta$. Equation (8.10) then gives:

$$\left|\text{MoM}_X[\Psi] - \mathbb{E}_P\left[\Psi\right]\right| \leq 2\zeta + \left|\text{MoM}_X[\Phi_i] - \mathbb{E}_P\left[\Phi_i\right]\right|. \tag{8.12}$$

Applying (8.9) to every Φ_i, the union bound gives that with probability at least $1 - \delta$ it holds:

$$\sup_{i \leq N} \left|\text{MoM}_X[\Phi_i] - \mathbb{E}_P\left[\Phi_i\right]\right| \leq 4\ \text{diam}(\mathcal{K}^d)\ \Gamma(\tau_X)\ \sqrt{\frac{\log(N/\delta)}{n}}.$$

Taking the supremum in both sides of (8.12), it holds with probability at least $1-\delta$:

$$\sup_{\Psi\in\mathcal{F}_{\mathrm{Lip}}}\left|\mathrm{MoM}_X[\Psi]-\mathbb{E}_P\left[\Psi\right]\right|\leq 2\zeta+4\ \mathrm{diam}(\mathcal{K}^d)\ \Gamma(\tau_X)\ \sqrt{\frac{C_{\mathrm{Lip}}^2\zeta^{-d}+\log(1/\delta)}{n}}.$$

Choosing $\zeta\sim 1/n^{1/(d+2)}$ and breaking the square root finally gives with probability at least $1-\delta$:

$$\sup_{\Psi\in\mathcal{F}_{\mathrm{Lip}}}\left|\mathrm{MoM}_X[\Psi]-\mathbb{E}_P\left[\Psi\right]\right|\leq\frac{C_1(\tau_X)}{n^{1/(d+2)}}+C_2(\tau_X)\sqrt{\frac{\log(1/\delta)}{n}},$$

with $C_1(\tau_X)=2+C_{\mathrm{Lip}}C_2(\tau_X)$, and $C_2(\tau_X)=4\ \mathrm{diam}(\mathcal{K}^d)\ \Gamma(\tau_X)$.

Adaptation to MoU. From (8.8), we get that the kernel $h_\Psi\colon(X,Y)\mapsto\Psi(X)-\Psi(Y)$ has finite essential supremum $\|h_\Psi(X,Y)\|_{ess,\infty}\leq\mathrm{diam}(\mathcal{K}^d)$. Using Proposition 8.4 with the same reasoning as above leads to the desired result, multiplying constants by a 2 factor.

8.5.5 Proof of Theorem 8.10

Since $n^{\frac{1}{d+2}+\frac{1-\beta}{2}}\geq C_1(\tau_X)/(2C_2(\tau_X)(2\tau_X)^{\frac{1}{4}})$, then for all $\delta\in]0,e^{-4n\sqrt{2\tau_X}}]$, it holds:

$$\frac{C_1(\tau_X)}{n^{1/(d+2)}}\leq C_2(\tau_X)\sqrt{\frac{4n\sqrt{2\tau_X}}{n^\beta}}\leq C_2(\tau_X)\sqrt{\frac{\log(1/\delta)}{n^\beta}}.$$

One then has:

$$\mathcal{W}(P_n^{\mathrm{MoM}},P)\geq 0\geq\frac{C_1(\tau_X)}{n^{1/(d+2)}}-C_2(\tau_X)\sqrt{\frac{\log(1/\delta)}{n^\beta}}.$$

Combining with the first results of Proposition 8.9, for all $\delta\in]0,e^{-4n\sqrt{2\tau_X}}]$, it holds with probability at least $1-\delta$:

$$\left|\mathcal{W}(P_n^{\mathrm{MoM}},P)-\frac{C_1(\tau_X)}{n^{1/(d+2)}}\right|\leq C_2(\tau_X)\sqrt{\frac{\log(1/\delta)}{n^\beta}}.$$

Reverting the inequality gives that it holds:

$$\mathbb{P}\left\{\left|\mathcal{W}(P_n^{\mathrm{MoM}},P)-\frac{C_1(\tau_X)}{n^{1/(d+2)}}\right|>t\right\}\leq e^{-n^\beta(t/C_2(\tau_X))^2},\tag{8.13}$$

for all t such that:

$$t\geq(32\ \tau_X)^{1/4}C_2(\tau_X)\ \sqrt{n^{1-\beta}}=\frac{(32\ \tau_X)^{1/4}}{\sqrt{\tau_X}}C_2(\tau_X)\sqrt{n^{1-\beta}\ \frac{n_{\mathcal{O}}}{n}}.\tag{8.14}$$

One may finally use that for a nonnegative random variable it holds:

$$\mathbb{E}\left|\mathcal{W}(P_n^{\mathrm{MoM}}, P) - \frac{C_1(\tau_X)}{n^{1/(d+2)}}\right| = \int_0^\infty \mathbb{P}\left\{\left|\mathcal{W}(P_n^{\mathrm{MoM}}, P) - \frac{C_1(\tau_X)}{n^{1/(d+2)}}\right| > t\right\} \mathrm{d}t$$

$$\leq \int_0^{\frac{(32\,\tau_X)^{1/4}}{\sqrt{\tau_X}} C_{\mathcal{O}} C_2(\tau_X) \sqrt{n^{\alpha_{\mathcal{O}} - \beta}}} 1 \,\mathrm{d}t + \int_0^\infty e^{-n^\beta (t/C_2(\tau_X))^2} \,\mathrm{d}t$$

$$\leq \frac{(32\,\tau_X)^{1/4}}{\sqrt{\tau_X}} \frac{C_{\mathcal{O}} C_2(\tau_X)}{n^{(\beta-\alpha_{\mathcal{O}})/2}} + \frac{\sqrt{\pi}\, C_2(\tau_X)}{2\, n^{\beta/2}}$$

$$= 2\,(2/\tau_X)^{1/4} \frac{C_{\mathcal{O}} C_2(\tau_X)}{n^{(\beta-\alpha_{\mathcal{O}})/2}} + \frac{\sqrt{\pi}\, C_2(\tau_X)}{2\, n^{\beta/2}}.$$

Where the second line holds thanks to Assumption 8.7.

Adaptation to MoU. The adaptation is straightforward, up to (8.14), that now writes:

$$t \geq 2 \times (32(\tau_X + \tau_Y))^{1/4} C_2(\tau_X + \tau_Y) \sqrt{n^{1-\beta}}$$

$$= 2 \times \frac{(32(\tau_X + \tau_Y))^{1/4}}{\sqrt{\tau_X + \tau_Y}} C_2(\tau_X + \tau_Y) \sqrt{n^{1-\beta} \left(\frac{n_{\mathcal{O}}}{n} + \frac{m_{\mathcal{O}}}{m}\right)}.$$

Using Assumption 8.7 on both samples from X and Y leads to the desired results.

9

Affine-Invariant Integrated Rank-Weighted Depth: Definition, Properties and Finite Sample Analysis

Contents

It is the main purpose of this chapter to overcome the lack of affine invariance of the integrated rank-weighted (IRW) depth by proposing a modified version of it, named AI-IRW. we show that the AI-IRW depth inherits all the properties and computational advantages of the IRW depth and satisfies the *affine-invariance* property in addition. Because its statistical counterpart based on a sample composed of independent copies of

the random variable X is a complex functional of the data, involving the square root of the empirical precision matrix, a finite-sample analysis is carried out here. Precisely, a concentration result for the sampling version of the AI-IRW depth is established. Beyond this theoretical analysis, the relevance of the AI-IRW depth notion is also supported by experimental results, showing its advantages over the IRW depth and other depth proposals standing as natural competitors when applied to anomaly detection.

This chapter is structured as follows. In Section 9.1, the AI-IRW depth is introduced, its properties are studied and approximation/estimation issues are discussed at length. The accuracy of the empirical version is investigated in Section 9.2 from a nonasymptotic perspective. Section 9.3 describes experimental results illustrating empirically the advantages of the AI-IRW depth. Some concluding remarks are collected in Section 9.4. Eventually, technical proofs are deferred to the Section 9.5. This chapter covers the contribution of:

- **G. Staerman**, P. Mozharovskyi, S. Clémençon. Affine-Invariant Integrated Rank-Weighted Depth: Definition, Properties and Finite Sample Analysis. *arXiv preprint arXiv:2106.11068*, 2021.

9.1 Affine-Invariant IRW Depth - Definition and Properties

9.1.1 Motivations

Many data depths are constructed as an infimum over unit-sphere projections of a univariate non parametric statistic such as the halfspace depth, the projection depth, or those introduced in Zhang (2002) or Zuo (2003). From a practical perspective, computing these projection-based depths involves the use of tools such as manifold optimization algorithms, facing various numerical difficulties as the dimension d increases, see Dyckerhoff et al. (2021). In addition, the halfspace depth suffers from two major problems: (i) for each data point, taking the direction achieving the minimum to assign a score to it possibly creates a significant sensitivity to noisy directions (ii) the null score assigned to each new data point outside of the convex hull of the support of the distribution P makes the score of such points indistinguishable. A remedy based on Extreme Value Theory has been proposed in Einmahl et al. (2015), which consists in smoothing the halfspace depth beyond the convex hull of the data. However, this variant relies on rather rigid parametric assumptions, is only approximately affine invariant and confronted with the limitation aforementioned regarding the non-smoothed part of the data. Recently, alternative depth functions have been proposed, obtained by replacing the infimum over all possible directions by an integral, see Cuevas and Fraiman (2009). In Ramsay et al. (2019), a new data depth, referred to as the Integral Rank-Weighted (IRW) depth, is defined by substituting an integral over the sphere $\mathbb{S}^{d-1}$ for the infimum in the halfspace depth. Here and throughout, the spherical probability measure on $\mathbb{S}^{d-1}$ is denoted by ω_{d-1}.

Definition 9.1 (Ramsay et al. (2019))**.** *The Integrated Rank-Weighted (IRW) depth of $x \in \mathbb{R}^d$ relative to a probability distribution $P \in \mathcal{P}(\mathbb{R}^d)$ is given by:*

$$D_{\mathrm{IRW}}(x,P) = \int_{\mathbb{S}^{d-1}} D_{\mathrm{H},1}(\langle u,\ x\rangle, P_u)\ \omega_{d-1}(\mathrm{d}u) \\ = \mathbb{E}\left[D_{\mathrm{H},1}(\langle U,\ x\rangle, P_U)\right], \tag{9.1}$$

where P_u is the pushforward distribution of P defined by the projection $x \in \mathbb{R}^d \mapsto \langle u, x\rangle$ and U is a r.v. uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$.

As explained at length in Ramsay et al. (2019), the name of the data depth (9.1) originates from the fact that it can be represented as a weighted average of a finite set of normalized center-outward ranks. It has many advantages over the original halfspace depth. First, by construction it is *robust* to noisy directions and *sensitive* to new data point outside of the convex hull of the training data set both at the same time, fixing then the two problems mentioned above. Moreover, concerning numerical feasibility, the computation of the IRW depth does not require to implement any manifold optimization algorithm and can be approximately made by means of basic Monte-Carlo techniques, providing in addition confidence intervals as a by-product, see Remark 9.2 below. Its contours $\{D_{\mathrm{IRW}}(x,P) = \alpha\}$, $\alpha \in [0,1]$, also exhibits a higher degree of smoothness in general (the depth function (9.1) is continuous at any point $x \in \mathbb{R}^d$ that is not an atom for P, cf. Proposition 1 in Ramsay et al. (2019)) and properties ($\mathbf{D}_2$, $\mathbf{D}_3$, $\mathbf{D}_4$, $\mathbf{D}_5$) have been proved to be satisfied by (9.1) under mild assumptions, see Theorem 2 in Ramsay et al. (2019).

Remark 9.2. *(Monte-Carlo approximation) Recall that a r.v. uniformly distributed on the hypersphere $\mathbb{S}^{d-1}$ can be generated from a d-dimensional centered Gaussian random vector W with the identity I_d as covariance matrix: if $W \sim \mathcal{N}(0, \mathrm{I}_d)$, then $W/||W|| \sim \omega_{d-1}$, see Krantz and Parks (2008). Hence, a basic Monte-Carlo method to approximate (9.1) would consist in generating $n_{\mathrm{proj}} \geq 1$ independent realizations $W_1, \ldots, W_m$ of $\mathcal{N}(0, \mathrm{I}_d)$ and compute*

$$\frac{1}{n_{\mathrm{proj}}} \sum_{j=1}^{n_{\mathrm{proj}}} D_{\mathrm{H},1}(\langle W_j/||W_j||,\ x\rangle, P_{W_j/||W_j||}), \tag{9.2}$$

refer to e.g. Kalos and Whitlock (2008) for an account of Monte-Carlo integration methods.

However, it does not satisfy the key property ($\mathbf{D}_1$) (affine-invariance) in general. The fact that it is affected by non-uniform scaling is problematic in practice (regarding its interpretability in particular or its use for anomaly detection tasks for instance, see Section 9.3.4) and is the main flaw of this approach, as pointed out in Cuevas and Fraiman (2009) and Ramsay et al. (2019). An analytical counter-example is provided below.

Consider the discrete probability measure P assigning the weight $1/3$ to the bivariate points in $\mathcal{S}_3 = \{(-1,2),(3,3),(2,1)\}$ and let us compute the IRW depth of $x = (0,1)$ and $y = (3,2)$ relative to P. It is easy to see that the mappings $u \in \mathbb{S}^1 \mapsto D_{\mathrm{H},1}(\langle u,\ x\rangle, P_u)$ and $u \in \mathbb{S}^1 \mapsto D_{\mathrm{H},1}(\langle u,\ y\rangle, P_u)$ take only two values, 0 or $1/3$. Identifying $\mathbb{S}^1$ as $[0, 2\pi[$, the univariate halfspace depth of x relative to P is then null for any $u \in [\pi/4, \pi/2] \cup [5\pi/4, 3\pi/2]$ and equal to $1/3$ if u belongs to the complementary set. In addition, $D_{\mathrm{H},1}(\langle u,\ y\rangle, P_u)$ is equal to 0 for any $u \in [3\pi/4, \pi] \cup [7\pi/4, 2\pi]$ and equal to $1/3$ on the complementary set. One may easily check that $D_{\mathrm{IRW}}(x,P) = D_{\mathrm{IRW}}(y,P) = 0.25$

and the same rank would be then assigned to each point by the IRW depth. Now, multiplying all ordinate values by 2, which is an affine transformation, the univariate halfspace depth of $\widetilde{x} = (0, 2)$ is now null for all u in $[\pi/8, \pi/2] \cup [9\pi/8, 3\pi/2]$ while it remains equal to $1/3$ on the complementary set of this region. The depth of $\widetilde{x}$ is thus lower than 0.25. On the other hand, the univariate depth of $\widetilde{y} = (3, 4)$ is now null on $[7\pi/8, \pi] \cup [15\pi/8, 2\pi]$ while it remains equal to $1/3$ on the complementary set of this interval. It follows that $D_{\text{IRW}}(\widetilde{x}) = 5/24 < 0.25 < 7/24 = D_{\text{IRW}}(\widetilde{y})$.

9.1.2 The AI-IRW Depth

Here we propose to modify the depth function (9.1) in order to ensure that property $(\mathbf{D}_1)$ is always satisfied when the random vector X with distribution P under study is assumed to be square integrable with positive definite covariance matrix Σ. Precisely, rather than taking the expectation w.r.t. a random direction U uniformly distributed on $\mathbb{S}^{d-1}$ (i.e. integrating over all possible directions $u \in \mathbb{S}^{d-1}$), one considers the random projections defined by the eigenfunctions of the matrix Σ, i.e. the principal components of the r.v. X. In other words, the expectation is taken w.r.t. the distribution of the random vector $V = \Sigma^{-\top/2}U/||\Sigma^{-\top/2}U||$ valued in $\mathbb{S}^{d-1}$, yielding the definition below.

Definition 9.3 (Affine-invariant IRW depth)**.** *The Affine-Invariant Integrated Rank-Weighted (AI-IRW) depth relative to a square integrable random vector X with probability distribution P on $\mathbb{R}^d$ and positive definite covariance matrix Σ is given by:*

$$\forall x \in \mathbb{R}^d, \quad D_{\text{AI-IRW}}(x, P) = \mathbb{E}\left[D_{\text{H},1}(\langle V,\ x\rangle, P_V)\right], \tag{9.3}$$

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

The matrix $\Sigma^{-1/2}$ can be obtained either by singular value decomposition or by the Cholesky decomposition and thus corresponds to a "whitening" matrix rather than the true square root matrix. We keep this small abuse of notation for for a better understanding of the approach. Of course, in the case where the covariance matrix Σ of the supposedly square integrable r.v. X is not invertible, the AI-IRW depth notion should be applied to an orthogonal projection, after an appropriate dimensionality reduction step. From a computational perspective, The AI-IRW depth can be approximated by Monte-Carlo methods in the same way as (9.1), see Remark 9.2. As revealed by the proposition stated below, the depth function (9.3) inherits all the properties of (9.1) under similar assumptions and is remarkably invariant under any affine transformation in addition.

Proposition 9.4. *The assertions below hold true for any probability distribution P of a square integrable r.v. X valued in $\mathbb{R}^d$ with positive definite covariance matrix.*

(*i*) *The AI-IRW depth satisfies the properties $\mathbf{D}_1$ and $\mathbf{D}_4$. In addition, $\mathbf{D}_2$ and $\mathbf{D}_3$ hold for halfspace symmetric distributions.*

(*ii*) *The AI-IRW depth function is continuous at each point x that is not an atom for P.*

The proof is detailed in Section 9.5.1. It is known that for elliptical distributions, affine invariant data depth level sets are concentric ellipsoids with the same center and orientation as the density level sets (Liu and Singh, 1993). Therefore, the ordering

returned by affine-invariant data depths should be equal to that of the density function. Thus, in order to highlight the discrepancy between AI-IRW and IRW w.r.t. affine-invariance, we propose to compare the ordering returned by AI-IRW and IRW to that of the density function on the Gaussian distribution which belongs to the family of elliptical distributions. As illustrated by the Rank-Rank plots in Figure 9.1, the ordering defined by the (empirical) AI-IRW depth is generally much closer to that induced by the underlying density than the order defined by the original (IRW depth) version. See also the Figure 9.2 that illustrates the non affine-invariance of the IRW and the affine-invariance of the AI-IRW. Indeed, the IRW contours are spherical while the AI-IRW contours are ellipsoidal like those of the underlying Student-10 density.

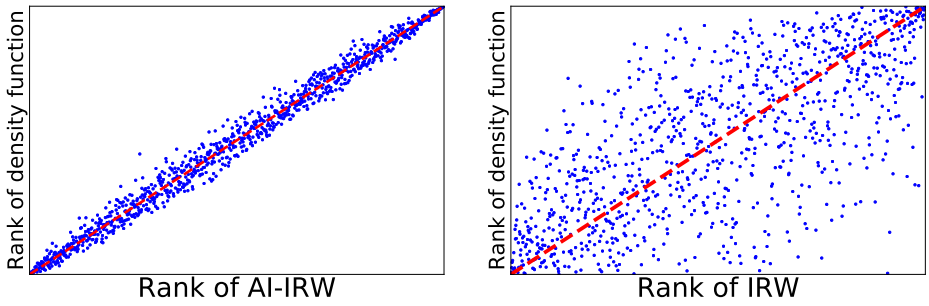


Figure 9.1 – Rank-Rank plots comparing the ranks of 1000 points sampled from a 10-d (anisotropic) Gaussian distribution with covariance matrix drawn at random from a Wishart distribution (with parameters (d, I_d)) induced by the depth (AI-IRW on the left, IRW on the right) and those induced by the Gaussian density.

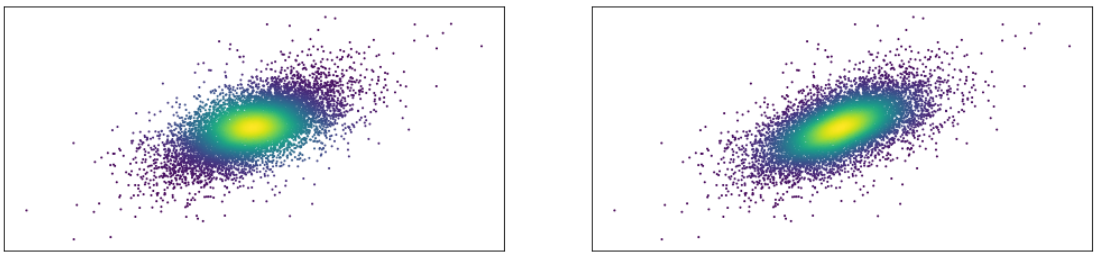


Figure 9.2 – The IRW depth (left) and the AI-IRW (right) depth on a Student-10 distribution. The darker the point, the lower the depth.

In practice, the distribution P is generally unknown as well as the covariance matrix Σ and only a sample $\mathcal{S}_n = \{X_1, \ldots, X_n\}$ composed of $n \geq 1$ independent realizations of the distribution P is available. A statistical counterpart of the AI-IRW depth can be obtained by replacing P with the empirical measure $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ and $\Sigma^{-\top/2}$ with an estimator $\widehat{\Sigma}^{-\top/2}$ based on $\mathcal{S}_n$ and plugging them next into formula (9.3) when $\widehat{\Sigma}$ is invertible, yielding: $\forall x \in \mathbb{R}^d$,

$$D_{\mathrm{AI-IRW}}(x, P_n) = \mathbb{E}\left[D_{\mathrm{H},1}(\langle \widehat{V}, x\rangle, P_{\widehat{V},n}) \mid \mathcal{S}_n\right], \tag{9.4}$$

where $\widehat{V} = \widehat{\Sigma}^{-\top/2}U/||\widehat{\Sigma}^{-\top/2}U||$ and U is a r.v. uniformly distributed on $\mathbb{S}^{d-1}$ independent from the X_i's. The two notations $D_{\mathrm{AI-IRW}}(x, P_n)$ and $\widehat{D}_{\mathrm{AI-IRW}}(x)$ are used throughout this section and have the same meaning. From a practical perspective, the (conditional) expectation (9.4) can also be approximated by means of a basic Monte-Carlo scheme, generating $n_{\mathrm{proj}} \geq 1$ i.i.d. random directions $U_1, \ldots, U_{n_{\mathrm{proj}}}$, copies of the generic r.v. U and independent from the original data $\mathcal{S}_n$: $\forall x \in \mathbb{R}^d$,

$$\widetilde{D}^{\mathrm{MC}}_{\mathrm{AI-IRW}}(x, P_n) = \frac{1}{n_{\mathrm{proj}}} \sum_{j=1}^{n_{\mathrm{proj}}} \min\left\{ F_{\widehat{V}_j,n}\left(\langle \widehat{V}_j,\ x\rangle\right), 1 - F_{\widehat{V}_j,n}\left(\langle \widehat{V}_j,\ x\rangle\right)\right\}, \tag{9.5}$$

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

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

Putting aside the issue of estimating $\Sigma^{-\top/2}$ (discussed below), attention should be paid to the fact that the approximate sample version (9.5) is very easy to compute (see the Algorithm 9.1 where the AI-IRW approximation is computed over the whole sample) and involves no optimization procedure, in contrast to many other notions of depth function.

Algorithm 9.1 Approximation of the AI-IRW depth.

input: $\mathcal{S}_n$,n_{proj}.

Construct $\mathbf{U} \in \mathbb{R}^{d\times n_{\mathrm{proj}}}$ by sampling uniformly n_{proj} vectors $U_1, \ldots, U_{n_{\mathrm{proj}}}$ in $\mathbb{S}^{d-1}$

Compute $\widehat{\Sigma}$ using any estimator

Perform Cholesky or SVD on $\widehat{\Sigma}$ to obtain $\widehat{\Sigma}^{-1/2}$

Compute $\mathbf{V} = \widehat{\Sigma}^{-1/2}\mathbf{U}/||\widehat{\Sigma}^{-1/2}\mathbf{U}||$

Compute $\mathbf{M} = \mathbf{X}\mathbf{V}$, where $\mathbf{X} = [X_1, \ldots, X_n]^\top \in \mathbb{R}^{n\times d}$

Compute the rank value $\sigma(i, j)$, the rank of index i in $\mathbf{M}_{:,j}$ for every $i \leq n$ and $j \leq n_{\mathrm{proj}}$

Set $\widetilde{D}^{\mathrm{MC}}_{\mathrm{AI-IRW}}(X_i, \mathcal{S}_n) = \frac{1}{n_{\mathrm{proj}}}\sum_{j=1}^{n_{\mathrm{proj}}} \sigma(i, j)$ for every $i \leq n$

return $\left\{\widetilde{D}^{\mathrm{MC}}_{\mathrm{AI-IRW}}(X_i, \mathcal{S}_n),\ \ 1 \leq i \leq n\right\}$

On estimating the square root of the precision matrix. Consider the $d \times n$ matrix $(X_1, \ldots, X_n)$ with the X_i's as columns. The simplest way of building an estimate $\widehat{\Sigma}^{-\top/2}$ consists in computing the empirical version of the covariance matrix $\widehat{\Sigma} = (1/n)\sum_{i=1}^{n} X_i X_i^\top$, which is a natural and nearly unbiased estimator, and inverting next its square root, when the latter is positive definite (which happens with overwhelming probability). For simplicity, this is the estimation we consider in the finite-sample study presented in the next section. However, alternative techniques can be used, yielding possibly more efficient estimators under specific assumptions, in high-dimension especially. Shrinkage procedures for covariance estimation under sparsity conditions have been investigated in e.g. Ledoit and Wolf (2004); Chen et al. (2010) and Schäfer

and Strimmer (2005), while a lasso method for direct estimation of the precision matrix, avoiding matrix inversion, is proposed in Friedman et al. (2008). Robust covariance estimation techniques, tailored to situations where the data are possibly contaminated or heavy-tailed, have also been documented in the literature, see e.g. Rousseeuw (1984) and Rousseeuw and van Driessen (1999). Classically, from a symmetric definite positive estimator of the covariance matrix, one can easily build an estimator of the square root of the precision matrix by inverting a triangular/diagonal matrix. Due to the presence of $\widehat{\Sigma}^{-\top/2}$ in (9.4) (respectively, in (9.5)), it is far from straightforward to assess the accuracy of the estimators of the AI-IRW depth proposed above. It is the purpose of the next section to study the uniform deviations between (9.3) and its empirical versions from a nonasymptotic perspective.

9.2 Finite-Sample Analysis - Concentration Bounds

We now investigate the accuracy of the statistical version, as well as that of its Monte-Carlo approximation, of the AI-IRW depth function introduced in the previous section in a nonasymptotic fashion. Precisely, we establish a concentration bound for the maximal deviations between the true and estimated AI-IRW depth functions. We assume here that the estimator of the square root of the precision matrix is given by the inverse of the square root of the empirical covariance, when the latter is definite positive (which happens with overwhelming probability), and by that of any definite positive regularized version (e.g. Tikhonov) of the latter otherwise. The subsequent analysis requires additional hypotheses, listed in the next section.

9.2.1 Assumptions

The first assumption, classical when estimating the precision matrix (see e.g. Cai and Zhou, 2013 or Fan et al., 2016), stipulates that the eigenvalues $\sigma_1, \ldots, \sigma_d$ of the covariance matrix Σ of the square integrable random vector X considered are bounded away from zero.

Assumption 9.5. *There exists $\varsigma > 0$ such that: $\forall k \in \{1, \ldots, d\}$, $\varsigma \leq \sigma_k$.*

The second assumption is technical, see Davis and Kahan (1970). It stipulates that Σ's eigenvalues are all of multiplicity 1 and that Σ's minimum eigengap is bounded away from zero.

Assumption 9.6. *There exists $\gamma > 0$ such that: $\forall k \in \{1, \ldots, d-1\}$, $\gamma \leq \sigma_{(k)} - \sigma_{(k+1)}$, where $\sigma_{(1)} > \ldots > \sigma_{(d)}$ are Σ's eigenvalues sorted by decreasing order of magnitude.*

We point out that, just like when Σ is not invertible, one can always reduce the analysis to a situation where Assumption 9.6 is fulfilled by means of a preliminary dimensionality reduction step. Notice incidentally that, when $\Sigma = \sigma \, \mathrm{I}_d$, with $\sigma > 0$, the AI-IRW reduces to IRW. The other assumptions correspond to smoothness conditions of Lipschitz type for the function $\phi : (u, x) \in \mathbb{S}^{d-1} \times \mathbb{R}^d \mapsto \mathbb{P}\left\{\langle u, X\rangle \leq \langle u, x\rangle\right\}$.

Assumption 9.7. (Uniform Lipschitz condition in projection) *For all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, there exists $L_p < +\infty$ such that:*

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

Assumption 9.8. (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:*

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

Notice that the same assumptions are involved in the non-asymptotic rate bound analysis carried out for the halfspace depth estimator in Burr and Fabrizio (2017) and are used to establish limit results related to its approximation in Nagy et al. (2020b). The Lipschitz conditions are satisfied by a large class of probability distributions, for which Lipschitz constants L_R and L_p can be both explicitly derived. For instance, if the distribution P of X has compact support included in the ball $\mathcal{B}(0, r) = \{x \in \mathbb{R}^d : ||x|| \leq r\}$ (relative to the Euclidean norm $||\cdot||$) with $r > 0$ and is absolutely continuous w.r.t. the Lebesgue measure with a density bounded by $M > 0$, the uniform Lipschitz conditions are then fulfilled with $L_R = MV_{d,r}$ and $L_p = MV_{d-1,r}$, where $V_{d,r} = \pi^{d/2} r^d / \Gamma(d/2 + 1)$ is the volume of the ball $\mathcal{B}(0, r)$ and $z \geq 0 \mapsto \Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \mathrm{d}t$ means the Gamma function, refer to Burr and Fabrizio (2017) for additional examples. In contrast, a necessary condition for Assumption 9.7 to be satisfied is the absolute continuity of the measure P w.r.t. the Lebesgue measure, see Section 4 in Nagy et al. (2020b).

Lemma 9.9. *Let* $r > 0$ *and denote by* $V_{d,r}$ *the volume of the d-ball* $\mathcal{B}(0, r)$. *Assume that* X *takes its values in* $\mathcal{B}(0, r)$ *and has an* M*- bounded density w.r.t. the Lebesgue measure* λ_d. *The r.v.* X *is uniformly* Radially Lipschitz Continuous *with constant* $L_R = MV_{d,r}$.

Proof. Let $x \in \mathbb{R}^d$. By $||.||_g$, we mean the geodesic norm on the unit sphere of $\mathbb{R}^d$. It holds:

$$\begin{aligned}
|\phi(u, x) - \phi(v, x)| &\leq \mathbb{P}\Big\{X \in \mathcal{B}(0, r) \,:\, \langle u, X - x\rangle \text{ and } \langle v, X - x\rangle \text{ are of opposite sign}\Big\} \\
&\leq M\, \lambda_d \Big\{z \in \mathcal{B}(-x, r) \,:\, \langle u, z\rangle \text{ and } \langle v, z\rangle \text{ are of opposite sign}\Big\} \\
&\overset{(i)}{\leq} M\, V_{d,r} \times \frac{2}{\pi} \arccos\Big(\langle u, v\rangle\Big) \\
&= M\, V_{d,r} \times \frac{2}{\pi}\, ||u - v||_g \\
&\leq M\, V_{d,r}\, ||u - v||,
\end{aligned}$$

Where (i) arises from the fact that the volume of $\mathcal{E}_{u,x,y}$ defined as:
$\mathcal{E}_{u,x,y} = \{z \in \mathcal{B}(-x, r) : \langle u, z\rangle$ and $\langle v, z\rangle$ are of opposite sign$\}$ is the volume of two cones of angle $||u - v||_g$, as depicted in Figure 9.3. ■

Lemma 9.10. *Let* $r > 0$ *and assume that* X *takes its values in* $\mathcal{B}(0, r)$ *and has* M*-bounded density w.r.t. the Lebesgue measure* λ_d. *Thus* X *is uniformly* Lipschitz continuous in projection *with constant* $L_p = MV_{d-1,r}$.

Proof. Let $u \in \mathbb{S}^{d-1}$. It holds:

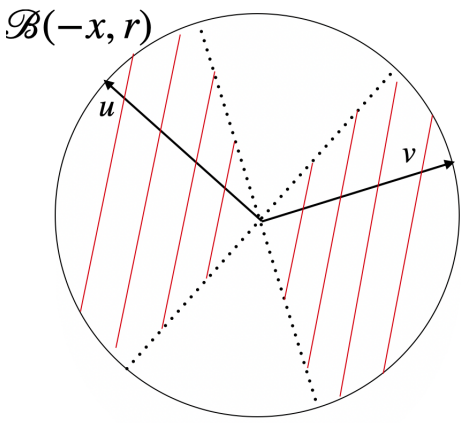


Figure 9.3 – Illustration of the set $\mathcal{E}_{u,x,y}$ in $\mathbb{R}^2$. It corresponds to the portion of $\mathcal{B}(-x, r)$ hatched in red.

$$
\begin{aligned}
|\phi(u,x) - \phi(u,y)| &\leq \mathbb{P}\Big\{X \in \mathcal{B}(0,r) \,:\, \langle u, X - x\rangle \text{ and } \langle u, X - y\rangle \text{ are of opposite sign}\Big\} \\
&\leq M\ \lambda_d\Big\{z \in \mathcal{B}(0,r) \,:\, \langle u, z - x\rangle \text{ and } \langle u, z - y\rangle \text{ are of opposite sign}\Big\} \\
&\overset{(i)}{\leq} M\ V_{d-1,r} \times |\langle u, x\rangle - \langle u, y\rangle| \\
&\leq M\ V_{d-1,r}\ ||x - y||.
\end{aligned}
$$

Where (i) arises from the fact that we encompass $\mathcal{F}_{u,x,y}$ by an hyper-cylinder of length $|\langle u, x\rangle - \langle u, y\rangle|$ where $\mathcal{F}_{u,x,y} = \{z \in \mathcal{B}(0,r) \,:\, \langle u, z - x\rangle \text{ and } \langle u, z - y\rangle \text{ are of opposite sign}\}$, as illustrated in Figure 9.4. ■

9.2.2 Intermediate Results

In order to prove our main results, we first recall useful results on maximum deviations of the halfspace depth estimator as well as the sample covariance matrix in the case of sub-Gaussian distributions.

Lemma 9.11 (Shorack and Wellner (1986), Chapter 26)**.** *Let $P \in \mathcal{P}(\mathbb{R}^d)$. Let $X_1, \ldots, X_n$ a sample from P with empirical measure $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$. Denote by F_u and $F_{u,n}$ the cdf of P_u and $P_{u,n}$ respectively. Then, for any $t > 0$, it holds:*

$$
\mathbb{P}\left(\sup_{\substack{x \in \mathbb{R}^d \\ u \in \mathbb{S}^{d-1}}} \left|F_{u,n}(u^\top x) - F_u(u^\top x)\right| > t\right) \leq \frac{6(2n)^{d+1}}{(d+1)!} \exp(-nt^2/8).
$$

Proof. Combining the bound in Vapnik and Chervonenkis (1974), page 215, with the fact that the set of halfspaces of $\mathbb{R}^d$ is a Vapnik-Chervonenkis class (VC) with dimension equal to $d + 2$ (Dudley, 1979), and applying this to the classical Vapnik-Chervonenkis inequality (Vapnik and Chervonenkis, 1971, Theorem 3) lead to the desired result. ■

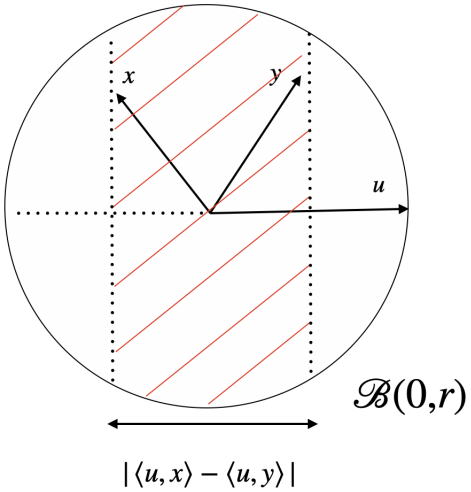


Figure 9.4 – Illustration of the set $\mathcal{F}_{u,x,y}$ in $\mathbb{R}^2$. It corresponds to the portion of $\mathcal{B}(0,r)$ hatched in red.

Lemma 9.12 (Variant of Vershynin, 2012, Proposition 2.1)**.** *Let Σ be the covariance matrix of a ϱ sub-Gaussian random variables X that takes its values in $\mathbb{R}^d$. Let $X_1, \ldots, X_n$ be a sample from X and denote by $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ the sample covariance (SC) estimator of Σ. Then it holds:*

$$\mathbb{P}\left(||\widehat{\Sigma} - \Sigma||_{\text{op}} > t\right) \leq 2 \times 9^d \exp\left\{-\frac{n}{2}\min\left\{\frac{t^2}{(32\varrho^2)^2}, \frac{t}{32\varrho^2}\right\}\right\}.$$

Let $\sigma_d > \ldots > \sigma_1$ and $\widehat{\sigma}_d > \ldots > \sigma_1$ be respectively the ordered eigeinvalues of Σ and $\widehat{\Sigma}$. Using Weyl's Theoreom (Weyl, 1912), it holds:

$$\mathbb{P}\left(\max_{1\leq k\leq d}|\widehat{\sigma}_k - \sigma_k| \ > \ t\right) \leq 2 \times 9^d \exp\left\{-\frac{n}{2}\min\left\{\frac{t^2}{(32\varrho^2)^2}, \frac{t}{32\varrho^2}\right\}\right\}.$$

Proof. Let N_ρ be an ρ-net of the sphere $\mathbb{S}^{d-1}$. Applying Lemma 2.2 in Vershynin (2012) on $\widehat{\Sigma} - \Sigma$, for any $t, \rho > 0$, we have:

$$\begin{aligned}\mathbb{P}\left(||\widehat{\Sigma} - \Sigma||_{\text{op}} > t\right) &\leq \mathbb{P}\left(\frac{1}{1-2\rho}\max_{v\in N_\rho}|v^\top(\widehat{\Sigma} - \Sigma)v| > t\right) \\ &\leq |N_\rho|\,\mathbb{P}\left(|v^\top(\widehat{\Sigma} - \Sigma)v| > (1-2\rho)\,t\right),\end{aligned}$$

where $|N_\rho|$ stands for the cardinal of the set N_ρ. Noticing that $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ is a sum of independent matrices we have:

$$v^\top(\widehat{\Sigma}-\Sigma)v = \frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}Z_i,$$

where $Z_i = (v^\top X_i)^2$ for every $1 \le i \le n$ and $Z_i - \mathbb{E}Z_i$ are i.i.d random variables that are $((16\varrho)^2, 16\varrho^2)$ sub-exponential.

Choosing $\rho = 1/4$, noticing that $N_{1/4} \le 9^d$ and applying the sub-exponential tail bound lead to the desired result.

■

9.2.3 Main Results

We are now ready to state the main theoretical contribution of this chapter. The bounds stated in the theorem below reveal the accuracy of the statistical estimates (9.4) and (9.5) and highlight their behavior through explicit constants depending on the parameters of the hypotheses involved.

Theorem 9.13. *Suppose that the distribution P of the r.v. X is ϱ sub-Gaussian and satisfies Assumptions 9.5, 9.6, 9.7 and 9.8. The following assertions hold true.*

(i) For any $\delta \in \left(\max\{\Theta, 12.9^d\}\, e^{-\frac{n}{2}\min\left\{\alpha,\alpha^2,\alpha\Delta/8\right\}}, 1\right)$, we have with probability at least $1-\delta$:

$$\sup_{x\in\mathbb{R}^d}\left|D_{\text{AI-IRW}}(x,P_n) - D_{\text{AI-IRW}}(x,P)\right| \le \Delta(L_R,d,\gamma,\varsigma,\varrho)\max_{s=1,2}\left(\frac{d+\log(2/\delta)}{n}\right)^{1/s} + \sqrt{\frac{8\log(\Theta/\delta)}{n}},$$

where $\Delta = 512 L_R \varrho^2 \max\{1/\xi,\ 2\sqrt{2d}/\gamma\}$ with $\xi \in (0,\varsigma)$, $\alpha(\varsigma,\varrho) = (\varsigma-\xi)/(32\varrho^2)$ and $\Theta = 12(2n)^{d+1}/(d+1)!$.

(ii) Let $r > 0$. For any $\delta \in \left(\max\{\Theta,\ 12.9^d\}\, e^{-n\min\left\{\alpha,\alpha^2,\alpha\Delta/8\right\}}, 1\right)$, we have with probability at least $1-\delta$:

$$\sup_{x\in\mathcal{B}(0,r)}\left|\widetilde{D}^{MC}_{\text{AI-IRW}}(x,P_n) - D_{\text{AI-IRW}}(x,P)\right| \le 2\sqrt{\frac{d\log\left(3r\ n_{\text{proj}}\right)+\log(6/\delta)}{18 n_{\text{proj}}}} + \frac{4L_p}{3n_{\text{proj}}} + \frac{8\Delta}{3}\max_{s=1,2}\left(\frac{d+\log(2/\delta)}{n}\right)^{1/s} + \sqrt{\frac{128\log(3\Theta/2\delta)}{9n}}$$

where the constants Θ, Δ, α and the parameter $\xi \in (0,\varsigma)$ are the same as those involved in (i).

The detailed proof is postponed to the Section 9.5.2. The upper confidence bound in assertion (i) is decomposed into two terms. The first term, of order $O(n^{-1/2})$, owes its presence to the replacement of $\Sigma^{-1/2}$ by its estimator. The second term, of order $O(\sqrt{\log(n)/n})$ and exhibiting a sublinear dependence in the dimension d , corresponds to the bound that would be obtained if $\Sigma^{-1/2}$ were known (it is then derived by means of the arguments used to study the concentration properties of the empirical halfspace depth, see Chapter 26 in Shorack and Wellner, 1986). The upper confidence bound in assertion (ii) differs from that in assertion (i) in two respects. First, the additional terms clearly show the effect of the Monte-Carlo approximation, which is negligible when $n \gg n_{\text{proj}}$. Second, the maximal deviation is taken over a compact subset of $\mathbb{R}^d$. Furthermore, our theoretical analysis can be easily extended to the deviations of the sample version of IRW by simply omitting the term involving the square root of the precision matrix corresponding to the first term of (i) and the last term of (ii) leading to faster rates (see Corollary 9.15).

A limited confidence interval. The proof of the assertion (i) relies on controlling the deviations between the eigenvectors (resp., the inverses of the square-root eigenvalues) of $\widehat{\Sigma}$ and those of the true covariance matrix. The lower bound of the δ-range results from this control and is not limiting in practice since it decreases exponentially fast when the sample size increases.

About the constants. Both upper bounds are provided with explicit constants. The explicit linear dependency in d is due to the operator norm that appears in the proof when controlling the eigenvectors of $\widehat{\Sigma} - \Sigma$. It implies an additional square root of d in the constant Δ following the classical inequality $||A||_{\text{op}} \leq \sqrt{d}||A||_1$ for any matrix $A \in \mathbb{R}^{d\times d}$ of rank d. However, Lipschitz constants L_p and L_R, that are mandatory in order to derive bounds uniformly on $\mathbb{R}^d$ (or $\mathcal{B}(0, r)$), appear to exhibit an implicit dependence on the dimension d. Indeed, these constants can be derived for r.v. valued in a compact support with bounded density exhibiting an exponential dependence on d. Unfortunately, this concern cannot be avoided unless removing the supremum involved in (i) and (ii). While the depth value at a single point $x \in \mathbb{R}^d$ is usually of limited importance, it is often more relevant in practice that an ensemble of depth values, i.e. the set $\{D(x, P),\ x \in \mathbb{R}^d\}$, are simultaneously well approximated by their empirical versions for comparison purposes. This implies estimation guarantees for the ranks induced by the depth function when computed on the whole sample $X_1, \ldots, X_n$, on which several applications such as anomaly detection fully rely on. The eigengap γ appears in the denominator due to the use of a variant of the Davis-Kahan theorem (Davis and Kahan, 1970), so as to control the deviations between the eigenvectors of $\widehat{\Sigma}$ and those of Σ, and can not be avoided. Observe that both upper-bounds explode as γ or ς vanish. These constants, related to the covariance matrix estimation, are often small in practice (see the Appendix where they are computed on the benchmarked datasets used in Section 9.3.4). However, they are often negligible w.r.t. the Lipschitz constant in the numerator that is $O(e^d)$ as mentioned above and is thus not limiting. Notice finally the presence of the free parameter $\xi \in (0,\ \varsigma)$ in the bound: the larger ξ, the smaller the constant Δ and the shorter the range of acceptable confidence levels δ.

On optimality. In absence of lower bound (and to the best of our knowledge, no such result is documented in the statistical depth literature yet), the optimality of the bounds above cannot be claimed of course. However, the proof partly consists in bounding the risk of the estimator of the covariance matrix Σ and involves the estimation rates given in Lemma 9.12, which are known to be optimal for sub-Gaussian distributions

(Vershynin, 2012). It has been shown that faster rates for the estimation of the inverse of the covariance matrix can be established under additional sparsity assumptions (see e.g. Theorem 5 in Cai et al., 2010).

Choice of n_{proj}. The difficulty of approximating an integral over $\mathbb{R}^d$ by means of Monte-Carlo techniques grows with d. Our theoretical results, such as the upper bound in (ii), shed light on the behaviour of n_{proj} w.r.t. the dimension d. Indeed, focusing on the term $4L_p/(3n_{\text{proj}})$, L_p can be made explicit for density bounded distributions involving the volume of the unit sphere $\mathbb{S}^{d-1}$ that depends exponentially on d (see the paragraph above Theorem 9.13). Thus, n_{proj} should be higher than $O(e^d)$ to yield a good statistical approximation. However, in practice, since computation times depend on n_{proj}, a trade-off between statistical accuracy (the higher n_{proj}, the better) and computational burden (the higher n_{proj}, the heavier) must be found in practice, see Section 9.3.

Remark 9.14. (Related work) *We point out that nonasymptotic results about the accuracy of sample versions of statistical depths, such as those stated above, are seldom in the literature. To the best of our knowledge, rate bounds have only been derived in the halfspace depth case before. The first result (see Shorack and Wellner, 1986, Chapter 26), where uniform rates of the sample version are provided, uses the fact that the set of halfspaces in $\mathbb{R}^d$ is of finite* VC *dimension. Recently, this result has been refined under the Assumptions 9.7 and 9.8 in Burr and Fabrizio (2017). Asymptotic rates of convergence for the Monte-Carlo approximation of the halfspace depth, i.e. when the minimum over the unit hypersphere is approximated from a finite number of directions, have been recently established in Nagy et al. (2020b). In contrast to the finite-sample framework, uniform asymptotic rates have been proved in several settings. Unfortunately, approximating a minimum over the unit sphere $\mathbb{S}^{d-1}$ using a Monte-Carlo scheme is not optimal. Indeed when the distribution is assumed to belong to a bounded subset of $\mathbb{R}^d$ with bounded density, the authors obtain slow rates of order $O((\log(n_{\text{proj}})/n_{\text{proj}})^{1/(d-1)})$ suffering from the curse of dimensionality. Futhermore, they show that obtaining uniform rates of the halfspace depth approximation is not possible in absence of the bounded density assumption (see Section 4.2 in Nagy et al., 2020b).*

A finite sample analysis on the approximation of IRW can be derived from our results on AI-IRW as it is described in the next corollary. The proof, which follows from that of the theorem, is detailed in Section 9.5.3 for completeness.

Corollary 9.15. *Suppose that the distribution P of the r.v. X satisfies Assumptions 9.7 and 9.8. Then, for any $\delta \in (0,1)$, it holds:*

$$\sup_{x \in \mathcal{B}(0,r)} \left| \widetilde{D}^{MC}_{\text{IRW}}(x) - D_{\text{IRW}}(x,P) \right| \leq \sqrt{\frac{8\log(\Theta/\delta)}{n}} + 2\sqrt{\frac{d\log\left(3r_{\text{proj}}\right) + \log(6/\delta)}{8n_{\text{proj}}}} + \frac{2L_p}{n_{\text{proj}}},$$

where $\Theta = 12(2n)^{d+1}/(d+1)!$.

9.3 Numerical Experiments

The advantages of the novel notion of depth introduced in Section 9.1 are supported by various experimental results in this part. First, we explore empirically the behavior of the returned ranks as the number of sampled projections increases in order to provide

insights on the choice of n_{proj} in practice. Second, the robustness of the AI-IRW depth is explored when using sample covariance (SC) and Minimum Covariance Determinant (MCD; Rousseeuw, 1984) estimators. Further, the variance of the returned score w.r.t. to both sample realizations and noisy projections is investigated. Eventualy, the application of the AI-IRW depth to anomaly detection is considered, illustrating clearly the improvement on the performance attained, compared to the IRW depth.

9.3.1 On Approximating the AI-IRW Depth

The accuracy of Monte-Carlo approximation, depending on the number n_{proj} of random directions uniformly sampled, is evaluated for the empirical versions of the AI-IRW depth. A robust estimator of the AI-IRW is also introduced using the well-known Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1984) of the sample covariance matrix. The experiment is based on samples of size $n = 1000$ drawn from a centered Gaussian distribution with covariance matrix sampled from a Wishart distribution (with parameters (d, I_d)) where the dimension d varies in $\{2, 5, 10, 15, 20, 30, 40, 50\}$. We compute $\widetilde{D}^{\text{MC}}_{\text{AI-IRW}}$ on these samples by varying the number of projections n_{proj} between 100 and 7000. As AI-IRW does not possess any closed-form, we propose to evaluate the quality of the returned ranks considering $\widetilde{D}^{\text{MC}}_{\text{AI-IRW}}$ computed with $n_{\text{proj}} = 200000$ projections as the "true" depth. The coherence between ranks is assessed using the popular Kendall τ correlation coefficient, see Kendall (1938). This whole procedure is repeated 10 times and the averaged results are reported in Figure 9.5. As expected, the quality of the approximation increases with n_{proj} and decreases with d. Interestingly, sharp approximations are obtained with far less than $O(e^d)$ projections. Indeed, in the worst case, i.e. when $d = 50$, a correlation of 0.93 is attained for AI-IRW, using both SC and MCD (with support fraction set to $(n+d+1)/2$) estimators, with only 5500 directions which is roughly $100 \times d$ while $e^{50} \approx 10^{21}$. In low dimension, few projections are needed to obtain correlation higher than 0.98. In view of these results and because of the computation time of the approximations depicted in Figure 9.6, choosing $n_{\text{proj}} = 100d$ appears as a good compromise between statistical accuracy and computation time, as done in the next experiment.

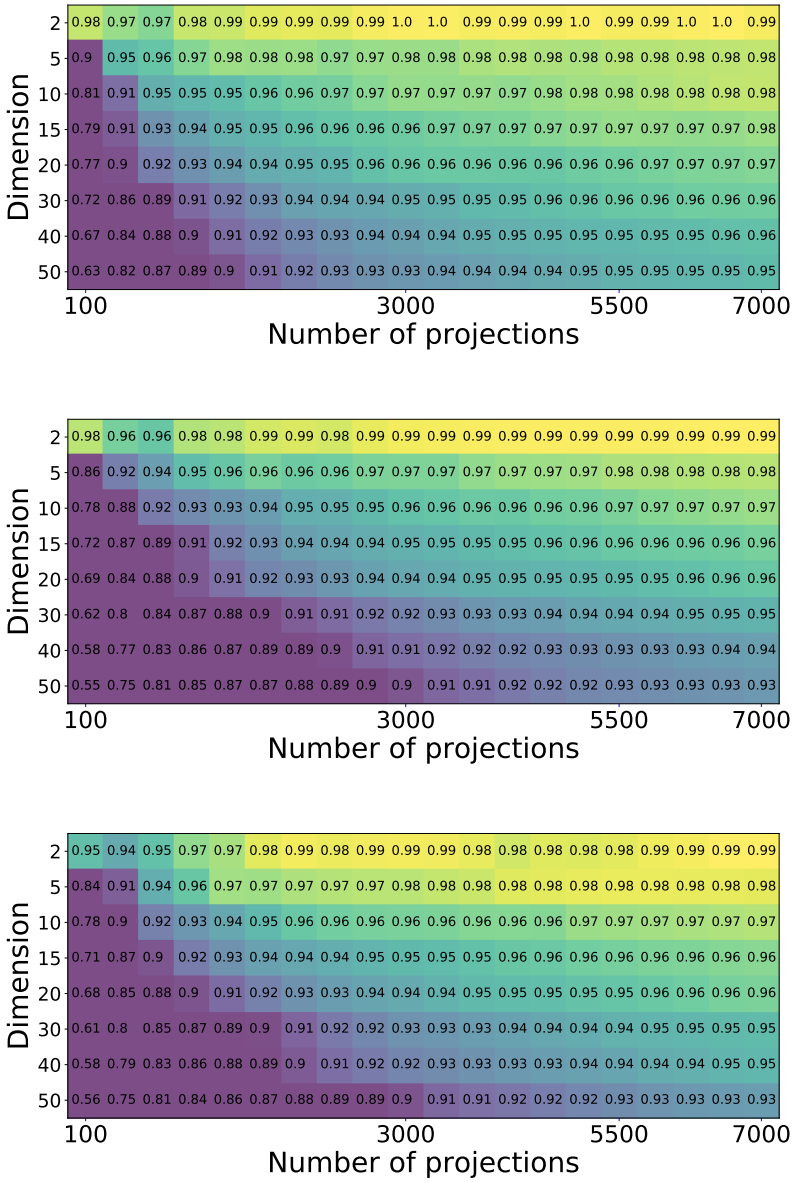


Figure 9.5 – Kendall correlation between the approximated ranks of the IRW depth (top), AI-IRW using SC (middle), AI-IRW using MCD (bottom) and their true ranks depending on the number of approximating projections n_{proj} for a Gaussian distribution.

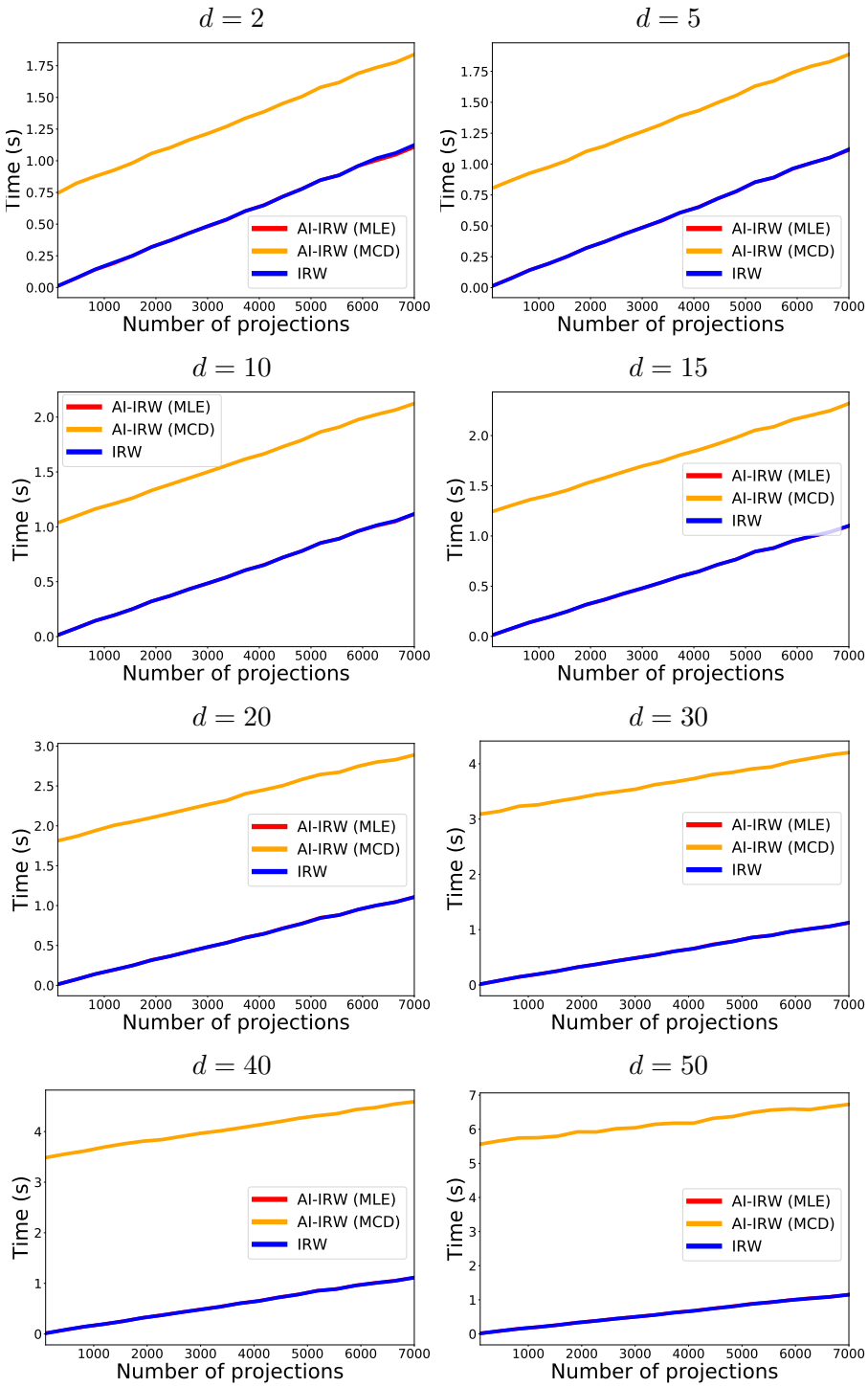


Figure 9.6 – Computation time of the AI-IRW depth using both SC and MCD estimators and the IRW depth depending on the number of projections for various dimensions. AI-IRW and IRW have the same computation time since the computation of the sample covariance matrix is negligible w.r.t. the computation of the IRW depth.

9.3.2 Exploring AI-IRW with the MCD Estimator

In this section, we investigate the quality of the approximation as well as the robustness of the AI-IRW depth using both SC and MCD estimators with two experiments.

9.3.2.1 Approximation and Robustness

The first experiment is conducted as follows. The accuracy of Monte-Carlo approximation, depending on the number m of random directions uniformly sampled, is evaluated for the empirical versions of the AI-IRW depth using SC and MCD estimators as well as the IRW depth. The experiment is based on samples of size $n = 1000$ drawn from the multivariate standard Gaussian distribution (standard, so that non affine invariant depth are not disadvantaged) in dimension $d = 5$. The classical Kendall τ distance, given by

$$d_\tau(\text{per}, \text{per}') = \frac{2}{n(n-1)} \sum_{i<j} \mathbb{I}_{\{(\text{per}(i)-\text{per}(j))(\text{per}'(i)-\text{per}'(j))<0\}},$$

for all permutations per and per$'$ of the index set $\{1, \dots, n\}$, is used to measure the deviation between the ranks induced by the "true" depth (approximated with $n_{\text{proj}} = 200000$ projections since there exists no closed-form) and those defined by the Monte-Carlo approximation of the sampling version. The averaged Kendall τ's (over 10 runs), that correspond to one minus the Kendall correlations, are displayed in Figure 9.7. One observes that the approximate empirical AI-IRW depth is not affected by the covariance estimation step, its behavior is similar to that of the approximate empirical IRW depth for the Gaussian distribution when using both covariance estimators. On the other hand, a slight advantage is awarded to MCD under the heavy-tailed Student-3 model.

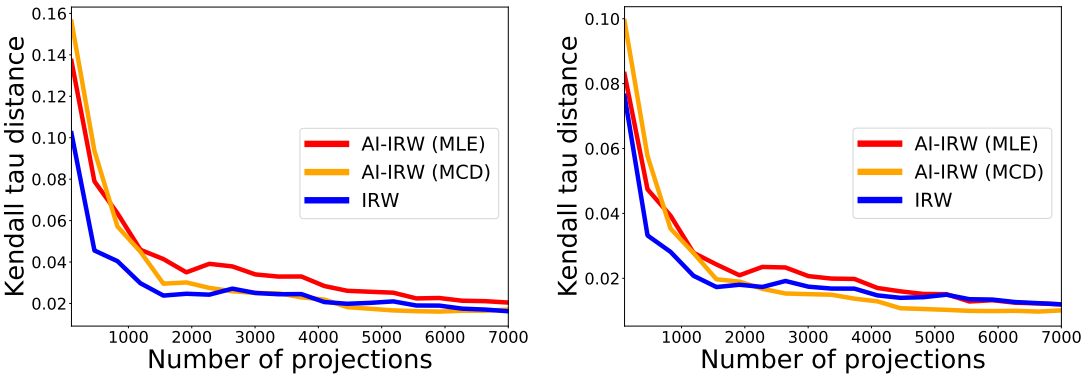


Figure 9.7 – Coherence of the returned rank measured by Kendall τ distance depending on the number of approximating projections for Gaussian (left) and Student-3 (right) distributions for AI-IRW (using SC and MCD estimates) and IRW.

9.3.2.2 Robustness w.r.t. Increasing Proportion of Outliers

In the second experiment, we examine the robustness of the returned ordering. It is based on the construction of two contaminated data sets from samples of size $n = 100$ drawn from the multivariate standard Gaussian distribution (standard, so that non affine invariant depths are not disadvantaged) in dimension $d = 2$. To build corrupted data set, the two following contaminated models are used. The first is based on adding "isolated outliers" where each of them is defined as $(0, a)$ where a is sampled uniformly

between $[4, 400]$. The second is based on adding "aggregated outliers" by randomly and uniformly drawing a location a' in $[4, 400]$ and then drawing anomalies following the Gaussian distribution $\mathcal{N}((a', a'), \mathrm{I}_2)$. Therefore, each data set is constructed as follows: a proportion of outliers $\epsilon \in [0, 0.15]$ is added to the normal data, represented by the standard Gaussian distribution, following one of the two aforementioned contamination models and thus yields two settings. The AI-IRW depth using SC and MCD estimators as well as the IRW depth are computed on these contaminated data sets. The Kendall τ distance is used to measure the deviation between the "true" ranks that are computed on samples without corruption and those computed on samples with corruption w.r.t. a proportion of anomalies ϵ. The averaged Kendall τ's (over 100 runs) are displayed in Figure 9.8. As expected, results show that the MCD estimator provides robustness to the AI-IRW depth while the sample covariance estimator breaks down after only 1% of anomalies. Interestingly, the MCD estimator does not bring more robustness than the underlying robustness of the IRW depth. It highlights somehow a "worst case" robustness between the estimator of the covariance matrix and the underlying IRW depth which is reached by the latter.

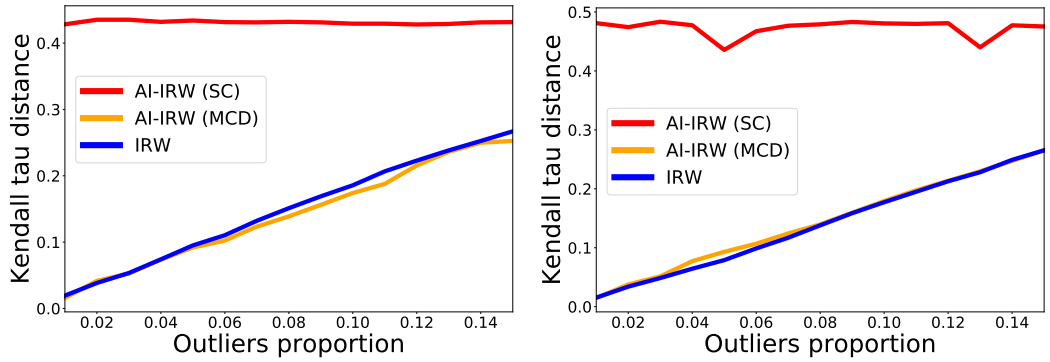


Figure 9.8 – Coherence of the returned rank measured by Kendall τ depending on outliers proportion for Student-3 (left) and Gaussian (right) distributions for AI-IRW (using SC and MCD estimates) and IRW.

9.3.3 Variance of AI-IRW Score

9.3.3.1 Variance w.r.t. Sample Realizations

We compare the stability of the approximation estimator AI-IRW measuring its variance. For 100 points stemming from a 10-dimensional Gaussian distribution with zero mean and covariance matrix drawn from the Wishart distribution (with parameters (d, I_d)) on the space of definite matrices, the variance of the returned score is computed on two points, denoted by x_1 and x_2, drawn randomly from the 100 points previous points. The score is computed for AI-IRW, IRW, halfspace mass (Chen et al., 2015a) and halfspace depths each approximated using $n_{\text{proj}} = 100$ directions. Figure 9.9 illustrates that (1) no additional variance is introduced by the affine-invariant version, (2) closeness of the three scores (due to absence of correlation), as well as (3) their higher concentrations compared to halfspace mass and halfspace depth.

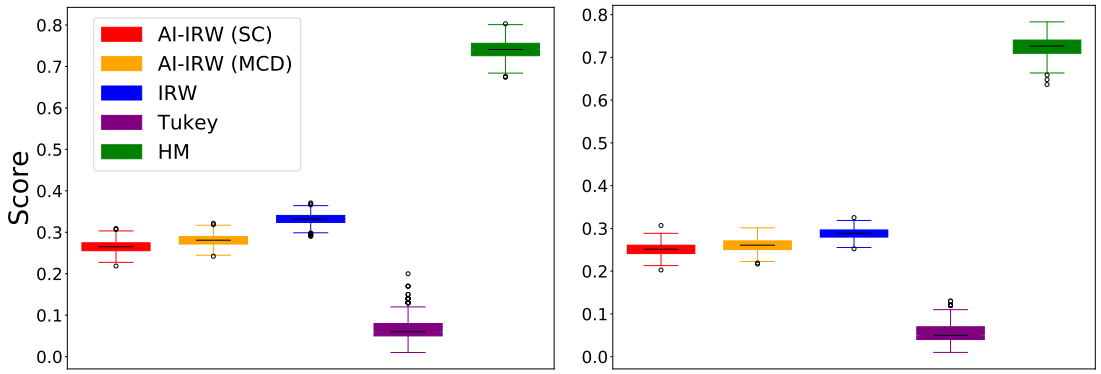


Figure 9.9 – Variance of the score of x_1, x_2 (from left to right) over 1000 repetitions for the AI-IRW, IRW, halfspace mass (HM) and halfspace (Tukey) depths.

9.3.3.2 Variance w.r.t. Noisy Directions

The previous experiment is repeated with different level of Gaussian noise that are added to sampled directions, i.e. $U = \frac{W+\epsilon\mathcal{N}(\mathbf{0},\mathrm{I}_d)}{||W+\epsilon\mathcal{N}(\mathbf{0},\mathrm{I}_d||}$. This experiment is conducted with AI-IRW, IRW, HM and halfspace depth using $n_{\text{proj}} = 100$ sampled directions. The root mean square variance (over 100 repetitions) between the returned score and the original score (without noise) are computed for x_1, x_2 (same as those in Section 9.3.3.1), see Figure 9.10. Results show that AI-IRW (using the SC estimator) shares very few differences with IRW while the superiority of AI-IRW (and IRW) over the existing methods depth such as haflspace and halfspace mass is highlighted.

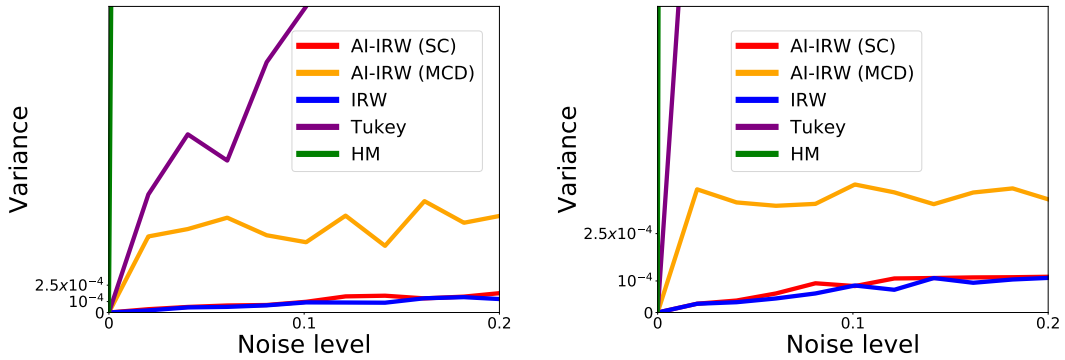


Figure 9.10 – Variance of the score of x_1, x_2 (from left to right) over the noise level induced in sampled directions with 1000 repetitions for the AI-IRW, IRW, Tukey depth.

9.3.4 Application to Anomaly Detection

9.3.4.1 Anomaly Detection: a Comparison on a Toy Data Set

In this part, a comparison between AI-IRW (with MCD), IRW and the halfspace depth is provided. To conduct this experiment, we construct a toy contaminated data set (see Figure 9.11, left) where aggregated outliers (green points) and some independent outliers (red points) are added to 1000 points stemming from a 2-dimensional *Gaussian*

distribution. The 100 lowest scores are depicted (see Figure 9.11, right) for the three benchmarked data depths. Results show that AI-IRW is able to assign the lowest depth to these anomalies while IRW and Tukey both fail to identify them.

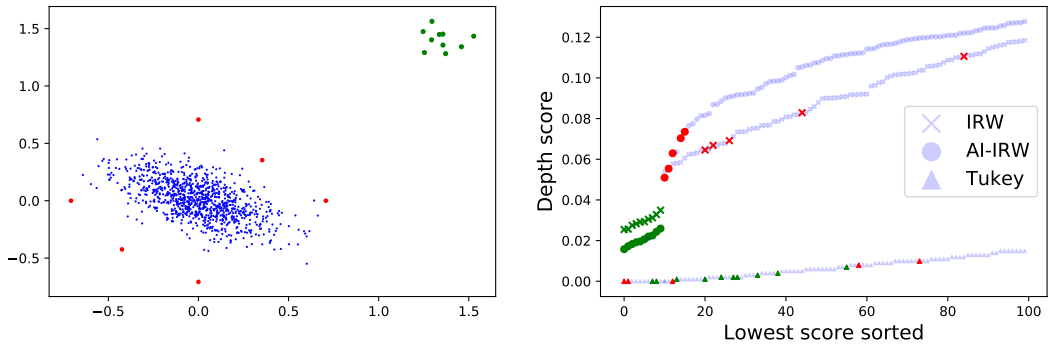


Figure 9.11 – Toy data set with outliers (left) and the AI-IRW, IRW and Tukey sorted scores (right).

9.3.4.2 Real Data Benchmarking

To illustrate the performance improvement due to introduction of affine invariance to the IRW, we conduct a comprehensive comparative study of anomaly detection on 10 widely used data sets in the literature [1]: *Mulcross*, *Shuttle*, *Thyroid*, *Wine*, *Http*, *Smtp*, *Ecoli*, *Breastw*, *Musk* and *Satimage* varying in size and dimension, see Table 9.1 for the details. In this unsupervised setting (we train all methods on unlabeled data), we use labels only to asses the performance of the methods by Area Under the Receiver Operation Characteristic curve (AUROC). We contrast the proposed approach with the non affine-invariant version, the original halfspace depth (T), halfspace mass depth (HM; Chen et al., 2015a), the AutoEncoder (AE; Zhou and Paffenroth, 2017) where the reconstruction error is used as anomaly score and one of the most used multivariate anomaly detection algorithms: Isolation Forest (IF; Liu et al., 2008). The performance of these methods being relatively insensitive to their parameters, they are set by default. Based on the previous experiment, AI-IRW, IRW, and halfspace depths are calibrated with $n_{\text{proj}} = 100 \times d$. From Table 9.1 one observes that AI-IRW uniformly (and significantly in many cases) improves on standard IRW that is rather comparable with Isolation Forest and the halfspace mass depth. Additional information on the data sets are given in Table 9.2. AI-IRW, IRW, HM and Tukey are implemented from scratch in Python using `numpy` python library. Isolation Forest implementation comes from `scikit-learn` Python library (Pedregosa et al., 2011) while the AutoEncoder implementation comes from `pyod` Python library (Zhao et al., 2019b). All the computations are done on a computer with 3.2 GHz Intel processor with 32 GB of RAM. The computation time of methods is given in Table 9.3.

[1] http://odds.cs.stonybrook.edu/

	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 9.1 – AUROCs of benchmarked anomaly detection methods.

	n	d	% of anomaly	$\hat{\gamma}$ ($\times 0.01$)	$\hat{\varsigma}$ ($\times 0.01$)
Ecoli	195	5	26	0.3	0.2
Shuttle	49097	9	7	9	5.7
Mulcross	262144	4	10	100	10^{-10}
Thyroid	3772	6	2.5	0.01	0.1
Wine	129	13	7.7	0.9	0.9
Http	567479	3	0.4	19	2.9
Smtp	95156	3	0.03	3.9	36
Breastw	683	9	35	80	20
Musk	3062	166	3.2	9.4	6
Satimage	5803	36	1.2	283	2.6

Table 9.2 – Left: Data sets considered for the performance comparison: n is the number of instances, d is the number of attributes, $\hat{\gamma}$ and $\hat{\varsigma}$ are the eigengap and the smallest eigenvalue of the SC estimator respectively.

	AI-IRW	IRW	HM	T	IF	AE
Ecoli	0.04	0.005	0.02	0.005	0.13	9
Shuttle	20	6.8	1.5	6.8	1.4	469
Mulcross	75	27	6.2	27	5.9	2383
Thyroid	1	0.21	0.2	0.2	0.18	42
Wine	0.05	0.01	0.06	0.008	0.12	8.1
Http	97	45	11	45	11	5197
Smtp	22	4.5	1	4.5	1.88	903
Breastw	0.46	0.04	0.06	0.04	0.14	17.2
Musk	20.5	5.23	2.5	5.2	0.43	103
Satimage	6.3	2.63	0.9	2.6	0.31	76

Table 9.3 – Computation time of benchmarked anomaly detection methods in seconds.

9.4 Conclusion

In this chapter, we have introduced a novel notion of statistical depth (AI-IRW), modifying the original Integrated Rank-Weighted (IRW) depth proposal in Ramsay et al. (2019). The statistical depth we have introduced has been shown not only to inherit all the compelling features of the IRW depth, its theoretical properties and its computational advantages (no optimization problem solving is required to compute it), but also to fulfill in addition the affine invariance property, crucial regarding interpretability/reliability issues. The natural idea at work consists in averaging univariate Tukey halfspace depths computed from random projections of the data onto (nearly) uncorrelated lines, defined by the (empirical) covariance structure of the data, rather than projections onto lines fully generated at random. Though the AI-IRW sample version exhibits a complex probabilistic structure, an estimator of the precision matrix being involved in its definition, a nonasymptotic analysis has been carried out here, revealing its good concentration properties around the true AI-IRW depth. The merits of the AI-IRW depth have been illustrated by encouraging numerical experiments, for anomaly detection purpose in particular, offering the perspective of a widespread use for various statistical learning tasks.

9.5 Proofs

9.5.1 Proof of Proposition 9.4

9.5.1.1 Affine-Invariance

Let $A \in \mathbb{R}^{d\times d}$ be a non-singular matrix and $b \in \mathbb{R}^d$. Let Σ_X and Σ_{AX} the covariance matrix of X and AX respectively. Defines the Cholesky decomposition as $\Sigma_X = \Lambda_X \Lambda_X^\top$ and $\Sigma_{AX} = A\Lambda_X\Lambda_X^\top A^\top = \Lambda_{AX}\Lambda_{AX}^\top$. It holds:

$$
\begin{aligned}
D_{\mathrm{AI-IRW}}(Ax+b, AX+b) &= \frac{1}{V_d}\int_{\mathbb{S}^{d-1}} D_{\mathrm{H},1}(\langle \frac{\Lambda_{AX+b}^{-\top}u}{||\Lambda_{AX+b}^{-\top}u||}, Ax+b\rangle,\ \langle \frac{\Lambda_{AX+b}^{-\top}u}{||\Lambda_{AX+b}^{-\top}u||}, AX+b\rangle)\ \mathrm{d}u \\
&= \frac{1}{V_d}\int_{\mathbb{S}^{d-1}} D_{\mathrm{H},1}(\langle \Lambda_{AX+b}^{-\top}u, Ax+b\rangle,\ \langle \Lambda_{AX+b}^{-\top}u, AX+b\rangle)\ \mathrm{d}u \\
&= \frac{1}{V_d}\int_{\mathbb{S}^{d-1}} D_{\mathrm{H},1}(\langle \Lambda_{AX}^{-\top}u, Ax\rangle,\ \langle \Lambda_{AX}^{-\top}u, AX\rangle)\ \mathrm{d}u \\
&= \frac{1}{V_d}\int_{\mathbb{S}^{d-1}} D_{\mathrm{H},1}(\langle u, \Lambda_X^{-1}x\rangle, \langle u, \Lambda_X^{-1}X\rangle)\ \mathrm{d}u \\
&= \frac{1}{V_d}\int_{\mathbb{S}^{d-1}} D_{\mathrm{H},1}(\langle \frac{\Lambda_X^{-\top}u}{||\Lambda_X^{-\top}u||}, x\rangle, \langle \frac{\Lambda_X^{-\top}u}{||\Lambda_X^{-\top}u||}, X\rangle)\ \mathrm{d}u \\
&= D_{\mathrm{AI-IRW}}(x, P).
\end{aligned}
$$

The same reasoning applies if the square matrix is given by the SVD decomposition.

9.5.1.2 Maximality at the center

Assume that P is halfspace symmetric about a unique β, i.e. $\mathbb{P}\left(X \in H_\beta\right) \geq \frac{1}{2}$ for every closed halfspace H_β such that $\beta \in \partial H$ with ∂H the boundary of H. Thus, it is easy to see that $D_{\mathrm{AI-IRW}}(\beta, P) \geq \frac{1}{2}$. The uniqueness of β and the fact that $D_{\mathrm{AI-IRW}}$ is lower

than 1/2 for any element in $\mathbb{R}^d$ by definition imply that

$$\beta = \operatorname*{argsup}_{x \in \mathbb{R}^d} D_{\text{AI-IRW}}(x, P).$$

9.5.1.3 Vanishing at Infinity

The proof is a particular case of the proof of Theorem 1 in Cuevas and Fraiman (2009). We detail it for the sake of clarity. Let U be a random variable following ω_{d-1}, the uniform measure on the unit sphere $\mathbb{S}^{d-1}$. Defines $V = \Sigma^{-\top/2}U/||\Sigma^{-\top/2}U||$ and ν_{d-1} its probability distribution. Let $\theta > 0$ and $x \in \mathbb{R}^d$, then $r(\theta) := \nu_{d-1}\{v : \frac{|\langle v,x\rangle|}{||x||} \leq \theta\}$ goes to zero when $\theta \to 0$. For any $x \in \mathbb{R}^d \backslash \{0\}$, we have

$$\begin{aligned}
D_{\text{AI-IRW}}(x, P) &= \int_{\mathbb{R}^d} \min\left\{F_v(v^\top x), 1 - F_v(v^\top x)\right\} \, \mathrm{d}\nu_{d-1}(v) \\
&\leq \int_{\mathbb{R}^d} \mathbb{I}\left\{v : \frac{|\langle v, x\rangle|}{||x||} \leq \theta\right\} \, \mathrm{d}\nu_{d-1}(v) \\
&+ \int_{\mathbb{R}^d} F_v(v^\top x) \, \mathbb{I}\left\{v : \frac{|\langle v, x\rangle|}{||x||} > \theta, \ \langle v, x\rangle \leq 0\right\} \, \mathrm{d}\nu_{d-1}(v) \\
&+ \int_{\mathbb{R}^d} (1 - F_v(v^\top x)) \, \mathbb{I}\left\{v : \frac{|\langle v, x\rangle|}{||x||} > \theta, \ \langle v, x\rangle > 0\right\} \, \mathrm{d}\nu_{d-1}(v) \\
&\leq r(\theta) + \int_{\mathbb{R}^d} F_v(-\theta||x||) \, \mathbb{I}\left\{v : \frac{|\langle v, x\rangle|}{||x||} > \theta, \ \langle v, x\rangle \leq 0\right\} \, \mathrm{d}\nu_{d-1}(v) \\
&+ \int_{\mathbb{R}^d} (1 - F_v(\theta||x||)) \, \mathbb{I}\left\{v : \frac{|\langle v, x\rangle|}{||x||} > \theta, \ \langle v, x\rangle > 0\right\} \, \mathrm{d}\nu_{d-1}(v).
\end{aligned}$$

Now, when $||x|| \to \infty$, the dominated convergence theorem ensures that

$$\lim_{||x|| \to \infty} \sup D_{\text{AI-IRW}}(x, P) \leq r(\theta) \underset{\theta \to 0}{\to} 0.$$

9.5.1.4 Decreasing Along Rays

The proof is a slight modification of the proof of Assertion (iii) of Theorem 2 in Ramsay et al. (2019). Details are left to the reader.

9.5.1.5 Continuity

For any $P \in \mathcal{P}(\mathbb{R}^d)$, the continuity of the inner product and the cdf ensure continuity of $D_{\text{H}}(v^\top x, v^\top X)$ for any $v \in \mathbb{S}^{d-1}$. Therefore, the continuity of $x \mapsto D_{\text{AI-IRW}}(x, P)$ follows from dominated convergence.

9.5.2 Proof of Theorem 9.13

9.5.2.1 Assertion (i)

Introducing terms and using triangle inequality, it holds:

$$\sup_{x\in\mathbb{R}^d}\left|\widehat{D}_{\text{AI-IRW}}(x)-D_{\text{AI-IRW}}(x,P)\right|\leq \underbrace{\sup_{x\in\mathbb{R}^d}\left|F_{\widehat{V},n}(\widehat{V}^\top x)-F_{\widehat{V}}(\widehat{V}^\top x)\right|}_{(1)}+\underbrace{\sup_{x\in\mathbb{R}^d}\left|F_{\widehat{V}}(\widehat{V}^\top x)-F_V(V^\top x)\right|}_{(2)}.$$

Now, the first term (1) can be controlled using the bound for the deviations of halfspace depth deferred in Lemma 9.11. Thus, for any $t>0$ it holds:

$$\begin{aligned}\mathbb{P}\left(\sup_{x\in\mathbb{R}^d}\left|F_{\widehat{V},n}(\widehat{V}^\top x)-F_{\widehat{V}}(\widehat{V}^\top x)\right|\ >\ t/2\right)&\leq \mathbb{P}\left(\sup_{\substack{y\in\mathbb{R}^d\\ u\in\mathbb{S}^{d-1}}}\left|F_{u,n}(u^\top y)-F_u(u^\top y)\right|\ >\ t/2\right)\\ &\leq \frac{6(2n)^{d+1}}{(d+1)!}\exp(-nt^2/32).\end{aligned}\tag{9.6}$$

The second term (2) relies on the influence of the deviations of the sample covariance matrix. First remark that:

$$\begin{aligned}\sup_{x\in\mathbb{R}^d}\left|F_{\widehat{V}}(\widehat{V}^\top x)-F_V(V^\top x)\right|\leq \sup_{\substack{x\in\mathbb{R}^d\\ u\in\mathbb{S}^{d-1}}}\Bigg|&\mathbb{P}\left(\langle\frac{\widehat{\Sigma}^{-\top/2}u}{||\widehat{\Sigma}^{-\top/2}u||},X-x\rangle\leq 0\ \middle|\ \mathcal{S}_n\right)\\ &-\mathbb{P}\left(\langle\frac{\Sigma^{-\top/2}u}{||\Sigma^{-\top/2}u||},X-x\rangle\leq 0\right)\Bigg|.\end{aligned}$$

Now, since X is radially Lipschitz continuous, we have:

$$\left|\mathbb{P}\left(\langle\frac{\widehat{\Sigma}^{-\top/2}u}{||\widehat{\Sigma}^{-\top/2}u||},X-x\rangle\leq 0\ \middle|\mathcal{S}_n\right)-\mathbb{P}\left(\langle\frac{\Sigma^{-\top/2}u}{||\Sigma^{-\top/2}u||},X-x\rangle\leq 0\right)\right|\leq L_R\left|\left|\frac{\widehat{\Sigma}^{-\top/2}u}{||\widehat{\Sigma}^{-\top/2}u||}-\frac{\Sigma^{-\top/2}u}{||\Sigma^{-\top/2}u||}\right|\right|.$$

Introducing terms and using triangle inequality leads to:

$$\begin{aligned}\left|\left|\frac{\widehat{\Sigma}^{-\top/2}u}{||\widehat{\Sigma}^{-\top/2}u||}-\frac{\Sigma^{-\top/2}u}{||\Sigma^{-\top/2}u||}\right|\right|&\leq\frac{||\widehat{\Sigma}^{-1/2}-\Sigma^{-1/2}||_{\text{op}}}{||\Sigma^{-1/2}u||}+||\widehat{\Sigma}^{-1/2}u||\left(\frac{1}{||\widehat{\Sigma}^{-1/2}u||}-\frac{1}{||\Sigma^{-1/2}u||}\right)\\ &\leq\frac{2||\widehat{\Sigma}^{-1/2}-\Sigma^{-1/2}||_{\text{op}}}{||\Sigma^{-1/2}u||},\end{aligned}$$

yielding

$$\sup_{x\in\mathbb{R}^d}\left|F_{\widehat{V}}(\widehat{V}^\top x)-F_V(V^\top x)\right|\leq\frac{2L_R}{||\Sigma^{-1/2}||_{\text{op}}}\,||\widehat{\Sigma}^{-1/2}-\Sigma^{-1/2}||_{\text{op}}.\tag{9.7}$$

Assume that $ODO^\top$ and $\widehat{O}\widehat{D}\widehat{O}^\top$ are the eigenvalues decomposition of Σ and $\widehat{\Sigma}$ in orthonormal bases. Thus, thanks to Theorem 4.1 in Wedin (1973), we have:

$$||\widehat{\Sigma}^{-1/2} - \Sigma^{-1/2}||_{\text{op}} \leq ||\widehat{D}^{-1/2} - D^{-1/2}||_{\text{op}} + ||D^{-1/2}||_{\text{op}}||\widehat{O} - O||_{\text{op}}$$
$$\leq ||\Sigma^{-1/2}||_{\text{op}} \left(||\widehat{D}^{1/2} - D^{1/2}||_{\text{op}} \, ||\widehat{D}^{-1/2}||_{\text{op}} + ||\widehat{O} - O||_{\text{op}} \right).$$

Now, since $\min_{i \leq d} \sqrt{\widehat{\sigma}_k} \geq \sqrt{\varsigma} - \max_{k \leq d} |\sqrt{\widehat{\sigma}_k} - \sqrt{\sigma}_k|$ and $\max_{k \leq d} |\sqrt{\widehat{\sigma}_k} - \sqrt{\sigma}_k| \leq \frac{1}{\sqrt{\varsigma}} \max_{1 \leq k \leq d} |\widehat{\sigma}_k - \sigma_k|$, using Weyl's inequality leads to

$$\sup_{x \in \mathbb{R}^d} \left| F_{\widehat{V}}(\widehat{V}^\top x) - F_V(V^\top x) \right| \leq 2L_R \left(\frac{||\widehat{\Sigma} - \Sigma||_{\text{op}}}{\varsigma - ||\widehat{\Sigma} - \Sigma||_{\text{op}}} + ||\widehat{O} - O||_{\text{op}} \right).$$

Let $\mathcal{A}_\xi = \left\{ ||\widehat{\Sigma} - \Sigma||_{\text{op}} < \varsigma - \xi \right\}$ for any $\xi \in [0, \varsigma)$. Using union bound and combining (9.7) with the previous equation, for any $t > 0$ and $\xi \in (0, \varsigma)$, it holds:

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}^d} \left| F_{\widehat{V}}(\widehat{V}^\top x) - F_V(V^\top x) \right| > t/2 \right) \leq \mathbb{P}\left(\frac{2L_R}{\xi} ||\widehat{\Sigma} - \Sigma||_{\text{op}} > t/4 \right) + \mathbb{P}\left(\mathcal{A}_\xi^c \right)$$
$$+ \mathbb{P}\left(2L_R \, ||\widehat{O} - O||_{\text{op}} > t/4 \right),$$

where $\mathcal{A}_\xi^c$ stands for the complementary event of $\mathcal{A}_\xi$. Applying Lemma 9.12 gives:

$$\mathbb{P}\left(||\widehat{\Sigma} - \Sigma||_{\text{op}} > \frac{\xi t}{8L_R} \right) \leq 2 \times 9^d \exp\left\{ -\frac{n}{2} \min\left\{ \frac{(\xi t)^2}{(256 L_R \varrho^2)^2}, \frac{\xi t}{256 L_R \varrho^2} \right\} \right\}, \tag{9.8}$$

and

$$\mathbb{P}\left(\mathcal{A}_\xi^c \right) \leq 2 \times 9^d \exp\left\{ -\frac{n}{2} \min\left\{ \frac{(\varsigma - \xi)^2}{(32\varrho^2)^2}, \frac{\varsigma - \xi}{32\varrho^2} \right\} \right\}. \tag{9.9}$$

Furthermore, it is easy to see that $||\widehat{O} - O||_{\text{op}} \leq \sqrt{d} \max_{k \leq d} ||\widehat{O}_k - O_k||$ where O_k is the k-th column of the matrix O. Let γ be the minimum eigengap, following a variant of the Davis-Kahan theorem (Davis and Kahan, 1970) (see Corollary 1 in Yu et al., 2014), it holds:

$$||\widehat{O} - O||_{\text{op}} \leq \frac{2\sqrt{2d}||\widehat{\Sigma} - \Sigma||_{\text{op}}}{\gamma}.$$

Using Lemma 9.12 again leads to:

$$\mathbb{P}\left(\frac{4L_R\sqrt{2d}||\widehat{\Sigma} - \Sigma||_{\text{op}}}{\gamma} > t/4 \right) \leq 2 \times 9^d \exp\left\{ -\frac{n}{2} \min\left\{ \frac{(\gamma t)^2}{(512 L_R \sqrt{2d} \varrho^2)^2}, \frac{\gamma t}{512 L_R \sqrt{2d} \varrho^2} \right\} \right\}. \tag{9.10}$$

Combining (9.8), (9.9) and (9.10), it holds:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}^d}\left|F_{\widehat{V}}(\widehat{V}^\top x)-F_V(V^\top x)\right| \;>\; t/2\right)\leq 6\times 9^d\exp\left(-\frac{n}{2}\min\left\{\left(\kappa t\right)^2,\kappa t\right\}\right),$$

for any $t\leq(\varsigma-\xi)/(32\varrho^2\kappa)$ where $\kappa=\dfrac{1}{256L_R\varrho^2}\left(\xi\wedge\dfrac{\gamma}{2\sqrt{2d}}\right)$. Finally, for any $t\leq(\varsigma-\xi)/(32\varrho^2\kappa)$ it holds:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}^d}\left|\widehat{D}_{\text{AI-IRW}}(x)-D_{\text{AI-IRW}}(x,P)\right| \;>\; t\right)\leq 6.9^d\exp\left(-\frac{n}{2}\min\left\{\left(\kappa t\right)^2,\kappa t\right\}\right)+\frac{6(2n)^{d+1}}{(d+1)!}\exp(-nt^2/32). \tag{9.11}$$

Bounding each term in the right side by $\delta/2$ and reverting the equation lead to the desired result.

9.5.2.2 Assertion (ii)

Let $\mathcal{B}(0,r)$ a centered ball of $\mathbb{R}^d$ with radius $r>0$ and assume that X satisfies assumption 2 for any $x\in\mathcal{B}(0,r)$. Introducing terms and using triangle inequality, it holds:

$$\sup_{x\in\mathcal{B}(0,r)}\left|\widetilde{D}^{\text{MC}}_{\text{AI-IRW}}(x)-D_{\text{AI-IRW}}(x,P)\right|\leq\underbrace{\sup_{x\in\mathbb{R}^d}\left|\widehat{D}_{\text{AI-IRW}}(x)-D_{\text{AI-IRW}}(x,P)\right|}_{(1)}+\underbrace{\sup_{x\in\mathcal{B}(0,r)}\left|D^{\text{MC}}_{\text{AI-IRW}}(x,P)-D_{\text{AI-IRW}}(x,P)\right|}_{(2)}.$$

The first term (1) can be bounded using assertion (i) while controlling the approximation term (2) relies on classical chaining arguments. As the function $z\mapsto\min(z,1-z)$ is 1-Lipschitz for any $z\in(0,1)$ and by triangle inequality, for any y in $\mathcal{B}(0,r)$, we have:

$$\left|D^{\text{MC}}_{\text{AI-IRW}}(y,P)-D_{\text{AI-IRW}}(y,P)\right|\leq\frac{1}{n_{\text{proj}}}\sum_{j=1}^{n_{\text{proj}}}\left|\mathbb{P}\left\{\langle V_j,y\rangle\;\middle|\;V_j\right\}-\mathbb{P}\left\{\langle V,y\rangle\right\}\right|.$$

Since it is an average of bounded and i.i.d random variables, combining Hoeffding inequality and union bound, for any $t>0$ and any y in $\mathcal{B}(0,r)$, it holds:

$$\mathbb{P}\left(\left|D^{\text{MC}}_{\text{AI-IRW}}(y,P)-D_{\text{AI-IRW}}(y,P)\right| \;>\; t/2\right)\leq 2\exp\left(-n_{\text{proj}}t^2/2\right). \tag{9.12}$$

As X is uniformly continuous Lipschitz in projection for any $u\in\mathbb{S}^{d-1}$, observe that $\forall(x,y)\in\mathcal{B}(0,r)^2$ it holds:

$$\begin{aligned}\left|D^{\text{MC}}_{\text{AI-IRW}}(x,P) - D_{\text{AI-IRW}}(x,P)\right| &\leq \left|D^{\text{MC}}_{\text{AI-IRW}}(x,P) - D^{\text{MC}}_{\text{AI-IRW}}(y,P)\right| + \left|D^{\text{MC}}_{\text{AI-IRW}}(y,P) - D_{\text{AI-IRW}}(y,P)\right| \\ &\quad + \left|D_{\text{AI-IRW}}(x,P) - D_{\text{AI-IRW}}(y,P)\right| \\ &\leq 2L_p\ ||x-y|| + \left|D^{\text{MC}}_{\text{AI-IRW}}(y,P) - D_{\text{AI-IRW}}(y,P)\right|. \end{aligned} \tag{9.13}$$

Now let $\zeta > 0$ and $y_1, \ldots, y_{N(\zeta,\mathcal{B}(0,r),||.||_2)}$ be a ζ-coverage of $\mathcal{B}(0,r)$ with respect to $||.||_2$. We have:

$$\log\left(N(\zeta, \mathcal{B}(0,r), ||.||_2)\right) \leq d \log\left(3r/\zeta\right). \tag{9.14}$$

Set $N = N\left(\zeta, \mathcal{B}(0,r), ||.||_2)\right)$ for simplicity. There exists $\ell \leq N$ such that $||x - y_\ell||_2 \leq \zeta$. Thus, (9.13) leads to:

$$\left|D^{\text{MC}}_{\text{AI-IRW}}(x,P) - D_{\text{AI-IRW}}(x,P)\right| \leq 2L_p\ \zeta + \left|D^{\text{MC}}_{\text{AI-IRW}}(y_\ell,P) - D_{\text{AI-IRW}}(y_\ell,P)\right|.$$

Applying (9.12) to every y_ℓ and the union bound, for any $t > 0$, we get:

$$\mathbb{P}\left(\sup_{\ell \leq N}\left|D^{\text{MC}}_{\text{AI-IRW}}(y_\ell,P) - D_{\text{AI-IRW}}(y_\ell,P)\right| \ >\ t/2\right) \leq 2N \exp\left(-n_{\text{proj}} t^2/2\right),$$

yielding

$$\mathbb{P}\left(\sup_{x \in \mathcal{B}(0,r)}\left|D^{\text{MC}}_{\text{AI-IRW}}(x,P) - D_{\text{AI-IRW}}(x,P)\right| > t/2\right) \leq 2N \exp\left(-2n_{\text{proj}}\left(t/2 - 2L_p\zeta\right)^2\right).$$

Using (9.11), the union bound and (9.14), we obtain:

$$\begin{aligned}&\mathbb{P}\left(\sup_{x \in \mathcal{B}(0,r)}\left|\widetilde{D}^{\text{MC}}_{\text{AI-IRW}}(x) - D_{\text{AI-IRW}}(x,P)\right| \ >\ t\right) \\ &\leq \mathbb{P}\left(\sup_{x \in \mathcal{B}(0,r)}\left|\widehat{D}_{\text{AI-IRW}}(x) - D_{\text{AI-IRW}}(x,P)\right| \ >\ t/2\right) \\ &+ \mathbb{P}\left(\sup_{x \in \mathcal{B}(0,r)}\left|D^{\text{MC}}_{\text{AI-IRW}}(x,P) - D_{\text{AI-IRW}}(x,P)\right| \ >\ t/2\right) \\ &\leq 6.9^d \exp\left(-\frac{n}{2}\min\left\{\left(\kappa t/2\right)^2, \kappa t/2\right\}\right) + \frac{6(2n)^{d+1}}{(d+1)!}\exp(-nt^2/128) \\ &+ 2\left(\frac{3r}{\zeta}\right)^d \exp\left(-2n_{\text{proj}}\left(t/2 - 2L_p\zeta\right)^2\right).\end{aligned}$$

Choosing $\zeta \sim n_{\text{proj}}^{-1}$, bounding each term on the right-hand side by $\delta/3$ and reverting the previous equation lead to the desired result.

9.5.3 Proof of Corollary 9.15

First notice that

$$\sup_{x\in\mathcal{B}(0,r)}\left|\widetilde{D}^{\text{MC}}_{\text{IRW}}(x,P_n)-D_{\text{IRW}}(x,P)\right| \leq \underbrace{\sup_{x\in\mathbb{R}^d}\left|\widehat{D}_{\text{IRW}}(x)-D_{\text{IRW}}(x,P)\right|}_{(1)} + \underbrace{\sup_{x\in\mathcal{B}(0,r)}\left|D^{\text{MC}}_{\text{IRW}}(x,P)-D_{\text{IRW}}(x,P)\right|}_{(2)}.$$

Now, the first term (1) can be controlled using the bound for the deviations of Halfspace Depth described in Lemma 9.11. Thus, for any $t > 0$, it holds:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}^d}\left|\widehat{D}_{\text{IRW}}(x) - D_{\text{IRW}}(x,P)\right| \;>\; t/2\right) \leq \frac{6(2n)^{d+1}}{(d+1)!}\exp(-nt^2/32). \tag{9.15}$$

The second term can be bounded following the same reasoning than for the Monte-Carlo approximated term of AI-IRW described in Section 9.5.2. Thus, with the same notations, for any $t > 0$, we have

$$\mathbb{P}\left(\sup_{x\in\mathcal{B}(0,r)}\left|D^{\text{MC}}_{\text{IRW}}(x,P) - D_{\text{IRW}}(x,P)\right| \;>\; t/2\right) \leq 2N\exp\left(-2n_{\text{proj}}\left(t/2 - 2L_p\zeta\right)^2\right). \tag{9.16}$$

Using (9.15) and (9.16), one gets:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathcal{B}(0,r)}\left|\widetilde{D}^{\text{MC}}_{\text{IRW}}(x,P_n) - D_{\text{IRW}}(x,P)\right| \;>\; t\right)\\
&\leq \mathbb{P}\left(\sup_{x\in\mathcal{B}(0,r)}\left|\widehat{D}_{\text{IRW}}(x) - D_{\text{IRW}}(x,P)\right| > t/2\right) + \mathbb{P}\left(\sup_{x\in\mathcal{B}(0,r)}\left|D^{\text{MC}}_{\text{IRW}}(x,P) - D_{\text{IRW}}(x,P)\right| > t/2\right)\\
&\leq \frac{6(2n)^{d+1}}{(d+1)!}\exp(-nt^2/32) + 2\left(\frac{3r}{\zeta}\right)^d \exp\left(-2n_{\text{proj}}\left(t/2 - 2L_p\zeta\right)^2\right).
\end{aligned}$$

Choosing $\zeta \sim n_{\text{proj}}^{-1}$, bounding each term on the right-hand side by $\delta/2$ and reverting the previous equation lead to the desired result.

10

A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions

Contents

This chapter presents a new discrepancy measure between probability distributions, well-defined for non-overlapping supports, that leverages the interesting features of data depths. This measure relies on the average of the Hausdorff distance between the depth-trimmed regions w.r.t. each distribution and defines a pseudo-metric in general. Its good behavior w.r.t. major transformation groups as well as its ability to factor out translations are depicted. Its robustness is investigated through the concept of finite sample breakdown point. Moreover, we propose an efficient approximation method with linear time complexity w.r.t. the size of the data set and its dimension. The quality of this approximation as well as the performance of the proposed approach are illustrated in numerical experiments.

This chapter is organized as follows. In Section 10.1, the pseudo-metric is introduced and a theoretical analysis of its properties is investigated. In Section 10.2, an efficient approximation of the depth-trimmed regions based pseudo-metric, relying on a nice feature of the Hausdorff distance when computed between convex bodies, is proposed for convex depth functions. In Section 10.3, the behavior of this algorithm w.r.t. its parameters is studied through numerical experiments which also highlight the by-design robustness of this pseudo-metric. In addition, an application to robust clustering of images is described. In Section 10.4, specific attention is devoted to automatic evaluation of natural language generation (NLG) showing benefits of this approach when bench-

marked with state-of-the-art probability metrics. Concluding remarks are collected in Section 10.5. Eventually, a technical proof is deferred to Section 10.6. This chapter covers the contribution of:

- **G. Staerman**, P. Mozharovskyi, P. Colombo, S. Clémençon, F. d'Alché-Buc. A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions. *arXiv preprint arXiv:2103.12711*, 2021.

10.1 A Pseudo-Metric based on Depth-Trimmed Regions

In this section, we introduce the depth-based pseudo-metric and study its properties. To fairly compare depth regions from different probability distributions, we consider depth regions possessing the same mass of probability (see e.g. Paindaveine and bever, 2013). We denote by $\alpha : (\beta, P) \in [0,1] \times \mathcal{P}(\mathbb{R}^d) \longmapsto \alpha(\beta, P) \in [0,1]$ the highest level such that the probability mass of the depth-trimmed region at this level is at least β. Precisely, for any pair $(\beta, P) \in [0,1) \times \mathcal{P}(\mathbb{R}^d)$:

$$\alpha(\beta, P) = \sup\{\gamma \in [0,1] \, : \, P(D^\gamma(P)) > \beta\}, \tag{10.1}$$

where $D^\gamma(P)$ is the depth region at level γ defined as $\{x \in \mathbb{R}^d \, : \, D(x,P) \geq \gamma\}$ for any data depth function $D(\cdot,\cdot) : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to [0,1]$. In the remainder of this chapter, when the quantity $\alpha(\beta, P)$ will be associated with depth regions of P, the second argument of the function $\alpha(\cdot,\cdot)$ will be omitted, for notation simplicity. It is worth mentioning that $D^{\alpha(\beta')}(P) \subseteq D^{\alpha(\beta)}(P)$ for any $\beta > \beta'$, since $\beta \mapsto \alpha(\beta, P)$ is a monotone decreasing function. Thus, $D^{\alpha(\beta)}(P)$ is the smallest depth region with probability larger than or equal to β and can be defined in an equivalent way as:

$$D^{\alpha(\beta)}(P) = \bigcap_{\gamma \in \Gamma_P(\beta)} D^\gamma(P),$$

where $\Gamma_P(\beta) = \{\zeta \in [0,1] \, : \, P(D^\zeta(P)) > \beta\}$. The strict inequalities in (10.1) and in the definition of $\Gamma_P(\beta)$ eliminate cases where the supremum does not exist. Indeed, when $\beta = 0$, the depth region is then an infinitesimal set with probability strictly higher than zero. Although the supremum exists (without necessarily being unique) such as in the case of the halfspace depth (Rousseeuw and Ruts, 1999) and the projection depth (Zuo, 2003) under mild assumptions, no universal results have been derived for data depths. The set $\{D^{\alpha(\beta)}(P) \, : \, \beta \in [0, 1-\varepsilon], \; \varepsilon \in (0,1]\}$ where each region probability mass is equal to β then defines quantile regions of P.

Let P, Q be two absolutely continuous probability measures (w.r.t. the Lebesgue measure) on $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ respectively. Denote by $d_{\mathcal{H}}(S_1, S_2)$ the Hausdorff distance between the sets S_1 and S_2. The pseudo-metric between probability distributions P and Q based on the depth-trimmed regions is defined as follows:

Definition 10.1. *Let $\varepsilon \in (0,1]$ and $p \in (0, \infty)$, for all pairs (P, Q) in $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, the depth-trimmed regions ($DR_{p,\varepsilon}$) discrepancy measure between P and Q is defined as:*

$$DR^p_{p,\varepsilon}(P, Q) = \int_0^{1-\varepsilon} d_{\mathcal{H}} \left(D^{\alpha(\beta)}(P), D^{\alpha(\beta)}(Q) \right)^p \, \mathrm{d}\beta. \tag{10.2}$$

Our discrepancy measure relies on the Hausdorff distance averaged over depth-trimmed regions with the same probability mass w.r.t. each distribution. Properties ($\mathbf{D}_4$-$\mathbf{D}_5$) of data depths ensure that for every $0 \leq \beta < 1$, $D^{\alpha(\beta)}(P)$ is a non-empty compact subset of $\mathbb{R}^d$ leading to a well-defined discrepancy measure. Observe that the parameter ε can be considered as a robustness tuning parameter. Indeed, choosing higher ε amounts to ignoring the larger upper-level sets of data depth function, i.e. the tails of the distributions.

Remark 10.2. *The use of data depths together with the ε-trimming provide robustness in (10.2). Several data depths exhibit attractive robustness properties. Indeed, the asymptotic breakdown point of the halfspace and the integrated rank-weighted medians have been shown to be higher than $1/(d+1)$ while the projection median is known to have a breakdown point equal to $1/2$ (Donoho and Gasko, 1992; Ramsay et al., 2019). See Section 2.4.3 in Chapter 2 for further details.*

10.1.1 Connection with Wasserstein Distance

When $d = 1$, the L_p-Wasserstein distance enjoys an explicit expression involving quantile and distribution functions. Let $X \sim P_1$, $Y \sim Q_1$ be two random variables where $P_1, Q_1 \in \mathcal{P}(\mathbb{R})$ are univariate probability distributions. Denoting by $F^{-1}_{P_1}$ the quantile function of X, the L_p-Wasserstein distance can be written as

$$\mathcal{W}_p(P_1, Q_1) = \left(\int_0^1 |F^{-1}_{P_1}(t) - F^{-1}_{Q_1}(t)|^p \, \mathrm{d}t \right)^{1/p}. \tag{10.3}$$

Since data depth and its central regions are extensions of cdf and quantiles to dimension $d > 1$, $DR_{p,\varepsilon}$ is then a possible (center-outward) generalization of an ε-trimmed version of (10.3) to higher dimensions. When $DR_{p,\varepsilon}$ is associated with the halfspace depth, a simple calculus (see below) leads to:

$$DR^p_{p,\varepsilon}(P_1, Q_1) = 2 \int_{\varepsilon/2}^{1/2} \max \left\{ \left| F^{-1}_{P_1}(t) - F^{-1}_{Q_1}(t) \right|^p, \left| F^{-1}_{P_1}(1-t) - F^{-1}_{Q_1}(1-t) \right|^p \right\} \mathrm{d}t.$$

Thus, $\mathcal{W}^p_p(P_1, Q_1) \leq \lim_{\varepsilon \to 0} DR^p_{p,\varepsilon}(P_1, Q_1)$ in general where the equality holds for symmetric distributions.

Proof. In dimension one, the halfspace depth of any $x \in \mathbb{R}$ w.r.t. P_1 and Q_1 boils down to:

$$D_{\mathrm{H},1}(x, P_1) = \min \Big\{ F_{P_1}(x), 1 - F_{P_1}(x) \Big\}, \quad D_{\mathrm{H},1}(x, Q_1) = \min \Big\{ F_{Q_1}(x), 1 - F_{Q_1}(x) \Big\},$$

and for any $\gamma \in [0, 1]$, its upper-level sets to intervals:

$$D^\gamma_{\mathrm{H},1}(P_1) = [F^{-1}_{P_1}(\gamma),\ F^{-1}_{P_1}(1-\gamma)] \quad \text{and} \quad D^\gamma_{\mathrm{H},1}\left(Q_1\right) = [F^{-1}_{Q_1}(\gamma),\ F^{-1}_{Q_1}(1-\gamma)]. \tag{10.4}$$

Now, the quantile function $\alpha(\beta, .)$ can be explicitly derived as function of $\beta \in [0, 1]$:

$$\begin{aligned}\alpha(\beta, P_1) &= \sup\left\{\gamma \in [0,1] \,:\, P_1\Big([F_{P_1}^{-1}(\gamma),\ F_{P_1}^{-1}(1-\gamma)]\Big) \geq \beta\right\} \\ &= \sup\left\{\gamma \in [0,1] \,:\, 1 - 2\gamma \geq \beta\right\} \\ &= \frac{1-\beta}{2}.\end{aligned}$$

Following the same reasoning, it holds $\alpha(\beta, Q_1) = \frac{1-\beta}{2}$. Further, by change of variables

$$\int_0^{1-\varepsilon} d_{\mathcal{H}}\left(D_{\mathrm{H},1}^{(1-\beta)/2}(P_1), D_{\mathrm{H},1}^{(1-\beta)/2}(Q_1)\right)^p \,\mathrm{d}\beta = 2\int_{\varepsilon/2}^{1/2} d_{\mathcal{H}}\left(D_{\mathrm{H},1}^{t}(P_1), D_{\mathrm{H},1}^{t}(Q_1)\right)^p \,\mathrm{d}t.$$

Recalling that the Hausdorff distance between two bounded subsets S_1, S_2 of $\mathbb{R}$ is defined as:

$$d_{\mathcal{H}}(S_1, S_2) = \max\left\{\sup_{x\in S_1}\inf_{y\in S_2}|x-y|,\ \sup_{y\in S_2}\inf_{x\in S_1}|x-y|\right\},$$

it leads to the desired result.

■

10.1.2 Metric Properties

We now investigate to which extent the proposed discrepancy measure satisfies the metric axioms. As a first go, we show that $DR_{p,\varepsilon}$ fulfills most conditions. However it does not define a distance in general.

Proposition 10.3 (Metric properties)**.** *For any convex data depth (see Section 2.1), $DR_{p,\varepsilon}$ is positive, symmetric and satisfies triangular inequality but the entailment $DR_{p,\varepsilon}(P,Q) = 0 \Longrightarrow P \overset{\mathcal{L}}{=} Q$ does not hold in general.*

Proof. For any $0 \leq \beta \leq 1-\varepsilon$ with $\varepsilon \in (0,1]$, and any $P \in \mathcal{P}(\mathcal{X})$, $Q \in \mathcal{P}(\mathcal{Y})$, $D^{\alpha(\beta)}(P), D^{\alpha(\beta)}(Q)$ are non-empty compact subsets of $\mathbb{R}^d$ due to the properties ($\mathbf{D}_4$-$\mathbf{D}_5$). The Hausdorff distance $d_{\mathcal{H}}$ is known to be a distance on the space of non-empty compact sets which implies that $DR_{p,\varepsilon}$ satisfies positivity, symmetry and the triangle inequality (thanks to Minkowski inequality). If $P \overset{\mathcal{L}}{=} Q$ then $D^{\alpha(\beta)}(P) = D^{\alpha(\beta)}(Q), \quad \forall\ \beta \in [0, 1-\varepsilon]$ which leads to $DR_{p,\varepsilon}(P,Q) = 0$. The reverse is not true. $DR_{p,\varepsilon}(P,Q) = 0$ implies $D^{\alpha(\beta)}(P) = D^{\alpha(\beta)}(Q) \quad \forall\ \beta \in [0, 1-\varepsilon]$ that not leads to $P \overset{\mathcal{L}}{=} Q$. Indeed, convex depth regions do not characterize probability distributions in general (see Nagy (2021) for the halfspace depth) that would be the first step in order to prove the previous entailment. ■

Thus, $DR_{p,\varepsilon}$ defines a pseudo-metric rather than a distance. Being based on distance, the proposed discrepancy measure preserves isometry invariance as stated in the following proposition.

Proposition 10.4 (Isometry invariance)**.** *Let $A \in \mathbb{R}^{d\times d}$ be a non-singular matrix and $b \in \mathbb{R}^d$. Define the isometry mapping $g : x \in \mathbb{R}^d \mapsto Ax + b$ with $AA^\top = \mathrm{I}_d$, then it holds:*

$$DR_{p,\varepsilon}(g_\sharp P, g_\sharp Q) = DR_{p,\varepsilon}(P, Q),$$

where $g_\sharp P$ is the push-forward of P by g. In particular, it ensures invariance of $DR_{p,\varepsilon}$ under translations and rotations.

Proof. Let $A \in \mathbb{R}^{d\times d}$ be a non-singular matrix and $b \in \mathbb{R}^d$ such that $g : x \mapsto Ax + b$. Then, it holds:

$$\begin{aligned} DR^p_{p,\varepsilon}\left(g_\sharp P, g_\sharp Q\right) &= \int_0^{1-\varepsilon} \left[d_{\mathcal{H}}\left(D^{\alpha(\beta)}(g_\sharp P), D^{\alpha(\beta)}(g_\sharp Q)\right)\right]^p \, \mathrm{d}\beta \\ &\overset{(i)}{=} \int_0^{1-\varepsilon} \left[d_{\mathcal{H}}\left(AD^{\alpha(\beta)}(P) + b, AD^{\alpha(\beta)}(Q) + b\right)\right]^p \, \mathrm{d}\beta, \end{aligned} \tag{10.5}$$

where (i) holds because any data depth satisfies $(\mathbf{D}_1)$ by definition. Furthermore,

$$\begin{aligned} &d_{\mathcal{H}}\left(AD^{\alpha(\beta)}(P) + b, AD^{\alpha(\beta)}(Q) + b\right) \\ &= \max\left\{\sup_{x\in D^{\alpha(\beta)}(P)} \inf_{y\in D^{\alpha(\beta)}(Q)} ||Ax - Ay||, \sup_{y\in D^{\alpha(\beta)}(Q)} \inf_{x\in D^{\alpha(\beta)}(P)} ||Ax - Ay||\right\} \\ &\overset{(ii)}{=} \max\left\{\sup_{x\in D^{\alpha(\beta)}(P)} \inf_{y\in D^{\alpha(\beta)}(Q)} ||x - y||, \sup_{y\in D^{\alpha(\beta)}(Q)} \inf_{x\in D^{\alpha(\beta)}(P)} ||x - y||\right\} \\ &= d_{\mathcal{H}}\left(D^{\alpha(\beta)}(P), D^{\alpha(\beta)}(Q)\right), \end{aligned}$$

where (ii) holds by virtue of hypothesis $AA^\top = \mathrm{I}_d$. Replacing it in (10.5) yields the desired results. ∎

Although formulas (10.2) and (10.3) are based on the same spirit, there are no apparent reasons why the proposed pseudo-metric should have the same behavior as the Wasserstein distance. It is the purpose of Proposition 10.5 to investigate the ability to factor out translations, for $DR_{2,\varepsilon}$ associated with the halfspace depth, giving a positive answer for the case of two Gaussian distributions with equal covariance matrices.

Proposition 10.5 (Translation characterization)**.** *Consider X, Y two random variables following $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$ with expectations μ_1, μ_2 and variance-covariance matrices Σ_1, Σ_2 respectively. Denoting by P^*, Q^* the centered versions of P, Q, for any $\varepsilon \in (0, 1]$, it holds:*

$$\left|DR^2_{2,\varepsilon}(P, Q) - DR^2_{2,\varepsilon}(P^*, Q^*) - ||\mu_1 - \mu_2||^2\right| \leq 2\ DR_{1,\varepsilon}(P^*, Q^*)\ ||\mu_1 - \mu_2||.$$

Now, let $P \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $Q \sim \mathcal{N}(\mu_2, \Sigma_2)$. Then it holds:

$$\left| DR_{1,\varepsilon}(P,Q) - ||\mu_1 - \mu_2|| \right| \leq C_\varepsilon \sup_{u \in \mathbb{S}^{d-1}} \left| \sqrt{u^\top \Sigma_1 u} - \sqrt{u^\top \Sigma_2 u} \right|,$$

where $C_\varepsilon = \int_0^{1-\varepsilon} \left| \Phi^{-1}(1-\alpha(\beta)) \right| \mathrm{d}\beta$ *with* Φ *the cdf of the univariate standard Gaussian distribution.*

The proof is detailed in Section 10.6 for the clarity of the reading. Following Proposition 10.5: when $\Sigma_1 = \Sigma_2$, one has $DR_{2,\varepsilon}(P,Q) = DR_{1,\varepsilon}(P,Q) = ||\mu_1 - \mu_2||$ for any $P \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $Q \sim \mathcal{N}(\mu_2, \Sigma_2)$ providing a closed-form expression in the Gaussian case.

10.1.3 Robustness

In this part, we explore the robustness of the proposed distance, associated with the halfspace depth, in view of the finite sample breakdown point (BP) (Donoho, 1982; Donoho and Huber, 1983). This notion investigates the smallest contamination fraction under which the estimation breaks down in the worst case. Considering a sample $\mathcal{S}_n = \{X_1, \ldots, X_n\}$ composed of i.i.d. observations drawn from a distribution $P \in \mathcal{P}(\mathcal{X})$ with empirical measure $P_n = (1/n)\sum_{i=1}^n \delta_{X_i}$, the finite sample breakdown point of $DR_{p,\varepsilon}$ w.r.t. $\mathcal{S}_n$, denoted by $BP(DR_{p,\varepsilon}, \mathcal{S}_n)$ is defined as

$$\min \left\{ \frac{o}{n+o} \ : \ \sup_{Z_1,\ldots,Z_o} DR_{p,\varepsilon}(P_{n+o}, P_n) = +\infty \right\},$$

where $P_{n+o} = \frac{1}{n+o}\left(\sum_{i=1}^n \delta_{X_i} + \sum_{j=1}^o \delta_{Z_j}\right)$ is the "concatenate" empirical measure between $X_1, \ldots, X_n$ and the contamination sample $Z_1, \ldots, Z_o$ with $o \in \mathbb{N}^*$. It is well known that the extremal regions of the halfspace depth are not robust while its central regions are rather stable under contamination (Donoho and Gasko, 1992). Fortunately, by construction, the parameter ε allows to ignore these extremal depth regions and thus to ensure robustness of the depth-trimmed regions distance. Based on results of Donoho and Gasko (1992) and Nagy and Dvořák (2021), the following proposition provides a lower bound on the finite sample breakdown point of $DR_{p,\varepsilon}$ which highlights the robustness of the proposed distance (as well as its dependence on ε).

Proposition 10.6 (Breakdown Point)**.** *For the halfspace depth function, for any* $\beta \in [0, 1-\varepsilon]$ *such that* $\alpha(\beta, P_n) < \alpha_{\max}(P_n)$*, it holds:*

$$BP(DR_{p,\varepsilon}, \mathcal{S}_n) \geq \begin{cases} \frac{\lceil n\alpha(1-\varepsilon,P_n)/(1-\alpha(1-\varepsilon,P_n))\rceil}{n+\lceil n\alpha(1-\varepsilon,P_n)/(1-\alpha(1-\varepsilon,P_n))\rceil} & \text{if } \alpha(1-\varepsilon,P_n) \leq \frac{\alpha_{\max}(P_n)}{1+\alpha_{\max}(P_n)}, \\ \frac{\alpha_{\max}(P_n)}{1+\alpha_{\max}(P_n)} & \text{otherwise,} \end{cases}$$

where $\alpha_{\max}(P_n) = \max_{x \in \mathbb{R}^d} D_{\mathrm{H}}(x, P_n)$.

Proof. For $DR_{p,\varepsilon}$ to break down at $\mathcal{S}_n$, it needs to have at least one trimmed-region that breaks down. Then the breakdown point of $DR_{p,\varepsilon}$ is higher than the minimum of the breakdown point of each region. Indeed, we have

$$
\begin{aligned}
BP(DR_{p,\varepsilon}, \mathcal{S}_n) &= \min\left\{\frac{o}{n+o} : \sup_{Z_1,\ldots,Z_o} DR_{p,\varepsilon}\left(P_{n+o}, P_n\right) = +\infty\right\} \\
&\geq \min_{\beta\in[0,1-\varepsilon]} \min\left\{\frac{o}{n+o} : \sup_{Z_1,\ldots,Z_o} d_{\mathcal{H}}\left(D_{\mathrm{H}}^{\alpha(\beta,P_{n+o})}(P_{n+o}), D_{\mathrm{H}}^{\alpha(\beta,P_n)}(P_n)\right) = +\infty\right\} \\
&= \min_{\beta\in[0,1-\varepsilon]} BP(D_{\mathrm{H}}^{\alpha(\beta,P_n)}(P_n), \mathcal{S}_n).
\end{aligned}
$$

Now applying Lemma 3.1 in Donoho and Gasko (1992) and Theorem 4 in Nagy and Dvořák (2021), a lower bound of the breakdown point of each halfspace region, for every $\beta \in [0, 1-\varepsilon]$, is given by

$$
BP(D_{\mathrm{H}}^{\alpha(\beta,P_n)}(P_n), \mathcal{S}_n) \geq \begin{cases} \frac{\lceil n\alpha(1-\varepsilon,P_n)/(1-\alpha(1-\varepsilon,P_n))\rceil}{n+\lceil n\alpha(1-\varepsilon,P_n)/(1-\alpha(1-\varepsilon,P_n))\rceil} & \text{if } \alpha(1-\varepsilon,P_n) \leq \frac{\alpha_{\max}(P_n)}{1+\alpha_{\max}(P_n)}, \\ \frac{\alpha_{\max}(P_n)}{1+\alpha_{\max}(P_n)} & \text{otherwise.} \end{cases}
$$

■

Thus, at least a proportion $\alpha(1-\varepsilon, P_n) \Big/ \left(1-\alpha(1-\varepsilon, P_n)\right)$ of outliers must be added to break down $DR_{p,\varepsilon}$ when considering larger regions, while central regions are robust independently of ε. For two data sets, $DR_{p,\varepsilon}$ breaks down if depth regions for at least one of the data sets do. The breakdown point is then the minimum between the breakdown points of each data set. However, the breakdown point considers the worst case, i.e. the supremum over all possible contaminations, and is often pessimistic. Indeed the proposed pseudo-metric can handle more outliers in certain cases as experimentally illustrated in Section 10.3.

10.2 Efficient Approximate Computation

Exact computation of $DR_{p,\varepsilon}$ can appear time-consuming, due to the high time complexity of the algorithms that calculate depth-trimmed regions (cf. Liu and Zuo (2014a); Liu et al. (2019) for projection and halfspace depths, respectively) rapidly growing with dimension. However, we design a universal approximate algorithm that achieves (log-) linear time complexity in n. Since properties $(\mathbf{D}_{3'}, \mathbf{D}_4, \mathbf{D}_5)$ ensure that depth regions are convex bodies in $\mathbb{R}^d$, they can be characterized by their support functions defined by $h_{\mathcal{K}_C^d}(u) = \sup\{\langle x, u\rangle : x \in \mathcal{K}_C^d\}$ for any $u \in \mathbb{S}^{d-1}$ where $\mathcal{K}_C^d$ is a convex compact of $\mathbb{R}^d$. Following Schneider (1993), for two (convex) regions $D^{\alpha(\beta)}(P)$ and $D^{\alpha(\beta)}(Q)$, the Hausdorff distance between them can be calculated as:

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

As we shall see in Section 10.3, mutual approximation of $h_{D^{\alpha(\beta)}}(u)$ by points from the sample and of sup by taking maximum over a finite set of directions allows for stable estimation quality. Recently, motivated by their numerous applications, a plethora

of algorithms have been developed for (exact and approximate) computation of data depths; see Section 2.4.2. Depths satisfying the projection property (which also include halfspace and projection depth, see Dyckerhoff (2004)) can be approximated by taking minimum over univariate depths; see e.g. Rousseeuw and Struyf (1998); Chen et al. (2013); Liu and Zuo (2014a), Nagy et al. (2020a) for theoretical guarantees, and Dyckerhoff et al. (2021) for an experimental validation. The case of AI-IRW is easier since its integral can be approximated by means of Monte-Carlo approximation.

Let $\mathcal{S}_n, \mathcal{S}'_m$ be two samples $\mathcal{S}_n = \{X_1, \dots, X_n\}$ and $\mathcal{S}'_m = \{Y_1, \dots, Y_m\}$ from P and Q respectively. When calculating approximated depth of sample points $D(\mathcal{S}_n) \triangleq \{D(X_i, P_n)\}_{i=1}^n$ (respectively $D(\mathcal{S}'_m)$), a matrix $M \in \mathbb{R}^{n \times n_{\text{proj}}}$ (respectively $M' \in \mathbb{R}^{m \times n_{\text{proj}}}$) of projections of sample points on (a common) set of $n_{\text{proj}} \in \mathbb{N}^*$ directions (with its element $M_{i,l} = \langle u_l, X_i \rangle$ for some $u_l \sim \mathcal{U}(\mathbb{S}^{d-1})$, where $\mathcal{U}(\cdot)$ is the uniform probability distribution) can be obtained as a side product. More precisely, $D(\mathcal{S}_n), D(\mathcal{S}'_m), M, M'$ are used in Algorithm 10.1, which implements the MC-approximation of the integral in (10.2). Time complexity of Algorithm 10.1 is $O\Big(n_{\text{proj}}(\Omega_\cdot(n \vee m, d) \vee n_\alpha(n \vee m))\Big)$, where $\Omega_\cdot(\cdot, \cdot)$ stands for the complete complexity of computing univariate depths—in projections on u—for all points of the sample. As a byproduct, projections on u can be saved to be reused after for the approximation of $h_{D^{\alpha(\beta)}(\cdot)}(u)$. e.g. for the halfspace depth $\Omega_{hsp}(n, d) = O\Big(n(d \vee \log n)\Big)$ composed of projection of the data onto u, ordering them, and passing to record the depths (see e.g. Mozharovskyi et al. (2015)). For the projection depth, $\Omega_{prj}(n, d) = O(nd)$, where after projecting the data onto u, univariate median and MAD can be computed with complexity $O(n)$ (see e.g. Liu and Zuo (2014a)). For the AI-IRW depth, $\Omega_{aiirw}(n, d) = O(d^3 \vee n(d \vee \log(n))$ since it involves the computation of the square root of the precision matrix. However, $O(d^3)$ may be improved, which depends on the algorithm employed for computing inverse of the covariance matrix.

10.3 Numerical Experiments

In this section, we first measure the quality of the approximation introduced in Section 10.2 and explore its dependency on the number of projections and to the n_α. Further, we present two studies on robustness of the proposed pseudo-metric $DR_{p,\varepsilon}$ to outliers. On synthetic data sets, we investigate how $DR_{p,\varepsilon}$ behaves under the presence of outliers using two different settings. On a real image data set extracted from Fashion-MNIST where images are seen as bags of pixels, we evaluate the robustness of spectral clustering based on $DR_{p,\varepsilon}$. Finally, we analyze the relevance of using $DR_{p,\varepsilon}$ as an evaluation metric in natural language generation to compare the empirical distributions of words of a pair of texts. Where applicable, we include state-of-the-art methods for comparison.

10.3.1 Approximation Error in Terms of the Number of Projections

Proposition 10.5 allows to derive a closed form expression for $DR_{2,\varepsilon}(P, Q)$ when P, Q are Gaussian distributions with the same variance-covariance matrix. In order to investigate the quality of the approximation on light-tailed and heavy-tailed distributions, we focus on computing $DR_{p,\varepsilon}$ (with $p = 2$, $\varepsilon = 0.3$, $n_\alpha = 20$ and using the halfspace depth) for varying number of random projections n_{proj} between a sample of 1000 points stemming from $P \sim \mathcal{N}(\mathbf{0}_d, \mathrm{I}_d)$ for $d = 5$ and two different samples. These two samples are

Algorithm 10.1 Approximation of $DR_{p,\varepsilon}$.

input: $\mathcal{S}_n$,$\mathcal{S}'_m$,n_α,n_{proj}.

init : $H = 0$.

Compute $D(\mathcal{S}_n), D(\mathcal{S}'_m), M, M'$

for $\ell = 1, \ldots, n_\alpha$ **do**

  Draw $\beta_\ell \sim \mathcal{U}([0, 1-\varepsilon])$

  Compute $\alpha_\ell(\cdot) := \alpha(\beta_\ell, \cdot)$

  Determine points inside $\alpha_\ell(\cdot)$-regions:
  $\mathcal{I}_\ell^{\mathcal{S}_n} = \{i : D_i^{\mathcal{S}_n} > \alpha_\ell(\mathcal{S}_n)\}$; $\mathcal{I}_\ell^{\mathcal{S}'_m} = \{j : D_j^{\mathcal{S}'_m} > \alpha_\ell(\mathcal{S}'_m)\}$

  for $l = 1, \ldots, n_{\text{proj}}$ **do**

    Compute approximation of support functions:
    $h_l^{\mathcal{S}_n} = \max\ M^{\mathcal{S}_n}_{(\mathcal{I}_\ell^{\mathcal{S}_n},\ l)}$; $h_l^{\mathcal{S}'_m} = \max\ M^{\mathcal{S}'_m}_{(\mathcal{I}_\ell^{\mathcal{S}'_m},\ l)}$

  Increase cumulative Hausdorff distance: $H \mathrel{+}= \max_{l \leq n_{\text{proj}}} |h_l^{\mathcal{S}_n} - h_l^{\mathcal{S}'_m}|^p$

return $\widehat{DR}_{p,\varepsilon} = (H/n_\alpha)^{1/p}$

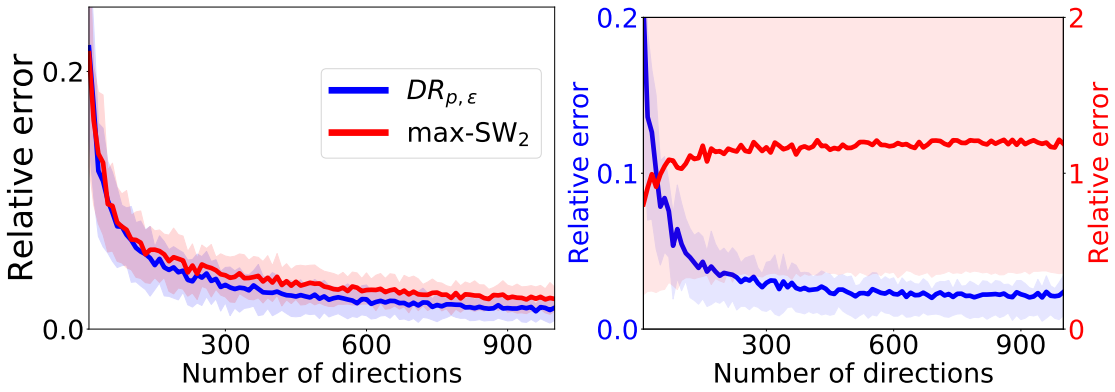


Figure 10.1 – Relative approximation error (averaged over 100 runs) of $DR_{p,\varepsilon}$ and the max Sliced-Wassserstein for *Gaussian* (left) and *Cauchy* (right) sample with dimension $d = 5$ for differing numbers of approximating directions.

constructed from 1000 observations stemming from *Gaussian* and symmetrical *Cauchy* distributions all with a center equal to $\mathbf{7}_d$. Comparison with the approximation of max Sliced-Wasserstein (max-SW) (see e.g. Kolouri et al., 2019), which shares the same closed-form as $DR_{2,\varepsilon}$, is also provided. Denoting by $\widehat{\text{max-SW}}$ the Monte-Carlo approximation of the max-SW, the relative approximation errors, i.e. $(\widehat{DR}_{p,\varepsilon} - ||\mathbf{7}_d||_2)/||\mathbf{7}_d||_2$ and $(\widehat{\text{max-SW}} - ||\mathbf{7}_d||_2)/||\mathbf{7}_d||_2$, are computed investigating both the quality of the approximation and the robustness of these discrepancy measures. Results, that report the averaged approximation error as well as the 25-75% empirical quantile intervals are depicted in Figure 10.1. They show that $DR_{p,\varepsilon}$ possesses the same behavior as max-SW

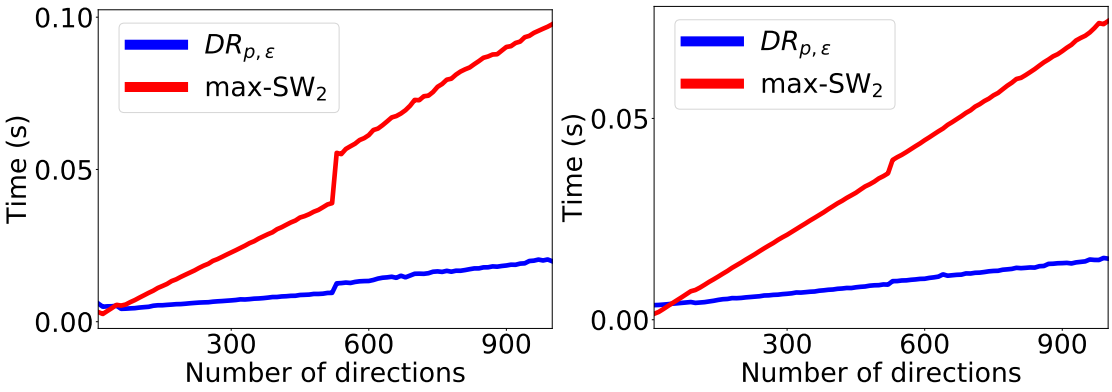


Figure 10.2 – Computation time (averaged over 100 runs) of $DR_{p,\varepsilon}$ and the max Sliced-Wassserstein for *Gaussian* (left) and *Cauchy* (right) sample with dimension $d = 5$ for differing numbers of approximating directions.

when considering *Gaussians* while it behaves advantageously for *Cauchy* distribution. Computation times are depicted in Figure 10.2 highlighting a constant-multiple improvement compared to the max-SW which is already computationally fast.

10.3.2 The Choice of the Parameter n_α

In order to investigate the quality of the approximation on light-tailed and heavy-tailed distributions, we focus on computing $DR_{2,0.1}$ (with $n_{\text{proj}} = 500$) for varying number of n_α between a sample of 1000 points stemming from $\mu \sim \mathcal{N}(\mathbf{0}_d, \Sigma)$ for $d \in \{2, 3, 10\}$, Σ drawn from the Wishart distribution (with parameters (d, I_d)) on the space of definite matrices and three different samples (which yields nine settings). These three samples are constructed from 1000 observations stemming from elliptically symmetric *Cauchy*, *Student-t_2* and *Gaussian* distributions all centered at $\mathbf{7}_d$. Results, that report the averaged approximation error as well as the 25-75% empirical quantile intervals are depicted in Figure 10.3. They show that $DR_{p,\varepsilon}$ converges slowly for *Cauchy* with growing n_α, while it converges with small n_α for *Gaussian* and *Student-t_2* distributions.

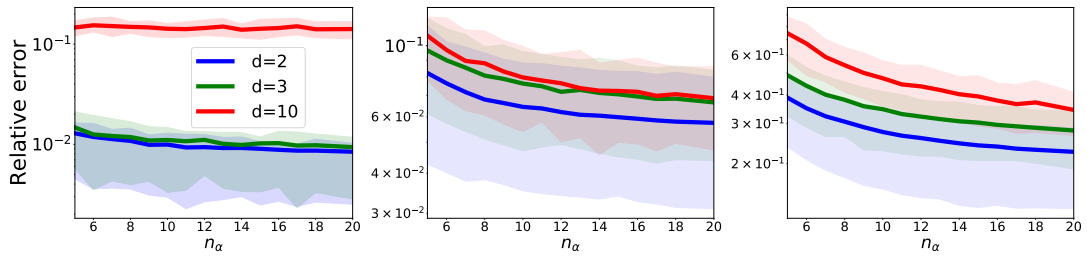


Figure 10.3 – Relative approximation error (averaged over 100 repetitions, y-axis in log scale) of $DR_{p,\varepsilon}$ for elliptically symmetric *Cauchy* (left), *Student-t_2* (middle) and *Gaussian* (right) distributions for differing numbers of n_α.

10.3.3 Robustness to Outliers

We analyze the robustness of $DR_{p,\varepsilon}$ by measuring its ability to overcome outliers. In this benchmark we naturally include existing robust extensions of the Wasserstein distance: Subspace Robust Wasserstein (SRW; Paty and Cuturi, 2019) searching for a maximal distance on lower-dimensional subspaces, ROBOT (Mukherjee et al., 2021) and RUOT (Balaji et al., 2020) being robust modifications of the unbalanced optimal transport (Chizat et al., 2018). Further, for completeness, we add the standard Wasserstein distance (W) and its approximation, the Sliced-Wasserstein (Sliced-W; Rabin et al., 2012) distance, with the same number of projections ($n_{\text{proj}} = 1000$) as $DR_{p,\varepsilon}$. Since the scales of the compared methods differ, *relative error* is used as a performance metric, i.e. the ratio of the absolute difference of the computed distance with and without anomalies divided by the latter. Two settings for a pair of distributions are addressed: (a) *Fragmented hypercube* precedently studied in Paty and Cuturi (2019), where the source distribution is uniform in the hypercube $[-1, 1]^2$ and the target distribution is transformed from the source via the map $T : x \mapsto x + 2\text{sign}(x)$ where $sign(.)$ is taken element-wisely. Outliers are drawn uniformly from $[-4, 4]^2$. (b) Two multivariate standard *Gaussian* distributions, one shifted by $\mathbf{10}_2$, with outliers drawn uniformly from $[-10, 20]^2$. Our analysis is conducted over 500 sampled points from the distributions described above.

In order to investigate the robustness of $DR_{p,\varepsilon}$, we consider the three following values of ε: $0.1, 0.2, 0.3$ computed with the projection depth. Thus, data depths are computed on source and target distributions such that 10%, 20%, 30% of data with lower depth values w.r.t. each distribution are not used in computation of $DR_{p,0.1}, DR_{p,0.2}, DR_{p,0.3}$, respectively. Figure 10.4, which plots the relative error depending on the portion of outliers varying up to 20%, illustrates advantageous behavior of $DR_{p,\varepsilon}$ (for $\varepsilon = 0.1, 0.2, 0.3$) for reasonable (starting with $\approx 2.5\%$) contamination. It also confirms the pessimism of the breakdown point provided in Proposition 10.6 since $DR_{p,0.1}$ (represented by the blue curve) show robustness to at least 20 % of outliers.

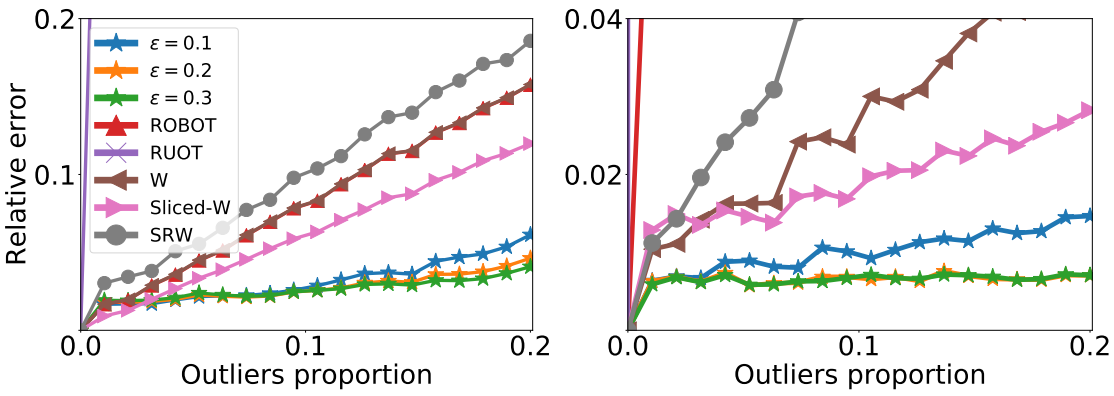


Figure 10.4 – Relative error (averaged over 100 runs) of different distances for increasing outliers proportion on *fragmented hypercube* (left) and *Gaussian* (right) data.

10.3.4 (Robust) Clustering on Bags of Pixels

We demonstrate the relevance of the proposed pseudo-metric through an application to (robust) clustering. To that end, we perform spectral clustering (Shi and Malik, 2000) on two data sets derived from Fashion-MNIST (FM). Each gray scale image is seen as a bag of pixels (Jebara, 2003), i.e. as an empirical probability distribution over a 3-dimensional space (the two first dimensions indicate the pixel position and the third one, its intensity). The first data set (FM) is constructed taking the 100 first images in each class of the Fashion-MNIST data set. The second data set (Cont. FM), considered as contaminated, is designed by introducing white patches on the left corner of 50 images drawn uniformly in the first data set, which yields 5% of contamination. We benchmark $DR_{p,\varepsilon}$ (using the projection depth) setting $p = 2$ and $\varepsilon = 0.1$ with the Wasserstein (W), the Sliced-Wasserstein (Sliced-W) and the Maximum Mean Discrepancy (MMD; Gretton et al., 2007) distances. $DR_{p,\varepsilon}$ and the Sliced-Wasserstein are approximated by Monte-Carlo using 100 directions while the MMD distance is computed using a Gaussian kernel with bandwidth equal to 1. As a baseline method, spectral clustering is also applied on images considered as vectors using Euclidean distance. Standard parameters of the `scikit-learn` spectral clustering implementation are employed with a number of clusters fixed to 10. Performance of the benchmarked metrics are assessed by measuring the normalized mutual information (NMI; Shannon, 1948) and the adjusted rank index (ARI; Hubert and Arabie, 1985), which are standard clustering evaluation measures when the ground truth class labels are available. Results presented in Table 10.1 show that for both cases, i.e. with or without contamination, spectral clustering based on $DR_{p,\varepsilon}$ outperforms spectral clustering based on the other metrics.

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

Table 10.1 – Spectral clustering performances.

10.4 Automatic Evaluation of Natural Language Generation (NLG)

There has been a recent surge towards NLG from the NLP community (Jalalzai et al., 2020; Colombo et al., 2019, 2021c,e). However, collecting human annotations to evaluate NLG systems is both expensive and time-consuming. Thus, automatically assessing the similarity between two texts is of high interest for the NLP community (Specia et al., 2010). This task aims to build an evaluation metric that achieves a high correlation with the score given by a human annotator. String-based metrics (i.e. that compare the string representations of texts) such as BLEU (Papineni et al., 2002), METEOR (MET.; (Banerjee and Lavie, 2005)), ROUGE (Lin, 2004), TER (Snover et al., 2006), have been outperformed in many tasks by embedding-based metrics (i.e.

that rely on continuous representations (Devlin et al., 2019)). Embedding-based metrics (e.g BertScore (BertS; Zhang et al., 2019), MoverScore (MoverS; Zhao et al., 2019a, Baryscore (Colombo et al., 2021d), InfoLM (Colombo et al., 2021b)), that are now the state-of-the-art of the domain, compare an input and a reference text both represented as probability distributions and are both constructed in a similar way. The first step relies on a deep contextualized encoder (BERT (Devlin et al., 2019) and its variants Dinkar et al., 2020; Chapuis et al., 2020, 2021) that maps texts into elements of a finite dimensional space. Precisely, each text corresponds to a collection of words, where each word is represented by an element in $\mathbb{R}^d$, where d is fixed by the encoder. The second step involves the use of a function that measures the similarity between the embedded texts.

For our purpose, we follow previous BERT-based metrics and evaluate performances of $DR_{p,\varepsilon}$ (with $p = 2$, $\varepsilon = 0.01$ and using the AI-IRW depth) on two different NLG tasks namely: data2text generation (using the WebNLG 2020 data set Ferreira et al., 2020) and summarization.

10.4.1 Data2text Experiment

Task description. In WebNLG 2020, the goal is to create new efficient generation algorithms that can verbalise knowledge based fragments. These algorithms are called Knowledge Base Verbalizers (Gardent et al., 2017) and are used during the micro-planning phase of NLG systems (Ferreira et al., 2018). WebNLG has been gathered to be more representative of the progress of recent NLG systems than previously existing task-oriented dialogue data sets (see e.g. SFHOTEL (Wen et al., 2015) and BAGEL (Mairesse et al., 2010)). Data and system performances can be found here[1]. The task consists in mapping Resource Description Framework (RDF) triplet to natural language (RDF format is used for many application including FOAF[2]. For WebNLG 2020, the triplets are extracted from DBpedia (Auer et al., 2007). For example, given the following triplet `(John_Blaha birthDate 1942_08_26)`, `(John_Blaha birthPlace San_Antonio)` and `(John_Blaha job Pilot)` the ground-truth reference is `John Blaha, born in San Antonio on 1942-08-26, worked as a pilot`.

Setting. Data have been made freely available from the authors here[3]. To compose this data set, 15 systems (both symbolic and neural-based) have been used. The final data set is composed of over 3k samples of human annotations[4]. We follow standard methods to assess performance of NLG metrics (see e.g. Zhao et al., 2019a). We compute the correlation with the following annotation scores: *correctness*, *data coverage*, and *relevance*. We report in Table 10.2 correlation results on the WebNLG task using Pearson (r), Spearman (ρ) and Kendall (τ) correlation coefficients. When performing a fair comparison between metrics, i.e. when $DR_{p,\varepsilon}$, W, Sliced-W, MMD are directly used on the output of BERT, we observe that $DR_{p,\varepsilon}$ achieves the best results on all configurations. It is worth noting that $DR_{p,\varepsilon}$ also compares favorably against existing state-of-the-art NLG methods in many different scenarios and shows promising results.

Results. We gather in Table 10.2 results on the WebNLG task. To compare $DR_{p,\varepsilon}$ (with $\varepsilon = 0.01$, $n_\alpha = 5$, $p = 2$) with the different metrics (i.e. Wasserstein, Sliced-Wasserstein, MMD), we work on Roberta-based model from the HuggingFace hub (Wolf

[1] https://webnlg-challenge.loria.fr/
[2] http://www.foaf-project.org/
[3] https://gitlab.com/shimorina/webnlg-dataset/-/tree/master/release_v3.0
[4] https://webnlg-challenge.loria.fr/files/WebNLG-2020-Presentation.pdf

	Correctness			Data Coverage			Relevance		
	r	τ	ρ	r	τ	ρ	r	τ	ρ
$DR_{p,\varepsilon}$	**89.4**	**80.0**	**92.6**	**84.2**	58.3	72.3	**86.2**	62.7	72.9
Wasserstein	86.2	73.0	86.7	80.4	45.3	62.3	83.8	51.3	67.6
Sliced-Wasserstein	86.1	73.0	85.8	80.9	45.5	60.0	82.0	51.3	68.2
MMD	25.4	71.7	8.3	19.1	45.3	10.0	26.1	51.3	15.0
BertScore	85.5	73.3	83.4	74.7	53.3	68.2	83.3	65.0	**79.4**
MoverScore	84.1	73.3	84.1	78.7	53.3	66.2	82.1	65.0	77.4
BLEU	77.6	60.0	66.3	55.7	36.6	50.2	63.0	51.6	65.2
ROUGE-1	80.6	65.0	65.0	76.5	**60.3**	**76.3**	64.3	56.7	69.2
ROUGE-2	73.6	58.3	63.3	54.7	35.0	43.1	62.0	46.7	60.8
METEOR	86.5	70.0	66.3	77.3	46.6	50.2	82.1	58.6	65.2
TER	79.6	58.0	78.3	69.7	38.0	58.2	75.0	**77.6**	70.2

Table 10.2 – WebNLG 2020: Absolute correlation at the system level with three human judgment criteria. Best overall results are indicated in bold, best results in their group are underlined.

et al., 2019) and extract representation from the 11th layer. From Table 10.2, we observe a similar behavior from BertScore and MoverScore. This similarity has been also reported in a different setting in the previous work of Zhao et al. (2019a). Overall, we observe that $DR_{p,\varepsilon}$ is always among the top scoring metrics in its group and also achieves best overall results on several configurations. It is worth noting that $DR_{p,\varepsilon}$ only relies on information available in the candidate and the reference text where BertScore and MoverScore use Inverse Document Frequency (IDF) information computed on every texts of the data set.

10.4.2 Summarization Experiment

Task description. Text summarization has attracted a lot of attention in recent years (Zhang et al., 2020). Two types of models exist: *extractive* and *abstractive*. In extractive summarization, the system copies chunks of informative fragments from the input texts, whereas in abstractive summarization the system generates novel words. In this section, we describe our experimental setting. We present the tasks and the baseline metrics used for automatic evaluation of summarization. For this task, we work with the data set from Bhandari et al. (2020). This data set has been introduced to solve several flaws (Rankel et al., 2013) present in existing summarization data sets such as TAC (Dang and Owczarzak, 2008; McNamee and Dang, 2009). The data set has been annotated using the pyramid score (Nenkova et al., 2007; Nenkova and Passonneau, 2004) and automatically built from the CNN/Daily News (Bhandari et al., 2020). It gathers 11 490 summaries coming from 11 extractive systems (See et al., 2017; Chen and Bansal, 2018; Raffel et al., 2020; Gehrmann et al., 2018; Dong et al., 2019; Liu and Lapata, 2019; Lewis et al., 2020; Yoon et al., 2020) and 14 abstractive systems (Zhou et al., 2018; Narayan et al., 2018; Kedzie et al., 2018; Zhong et al., 2019; Liu and Lapata, 2019; Dong et al., 2019; Wang et al., 2020; Zhong et al., 2020).

	Abstractive			Extractive		
	r	τ	ρ	r	τ	ρ
$DR_{p,\varepsilon}$	<u>72.1</u>	<u>72.1</u>	70.1	<u>91.5</u>	**91.5**	<u>69.2</u>
Wasserstein	71.0	70.4	<u>71.1</u>	74.2	74.2	40.0
Sliced-Wasserstein	70.1	68.7	71.0	72.4	73.9	<u>69.2</u>
MMD	68.2	67.5	67.9	75.6	75.6	56.1
BertScore	71.7	<u>71.9</u>	72.0	70.9	72.9	**73.8**
MoverScore	<u>72.4</u>	<u>71.9</u>	<u>73.0</u>	<u>76.1</u>	<u>76.1</u>	47.4
ROUGE-1	**73.5**	73.0	**74.4**	72.2	<u>74.0</u>	<u>69.1</u>
ROUGE-2	73.0	**73.5**	73.0	55.1	53.2	<u>69.1</u>
JS-2	68.9	6.8	69.8	**92.9**	5.5	19.0

Table 10.3 – Summarization: absolute correlation coefficients between different metrics on text summarization. Best overall results are indicated in bold, best results in their group are underlined.

Example. The goal is to assign a similarity score between a reference text: "*Manchester United take on Manchester City on Sunday. Match will begin at 4 pm local time at United's Old Trafford home. Police have no objections to kick-off being so late in the afternoon. Last late afternoon weekend kick-off in the Manchester derby saw 34 fans arrested at Wembley in 2011 fa cup semi-final*" and the text generated by a NLG system: "*Manchester Derby takes place at Old Trafford on Sunday afternoon police have no objections to the late afternoon kick-off both sides are challenging for a top-four spot in the Premier League the man in charge of patrolling the sell-out clash has no such fears*".

Results. We gather in Table 10.3, the results on the summarization task. We use BERT-based uncased model and rely on the representations extracted from the 9th layer (similarly to BertScore). For this experiment the following parameters are used: $\varepsilon = 0.01$, $n_\alpha = 5$, $p = 2$. For this task, we are able to reproduce results from Bhandari et al. (2020) where the different behavior regarding the extractive and the abstractive systems is also observed. In this experiment, we observe that $DR_{p,\varepsilon}$ is able to achieve stronger results than other metrics based on Wasserstein, Sliced-Wasserstein and MMD. We also observe that $DR_{p,\varepsilon}$ outperforms MoverScore and BertScore on extractive systems (on r and τ). We believe these results support our approach.

It could be relevant to extend this study to multimodal cases (Garcia et al., 2019; Colombo et al., 2021a) and others tasks of generation such as dialog acts (Colombo et al., 2020; Witon et al., 2018).

10.5 Concluding Remarks

Leveraging the notion of statistical data depth function, a novel pseudo-metric between multivariate probability distributions was introduced. The developed framework exhibits an inherent versatility due to the existence of numerous data depths variants. The linear approximation algorithm and the robustness property make $DR_{p,\varepsilon}$ a promising tool for a large spectrum of applications beyond clustering and NLG, e.g. in generative adversarial networks (GANs) or in information retrieval. Moreover, recent works, extending the notion of data depth to further types of data such as functional and time-series data (Nieto-Reyes and Battey, 2016; Gijbels and Nagy, 2017), curves (or paths)

data (Lafaye de Micheaux et al., 2021), directional (or spherical) data (Ley et al., 2014), random matrices (Paindaveine and Van Bever, 2018) and random sets (Cascos et al., 2021) shall allow for the use of the proposed pseudo-metric for a wide range of applications. Finally, it is noteworthy that approximation was performed via random projections in the current work, while techniques similar to those by Dyckerhoff et al. (2021) could further accelerate the computation.

10.6 Proof

10.6.1 Proof of Proposition 10.5

Let $u \in \mathbb{S}^{d-1}$ and $X \sim P$ where $P \in \mathcal{P}(\mathcal{X})$ with $\mathcal{X} \subset \mathbb{R}^d$. We define the $(1-\beta)$ directional quantile of a distribution P in the direction u as:

$$q_{P,u}^{1-\beta} = \inf\left\{t \in \mathbb{R} \,:\, \mathbb{P}\Big(\langle u, X\rangle \leq t\Big) \geq 1-\beta\right\} \tag{10.6}$$

and the upper $(1-\beta)$ quantile set of P:

$$S_{P,u}^{1-\beta} = \left\{x \in \mathbb{R}^d \,:\, \langle u, x\rangle \leq q_{P,u}^{1-\beta}, \quad \forall\, u \in \mathbb{S}^{d-1}\right\}. \tag{10.7}$$

We first recall two useful results, so as to characterize the halfspace depth regions.

Lemma 10.7 (Brunel, 2019, Lemma 1)**.** *Let $P \in \mathcal{P}(\mathcal{X})$, for any $\beta \in (0,1)$, it holds: $D^{\beta}(P) = S_{P,u}^{1-\beta}$.*

Lemma 10.8 (Brunel, 2019, Proposition 1)**.** *Let $P \in \mathcal{P}(\mathcal{X})$ with a $(1-\beta)$ directional quantile $q_{P,u}^{1-\beta}$. Assume that $u \mapsto q_{P,u}^{1-\beta}$ are sublinear, i.e. $q_{P,u+\zeta v}^{1-\beta} \leq q_{P,u}^{1-\beta} + \zeta\, q_{P,v}^{1-\beta}$, $\forall\, \zeta > 0$. Then for any $u \in \mathbb{S}^{d-1}$, it holds $h_{S_{P,u}^{1-\beta}}(u) = q_{P,u}^{1-\beta}$.*

First assertion. Denote Z_1, Z_2 two random variables following P^*, Q^* respectively. For any $x \in \mathbb{R}^d$ and $\beta \in [0, 1-\varepsilon]$,

$$\begin{aligned}
x \in D_{\mathrm{H}}^{\alpha(\beta)}(P) \iff D_{\mathrm{H}}(x)(P) \geq \alpha(\beta) &\iff \forall\, u \in \mathbb{S}^{d-1},\ \mathbb{P}\Big(\langle u, X\rangle \ \leq \langle u, x\rangle\Big) \geq \alpha(\beta) \\
&\iff \forall\, u \in \mathbb{S}^{d-1},\ \mathbb{P}\Big(\langle u, Z_1 + \mu_1\rangle \ \leq \langle u, x\rangle\Big) \geq \alpha(\beta) \\
&\iff \forall\, u \in \mathbb{S}^{d-1},\ \mathbb{P}\Big(\langle u, Z_1\rangle \ \leq \langle u, x - \mu_1\rangle\Big) \geq \alpha(\beta) \\
&\iff x - \mu_1 \in D_{P^*}^{\alpha(\beta)},
\end{aligned}$$

where $D_{P^*}^{\alpha(\beta)}$ means the depth region at level $\alpha(\beta)$ of the halfspace depth w.r.t. P^* and will be used throughout this subsection for the sake of clarity. The same reasoning holds for Q and Q^*. Following this, for any $\beta \in [0,\ 1-\varepsilon]$ and $u \in \mathbb{S}^{d-1}$, it holds:

$$h_{D_P^{\alpha(\beta)}}(u) = h_{D_{P^*}^{\alpha(\beta)}}(u) - \langle u, \mu_1\rangle \qquad \text{and} \qquad h_{D_Q^{\alpha(\beta)}}(u) = h_{D_{Q^*}^{\alpha(\beta)}}(u) - \langle u, \mu_2\rangle$$

Thus, it holds:

$$\begin{aligned}
DR^2_{2,\varepsilon}(P,Q) &= \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D^{\alpha(\beta)}_{P^*}}(u) - \langle u,\mu_1\rangle - h_{D^{\alpha(\beta)}_{Q^*}}(u) + \langle u,\mu_2\rangle \right|^2 \mathrm{d}\beta \\
&\leq \sup_{u\in\mathbb{S}^{d-1}} \left|\langle u,\mu_1-\mu_2\rangle\right|^2 + \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D^{\alpha(\beta)}_{P^*}}(u) - h_{D^{\alpha(\beta)}_{Q^*}}(u)\right|^2 \mathrm{d}\beta \\
&\quad + 2 \sup_{u\in\mathbb{S}^{d-1}} \left|\langle u,\mu_1-\mu_2\rangle\right| \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D^{\alpha(\beta)}_{P^*}}(u) - h_{D^{\alpha(\beta)}_{Q^*}}(u)\right| \mathrm{d}\beta \\
&= ||\mu_1-\mu_2||^2 + DR^2_{2,\varepsilon}(P^*,Q^*) + 2||\mu_1-\mu_2|| DR_{1,\varepsilon}(P^*,Q^*). \qquad (10.8)
\end{aligned}$$

On the other side, we have:

$$\begin{aligned}
DR^2_{2,\varepsilon}(P,Q) &\geq \sup_{u\in\mathbb{S}^{d-1}} \left|\langle u,\mu_1-\mu_2\rangle\right|^2 + \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D^{\alpha(\beta)}_{P^*}}(u) - h_{D^{\alpha(\beta)}_{Q^*}}(u)\right|^2 \mathrm{d}\beta \\
&\quad - 2 \sup_{u\in\mathbb{S}^{d-1}} \left|\langle u,\mu_1-\mu_2\rangle\right| \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D^{\alpha(\beta)}_{P^*}}(u) - h_{D^{\alpha(\beta)}_{Q^*}}(u)\right| \mathrm{d}\beta \\
&= ||\mu_1-\mu_2||^2 + DR^2_{2,\varepsilon}(P^*,Q^*) - 2||\mu_1-\mu_2|| DR_{1,\varepsilon}(P^*,Q^*). \qquad (10.9)
\end{aligned}$$

Combining (10.8) and (10.9) leads to the desired result.

Second assertion. For any $u \in \mathbb{S}^{d-1}$, the $(1-\alpha(\beta))$ quantiles of random variables $\langle u, X\rangle$ and $\langle u, Y\rangle$ such that $\langle u, X\rangle \sim \mathcal{N}(\langle u,\mu_1\rangle, u^\top\Sigma_1 u)$ and $\langle u, Y\rangle \sim \mathcal{N}(\langle u,\mu_2\rangle, u^\top\Sigma_2 u)$ are defined by:

$$q^{1-\alpha(\beta)}_{P,u} = \langle u,\mu_1\rangle + \Phi^{-1}(1-\alpha(\beta))\sqrt{u^\top\Sigma_1 u} \qquad q^{1-\alpha(\beta)}_{Q,u} = \langle u,\mu_2\rangle + \Phi^{-1}(1-\alpha(\beta))\sqrt{u^\top\Sigma_2 u},$$

where Φ is the cumulative distribution function of the univariate standard Gaussian distribution. Now, to apply Lemma 10.8, it is sufficient to prove that directional quantiles are sublinear. It holds using subadditivity of the square root function. Indeed, for any $u, v \in \mathbb{S}^{d-1}$ and $\zeta > 0$, we have:

$$\begin{aligned}
&\langle u+\zeta v,\mu_1\rangle + \Phi^{-1}(1-\alpha(\beta))\sqrt{(u+\zeta v)^\top\Sigma_1(u+\zeta v)} \\
&= \langle u,\mu_1\rangle + \zeta\langle v,\mu_1\rangle + \Phi^{-1}(1-\alpha(\beta))\sqrt{(u+\zeta v)^\top\Sigma_1(u+\zeta v)} \\
&\leq \langle u,\mu_1\rangle + \zeta\langle v,\mu_1\rangle + \Phi^{-1}(1-\alpha(\beta))\left[\sqrt{u^\top\Sigma_1 u} + \zeta\sqrt{v^\top\Sigma_1 v}\right] \\
&= q^{1-\alpha(\beta)}_{P,u} + \zeta\, q^{1-\alpha(\beta)}_{P,v}.
\end{aligned}$$

The same reasoning holds for Q. Applying Lemma 10.7 and Lemma 10.8, for any $u \in \mathbb{S}^{d-1}$, we have $h_{D^{\alpha(\beta)}_P}(u) = q^{1-\alpha(\beta)}_{P,u}$ and $h_{D^{\alpha(\beta)}_Q}(u) = q^{1-\alpha(\beta)}_{Q,u}$. It follows:

$$
\begin{aligned}
DR_{1,\varepsilon}(P,Q) &= \int_0^{1-\varepsilon} d_{\mathcal{H}}\left(D_P^{\alpha(\beta)}, D_Q^{\alpha(\beta)}\right) \mathrm{d}\beta = \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| h_{D_P^{\alpha(\beta)}}(u) - h_{D_Q^{\alpha(\beta)}}(u) \right| \mathrm{d}\beta \\
&= \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| \langle u, \mu_1 - \mu_2 \rangle + \Phi^{-1}(1-\alpha(\beta)) \left[\sqrt{u^\top \Sigma_1 u} - \sqrt{u^\top \Sigma_2 u} \right] \right| \mathrm{d}\beta \\
&\leq ||\mu_1 - \mu_2|| + \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| \Phi^{-1}(1-\alpha(\beta)) \left[\sqrt{u^\top \Sigma_1 u} - \sqrt{u^\top \Sigma_2 u} \right] \right| \mathrm{d}\beta \\
&= ||\mu_1 - \mu_2|| + C_\varepsilon \sup_{u\in\mathbb{S}^{d-1}} \left| \sqrt{u^\top \Sigma_1 u} - \sqrt{u^\top \Sigma_2 u} \right|,
\end{aligned}
$$

with $C_\varepsilon = \int_0^{1-\varepsilon} \left| \Phi^{-1}(1-\alpha(\beta)) \right| \mathrm{d}\beta$. The lower bound is obtained by means the same reasoning. Notice that:

$$
||\mu_1 - \mu_2|| = \sup_{u\in\mathbb{S}^{d-1}} \left| \langle u, \mu_1 - \mu_2 \rangle \right| = \int_0^{1-\varepsilon} \sup_{u\in\mathbb{S}^{d-1}} \left| \langle u, \mu_1 - \mu_2 \rangle \right| \mathrm{d}\beta.
$$

Introducing $h_{D_P^{\alpha(\beta)}}(u), h_{D_Q^{\alpha(\beta)}}(u)$ and using triangular inequality, subadditivity of the supremum and linearity of the integral, we obtain:

$$
||\mu_1 - \mu_2|| \leq DR_{1,\varepsilon}(P,Q) + C_\varepsilon \sup_{u\in\mathbb{S}^{d-1}} \left| \sqrt{u^\top \Sigma_1 u} - \sqrt{u^\top \Sigma_2 u} \right|,
$$

which ends the proof.

Conclusion and Perspectives

In this dissertation, the work presented is focused around two main axes: the design of robust probability metrics and functional anomaly detection. We have introduced many efficient statistical procedures, both in terms of statistical accuracy and computational time, for both functional anomaly detection and robust learning.

With the ubiquity of sensors in the IoT era, statistical observations are becoming increasingly available in the form of massive time-series or functions. The case of functional data is thus of crucial interest in practice. Although unsupervised anomaly detection has been widely documented in the literature for multivariate data, the case of functional data remains understudied. Filling this gap is the angle embraced in the first part of this thesis. Indeed, in Part II, we have introduced two novel and generic techniques, avoiding dimensionality reduction steps, to perform (unsupervised) functional anomaly detection.

The Functional Isolation Forest algorithm has been proposed, which is an extension of Isolation Forest to functional data. The combined choice of the dictionary itself, the probability distribution used to pick a Split variable and the scalar product used for the projection enables FIF to exhibit a great flexibility in detecting anomalies for a variety of tasks. It is worth mentioning that FIF is easily extendable to multivariate functional data. However, no theory exists for either the multivariate or the functional version and would deserve more attention in the future.

Further, we have introduced the ACH depth, a novel functional depth function on the space $\mathcal{C}(\mathcal{T})$ of real valued continuous curves on $\mathcal{T}$ that presents various advantages. Regarding interpretability first, the depth computed at a query curve $\mathbf{x} \in \mathcal{C}(\mathcal{T})$ takes the form of an expected ratio, quantifying the relative increase of the area of the convex hull of i.i.d. random curves when adding $\mathbf{x}$ to the batch. We have shown that this depth satisfies several desirable properties and have explained how to solve approximation issues, concerning the sampled character of observations in practice and scalability namely.

The performance of recent functional anomaly detection techniques, involving FIF and ACH, is evaluated on two real-world data sets related to the monitoring of helicopters in flight and to the spectrometry of construction materials namely, in order to provide recommendation guidance for practitioners.

The second part of this thesis introduced general statistical procedures, relying on the concept of data depth and robust mean estimation, that are able to handle corrupted data during inference as highlighted by theoretical and numerical results. Indeed, in Part III, we have contributed to bridge the gap between robustness, probability metrics and data depths.

We have introduced three robust estimators of the Wasserstein distance based on MoM methodology. We have shown asymptotic and nonasymptotic results in the context of polluted data possibly designed by a malicious adversary. Sharper results may be obtained regarding recent advances exhibited in Weed and Bach (2019). Surpassing computational issues, we have designed an algorithm to compute, in a efficient way, these

estimators. In addition, we proposed to robustify WGANs using one of the introduced estimators and have shown its benefits on convincing numerical results. The theoretically well-founded MoM approaches to robustify the Wasserstein distance open the door to numerous applications beyond WGAN, including variational generative modeling. The promising MoMGAN deserves more attention and future work will concern the analysis of the estimator it provides.

We have introduced a novel notion of statistical depth (AI-IRW), modifying the original Integrated Rank-Weighted (IRW) depth proposal in Ramsay et al. (2019). The statistical depth we have introduced has been shown not only to inherit all the compelling features of the IRW depth, its theoretical properties and its computational advantages (no optimization problem solving is required to compute it), but also to fulfill in addition the affine invariance property, crucial regarding interpretability/reliability issues. Though the AI-IRW sample version exhibits a complex probabilistic structure, an estimator of the precision matrix being involved in its definition, a nonasymptotic analysis has been carried out here, revealing its good concentration properties around the true AI-IRW depth. The merits of the AI-IRW depth have been illustrated by encouraging numerical experiments, for anomaly detection purpose in particular, offering the perspective of a widespread use for various statistical learning tasks. Although highlighted in experiments, theoretical properties of the robustness of AI-IRW combined with MCD remains to be investigated.

Leveraging the notion of statistical data depth function, a novel pseudo-metric between multivariate probability distributions—that meets the aforementioned requirements—was introduced. The developed framework exhibits an inherent versatility due to the existence of numerous data depth variants. The linear approximation algorithm and the robustness property make this pseudo-metric a promising tool for a large spectrum of applications beyond clustering and NLG. Moreover, recent works, extending the notion of data depth to further types of data such as functional and time-series data (Nieto-Reyes and Battey, 2016; Gijbels and Nagy, 2017), curves (or paths) data (Lafaye de Micheaux et al., 2021), directional (or spherical) data (Ley et al., 2014), random matrices (Paindaveine and Van Bever, 2018), random sets (Cascos et al., 2021), and metric spaces (Dai et al., 2021) shall allow for the use of the proposed pseudo-metric for a wide range of applications. The main drawback of this approach to be widely used in Machine Learning is its non-differentiability w.r.t. to each distribution. This aspect, shared with most of data depths, would deserve further investigations in order to use our pseudo-metric in generative models involving neural networks. Recent developments in Dyckerhoff et al. (2021) may be employed to obtain better accuracy of the supremum over the unit sphere that involves in our approximation.

Safer Machine learning systems is fundamental towards a large scale adoption of AI in real-world. Even though deep learning algorithms have achieved state-of-the-art in several tasks, their adoption in critical system remains shy (Alves et al., 2018; Johnson, 2018). One of the reasons is that the training set does not reflect well enough the real-life environment and the lacking trustworthiness of results when these models are deployed (Amodei et al., 2016). The variability of such environments are prohibitively to modelling during training time, thus an intelligent agent should be able to detect if the data it encounters is adapted or not to what it was trained for (Sehwag et al., 2019). Also, it has been shown that a simple change in input data adversarially introduced by someone having full knowledge of the uncorrupted data distribution can drastically deteriorate the performance of the most sophisticated models (Szegedy et al., 2014).

Over time, a vast literature has been produced investigating defense methods against adversarial examples. On the one hand, techniques to train models with improved robustness to upcoming attacks have been proposed (Zheng et al., 2016; Madry et al., 2018; Athalye et al., 2018). On the other hand, effective methods to detect adversarial examples given a pre-trained model have also been studied (Meng and Chen, 2017; Lee et al., 2018). To that end, the work of this thesis may be applied either to build defense procedures, based on the proposed robust metrics, or to detect adversarial attacks inside of networks through the introduced anomaly detection approaches.

Appendices

Appendices

A Additional Materials for FIF

In this section, additional materials for the chapter 5 is depicted.

A.1 Additional Study of the Parameters of FIF

Here, we present results of a simulation study of the variance of the FIF algorithm. The experiments were conducted on the data sets (a) and (b) from Section 5.3.1, for each of the four specified observations $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ using the following settings (except varying parameter):

> *Dictionary*: Gaussian wavelets (negative second derivative of the standard Gaussian density) with random variance selected in an uniform way in $[0.2, 1]$ and a translation parameter selected randomly in $[-4, 4]$. We fixed the size of the dictionary to 1000.
>
> *Scalar product*: L_2 dot product.
>
> *Size of the data set*: $n = 500$.
>
> *Subsampling size*: $\psi = 64$.
>
> *The number of trees*: $N = 100$.
>
> *The height limit*: fixed to $l = \lceil \log_2(\psi) \rceil$.

The figures below indicate boxplots of the FIF anomaly score, over 100 runs. Empirical study of the FIF anomaly score and its variance when increasing the number of F-*i*trees is depicted in Figure A.5. Empirical study of the FIF anomaly score and its variance when increasing the subsample size is depicted in Figure A.6. Empirical study of the FIF anomaly score and its variance when increasing the height limit of the F-*i*tree is depicted in Figure A.7. Taking finite size versions of the infinite *gaussian wavelets* dictionary, an empirical study of the FIF anomaly score and its variance when increasing the size of the dictionary is depicted in Figure A.8. Empirical study of the FIF anomaly score for a variety of dictionaries with the L_2 scalar product of the derivatives is depicted in Figure A.9. Empirical study of the FIF anomaly score for a variety of dictionaries with the L_2 scalar product of the derivatives is depicted in Figure A.9.

Analysis of the results. In a first experiment, we show the boxplots of the score estimated by FIF when increasing the number of F-*i*trees and observe that, as expected, the variance diminishes when N grows (see Figure A.5). We also see in Figure A.6 that with an increasing subsample size ψ the FIF anomaly score increases for anomalies since these are more often present in the subsample and thus isolated faster (with shorter path length) when calculating the score than when they were absent in the subsample; this effect is reciprocal for normal observations. A similar behavior is observed with increasing height limit l in Figure A.7. The variance of the score tends to slightly increase with ψ and l because of more observations/branching possibilities. If the dictionary

is sufficiently rich, its size does not influence the FIF anomaly score and its variance stabilizes relatively fast while growing the size of the dictionary (see Figure A.8) which encourages the use of massive (and infinite size) dictionaries.

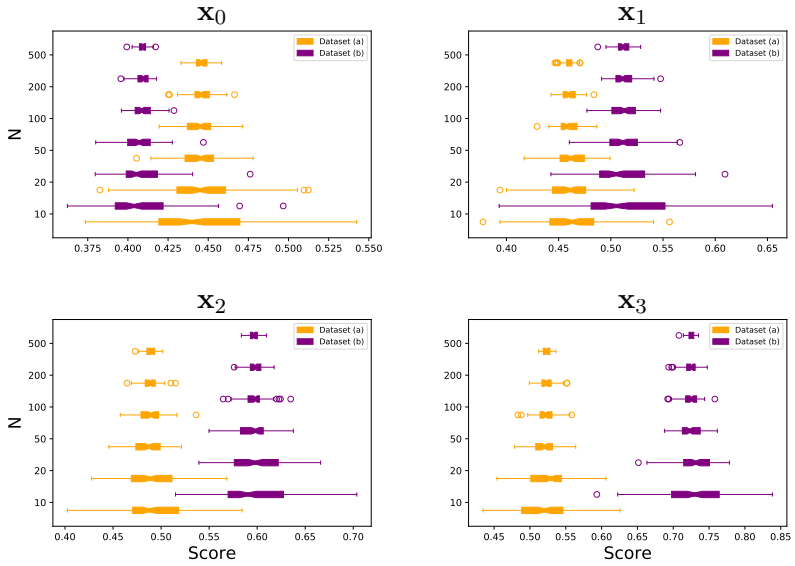


Figure A.5 – Boxplot (over 100 repetitions) of the FIF score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for different N. The orange boxplots represent the data set (a) while the purple boxplots represent the data set (b).

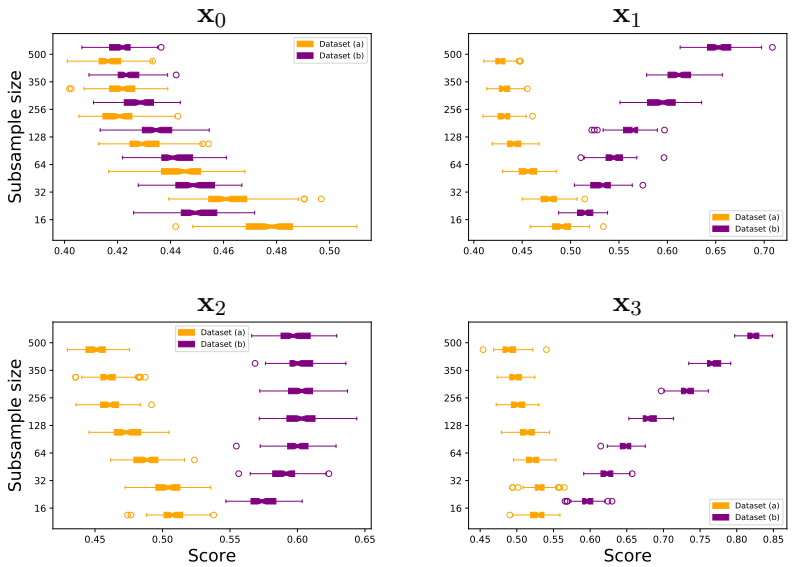


Figure A.6 – Boxplot (over 100 repetitions) of the FIF score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for different subsample sizes. The orange boxplots represent the data set (a) while the purple boxplots represent the data set (b).

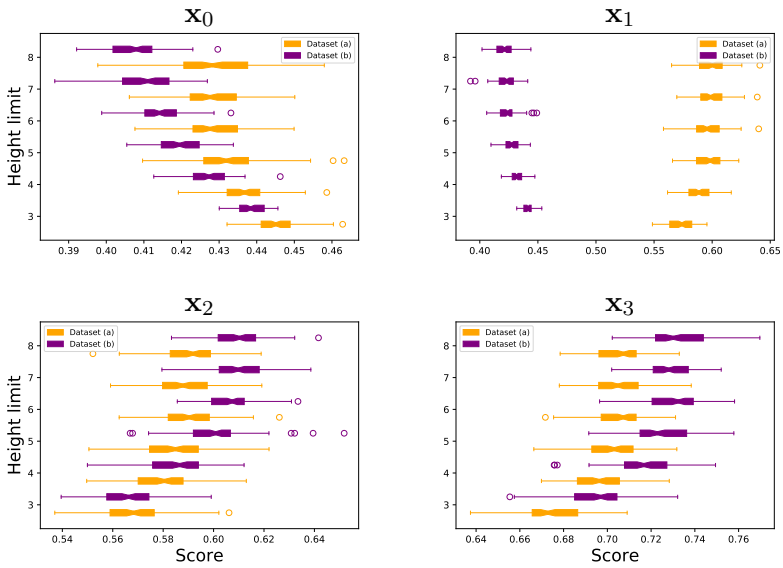


Figure A.7 – Boxplot (over 100 repetitions) of the FIF score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for different height limits. The orange boxplots represent the data set (a) while the purple boxplots represent the data set (b).

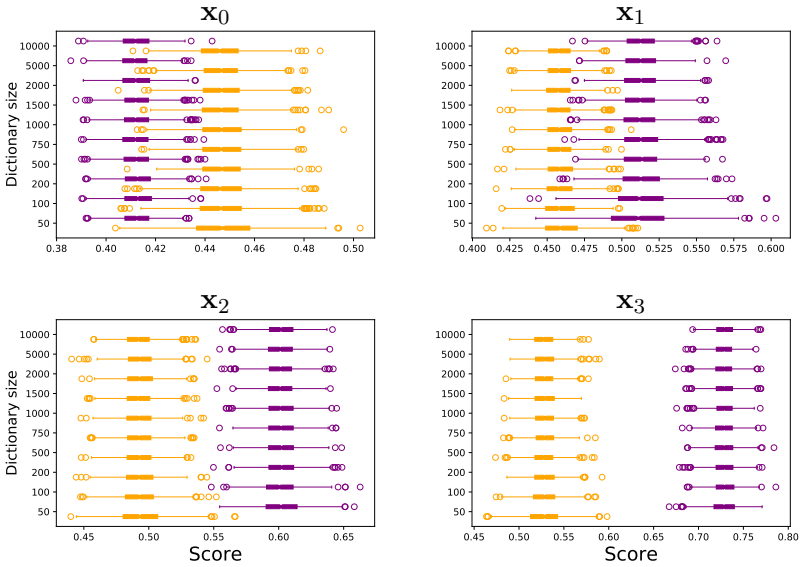


Figure A.8 – Boxplot (over 100 repetitions) of the FIF score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for different dictionary sizes. The orange boxplots represent the data set (a) while the purple boxplots represent the data set (b).

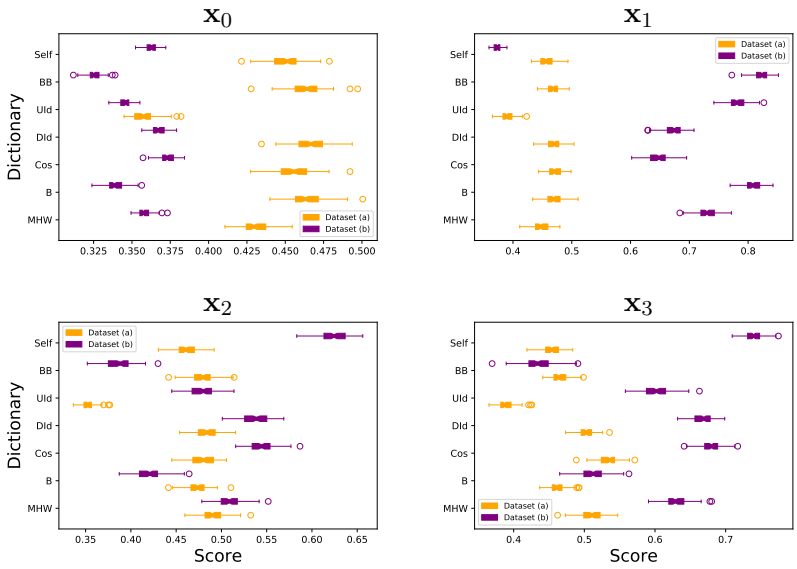


Figure A.9 – Boxplot (over 100 repetitions) of the FIF score for the observations $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for different dictionaries using the L_2 scalar product of the derivatives. The orange boxplots represent the data set (a) while the purple boxplots represent the data set (b).

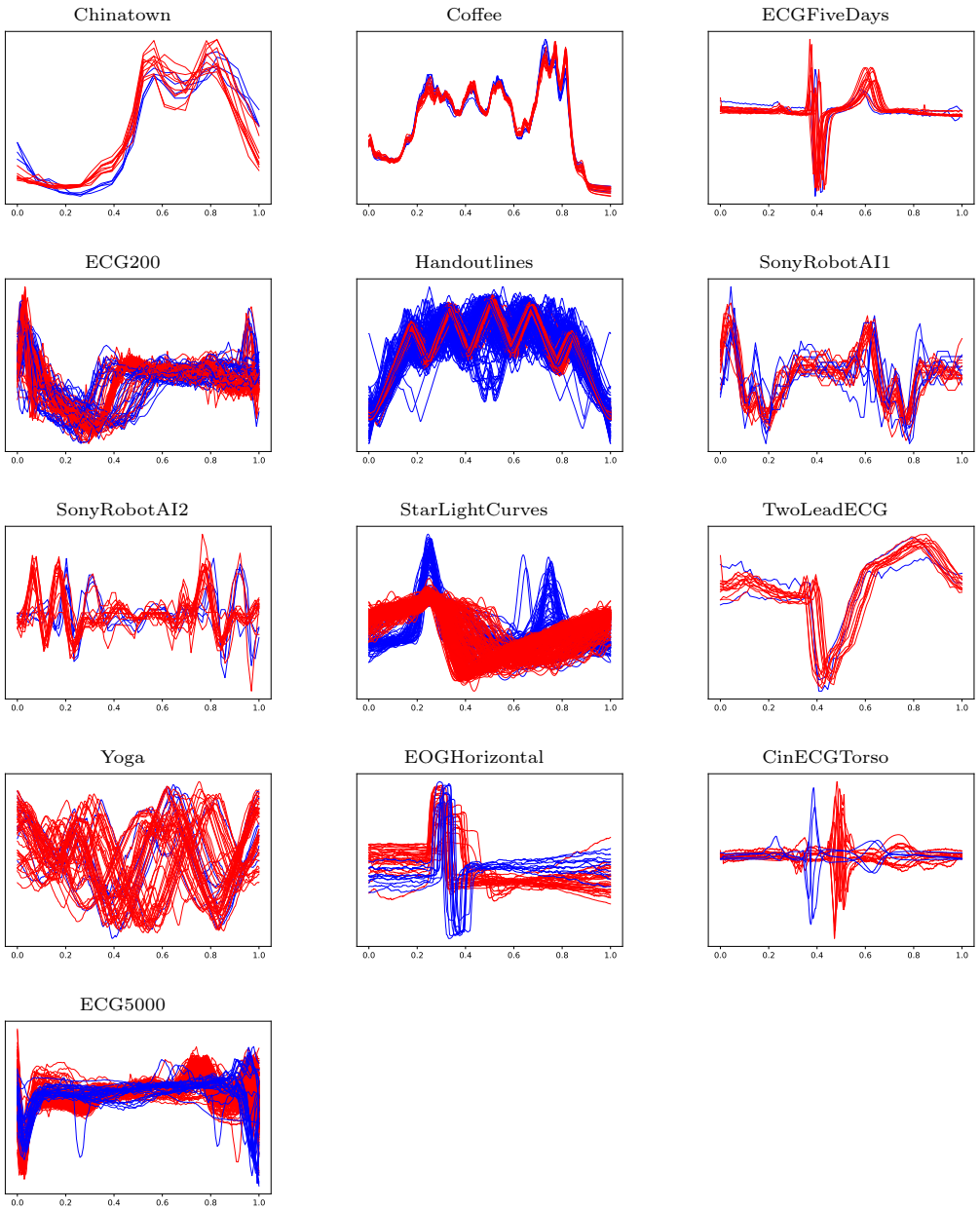


Figure A.10 – Benchmark data sets in the anomaly detection experiment for Chapter 5.

A.2 Complementary Results on the Performance Comparison

Here, complementary results of the benchmark displayed in Section 5.3.2 are provided.

A.2.1 Benchmark Data Sets

In Figure A.10, we plot the thirteen benchmark train data sets used in the experiment. Anomalies are represented by blue color while normal data are drawn in red.

A.2.2 Functional Depth

In this part, we present further 8 functional depths which are outperformed (on an average) by the two depth functions presented in Section 5.3.2 on the 13 real-world data sets and we display their AUC performance. SFD corresponds to *simplicial integrated depth*, HFD to *Halfspace integrated depth*, RP-SD to the *random projection method with simplicial depth*, RP-RHD to the *random projection method with random halfspace depth*, fAO to the *functional adjusted outlyingness*, fDO to the *functional directional outlyingness* to and fbd to *functional bagdistance*. The reader is referred to Cuevas et al. (2007); Fraiman and Muniz (2001); Hubert et al. (2015) for the bibliography on employed functional data depth notions.

Data sets	SFD	HFD	Modal	RP-SD	RP-RHD	fAO	fDO	fbd
Chinatown	0.74	0.77	0.75	0.77	0.73	0.70	0.83	0.83
Coffee	0.60	0.59	0.69	0.70	0.51	0.53	0.59	0.60
ECGFiveDays	0.65	0.64	0.60	0.64	0.56	0.72	0.76	0.80
ECG200	0.82	0.82	0.84	0.85	0.74	0.78	0.82	0.82
Handoutlines	0.70	0.70	0.75	0.72	0.63	0.60	0.71	0.73
SonyRobotAI1	0.89	0.89	0.94	0.83	0.62	0.90	0.93	0.93
SonyRobotAI2	0.82	0.82	0.92	0.86	0.71	0.80	0.82	0.80
StarLightCurves	0.80	0.80	0.85	0.78	0.68	0.80	0.82	0.83
TwoLeadECG	0.68	0.67	0.68	0.66	0.60	0.67	0.69	0.69
Yoga	0.55	0.53	0.57	0.57	0.54	0.54	0.55	0.56
EOGHorizontal	0.59	0.52	0.84	0.74	0.64	0.53	0.59	0.66
CinECGTorso	0.69	0.69	0.73	0.62	0.66	0.85	0.83	0.79
ECG5000	0.90	0.90	0.92	0.92	0.84	0.87	0.92	0.92

A.2.3 Isolation Forest after Dimension Reduction by Filtering Methods on the Benchmark Data Sets

Here, we show the results of the filtering approach using 106 bases from the `PyWavelets` python library and the Fourier basis. Afterwards, we apply (multivariate) Isolation Forest on the coefficients of the projections and display the AUC performance.

data sets	Fourier	bior1.1	bior1.3	bior1.5	bior2.2	bior2.4	bior2.6	bior2.8	bior3.1	bior3.3	bior3.5	bior3.7	bior3.9	bior4.4	bior5.5	bior6.8	coif1	coif2	coif3	coif4	coif5	coif6
Chinatow	0.77	0.92	0.87	0.89	0.93	0.90	0.92	0.95	0.64	0.94	0.90	0.95	0.93	0.93	0.97	0.93	0.91	0.94	0.97	0.96	0.97	0.93
Coffee	0.49	0.56	0.67	0.67	0.60	0.56	0.69	0.60	0.51	0.53	0.65	0.53	0.47	0.47	0.62	0.65	0.51	0.76	0.50	0.71	0.54	0.69
ECGFiveDays	0.58	0.78	0.78	0.80	0.73	0.72	0.73	0.80	0.58	0.68	0.67	0.82	0.70	0.67	0.75	0.75	0.75	0.75	0.82	0.69	0.75	0.69
ECG200	0.46	0.70	0.72	0.68	0.57	0.53	0.59	0.66	0.44	0.46	0.50	0.54	0.53	0.52	0.59	0.68	0.61	0.60	0.63	0.54	0.69	0.68
Handoutlines	0.50	0.74	0.77	0.77	0.56	0.59	0.55	0.56	0.57	0.51	0.52	0.54	0.51	0.58	0.47	0.49	0.57	0.52	0.53	0.56	0.56	0.55
SonyRobotAI1	0.98	0.95	0.95	0.97	0.98	0.97	0.98	0.97	0.96	0.96	0.97	0.96	0.97	0.97	0.94	0.96	0.97	0.98	0.98	0.95	0.96	0.97
SonyRobotAI2	0.89	0.81	0.80	0.78	0.81	0.84	0.81	0.84	0.83	0.81	0.81	0.80	0.84	0.83	0.83	0.84	0.82	0.83	0.87	0.83	0.85	0.78
StarLightCurves	0.46	0.68	0.69	0.70	0.58	0.55	0.58	0.58	0.53	0.54	0.55	0.54	0.62	0.57	0.54	0.60	0.55	0.57	0.57	0.62	0.63	0.69
TwoLeadECG	0.52	0.64	0.63	0.66	0.57	0.56	0.58	0.59	0.55	0.57	0.58	0.56	0.59	0.59	0.56	0.62	0.54	0.65	0.58	0.62	0.59	0.60
Yoga	0.63	0.59	0.59	0.61	0.58	0.59	0.58	0.60	0.61	0.61	0.61	0.61	0.62	0.59	0.59	0.60	0.58	0.57	0.60	0.61	0.61	0.61
EOGHorizontal	0.44	0.61	0.60	0.56	0.61	0.59	0.60	0.61	0.62	0.63	0.63	0.62	0.60	0.66	0.63	0.64	0.63	0.64	0.62	0.62	0.63	0.63
CinECGTorso	0.28	0.25	0.24	0.20	0.17	0.16	0.15	0.17	0.15	0.17	0.16	0.15	0.14	0.14	0.15	0.14	0.14	0.16	0.17	0.16	0.15	0.15
ECG5000	0.65	0.85	0.87	0.89	0.72	0.76	0.77	0.82	0.65	0.66	0.71	0.73	0.81	0.77	0.75	0.78	0.78	0.75	0.79	0.84	0.80	0.82

data sets	coif7	coif8	coif9	coif10	coif11	coif12	coif13	coif14	coif15	coif16	coif17	db1	db2	db3	db4	db5	db6	db7	db8	db9	db10	db11
Chinatow	0.95	0.95	0.94	0.93	0.93	0.95	0.98	0.95	0.90	0.93	0.91	0.94	0.90	0.92	0.97	0.89	0.92	0.94	0.95	0.97	0.94	0.95
Coffee	0.68	0.59	0.49	0.69	0.65	0.59	0.73	0.63	0.65	0.56	0.51	0.65	0.51	0.67	0.60	0.37	0.56	0.47	0.55	0.56	0.63	0.63
ECGFiveDays	0.67	0.71	0.68	0.71	0.70	0.77	0.83	0.77	0.79	0.79	0.80	0.70	0.78	0.80	0.82	0.70	0.79	0.73	0.74	0.74	0.73	0.66
ECG200	0.72	0.71	0.84	0.86	0.85	0.85	0.85	0.84	0.85	0.85	0.86	0.63	0.59	0.57	0.55	0.65	0.57	0.63	0.64	0.62	0.69	0.70
Handoutlines	0.47	0.56	0.55	0.57	0.57	0.55	0.60	0.57	0.59	0.51	0.61	0.74	0.55	0.58	0.58	0.54	0.52	0.57	0.52	0.51	0.53	0.60
SonyRobotAI1	0.92	0.94	0.96	0.91	0.91	0.92	0.90	0.93	0.89	0.94	0.90	0.97	0.98	0.98	0.97	0.98	0.96	0.97	0.97	0.95	0.96	0.97
SonyRobotAI2	0.79	0.80	0.82	0.79	0.79	0.79	0.80	0.78	0.83	0.78	0.83	0.79	0.82	0.81	0.87	0.81	0.86	0.83	0.87	0.87	0.85	0.90
StarLightCurves	0.68	0.71	0.64	0.72	0.76	0.77	0.76	0.75	0.75	0.77	0.75	0.70	0.57	0.53	0.53	0.60	0.52	0.54	0.61	0.61	0.65	0.57
TwoLeadECG	0.65	0.67	0.68	0.69	0.64	0.76	0.71	0.73	0.73	0.67	0.69	0.58	0.58	0.57	0.57	0.57	0.62	0.57	0.53	0.60	0.62	0.63
Yoga	0.62	0.59	0.59	0.59	0.62	0.62	0.61	0.61	0.61	0.60	0.59	0.58	0.57	0.58	0.58	0.60	0.60	0.60	0.59	0.60	0.61	0.61
EOGHorizontal	0.60	0.61	0.65	0.56	0.68	0.59	0.60	0.64	0.62	0.60	0.59	0.63	0.61	0.62	0.64	0.65	0.62	0.61	0.60	0.63	0.64	0.57
CinECGTorso	0.15	0.17	0.18	0.17	0.17	0.16	0.17	0.17	0.19	0.17	0.16	0.23	0.14	0.16	0.15	0.18	0.16	0.17	0.14	0.17	0.16	0.18
ECG5000	0.86	0.85	0.85	0.83	0.87	0.92	0.91	0.93	0.92	0.91	0.93	0.85	0.79	0.75	0.71	0.69	0.78	0.74	0.73	0.73	0.74	0.72

data sets	db12	db13	db14	db15	db16	db17	db18	db19	db20	db21	db22	db23	db24	db25	db26	db27	db28	db29	db30	db31	db32	db33
Chinatow	0.95	0.93	0.95	0.92	0.96	0.92	0.96	0.95	0.92	0.96	0.94	0.94	0.95	0.97	0.97	0.94	0.94	0.97	0.97	0.96	0.96	0.93
Coffee	0.53	0.65	0.62	0.71	0.55	0.63	0.58	0.51	0.51	0.64	0.51	0.63	0.53	0.45	0.73	0.55	0.47	0.62	0.68	0.50	0.54	0.56
ECGFiveDays	0.69	0.65	0.66	0.70	0.70	0.68	0.73	0.68	0.64	0.67	0.62	0.66	0.68	0.70	0.72	0.68	0.66	0.67	0.66	0.64	0.67	0.65
ECG200	0.65	0.72	0.72	0.72	0.75	0.72	0.71	0.73	0.74	0.70	0.74	0.73	0.71	0.84	0.84	0.84	0.84	0.84	0.85	0.84	0.84	0.82
Handoutlines	0.55	0.55	0.52	0.53	0.55	0.57	0.55	0.54	0.56	0.50	0.61	0.57	0.55	0.51	0.53	0.62	0.51	0.54	0.51	0.55	0.58	0.56
SonyRobotAI1	0.97	0.99	0.98	0.98	0.97	0.98	0.98	0.93	0.90	0.92	0.94	0.94	0.96	0.92	0.92	0.91	0.97	0.94	0.92	0.96	0.96	0.91
SonyRobotAI2	0.94	0.90	0.88	0.90	0.89	0.76	0.86	0.80	0.79	0.82	0.78	0.80	0.79	0.81	0.80	0.75	0.78	0.77	0.80	0.82	0.83	0.80
StarLightCurves	0.54	0.66	0.67	0.66	0.67	0.68	0.66	0.69	0.68	0.67	0.70	0.70	0.65	0.68	0.69	0.64	0.66	0.68	0.70	0.72	0.68	0.76
TwoLeadECG	0.62	0.66	0.63	0.62	0.62	0.62	0.62	0.66	0.62	0.59	0.65	0.66	0.67	0.65	0.66	0.67	0.70	0.61	0.74	0.74	0.68	0.68
Yoga	0.58	0.59	0.61	0.61	0.59	0.61	0.59	0.59	0.60	0.59	0.61	0.59	0.60	0.59	0.59	0.60	0.61	0.60	0.63	0.60	0.61	0.60
EOGHorizontal	0.63	0.63	0.65	0.67	0.57	0.59	0.61	0.68	0.63	0.60	0.61	0.63	0.64	0.65	0.65	0.65	0.65	0.61	0.61	0.66	0.63	0.59
CinECGTorso	0.16	0.16	0.19	0.17	0.18	0.16	0.14	0.17	0.16	0.17	0.17	0.15	0.17	0.16	0.16	0.18	0.18	0.16	0.18	0.17	0.17	0.18
ECG5000	0.77	0.73	0.79	0.76	0.77	0.79	0.78	0.86	0.81	0.85	0.83	0.77	0.84	0.84	0.85	0.82	0.84	0.82	0.83	0.81	0.81	0.81

data sets	db34	db35	db36	db37	db38	dmey	haar	rbio1.1	rbio1.3	rbio1.5	rbio2.2	rbio2.4	rbio2.6	rbio2.8	rbio3.1	rbio3.3	rbio3.5	rbio3.7	rbio3.9	rbio4.4	rbio5.5	rbio6.8
Chinatow	0.98	0.95	0.93	0.92	0.95	0.94	0.89	0.92	0.91	0.93	0.93	0.90	0.96	0.97	0.97	0.96	0.95	0.96	0.97	0.88	0.95	0.95
Coffee	0.55	0.62	0.55	0.62	0.51	0.55	0.56	0.74	0.64	0.47	0.64	0.56	0.56	0.50	0.72	0.59	0.60	0.53	0.67	0.41	0.55	0.60
ECGFiveDays	0.67	0.76	0.79	0.82	0.74	0.72	0.74	0.76	0.81	0.73	0.80	0.79	0.75	0.76	0.89	0.85	0.83	0.84	0.79	0.73	0.72	0.83
ECG200	0.84	0.85	0.83	0.82	0.86	0.84	0.64	0.71	0.55	0.60	0.67	0.58	0.66	0.62	0.75	0.73	0.62	0.64	0.65	0.63	0.60	0.57
Handoutlines	0.55	0.60	0.57	0.50	0.55	0.58	0.73	0.71	0.55	0.55	0.57	0.54	0.55	0.58	0.73	0.56	0.56	0.56	0.54	0.56	0.53	0.53
SonyRobotAI1	0.93	0.95	0.96	0.92	0.95	0.94	0.97	0.96	0.96	0.97	0.95	0.97	0.97	0.96	0.90	0.97	0.97	0.97	0.96	0.95	0.95	0.97
SonyRobotAI2	0.77	0.79	0.81	0.85	0.80	0.77	0.82	0.82	0.83	0.84	0.81	0.81	0.78	0.83	0.80	0.79	0.86	0.80	0.88	0.85	0.82	0.80
StarLightCurves	0.75	0.73	0.76	0.74	0.76	0.81	0.70	0.70	0.54	0.56	0.56	0.56	0.58	0.62	0.71	0.53	0.55	0.60	0.60	0.56	0.57	0.51
TwoLeadECG	0.70	0.73	0.69	0.69	0.69	0.67	0.53	0.59	0.61	0.59	0.64	0.58	0.58	0.58	0.68	0.63	0.59	0.57	0.58	0.60	0.60	0.60
Yoga	0.60	0.60	0.62	0.61	0.62	0.60	0.59	0.60	0.59	0.59	0.58	0.57	0.59	0.59	0.61	0.58	0.58	0.57	0.58	0.60	0.60	0.58
EOGHorizontal	0.63	0.62	0.59	0.64	0.62	0.67	0.60	0.55	0.63	0.58	0.59	0.64	0.64	0.63	0.63	0.64	0.56	0.63	0.64	0.57	0.61	0.59
CinECGTorso	0.19	0.17	0.18	0.16	0.19	0.18	0.25	0.23	0.16	0.18	0.15	0.16	0.15	0.15	0.28	0.14	0.16	0.17	0.19	0.15	0.15	0.17
ECG5000	0.83	0.79	0.93	0.92	0.93	0.84	0.86	0.85	0.75	0.73	0.82	0.81	0.79	0.83	0.90	0.82	0.83	0.79	0.83	0.78	0.83	0.83

data sets	sym2	sym3	sym4	sym5	sym6	sym7	sym8	sym9	sym10	sym11	sym12	sym13	sym14	sym15	sym16	sym17	sym18	sym19	sym20
Chinatow	0.86	0.90	0.95	0.89	0.83	0.97	0.96	0.97	0.88	0.94	0.95	0.89	0.94	0.96	0.94	0.95	0.94	0.94	0.95
Coffee	0.50	0.55	0.46	0.63	0.51	0.46	0.59	0.54	0.59	0.41	0.50	0.46	0.58	0.51	0.62	0.64	0.56	0.46	0.47
ECGFiveDays	0.80	0.75	0.75	0.76	0.77	0.76	0.80	0.87	0.74	0.68	0.75	0.74	0.72	0.77	0.72	0.75	0.72	0.71	0.69
ECG200	0.58	0.52	0.54	0.54	0.49	0.58	0.62	0.58	0.57	0.59	0.63	0.65	0.71	0.72	0.74	0.74	0.67	0.64	0.70
Handoutlines	0.63	0.59	0.49	0.55	0.56	0.57	0.57	0.53	0.53	0.55	0.56	0.56	0.51	0.60	0.53	0.51	0.51	0.54	0.54
SonyRobotAI1	0.97	0.98	0.94	0.98	0.97	0.97	0.96	0.95	0.96	0.97	0.98	0.95	0.96	0.96	0.94	0.96	0.96	0.94	0.96
SonyRobotAI2	0.86	0.86	0.81	0.87	0.86	0.84	0.86	0.86	0.83	0.86	0.89	0.89	0.87	0.86	0.89	0.81	0.79	0.83	0.80
StarLightCurves	0.57	0.53	0.55	0.58	0.59	0.53	0.58	0.66	0.62	0.64	0.64	0.63	0.59	0.58	0.59	0.71	0.67	0.71	0.65
TwoLeadECG	0.58	0.57	0.55	0.56	0.63	0.62	0.55	0.57	0.58	0.65	0.59	0.64	0.62	0.68	0.62	0.61	0.60	0.66	0.57
Yoga	0.58	0.58	0.59	0.60	0.60	0.59	0.59	0.58	0.60	0.60	0.61	0.59	0.60	0.59	0.60	0.62	0.62	0.61	0.61
EOGHorizontal	0.60	0.66	0.62	0.55	0.61	0.63	0.63	0.61	0.61	0.65	0.67	0.68	0.58	0.67	0.64	0.67	0.71	0.64	0.60
CinECGTorso	0.16	0.16	0.15	0.15	0.17	0.15	0.16	0.16	0.17	0.15	0.17	0.16	0.16	0.15	0.16	0.15	0.17	0.17	0.17
ECG5000	0.79	0.75	0.81	0.77	0.81	0.72	0.79	0.78	0.77	0.79	0.84	0.83	0.82	0.80	0.84	0.84	0.86	0.81	0.83

A.2.4 IF$_{FPCA}$ with Different Filtering Preliminary Step on the Benchmark Data Sets

Here, we show the results of the FPCA approach using the Fourier basis and 106 further bases from the `PyWavelets` python library as preliminary filtering stage. Afterwards, we apply (multivariate) Isolation Forest on the coefficients of the projections and display the AUC performance.

data sets	bior1.1	bior1.3	bior1.5	bior2.2	bior2.4	bior2.6	bior2.8	bior3.1	bior3.3	bior3.5	bior3.7	bior3.9	bior4.4	bior5.5	bior6.8	coif1	coif2	coif3	coif4	coif5	coif6	coif7
Chinatow	0.69	0.69	0.68	0.78	0.73	0.75	0.66	0.74	0.69	0.77	0.65	0.68	0.69	0.62	0.70	0.66	0.68	0.71	0.80	0.73	0.68	0.71
Coffee	0.60	0.54	0.47	0.53	0.45	0.62	0.44	0.38	0.49	0.55	0.64	0.49	0.58	0.56	0.45	0.53	0.51	0.53	0.45	0.42	0.59	0.58
ECGFiveDays	0.81	0.79	0.83	0.89	0.85	0.87	0.90	0.85	0.82	0.88	0.89	0.86	0.85	0.83	0.87	0.91	0.88	0.83	0.80	0.80	0.84	0.79
ECG200	0.80	0.81	0.78	0.76	0.75	0.71	0.78	0.78	0.73	0.70	0.75	0.77	0.73	0.71	0.71	0.81	0.78	0.80	0.76	0.78	0.77	0.81
Handoutline	0.68	0.68	0.68	0.72	0.66	0.70	0.72	0.69	0.78	0.70	0.68	0.69	0.70	0.68	0.70	0.70	0.72	0.73	0.68	0.71	0.71	0.70
SonyRobotAI1	0.79	0.71	0.81	0.77	0.80	0.76	0.76	0.77	0.82	0.84	0.85	0.85	0.82	0.81	0.82	0.82	0.87	0.79	0.85	0.80	0.79	0.88
SonyRobotAI2	0.86	0.78	0.89	0.80	0.80	0.85	0.89	0.80	0.82	0.81	0.84	0.87	0.84	0.87	0.89	0.83	0.85	0.84	0.85	0.87	0.85	0.88
StarLightCurves	0.76	0.78	0.76	0.74	0.73	0.77	0.76	0.62	0.69	0.78	0.75	0.76	0.77	0.74	0.75	0.75	0.74	0.76	0.77	0.75	0.74	0.70
TwoLeadECG	0.71	0.62	0.69	0.65	0.60	0.72	0.68	0.62	0.61	0.71	0.66	0.78	0.65	0.66	0.66	0.72	0.65	0.72	0.61	0.55	0.67	0.73
Yoga	0.57	0.55	0.57	0.57	0.56	0.57	0.58	0.57	0.55	0.54	0.56	0.59	0.58	0.58	0.57	0.57	0.56	0.59	0.59	0.57	0.57	0.58
EOGHorizontal	0.70	0.72	0.71	0.76	0.74	0.66	0.72	0.67	0.83	0.74	0.68	0.76	0.71	0.65	0.70	0.66	0.63	0.69	0.66	0.74	0.72	0.69
CinECGTorso	0.51	0.46	0.46	0.32	0.42	0.28	0.49	0.32	0.48	0.43	0.39	0.35	0.43	0.38	0.39	0.53	0.50	0.39	0.50	0.67	0.44	0.49
ECG5000	0.96	0.94	0.94	0.94	0.95	0.95	0.94	0.93	0.95	0.95	0.93	0.94	0.94	0.95	0.95	0.94	0.95	0.96	0.95	0.95	0.95	0.95

data sets	coif8	coif9	coif10	coif11	coif12	coif13	coif14	coif15	coif16	coif17	db1	db2	db3	db4	db5	db6	db7	db8	db9	db10	db11	db12
Chinatow	0.68	0.67	0.64	0.62	0.71	0.73	0.76	0.63	0.76	0.71	0.72	0.69	0.71	0.72	0.75	0.69	0.73	0.75	0.71	0.69	0.73	0.66
Coffee	0.74	0.58	0.46	0.46	0.54	0.49	0.38	0.41	0.44	0.53	0.54	0.44	0.58	0.47	0.41	0.56	0.51	0.58	0.53	0.55	0.49	0.44
ECGFiveDays	0.84	0.87	0.85	0.85	0.88	0.83	0.83	0.87	0.82	0.78	0.83	0.84	0.84	0.85	0.85	0.83	0.87	0.87	0.80	0.89	0.86	0.88
ECG200	0.79	0.79	0.84	0.82	0.82	0.84	0.80	0.81	0.82	0.82	0.83	0.80	0.81	0.77	0.84	0.81	0.75	0.75	0.75	0.76	0.81	0.76
Handoutlines	0.68	0.72	0.77	0.74	0.76	0.71	0.72	0.70	0.64	0.69	0.66	0.78	0.70	0.72	0.68	0.72	0.72	0.66	0.72	0.75	0.73	0.71
SonyRobotAI1	0.85	0.85	0.87	0.89	0.81	0.88	0.83	0.86	0.85	0.79	0.79	0.78	0.78	0.78	0.80	0.83	0.80	0.88	0.87	0.88	0.90	0.81
SonyRobotAI2	0.84	0.86	0.85	0.86	0.86	0.84	0.86	0.82	0.86	0.84	0.88	0.79	0.86	0.79	0.85	0.84	0.88	0.84	0.85	0.78	0.89	0.86
StarLightCurves	0.74	0.73	0.73	0.72	0.73	0.74	0.72	0.71	0.70	0.71	0.75	0.75	0.77	0.78	0.75	0.75	0.79	0.77	0.75	0.78	0.77	0.75
TwoLeadECG	0.60	0.66	0.52	0.63	0.60	0.59	0.62	0.65	0.60	0.55	0.67	0.68	0.67	0.66	0.67	0.72	0.66	0.64	0.63	0.55	0.57	0.66
Yoga	0.60	0.58	0.58	0.58	0.60	0.59	0.60	0.59	0.59	0.58	0.56	0.56	0.57	0.56	0.57	0.56	0.58	0.56	0.56	0.57	0.59	0.60
EOGHorizontal	0.71	0.68	0.70	0.70	0.70	0.68	0.66	0.69	0.67	0.68	0.69	0.78	0.72	0.71	0.69	0.72	0.67	0.72	0.72	0.75	0.75	0.70
CinECGTorso	0.52	0.55	0.46	0.34	0.54	0.45	0.55	0.46	0.48	0.46	0.47	0.43	0.50	0.62	0.39	0.49	0.44	0.40	0.46	0.40	0.46	0.39
ECG5000	0.94	0.94	0.94	0.95	0.95	0.95	0.96	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.94

data sets	db13	db14	db15	db16	db17	db18	db19	db20	db21	db22	db23	db24	db25	db26	db27	db28	db29	db30	db31	db32	db33	db34
Chinatow	0.72	0.62	0.74	0.71	0.76	0.75	0.68	0.76	0.69	0.76	0.78	0.70	0.74	0.65	0.80	0.69	0.65	0.74	0.72	0.74	0.73	0.76
Coffee	0.51	0.42	0.38	0.42	0.47	0.49	0.54	0.46	0.37	0.46	0.40	0.58	0.62	0.45	0.53	0.33	0.40	0.53	0.55	0.42	0.49	0.41
ECGFiveDays	0.83	0.86	0.83	0.89	0.80	0.87	0.81	0.77	0.81	0.79	0.82	0.79	0.84	0.85	0.86	0.85	0.72	0.84	0.86	0.80	0.81	0.79
ECG200	0.81	0.75	0.80	0.79	0.76	0.80	0.81	0.83	0.86	0.81	0.78	0.78	0.82	0.77	0.81	0.81	0.82	0.77	0.83	0.80	0.81	0.78
Handoutlines	0.74	0.76	0.73	0.77	0.73	0.74	0.71	0.72	0.75	0.72	0.74	0.74	0.74	0.78	0.75	0.70	0.69	0.73	0.74	0.74	0.72	0.72
SonyRobotAI1	0.82	0.83	0.79	0.82	0.84	0.84	0.87	0.85	0.84	0.82	0.84	0.84	0.84	0.85	0.83	0.86	0.87	0.83	0.81	0.79	0.82	0.82
SonyRobotAI2	0.86	0.88	0.87	0.88	0.82	0.83	0.85	0.85	0.85	0.83	0.83	0.87	0.87	0.86	0.84	0.86	0.86	0.83	0.82	0.85	0.87	0.86
StarLightCurves	0.78	0.77	0.73	0.77	0.74	0.75	0.74	0.75	0.73	0.72	0.73	0.72	0.72	0.73	0.73	0.70	0.71	0.73	0.74	0.70	0.71	0.72
TwoLeadECG	0.71	0.63	0.70	0.72	0.62	0.58	0.65	0.70	0.48	0.64	0.73	0.62	0.71	0.64	0.61	0.57	0.63	0.62	0.65	0.64	0.62	0.58
Yoga	0.59	0.57	0.59	0.59	0.58	0.59	0.59	0.58	0.59	0.59	0.59	0.59	0.61	0.58	0.59	0.59	0.58	0.58	0.57	0.60	0.58	0.59
EOGHorizontal	0.73	0.66	0.75	0.75	0.73	0.70	0.77	0.69	0.72	0.74	0.71	0.72	0.67	0.74	0.73	0.68	0.75	0.71	0.73	0.68	0.72	0.72
CinECGTorso	0.40	0.50	0.52	0.52	0.52	0.47	0.50	0.45	0.41	0.54	0.52	0.52	0.37	0.47	0.51	0.44	0.51	0.48	0.45	0.53	0.44	0.38
ECG5000	0.95	0.95	0.95	0.94	0.94	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95

data sets	db35	db36	db37	db38	dmey	haar	rbio1.1	rbio1.3	rbio1.5	rbio2.2	rbio2.4	rbio2.6	rbio2.8	rbio3.1	rbio3.3	rbio3.5	rbio3.7	rbio3.9	rbio4.4	rbio5.5	rbio6.8	sym2
Chinatow	0.71	0.69	0.77	0.74	0.74	0.71	0.71	0.69	0.72	0.78	0.74	0.72	0.81	0.75	0.73	0.74	0.61	0.72	0.79	0.67	0.81	0.66
Coffee	0.50	0.38	0.41	0.50	0.54	0.53	0.58	0.46	0.50	0.45	0.67	0.49	0.63	0.33	0.42	0.49	0.56	0.49	0.54	0.53	0.54	0.47
ECGFiveDays	0.84	0.85	0.83	0.84	0.82	0.88	0.86	0.88	0.81	0.85	0.65	0.90	0.82	0.92	0.82	0.79	0.87	0.84	0.88	0.77	0.78	0.81
ECG200	0.78	0.80	0.78	0.83	0.79	0.78	0.80	0.76	0.78	0.72	0.68	0.70	0.72	0.81	0.71	0.76	0.75	0.72	0.75	0.73	0.82	0.79
Handoutlines	0.72	0.72	0.70	0.74	0.75	0.70	0.65	0.70	0.71	0.75	0.69	0.71	0.73	0.74	0.71	0.73	0.73	0.67	0.72	0.70	0.68	0.75
SonyRobotAI1	0.86	0.87	0.83	0.84	0.87	0.74	0.80	0.87	0.80	0.86	0.76	0.81	0.79	0.81	0.77	0.85	0.79	0.84	0.79	0.80	0.86	0.80
SonyRobotAI2	0.82	0.85	0.86	0.81	0.84	0.88	0.86	0.86	0.85	0.83	0.85	0.86	0.85	0.81	0.88	0.92	0.87	0.88	0.83	0.85	0.81	0.86
StarLightCurves	0.75	0.70	0.73	0.73	0.73	0.77	0.76	0.79	0.77	0.77	0.75	0.74	0.78	0.72	0.76	0.73	0.74	0.75	0.77	0.76	0.73	0.74
TwoLeadECG	0.69	0.58	0.62	0.67	0.57	0.70	0.66	0.68	0.71	0.63	0.63	0.67	0.66	0.66	0.61	0.73	0.71	0.62	0.62	0.64	0.66	0.63
Yoga	0.57	0.60	0.59	0.56	0.59	0.58	0.59	0.59	0.59	0.58	0.58	0.58	0.59	0.56	0.60	0.58	0.60	0.57	0.55	0.56	0.59	0.56
EOGHorizontal	0.74	0.71	0.71	0.70	0.68	0.69	0.70	0.64	0.74	0.68	0.65	0.66	0.68	0.71	0.75	0.72	0.68	0.67	0.69	0.76	0.72	0.74
CinECGTorso	0.42	0.48	0.38	0.51	0.47	0.48	0.53	0.42	0.44	0.41	0.36	0.35	0.41	0.30	0.39	0.30	0.49	0.36	0.41	0.42	0.42	0.44
ECG5000	0.95	0.95	0.96	0.95	0.95	0.95	0.94	0.94	0.95	0.96	0.94	0.95	0.95	0.95	0.94	0.93	0.94	0.95	0.95	0.95	0.95	0.95

data sets	sym3	sym4	sym5	sym6	sym7	sym8	sym9	sym10	sym11	sym12	sym13	sym14	sym15	sym16	sym17	sym18	sym19	sym20
Chinatow	0.67	0.70	0.74	0.67	0.80	0.72	0.70	0.74	0.76	0.65	0.79	0.74	0.71	0.83	0.69	0.62	0.77	0.66
Coffee	0.65	0.62	0.64	0.63	0.49	0.50	0.41	0.38	0.60	0.46	0.45	0.56	0.55	0.56	0.46	0.46	0.56	0.55
ECGFiveDays	0.90	0.81	0.81	0.81	0.90	0.89	0.81	0.75	0.78	0.79	0.80	0.88	0.83	0.82	0.81	0.83	0.78	0.85
ECG200	0.76	0.76	0.81	0.72	0.76	0.77	0.72	0.78	0.77	0.78	0.82	0.79	0.81	0.78	0.77	0.81	0.83	0.80
Handoutlines	0.75	0.73	0.70	0.74	0.71	0.74	0.69	0.71	0.73	0.72	0.73	0.70	0.69	0.74	0.72	0.75	0.72	0.72
SonyRobotAI1	0.75	0.85	0.82	0.76	0.82	0.76	0.86	0.86	0.83	0.83	0.90	0.86	0.83	0.88	0.90	0.88	0.84	0.83
SonyRobotAI2	0.87	0.86	0.84	0.85	0.86	0.86	0.84	0.85	0.81	0.87	0.87	0.91	0.88	0.85	0.85	0.87	0.85	0.87
StarLightCurves	0.76	0.77	0.77	0.76	0.76	0.75	0.77	0.76	0.77	0.76	0.75	0.74	0.70	0.74	0.76	0.73	0.78	0.71
TwoLeadECG	0.66	0.68	0.58	0.62	0.71	0.69	0.69	0.64	0.66	0.69	0.64	0.67	0.65	0.64	0.68	0.63	0.67	0.70
Yoga	0.58	0.57	0.59	0.60	0.58	0.59	0.58	0.58	0.58	0.59	0.59	0.57	0.57	0.58	0.59	0.58	0.58	0.59
EOGHorizontal	0.72	0.68	0.68	0.70	0.70	0.71	0.67	0.70	0.69	0.71	0.70	0.70	0.69	0.65	0.69	0.68	0.72	0.65
CinECGTorso	0.45	0.44	0.51	0.41	0.49	0.56	0.39	0.48	0.49	0.37	0.46	0.48	0.43	0.58	0.36	0.41	0.48	0.40
ECG5000	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.94	0.95	0.94	0.95	0.95	0.95	0.95	0.95

B Additional Materials for the Benchmark Study

B.1 Additional Experiments on Simulated Anomalies

In this part, complementary experiments to the Section 7.2 are displayed. They are conducted with the same methodology but varying proportion of anomalies: 1% in Table B.4, 2% in Table B.5, 3% in Table B.6 and 4% in Table B.7.

	Isolated		**Magnitude I**		**Magnitude II**		**Shape**	
Methods	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC
FIF	0	0.20	**1**	**1**	0	0.32	**0.06**	**0.98**
fAO	0	0.44	**1**	**1**	0	0.54	0	0.67
fbd	0	0.44	**1**	**1**	0	0.54	0	0.68
fSDO	0	0.42	**1**	**1**	0	0.54	0	0.77
fT	0	0.43	**1**	**1**	0	0.44	0	0.72
ACH	0	**0.62**	**1**	**1**	**1**	**1**	0	0.61
Outliergram	0	0.55	**1**	**1**	0	0.54	0	0.51
MS + IF	0	0.05	**1**	**1**	0	0.77	0	0.77
FOM (fSDO) + IF	0	0.27	**1**	**1**	**1**	**1**	0	0.97
FOM (fAO) + IF	0	0.31	**1**	**1**	0.88	**1**	0	0.96
FPCA + IF	0	0.08	0	0.96	0	0.77	0	0.86
FPCA + LOF	0	0.31	0.18	0.86	0	0.42	0	0.67
FPCA + OC	0	0.03	0	0.93	0	0.78	0	0.88

Table B.4 – Methods considered in performance comparison with the TPR and the Area Under the Receiver Operating Characteristic (AUC) for the four simulated models with 1% of added anomalies. Bold numbers correspond to the best result.

	Isolated		**Magnitude I**		**Magnitude II**		**Shape**	
Methods	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC
FIF	0	0.23	0.97	**1**	0	0.32	**0.26**	**0.98**
fAO	0	0.44	**1**	**1**	0	0.54	0	0.67
fbd	0	0.44	**1**	**1**	0	0.54	0	0.68
fSDO	0	0.42	**1**	**1**	0	0.43	0	0.77
fT	0	0.43	**1**	**1**	0	0.44	0	0.72
ACH	0	**0.62**	**1**	0.83	**0.94**	**1**	0	0.61
Outliergram	0	0.55	**1**	**1**	0	0.54	0	0.49
MS + IF	0	0.05	**1**	**1**	0	0.77	0	0.77
FOM (fSDO) + IF	0	0.22	0.94	**1**	0.86	**1**	0	0.95
FOM (fAO) + IF	0	0.26	0.89	**1**	0.66	0.99	0	0.95
FPCA + IF	0	0.08	0	0.92	0	0.71	0	0.90
FPCA + LOF	0	0.35	0.09	0.81	0	0.48	0.09	0.77
FPCA + OC	0	0.03	0	0.94	0	0.79	0	0.91

Table B.5 – Methods considered in performance comparison with the TPR and the Area Under the Receiver Operating Characteristic (AUC) for the four simulated models with 2% of added anomalies. Bold numbers correspond to the best result.

C Illustration of the Work for Valeo

This section summarizes some aspects of the work done in the production line of Valeo's company without breaking confidentiality of data. This work has been done in the context of the PSPC Expresso project funded by BPI France.

	Isolated		**Magnitude I**		**Magnitude II**		**Shape**	
Methods	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC
FIF	0	0.21	0.98	**1**	0	0.33	**0.49**	**0.98**
fAO	0	0.44	**1**	**1**	0	0.54	0	0.67
fbd	0	0.44	**1**	**1**	0	0.54	0	0.68
fSDO	0	0.42	**1**	**1**	0	0.43	0	0.77
fT	0	0.43	**1**	**1**	0	0.44	0	0.72
ACH	0	**0.60**	**1**	0.85	**0.88**	**1**	0	0.60
Outliergram	0	0.55	**1**	**1**	0	0.54	0	0.49
MS + IF	0	0.05	**1**	**1**	0	0.75	0	0.77
FOM (fSDO) + IF	0	0.06	0.89	**1**	0.64	0.99	0	0.89
FOM (fAO) + IF	0	0.24	0.81	**1**	0.70	0.99	0	0.93
FPCA + IF	0	0.07	0	0.93	0	0.70	0	0.93
FPCA + LOF	0	0.39	0.16	0.87	0	0.57	0.17	0.70
FPCA + OC	0	0.04	0	0.94	0	0.78	0	0.93

Table B.6 – Methods considered in performance comparison with the TPR and the Area Under the Receiver Operating Characteristic (AUC) for the four simulated models with 3% of added anomalies. Bold numbers correspond to the best result.

	Isolated		**Magnitude I**		**Magnitude II**		**Shape**	
Methods	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC
FIF	0	0.23	0.94	**1**	0	0.34	**0.58**	**0.98**
fAO	0	0.44	**1**	**1**	0	0.54	0	0.67
fbd	0	0.44	**1**	**1**	0	0.54	0	0.68
fSDO	0	0.42	**1**	**1**	0	0.43	0	0.77
fT	0	0.43	**1**	**1**	0	0.44	0	0.72
ACH	0	**0.63**	**1**	0.85	**0.85**	**1**	0	0.60
Outliergram	0	0.55	**1**	**1**	0	0.54	0	0.46
MS + IF	0	0.05	**1**	**1**	0	0.74	0	0.77
FOM (fSDO) + IF	0	0	0.80	**1**	0.65	0.99	0	0.87
FOM (fAO) + IF	0	0.10	0.89	**1**	0.55	0.98	0	0.87
FPCA + IF	0	0.08	0	0.92	0	0.72	0.28	0.96
FPCA + LOF	0	0.43	0.26	0.82	0	0.59	0.18	0.70
FPCA + OC	0	0.03	0	0.93	0	0.78	0.08	0.95

Table B.7 – Methods considered in performance comparison with the TPR and the Area Under the Receiver Operating Characteristic (AUC) for the four simulated models with 4% of added anomalies. Bold numbers correspond to the best result.

C.1 Objectives and Context

The aim of the Expresso project was to provide state-of-the-art anomaly detection algorithm for a production line in a factory of the french company Valeo, world-leading global automotive supplier operating in 33 countries, and partnering with automakers worldwide. In a production line from a Valeo's factory, data are collected during the

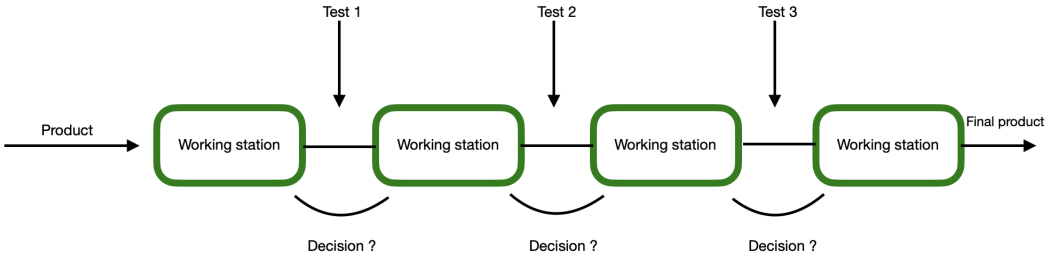


Figure C.11 – Illustration of a production line.

manufacturing process. Typically, a product is characterized by many quantitative variables measured by many "tests" performed on each product during its path on the production line, see Figure C.11 for an illustration. The amount of data from both the number of constructed products and the number of sensors used to measure features of products prevents a human-analysis and requires automatic Machine Learning tools to leverage the collected information. The aim of ML approaches is to help engineers to automatically monitor the production line in order to provide the following benefits: (i) to reduce the number of products with flaws that are considered as sane when leaving the production line and (ii) to detect as fast as possible flaws in a product to avoid useless and expensive operations and thus increase the number of intact items produced per day.

Two settings. Each test performed on the production line measures quantitative information of manufactured products. The number of measured features oscillates between 40 and 200 varying with the considered test. Machine Learning may help the decision-making from two different settings allowing (i) and (ii). On the one hand, data collected in each test can be processed separately leading to a multivariate data set at each step. Therefore, an unsupervised anomaly detection algorithm adapted to multivariate data can be performed at each stage of the production line. On the other hand, important features of products are measured through all tests, information can thus be summarized by a (possibly multivariate) time-series and a decision can be take at the end of the line. In both cases, the decision made in the production line is done by a pre-trained ML model due to resources constraints. ML models that have been used in this project, whether multivariate or functional, are described in Section C.2.

Resources constraints. Computers that operate on the production line possess low computational resources and low memory. In addition, one cannot stop the production line in order to wait a decision whether the product is abnormal or not. Computers have to provide a decision in less than one second. Therefore, the training of the models cannot be done on the computers of the production line, but rather on a computer outside of the production line with high resources, see Figure C.12 for an illustration. Data used for the training are collected and transferred on the high resources computer in order to build models. Roughly, information on one million of products were available without knowing normal or abnormal ones. Once the model is trained, it is transferred to computers that operate on the production line to test the new manufacturing products.

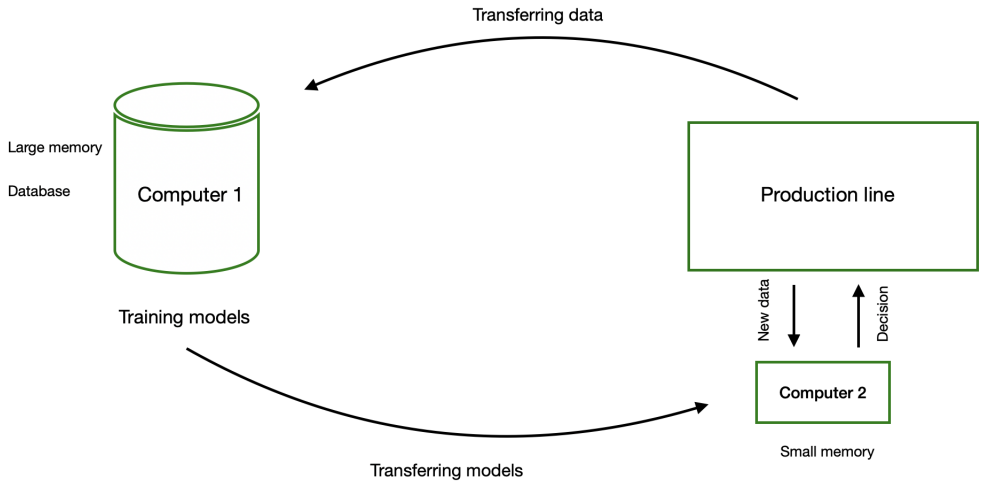


Figure C.12 – Illustration of computational constraints.

C.2 Employed Methods

In the case of functional data, methods employed are those designed in this dissertation and are fully described in Chapters 5 and 6 referring to Staerman et al. (2019) and Staerman et al. (2020) (see also Staerman et al., 2022). In the following, we present two of the used methods: One-Class SVM and Isolation Forest. The projection depth, is fully presented in Chapter 2 and the AI-IRW depth in Staerman et al. (2021b) (see also Staerman et al., 2021c).

One-Class SVM. The One Class Support Vector Machine (OCSVM; Schölkopf et al., 2001) is an algorithm similar to SVM (Support Vector Machine). OCSVM separates anomalies from the rest of the data in an unsupervised fashion. To that end, OCSVM exploits Reproducing Kernel Hilbert Space theory to handle non-linearity (kernel trick) and the Support Vector Machine fashion: once mapped into the feature space corresponding to the kernel, the data are separated from the origin with maximum margin. Thus, a new point x will be predicted as inlier or outlier depending on which side of the hyperplane it falls on, in feature space. Precisely, n being the training sample size, it solves the following optimization problem:

$$\begin{aligned} &\underset{w,\xi,\rho}{\text{minimize}} \quad ||w||_2^2 + \frac{1}{\nu l}\sum_{i=1}^{n}\xi_i - \rho \\ &\text{subject to} \quad \langle w, \phi(x_i)\rangle \leq \rho - \xi_i,\ \xi_i \geq 0, i = 1, \ldots, n \end{aligned}$$

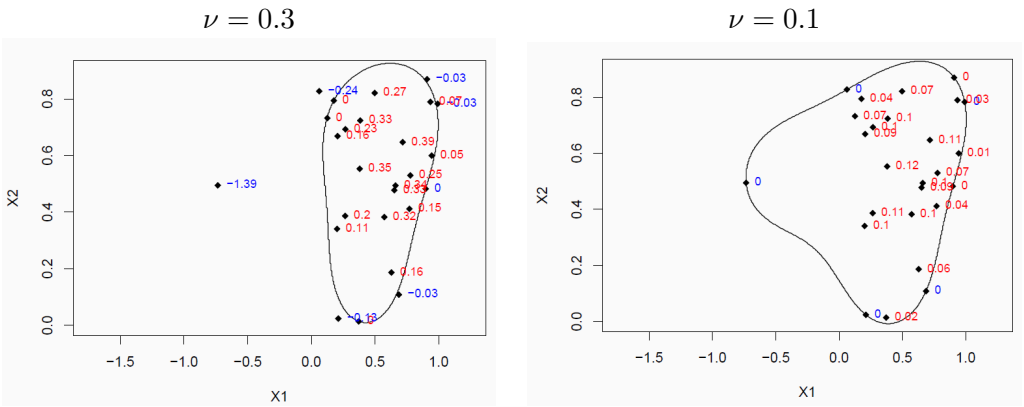


Figure C.13 – OCSVM example on simulated data.

where x_i are the training samples, w are the weights of our model we are looking for, and ϕ corresponds to the feature map but is not computed in practice. As we can see on figure C.13, OCSVM returns a score depending on which side of the hyperplane in feature space a point x is and its distance to this separator hyperplane. Parameter ν is an upper bound on the fraction of outliers and a lower bound on the fraction of Support Vectors, so the smaller ν is, the larger the volume of high density is and the fewer points are considered as outliers. However, this bound is not tight, i.e. if $\nu = 0.1$, it does not mean that 1% of data will be labeled as outliers. This ν works as a prior calibration of OCSVM but one needs to perform a posterior calibration to obtain a more rigorously calibrated detection.

Isolation Forest. Isolation Forest (Liu et al., 2008) is an Ensemble Learning approach (cf. Random Forest). Indeed, the idea here is to isolate outliers by building a binary tree recursively. At each step, a variable and a split value are drawn randomly and all samples are sorted in the left or right leaves whether each variable is greater or smaller than the split value. This process is repeated until all elements of the training set are isolated. The outliers appear then near to tree's root and the anomaly score computed is reciprocal to separation path length.

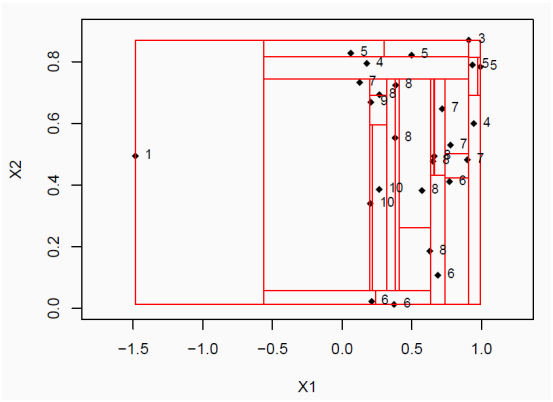


Figure C.14 – Isolation tree example on toy data set.

An advantage of this approach is its scalability since it benefits easily from subsampling and bootstrap aggregating: one can use many classifier trees trained on many subsampling of data and variables, and compute an anomaly score based on their aggreg-

ations. Nonetheless, this method suffers from a lack of explainability, since it relies on randomness and aggregation. Besides, it isolates anomalies based on different features available and so isolates them in hyper-rectangle as illustrated in figure C.14, but the actual distribution of data generally has a more complex shape than a cube and the discriminator factors between normal and anormal data could be combinations of available features and not the raw features. In this case, Extended Isolation Forest (Hariri et al., 2019) may be preferred.

C.3 Some Illustrations on Simulated Data

In this section, we illustrate some of the employed methods on simulated data that are similar to those involved in the project. First of all, we computed OCSVM, IF and the projection depth on the data set and obtained the following score, depicted in Figure C.15.

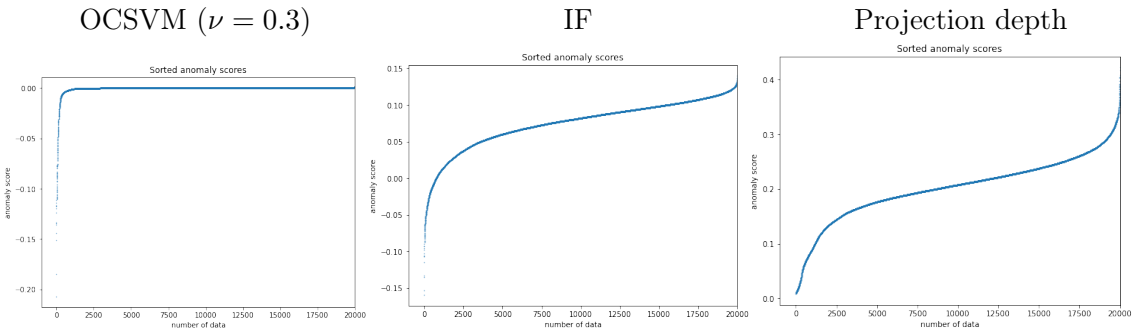


Figure C.15 – Sorted anomaly scores obtained on the data set.

The lower the score is, the more abnormal the point is. Having the score for each data, we need to assess a threshold to label anomalies as a posterior calibration. This choice is thus very important and should be tuned properly. Expert knowledge is primordial, but one can also use visualisation to choose this threshold. A very first idea is to use PCA to visualise data projected on first two components. An example with OCSVM is given with two threshold, as displayed in Figures C.16 and C.17.

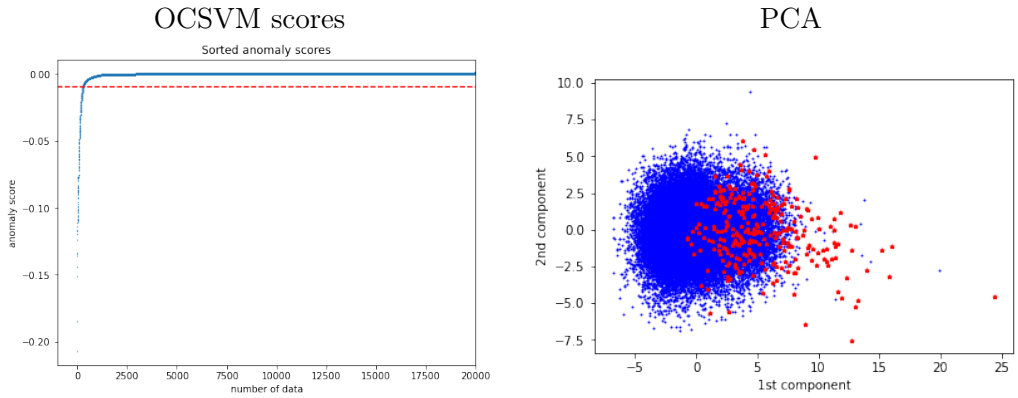


Figure C.16 – Threshold equal to -0.01 leading to 292 outliers.

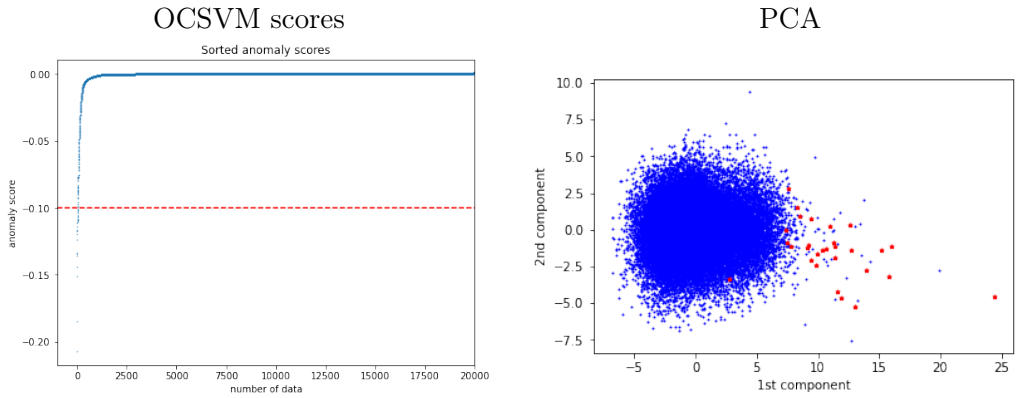


Figure C.17 – Threshold equal to -0.1 leading to 29 outliers.

We note that the most abnormal points in terms of OCSVM score lie outside of the central cloud and most of the points outside of this cloud are points with low scores (but not all of them). PCA gives us a first insight on tuning the threshold but as said earlier, it does not allow to separate well the outliers to distribution's center. Thus, we implemented a relevant visualisation tool for outlier identification and analysis: Anomaly Component Analysis.

Anomaly Component Analysis (ACA). The idea is to find 2 or 3 unit directions defining a relevant and robust space to highlight outliers. To do so, we use the projection depth. We recall the formula of projection depth of x w.r.t. a probability distribution $P \in \mathcal{P}(\mathbb{R}^d)$:

$$D_{\mathrm{P}}(x, P) = \left(1 + \max_{u \in \mathbb{S}^{d-1}} \frac{|x^\top u - \mathrm{med}(X^\top u)|}{\mathrm{MAD}(X^\top u)}\right)^{-1}, \tag{10}$$

We note $f(x,u) = \frac{|x^T u - \mathrm{med}(X^\top u)|}{\mathrm{MAD}(X^\top u)}$. Given a data set $\{x_1, \ldots, x_n\}$, we choose as first direction, the unit direction: $\underset{u \in \mathbb{S}^{d-1}}{\mathrm{argmax}} \ \underset{i \in \{1,\ldots,n\}}{\mathrm{argmax}} \ f(x_i, u)$. The intuition is to choose the most abnormal point of the training set under the projection depth and pick the unit sphere for which its projection is the most abnormal.

For the second direction, we project all data on the hyperspace of dimension $d-1$ orthogonal to the first direction, we sample n_{dir} unit directions on the unit sphere $\mathbb{S}^{d-2}$ in $\mathbb{R}^{d-1}$, and we repeat the exact same operation. We can do the same on the hyperspace of dimension $d-2$ orthogonal to first 2 directions for the 3rd one.

The Figure C.18 is an example of ACA on a data set, with colors corresponding to OCSVM scores (yellow are the most abnormal points, blue the most normal).

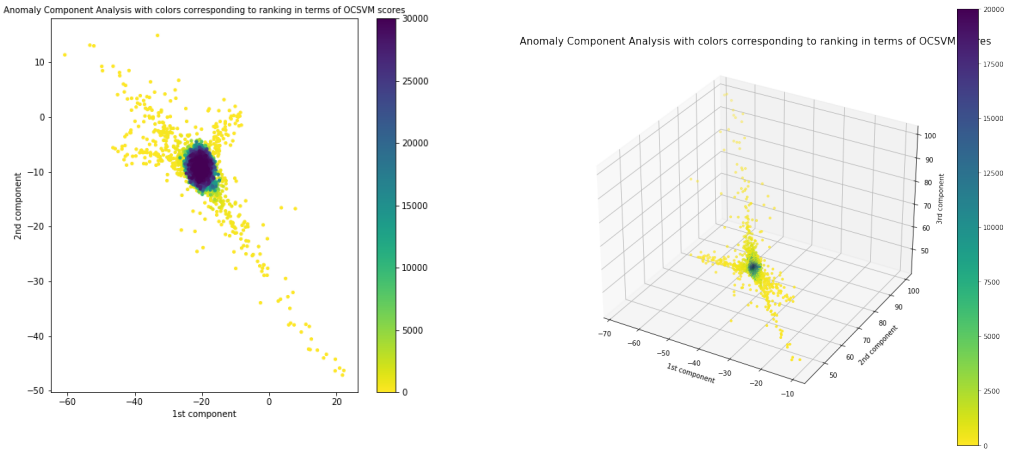


Figure C.18 – Anomaly Component Analysis on a data set.

We can see that this choice of space we call "anomaly space", highlights well the abnormal points which are outside of the main data cloud, even if here colors correspond to OCSVM scores, so we can see a clear concordance between OCSVM and the projection depth.

One important question is: is this visualisation robust to the choice of reference point to choose the unit directions of the anomaly space? If we choose the 2nd or 3rd most abnormal point in the computation of the directions, will anomaly space highlight both outliers and central points? We show in Figure C.19 an example of 2nd or 3rd most abnormal point, to show that they provide different anomaly spaces but still highlighting abnormal points.

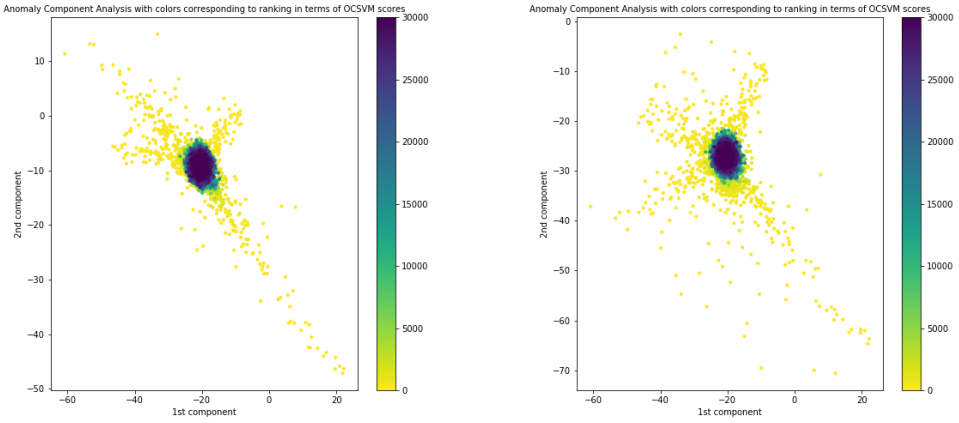


Figure C.19 – Anomaly Component Analysis on a data set with different reference point to compute anomaly space.

Threshold fine-tuning with ACA. Thus, one can use ACA as a visulation tool to tune threshold. An example with OCSVM is depicted in Figure C.20.

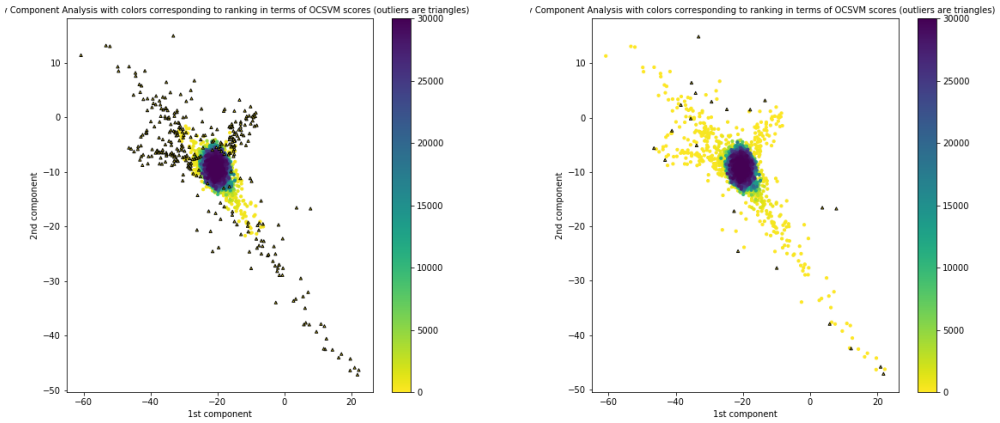


Figure C.20 – Anomaly Component Analysis on a data set with outliers as triangles depending different threshold choices: -0.01 (left) and -0.1 (right).

The goal is to find the optimal trade-off between targeting the most data considered as outliers in terms of scores (OCSVM here), and not targeting data in the main center data cloud. Here, -0.02 appears as a good choice. We can proceed similarly for the projection depth but it is more difficult in the case of IF as illustrated in Figure C.22.

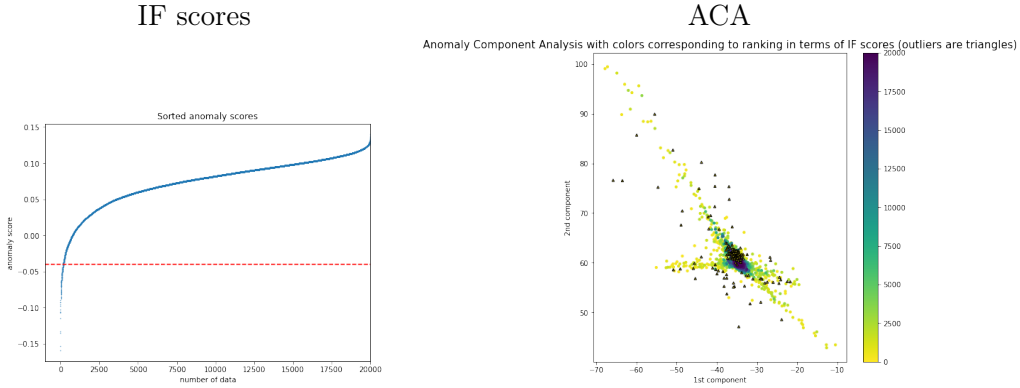
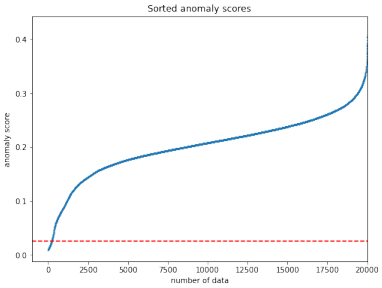


Figure C.21 – Threshold equal to -0.04 leading to 177 outliers.

Projection depth



ACA

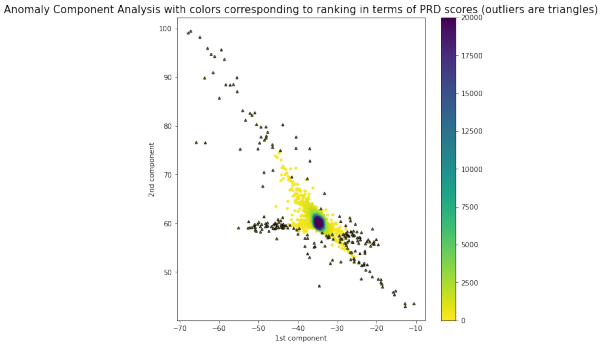


Figure C.22 – Threshold equal to 0.025, leading to 214 outliers.

Many points detected as abnormal by IF lie in the main data cloud. Unlike OCSVM, we do not find an important intersection between anomalies detected by IF and projection depth, making it harder to use ACA to tune IF's threshold.

Bibliography

P. Afshani and T. Chan. On approximate range counting and depth. *Discrete and Computational Geometry*, 42(1):3–21, 2009. page 63

C. Agostinelli. Local half-region depth for functional data. *Journal of Multivariate Analysis*, 163:67–79, 2018. page 77

C. Agostinelli and M. Romanazzi. Local depth. *Journal of Statistical Planning and Inference*, 141(2):817–830, 2011. page 48

C. Agostinelli and M. Romanazzi. Ordering curves by data depth. *Studies in Classification, Data Analysis, and Knowledge Organization*, pages 1–8, 2013. page 76

N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and system sciences*, 58(1):137–147, 1999. pages 36, 144

E. Alves, D. Bhatt, B. Hall, K. Driscoll, A. Murugesan, and J. Rushby. Considerations in assuring safety of increasingly autonomous systems. *NASA*, 2018. page 216

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005. page 88

D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, abs/1606.06565, 2016. URL http://arxiv.org/abs/1606.06565. page 216

J. An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015. page 27

M. A. Arcones and E. Gine. Limit theorems for u-processes. *The annals of probability*, 21(3):1494–1542, 1993. page 60

M. A. Arcones, Z. Chen, and E. Gine. Estimators Related to U-Processes with Applications to Multivariate Medians: Asymptotic Normality. *The Annals of Statistics*, 22 (3):1460–1477, 1994. page 61

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223, 2017. pages 35, 36, 82, 83, 84, 153, 157, 158

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. page 88

A. Arribas-Gil and J. Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619, 2014. pages 79, 129, 135

A. Athalye, N. Carlini, and D. A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, volume 80, pages 274–283. PMLR, 2018. page 217

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011. page 145

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007. page 209

Z.-D. Bai and X. He. Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *The Annals of Statistics*, 27(5):1616–1637, 1999. page 61

A. Baillo. Total error in a plug-in estimator of level sets. *Statistics & Probability Letters*, 65(4):411–417, 2003. page 26

A. Baillo, J. A. Cuesta-Albertos, and A. Cuevas. Convergence rates in nonparametric estimation of level sets. *Statistics & Probability Letters*, 53(1):27–35, 2001. page 25

Y. Balaji, R. Chellappa, and S. 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. pages 85, 207

S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. page 208

V. Barnett. The ordering of multivariate data. *Journal of the Royal Statistical Society. Series A (General)*, 139(3):318–355, 1976. page 50

F. Bassetti, A. Bodini, and E. Regazzini. On minimum kantorovich distance estimators. *Statistics & Probability Letters*, 76:1298–1302, 2006. pages 36, 84

E. Bayraktar and G. Guo. Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26:1 – 13, 2021. page 86

R. J. Beran and P. W. Millar. Multivariate symmetry models. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 13–42. Springer, 1997. page 47

M. Bhandari, P. Gour, A. Ashfaq, P. Liu, and G. Neubig. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*, 2020. pages 210, 211

P. Billingsley. *Convergence of probability measures (2nd ed.)*. John Wiley & Sons, 1999. pages 34, 81

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. pages 34, 81

M. Blondel, V. Seguy, and A. Rolet. Smooth and sparse optimal transport. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS 2018)*, 2018. page 85

E. Boissard. Simple bounds for the convergence of empirical and occupation measures in 1-wasserstein distance. *Electron. J. Probab.*, 16(83):2296–2333, 2011. page 84

E. Boissard and T. L. Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2):539 – 563, 2014. page 84

N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015. pages 84, 85

N. Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. Theses, Université Paris Sud - Paris XI ; Scuola normale superiore (Pise, Italie), 2013. page 86

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. page 101

D. Bremner, K. Fukuda, and V. Rosta. Primal-dual algorithms for data depth. *DIMACS series in discrete mathematics and theoretical computer science*, 72:171–194, 2006. page 63

D. Bremner, D. Chen, J. Iacono, S. Langerman, and P. Morin. Output-sensitive algorithms for tukey depth and related problems. *Statistics and Computing*, 18(3): 259–266, 2008. page 63

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991. page 58

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. 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. pages 27, 106, 129

C. Brownlees, E. Joly, and G. Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6):2507–2536, 2015. page 146

V.-E. Brunel. Concentration of the empirical level sets of tukey's halfspace depth. *Probability Theory and Related Fields*, 173(3):1165–1196, 2019. pages 62, 212

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013. pages 37, 144, 146

M. A. Burr and R. J. Fabrizio. Uniform convergence rates for halfspace depth. *Statistics & Probability Letters*, 124:33–40, 2017. pages 61, 176, 181

B. Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 97 (4):999–1023, 2006. page 26

T. T. Cai and H. H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5):2389–2420, 2013. page 175

T. T. Cai, C.-H. Zhang, and H. H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010. page 181

A. Cannon, J. Howse, D. Hush, and C. Scovel. Learning with the neyman-pearson and min-max criteria. *Los Alamos National Laboratory, Technical Repeport LA-UR*, pages 02–2951, 2002. page 25

I. Cascos. Data depth: multivariate statistics and geometry. *New perspectives in stochastic geometry*, pages 398–423, 2010. page 45

I. Cascos and M. López-Díaz. On the uniform consistency of the zonoid depth. *Journal of Multivariate Analysis*, 143:394–397, 2016. page 60

I. Cascos, Q. Li, and I. Molchanov. Depth and outliers for samples of sets and random sets distributions. *Australian & New Zealand Journal of Statistics*, 2021. in press. pages 212, 216

O. 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. pages 37, 144, 145

S. Cha and S. N. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35(6):1355–1370, 2002. pages 34, 81

A. Chakraborty and P. Chaudhuri. On data depth in infinite dimensional spaces. *Annals of the Institute of Statistical Mathematics*, 66(2):303–324, 2014a. pages 68, 69, 74, 76

A. Chakraborty and P. Chaudhuri. The spatial distribution in infinite dimensional spaces and related quantiles and depths. *The Annals of Statistics*, 42(3):1203–1231, 2014b. pages 32, 67, 78

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009. pages 24, 27, 154

E. Chapuis, P. Colombo, M. Manica, M. Labeau, and C. Clavel. Hierarchical pretraining for sequence labelling in spoken dialog. *arXiv preprint arXiv:2009.11152*, 2020. page 209

E. Chapuis, P. Colombo, M. Labeau, and C. Clavel. Code-switched inspired losses for generic spoken dialog representations. *arXiv preprint arXiv:2108.12465*, 2021. page 209

P. Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872, 1996. pages 29, 45, 55

B. Chen, K. M. Ting, T. Washio, and G. Haffari. Half-space mass: a maximally robust and efficient data depth method. *Machine Learning*, 100(2):677–699, 2015a. pages 186, 188

D. Chen, P. Morin, and U. Wagner. Absolute approximation of tukey depth: Theory and experiments. *Computational Geometry*, 46(5):566–573, 2013. pages 63, 204

Y. Chen, X. Dang, H. Peng, and H. L. Bart. Outlier detection with the kernelized spatial depth function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):288–305, 2009. pages 48, 55

Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010. page 174

Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The ucr time series classification archive, 2015b. URL www.cs.ucr.edu/~eamonn/time_series_data. pages 100, 105, 120

Y.-C. Chen and M. Bansal. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*, 2018. page 210

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112 (520):1684–1696, 2017. page 26

Z. Chen. Bounds for the Breakdown Point of the Simplicial Median. *Journal of Multivariate Analysis*, 55(1):1–13, 1995. page 65

Z. Chen and D. E. Tyler. The influence function and maximum bias of Tukey's median. *The Annals of Statistics*, 30(6):1737–1759, 2002. page 66

V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge–kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017. pages 29, 45, 58, 60

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018. pages 85, 207

L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. In *Advances in Neural Information Processing Systems*, volume 33, pages 2257–2269, 2020. page 84

G. Claeskens, M. Hubert, L. Slaets, and K. Vakili. Multivariate functional halfspace depth. *Journal of American Statistical Association*, 109(505):411–423, 2014. pages 71, 72, 103, 116, 119, 129, 131

S. Clémençon and N. Vayatis. Nonparametric estimation of the precision-recall curve. In *26-th International Conference in Machine Learning*, page 185–192, 2009. page 130

S. Clémençon and J. Jakubowicz. Scoring anomalies: a m-estimation formulation. In *Artificial Intelligence and Statistics*, pages 659–667. PMLR, 2013. page 28

S. Clémençon and A. Thomas. Mass volume curves and anomaly ranking. *Electronic Journal of Statistics*, 12(2):2806 – 2872, 2018. page 28

P. Colombo, W. Witon, A. Modi, J. Kennedy, and M. Kapadia. Affect-driven dialog generation. *arXiv preprint arXiv:1904.02793*, 2019. page 208

P. Colombo, E. Chapuis, M. Manica, E. Vignon, G. Varni, and C. Clavel. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7594–7601, 2020. page 211

P. Colombo, E. Chapuis, M. Labeau, and C. Clavel. Improving multimodal fusion via mutual dependency maximisation. *arXiv preprint arXiv:2109.00922*, 2021a. page 211

P. Colombo, C. Clave, and P. Piantanida. Infolm: A new metric to evaluate summarization & data2text generation. *arXiv preprint arXiv:2112.01589*, 2021b. page 209

P. Colombo, C. Clavel, and P. Piantanida. A novel estimator of mutual information for learning to disentangle textual representations. *arXiv preprint arXiv:2105.02685*, 2021c. page 208

P. Colombo, G. Staerman, C. Clavel, and P. Piantanida. Automatic text evaluation through the lens of Wasserstein barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10466, 2021d. page 209

P. Colombo, C. Yang, G. Varni, and C. Clavel. Beam search with bidirectional strategies for neural response generation. *arXiv preprint arXiv:2110.03389*, 2021e. page 208

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 274–289, 2014. page 83

I. Csiszár. 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. pages 34, 35, 81, 87

J. A. Cuesta-Albertos and A. Nieto-Reyes. The random tukey depth. *Computational Statistics & Data Analysis*, 52(11):4979–4988, 2008a. pages 63, 74

J. A. Cuesta-Albertos and A. Nieto-Reyes. The tukey and the random tukey depths characterize discrete distributions. *Journal of Multivariate Analysis*, 99(10):2304–2311, 2008b. page 52

A. Cuevas and R. Fraiman. On depth measures and dual statistics. a methodology for dealing with general data. *Journal of Multivariate Analysis*, 100(4):753–766, 2009. pages 54, 60, 61, 73, 74, 170, 171, 191

A. Cuevas, M. Febrero, and R. Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496, 2007. pages 78, 97, 101, 106, 226

M. Cuturi, O. Teboul, and J.-P. Vert. Sinkhorn distances: Lightspeed computation of optimal transportation. In *Advances in Neural Information Processing Systems (NeurIPS 2013)*, volume 26, pages 2292–2300, 2013. pages 84, 153

W. Dai and M. Genton. Multivariate functional data visualization and outlier detection. *Journal of Computational and Graphical Statistics*, 27(4):923–934, 2019a. pages 129, 135

W. Dai and M. G. Genton. Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis*, 131:50–65, 2019b. pages 71, 79

W. Dai, T. Mrkvička, Y. Sun, and M. G. Genton. Functional outlier detection and taxonomy by sequential transformations. *Computational Statistics & Data Analysis*, 149:106960, 2020. page 79

X. Dai, S. Lopez-Pintado, and for the Alzheimer's Disease Neuroimaging Initiative. Tukey's depth for object data. *Journal of the American Statistical Association*, 0(ja): 1–37, 2021. page 216

H. T. Dang and K. Owczarzak. Overview of the tac 2008 update summarization task. In *Proceedings of the Text Analysis Conference (TAC)*, 2008. page 210

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. pages 175, 180, 193

J. Depersin. Robust subgaussian estimation with vc-dimension. *arXiv preprint arXiv:2004.11734*, 2020. page 146

J. Depersin and G. Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019. pages 37, 144, 146

S. Dereich, M. Scheutzow, and R. Schottstedt. Constructive quantization: Approximation by empirical measures. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pages 1183–1203, 2013. page 84

I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. G. Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019. page 86

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. pages 34, 81, 209

L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016. pages 37, 56, 144, 145

T. Dinkar, P. Colombo, M. Labeau, and C. Clavel. The importance of fillers for text representations of speech transcripts. *arXiv preprint arXiv:2009.11340*, 2020. page 209

L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*, 2019. page 210

D. L. Donoho. Breakdown properties of location estimators. *P.h.D., qualifying paper, Dept. Statistics, Hardvard University*, 1982. pages 56, 65, 202

D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on half space depth and projected outlyingness. *The Annals of Statistics*, 20:1803–1827, 1992. pages 28, 51, 52, 60, 65, 199, 202, 203

D. L. Donoho and P. J. Huber. The notion of breakdown point. *A Festschrift for Erich Lehman*, pages 157–184, 1983. pages 65, 202

R. M. Dudley. The speed of mean glivenko-cantelli convergence. *Ann. Math. Statist.*, 40(1):40–50, 1969. pages 36, 84

R. M. Dudley. Balls in $\mathbb{R}^k$ do not cut all subsets of k + 2 points. *Advances in Mathematics*, 31(3):306–308, 1979. page 177

R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002. page 124

S. Dutta and A. K. Ghosh. On robust classification using projection depth. *Annals of the Institute of Statistical Mathematics*, 64(3):657–676, 2012. page 63

S. Dutta, A. K. Ghosh, and P. Chaudhuri. Some intriguing properties of Tukey's half-space depth. *Bernoulli*, 17(4):1420–1434, 2011. pages 32, 52, 67, 68, 69, 74

S. Dutta, S. Sarkar, and A. K. Ghosh. Multi-scale classification using localized spatial depth. *Journal of Machine Learning Research*, 17(217):1–30, 2016. page 48

R. Dyckerhoff. Data depth satisfying the projection property. *Allgemeines Statistisches Archiv*, 88(2):163–190, 2004. pages 46, 47, 48, 49, 51, 63, 204

R. Dyckerhoff. Convergence of depths and depth-trimmed regions. *arXiv preprint arXiv:1611.08721*, 2017. pages 52, 60

R. Dyckerhoff and K. Mosler. Weighted-mean trimming of multivariate data. *Journal of Multivariate Analysis*, 102(3):405–421, 2011. page 50

R. Dyckerhoff and K. Mosler. Weighted-mean regions of a probability distribution. *Statistics & Probability Letters*, 82(2):318–325, 2012. page 50

R. Dyckerhoff and P. Mozharovskyi. Exact computation of the halfspace depth. *Computational Statistics & Data Analysis*, 98:19–30, 2016. page 63

R. Dyckerhoff, K. Mosler, and G. A. Koshevoy. Zonoid data depth: Theory and computation. In *COMPSTAT*, pages 235–240, 1996. page 57

R. Dyckerhoff, P. Mozharovskyi, and S. Nagy. Approximate computation of projection depths. *Computational Statistics & Data Analysis*, 157(C):107166, 2021. pages 63, 170, 204, 212, 216

L. Dümbgen. Limit theorems for the simplicial depth. *Statistics & Probability Letters*, 14(2):119–128, 1992. pages 51, 60, 61

W. F. Eddy. Convex hull peeling. In *COMPSTAT 1982 5th Symposium held at Toulouse 1982*, pages 42–47, 1982. page 50

H. Edelsbrunner. *Algorithms in Combinatorial Geometry*. Springer-Verlag, 1987. page 62

J. H. Einmahl and D. M. Mason. Generalized quantile process. *The Annals of Statistics*, 20:1062–1078, 1992. page 26

J. H. Einmahl, J. Li, and R. Y. Liu. Bridging centrality and extremity: Refining empirical data depth using extreme value statistics. *The Annals of Statistics*, 43(6): 2738–2765, 2015. page 170

R. T. Elmore, T. P. Hettmansperger, and F. Xuan. Spherical data depth and a multivariate median. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72:87, 2006. page 57

E. Enqvist. *On sampling from sets of random variables with application to incomplete U-statistics*. PhD thesis, Lund University, 1978. page 115

K. Esbensen. Multivariate data analysis in practice. *Camo Software*, 2001. page 120

J. Fan, Y. Liao, and H. Liu. An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32, 2016. page 175

T. Fawcett. An Introduction to ROC Analysis. *Letters in Pattern Recognition*, 27(8): 861–874, 2006. page 129

M. Febrero-Bande and M. O. de la Fuente. Statistical computing in functional data analysis: The r package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012. page 64

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis*. Springer-Verlag, 2006. pages 30, 71

T. Ferreira, C. Gardent, N. Ilinykh, C. van der Lee, S. Mille, D. Moussallem, and A. Shimorina. The 2020 bilingual, bi-directional webnlg+ shared task overview and evaluation results (webnlg+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, 2020. page 209

T. C. Ferreira, D. Moussallem, E. Krahmer, and S. Wubben. Enriching the webnlg corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, 2018. page 209

A. Figalli. On the continuity of center-outward distribution and quantile functions. *Nonlinear Analysis*, 177:413–421, 2018. page 58

D. Fischer, K. Mosler, J. Möttönen, K. Nordhausen, O. Pokotylo, and D. Vogel. Computing the oja median in r: The package ojanp. *Journal of Statistical Software*, 92 (8):1–36, 2020. page 62

N. Fournier and A. Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015. page 84

R. Fraiman and J. Meloche. Multivariate l-estimation. *Test*, 8(2):255–317, 1999. page 48

R. Fraiman and G. Muniz. Trimmed means for functional data. *Test*, 10(2):419–440, 2001. pages 32, 67, 72, 226

J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23(9):881–890, 1974. page 56

J. H. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. page 175

C. Gao, J. Liu, Y. Yao, and W. Zhu. Robust estimation and generative adversarial nets, 2018. arXiv preprint arXiv:1810.02030. page 156

A. Garcia, P. Colombo, S. Essid, F. d'Alché Buc, and C. Clavel. From the token to the review: A hierarchical multimodal approach to opinion mining. *arXiv preprint arXiv:1908.11216*, 2019. page 211

C. Gardent, A. Shimorina, S. Narayan, and L. Perez-Beltrachini. Creating training corpora for nlg micro-planning. In *55th annual meeting of the Association for Computational Linguistics (ACL)*, 2017. page 209

S. Gehrmann, Y. Deng, and A. Rush. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, 2018. page 210

M. Genest, J.-C. Massé, and J.-F. Plante. depth: Nonparametric depth functions for multivariate analysis. *R package version 2.1-1.1*, 2019. page 64

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. page 101

I. Gijbels and S. Nagy. On a General Definition of Depth for Functional Data. *Statistical Science*, 32(4):630–639, 2017. pages 33, 68, 69, 70, 72, 74, 76, 77, 78, 211, 216

N. Goix. How to evaluate the quality of unsupervised anomaly detection algorithms? *arXiv preprint arXiv:1607.01152*, 2016. page 28

M. Goldstein and A. Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, pages 59–63, 2012. page 27

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS 2014)*, 2014. pages 34, 35, 81, 87

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19, pages 513–520, 2007. pages 88, 208

B. Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. *Pacific Journal of Mathematics*, 10(4):1257–1261, 1960. pages 51, 52

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777, 2017. pages 35, 82, 153, 157

M. Hallin, D. Paindaveine, and M. Šiman. Multivariate quantiles and multiple-output regression quantiles: From L1 optimization to halfspace depth. *The Annals of Statistics*, 38(2):635–669, 2010. pages 29, 63

M. Hallin, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165, 2021. pages 45, 58

F. R. Hampel. Contributions to the theory of robust estimation. *P.h.D., qualifying paper, Dept. Statistics, Hardvard University*, 1968. page 65

F. R. Hampel. A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971. page 66

S. Hariri, M. C. Kind, and R. J. Brunner. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, 2019. pages 27, 94, 95, 101, 234

T. Harris, J. D. Tucker, B. Li, and L. Shand. Elastic depths for detecting shape anomalies in functional data. *Technometrics*, 63(4):466–476, 2021. pages 67, 79

J. A. Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82(397):267–270, 1987. page 26

A. Hassairi and O. Regaieg. On the tukey depth of a continuous probability distribution. *Statistics & Probability Letters*, 78(15):2308–2313, 2008. page 52

X. He and G. Wang. Convergence of depth contours for multivariate datasets. *The Annals of Statistics*, 25(2):495–504, 1997. page 60

S. Helander, G. V. Bever, S. Rantala, and P. Ilmonen. Pareto depth for functional data. *Statistics*, 54(1):182–204, 2020. pages 67, 79

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637. 2017. page 159

W. Hoeffding. The strong law of large numbers for u-statistics. Technical report, North Carolina State University. Dept. of Statistics, 1961. page 113

H. Hotelling. Stability in competition. *The Economic Journal*, 39(153):41–57, 1929. page 28

D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016. pages 37, 144, 145

Y. Hu, Y. Wang, Y. Wu, Q. Li, and C. Hou. Generalized Mahalanobis depth in the reproducing kernel Hilbert space. *Statistical Papers*, 52(3):511–522, 2011. page 48

H. Huang and Y. Sun. A decomposition of total variation depth for understanding functional outliers. *Technometrics*, 61(4):445–458, 2019. page 79

P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. page 22

P. J. Huber and E. M. Ronchetti. *Robust Statistics (Second Edition)*. John Wiley & Sons, 2009. page 143

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. page 208

M. Hubert, P. J. Rousseeuw, and P. Segaert. Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2):177–202, 2015. pages 29, 30, 71, 72, 101, 106, 108, 119, 120, 128, 129, 130, 135, 226

R. J. Hyndman and H. L. Shang. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1):29–45, 2010. page 135

F. Ieva and A. Paganoni. Depth measures for multivariate functional data. *Communication in Statistics- Theory and Methods*, 42(7):1265–1276, 2013. page 76

H. Jalalzai, P. Colombo, C. Clavel, É. Gaussier, G. Varni, E. Vignon, and A. Sabourin. Heavy-tailed representations, text polarity classification & data augmentation. *Advances in Neural Information Processing Systems*, 33:4295–4307, 2020. page 208

T. Jebara. Images as bags of pixels. In *International Conference on Computer Vision*, volume 2, pages 265–272, 2003. page 208

M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986. pages 36, 144

C. Johnson. The increasing risks of risk assessment: On the rise of artificial intelligence and non-determinism in safety-critical systems. In *the 26th Safety-Critical Systems Symposium*, page 15. Safety-Critical Systems Club York, UK., 2018. page 216

T. Johnson, I. Kwok, and R. Ng. Fast computation of 2-dimensional depth contours. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, page 224–228, 1998. page 62

E. Joly and G. Lugosi. Robust estimation of u-statistics. *Stochastic Processes and their Applications*, 126(12):3760–3773, 2016. pages 37, 144, 145, 149

M. I. Jordan. *Learning in graphical models*, volume 89. Springer Science & Business Media, 1998. pages 34, 81

R. Jörnsten. Clustering and classification based on the l1 data depth. *Journal of Multivariate Analysis*, 90(1):67–89, 2004. page 29

M. H. Kalos and P. A. Whitlock. *Monte Carlo Methods*. Wiley-Blackwell, 2008. pages 64, 171

L. V. Kantorovich and G. S. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958. pages 36, 83

L. Kantorovitch. On the translocation of masses (in russian). In *Proceedings of the USSR Academy of Sciences.*, 1942. page 82

C. Kedzie, K. McKeown, and H. Daumé III. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, 2018. page 210

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. page 182

J. T. Kent, F. Er, and P. D. Constable. Algorithms for the spatial median. In *Modern Nonparametric, Robust and Multivariate Methods*, pages 205–224. 2015. page 62

J. Kim. Rate of convergence of depth contours: with application to a multivariate metrically trimmed mean. *Statistics & Probability Letters*, 49(4):393–400, 2000. page 62

S.-J. Kim. The Metrically Trimmed Mean as a Robust Estimator of Location. *The Annals of Statistics*, 20(3):1534–1547, 1992. page 61

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. pages 35, 87

D. P. Kingma and M. Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. pages 35, 87

A. N. Kolmogorov and V. M. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations: Series 2*, 17:277–364, 1961. page 164

S. Kolouri, K. Nadjahi, S. Umut, R. Badeau, and G. Rohde K. Generalized sliced wasserstein distance. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. pages 84, 86, 205

V. I. Koltchinskii. M-estimation, convexity and quantiles. *The Annals of Statistics*, 25 (2):435–477, 1997. page 55

V. I. Koltchinskii and R. M. Dudley. On spatial quantiles. *Unpublished manuscript*, 1996. page 55

L. Kong and I. Mizera. Quantile tomography: using quantiles with multivariate data. *Statistica Sinica*, 22(4):1589–1610, 2012. page 63

L. Kong and Y. Zuo. Smooth depth contours characterize the underlying distribution. *Journal of Multivariate Analysis*, 101(9):2222–2226, 2010. pages 51, 52

G. A. Koshevoy. The Tukey depth characterizes the atomic measure. *Journal of Multivariate Analysis*, 83:360–364, 2002. page 52

G. A. Koshevoy. Lift-zonoid and multivariate depths. In *Developments in Robust Statistics*, pages 194–202, 2003. pages 52, 55

G. A. Koshevoy and K. Mosler. Zonoid trimming for multivariate distributions. *The Annals of Statistics*, 25(5):1998–2017, 1997. pages 29, 45, 50, 57

D. Kosiorowski and Z. Zawadzki. Depthproc an r package for robust exploration of multidimensional economic phenomena. *arXiv preprint arXiv:1408.4542*, 2019. page 64

S. G. Krantz and H. R. Parks. *Geometric Integration Theory*. Birkhauser, 2008. pages 63, 171

J. Kuelbs and J. Zinn. Concerns with functional depth. *Latin American Journal of Probability and Mathematical Statistics*, 10:831–855, 2013. page 74

J. Kuelbs and J. Zinn. Half-region depth for stochastic processes. *Journal of Multivariate Analysis*, 142(C):86–105, 2015. pages 32, 77

J. Kuelbs and J. Zinn. Convergence of quantile and depth regions. *Stochastic Processes and their Applications*, 126(12):3681–3700, 2016. page 60

S. Kuhnt and A. Rehage. An angle-based multivariate functional pseudo-depth for shape outlier detection. *Journal of Multivariate Analysis*, 146:325–340, 2016. page 79

S. Kullback. *Information Theory and Statistics*. John Wiley, 1959. pages 34, 81

M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015. pages 34, 81

P. Lafaye de Micheaux, P. Mozharovskyi, and M. Vimond. Depth for curve data and applications. *Journal of the American Statistical Association*, 116(536):1881–1897, 2021. pages 212, 216

P. Laforgue, S. Clémençon, and P. Bertail. On medians of (randomized) pairwise means. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019. pages 37, 144, 145, 149

P. Laforgue, G. Staerman, and S. 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. page 146

T. Lange, K. Mosler, and P. Mozharovskyi. Fast nonparametric classification based on data depth. *Statistical Papers*, 55(1):49–69, 2014. pages 29, 79

S. Langerman and W. Steiger. The complexity of hyperplane depth in the plane. *Discrete & Computational Geometry*, 30(2):299–309, 2003. page 62

F. Larsen, F. Berg, and S. Engelsen. An exploratory chemometric study of 1 h nmr spectra of table wine. *Journal of Chemometrics*, 20:198–208, 2006. page 120

G. Lecué and M. Lerasle. Learning from mom's principles: Le cam's approach. *Stochastic Processes and their applications*, 129(11):4385–4410, 2019. page 146

G. Lecué, M. Lerasle, and T. Mathieu. Robust classification via mom minimization. *Machine Learning*, 109(8):1635–1665, 2020. pages 37, 144, 146, 150, 153

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. page 107

G. Lecué and M. Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020. pages 56, 143

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. page 174

A. J. Lee. *U-statistics: Theory and practice*. Marcel Dekker, 1990. pages 53, 112, 145

K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. page 217

M. Lerasle, Z. Szabo, T. Mathieu, and G. Lecué. Monk–outlier-robust mean embedding estimation by median-of-means. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019. pages 37, 144, 146, 147, 148, 149, 152

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. pages 7871–7880, 2020. page 210

C. Ley, C. Sabbah, and T. Verdebout. A new concept of quantiles for directional data and the angular Mahalanobis depth. *Electronic Journal of Statistics*, 8(1):795–816, 2014. pages 212, 216

J. Li, J. A. Cuesta-Albertos, and R. Y. Liu. Dd-classifier: Nonparametric classification procedure based on dd-plot. *Journal of the American Statistical Association*, 107 (498):737–753, 2012. pages 29, 108

Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu. Copod: copula-based outlier detection. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1118–1123, 2020. page 27

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004. page 208

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *In Proceedings 8th IEEE International Conference on Data Mining*, pages 413–422, 2008. pages 27, 94, 95, 101, 104, 106, 120, 129, 188, 233

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012. pages 27, 94, 101

R. Y. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990. pages 29, 45, 46, 47, 53, 54

R. Y. Liu. Data depth and multivariate rank tests. In L_1-*Statistical Analysis and Related Methods*, pages 279–294. 1992. pages 29, 45, 56

R. Y. Liu and K. Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260, 1993. pages 29, 45, 49, 56, 60, 172

R. Y. Liu and K. Singh. Notions of limiting p values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92(437):266–277, 1997. page 67

R. Y. Liu, J. M. Parelius, and K. Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics*, 27(3):783–858, 1999. page 48

X. Liu. Fast implementation of the Tukey depth. *Computational Statistics*, 32(4): 1395–1410, 2017. page 63

X. Liu and Y. Zuo. Computing projection depth and its associated estimators. *Statistics and Computing*, 24(1):51–63, 2014a. pages 62, 203, 204

X. Liu and Y. Zuo. Computing halfspace depth and regression depth. *Communications in Statistics-Simulation and Computation*, 43(5):969–985, 2014b. page 63

X. Liu, K. Mosler, and P. Mozharovskyi. Fast computation of tukey trimmed regions and median in dimension $p > 2$. *Journal of Computational and Graphical Statistics*, 28(3):682–697, 2019. pages 63, 203

X. Liu, S. Luo, and Y. Zuo. Some results on the computing of tukey's halfspace median. *Statistical Papers*, 61(1):303–316, 2020. page 52

Y. Liu and M. Lapata. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, 2019. page 210

Z. Liu and R. Modarres. Lens data depth and median. *Journal of Nonparametric Statistics*, 23(4):1063–1074, 2011. pages 57, 60, 61

W. S. Lok and S. M. Lee. A new statistical depth function with applications to multimodal data. *Journal of Nonparametric Statistics*, 23(3):617–631, 2011. page 48

J. P. Long and J. Z. Huang. A study of functional depths. *arXiv preprint arXiv:1506.01332*, 2016. page 78

S. López-Pintado and R. Jornsten. Functional analysis via extensions of the band depth. In *Complex datasets and inverse problems*, pages 103–120. Institute of Mathematical Statistics, 2007. page 76

S. López-Pintado and Y. Wei. Depth for sparse functional data. In *Recent advances in functional data analysis and related topics*, pages 209–212, 2011. page 76

S. López-Pintado, Y. Sun, J. Lin, and M. G. Genton. Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, 8:321–338, 2014. pages 71, 76

H. P. Lopuhaa and P. J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1): 229–248, 1991. pages 65, 66

G. Lugosi and S. Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019a. pages 37, 144, 145

G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 2019b. pages 37, 144, 146

S. López-Pintado and J. Romo. On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734, 2009. pages 32, 67, 68, 70, 76, 77, 115

S. López-Pintado and J. Romo. A half-region depth for functional data. *Computational Statistics & Data Analysis*, 55(4):1679–1695, 2011. pages 32, 67, 70, 77, 79

D. J. MacKay and D. J. Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. pages 34, 81

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. page 217

P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of National Academy of Science of India*, 2:49–55, 1936. page 56

F. Mairesse, M. Gasic, F. Jurcicek, S. Keizer, B. Thomson, K. Yu, and S. Young. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561, 2010. page 209

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993. page 96

E. Mammen and W. Polonik. Confidence regions for level sets. *Journal of Multivariate Analysis*, 122:202–214, 2013. page 26

T. Manole, S. Balakrishnan, and L. Wasserman. Minimax confidence intervals for the sliced wasserstein distance. *arXiv preprint arXiv:1909.07862*, 2019. page 86

R. A. Maronna and V. J. Yohai. The behavior of the stahel-donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341, 1995. page 61

D. M. Mason and W. Polonik. Asymptotic normality of plug-in level set estimates. *The Annals of Applied Probability*, 19(3):1108 – 1142, 2009. page 26

J.-C. Massé. Asymptotics for the tukey median. *Journal of Multivariate Analysis*, 81 (2):286–300, 2002. page 61

J.-C. Massé. Asymptotics for the Tukey depth process, with an application to a multivariate trimmed mean. *Bernoulli*, 10(3):397–419, 2004. page 61

J.-C. Massé. Multivariate trimmed means based on the tukey depth. *Journal of Statistical Planning and Inference*, 139(2):366–384, 2009. page 61

J.-C. Massé and R. Theodorescu. Halfplane trimming for bivariate distributions. *Journal of Multivariate Analysis*, 48(2):188–202, 1994. pages 49, 52, 60

V. Maz'ya. *Sobolev Spaces: with Applications to Elliptic Partial Differential Equations*. Springer-Verlag, 2011. page 101

R. J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309 – 323, 1995. page 58

P. McNamee and H. T. Dang. Overview of the tac 2009 knowledge base population track. In *Proceedings of the Text Analysis Conference (TAC)*, volume 17, pages 111–113, 2009. page 210

D. Meng and H. Chen. Magnet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 135–147. ACM, 2017. page 217

P. Milasevic and G. R. Ducharme. Uniqueness of the Spatial Median. *The Annals of Statistics*, 15(3):1332 – 1333, 1987. page 55

K. Miller, S. Ramaswami, P. Rousseeuw, J. A. Sellarès, D. Souvaine, I. Streinu, and A. Struyf. Efficient computation of location depth contours by methods of computational geometry. *Statistics and Computing*, 13(2):153–162, 2003. page 62

S. Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21 (4):2308–2335, 2015. pages 37, 144, 145

M. Mirzargar, R. T. Whitaker, and R. M. Kirby. Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2654–2663, 2014. page 76

I. Mizera. On depth and deep points: a calculus. *The Annals of Statistics*, 30(6): 1681–1736, 2002. page 51

I. Mizera and C. H. Müller. Location–scale depth. *Journal of the American Statistical Association*, 99(468):949–966, 2004. page 45

I. Mizera and M. Volauf. Continuity of halfspace depth contours and maximum depth estimators: Diagnostics of depth-related methods. *Journal of Multivariate Analysis*, 83(2):365–388, 2002. pages 51, 52

J. M. Moguerza and A. Muñoz. Support Vector Machines with Applications. *Statistical Science*, 21(3):322–336, 2006. page 27

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royal des Sciences*, pages 666–704, 1781. page 82

K. Mosler. *Multivariate Dispersion, Central Regions, and Depth*. Springer, 2002. page 57

K. Mosler. Depth statistics. In *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*, pages 17–34. Springer, 2013. pages 45, 49

K. Mosler and P. Mozharovskyi. Fast dd-classification of functional data. *Statistical Papers*, 58(4):1055–1089, 2017. page 101

K. Mosler and P. Mozharovskyi. Choosing among notions of multivariate depth statistics. *arXiv preprint arXiv:2004.01927*, 2020. pages 28, 45, 64

K. Mosler and Y. Polyakova. General notions of depth for functional data. *arXiv preprint arXiv 1208.1981*, 2012. pages 67, 68, 70, 72, 74, 75

K. Mosler, T. Lange, and P. Bazovkin. Computing zonoid trimmed regions of dimension d>2. *Computional Statistics and Data Analysis*, 53(7):2500–2510, 2009. page 63

P. Mozharovskyi, K. Mosler, and T. Lange. Classifying real-world data with the $DD\alpha$-procedure. *Advances in Data Analysis and Classification*, 9(3):287–314, 2015. pages 63, 204

D. Mukherjee, A. Guha, J. M. Solomon, Y. Sun, and M. Yurochkin. Outlier-robust optimal transport. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 7850–7860, 2021. pages 85, 207

M. Myllymäki, T. Mrkvička, P. Grabarnik, H. Seijo, and U. Hahn. Global envelope tests for spatial processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):381–404, 2017. page 79

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. pages 34, 35, 81, 87

K. Nadjahi, A. Durmus, U. Simsekli, and R. Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *Advances in Neural Information Processing Systems*, 32, 2019. page 86

K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Simsekli. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020. page 86

S. Nagy. Statistical depth for functional data. *PhD thesis KU Leuven and Charles University*, 2016. pages 71, 77

S. Nagy. Monotonicity properties of spatial depth. *Statistics & Probability Letters*, 129 (C):373–378, 2017. page 55

S. Nagy. Halfspace depth does not characterize probability distributions. *Statistical Papers*, 62(3):1135–1139, 2021. pages 52, 200

S. Nagy and J. Dvořák. Illumination depth. *Journal of Computational and Graphical Statistics*, 30(1):78–90, 2021. pages 65, 202, 203

S. Nagy, I. Gijbels, and D. Hlubinka. Weak convergence of discretely observed functional data with applications. *Journal of Multivariate Analysis*, 146:46–62, 2016a. pages 71, 114

S. Nagy, I. Gijbels, M. Omelka, and D. Hlubinka. Integrated depth for functional data: statistical properties and consistency. *ESAIM: PS*, 20:95–130, 2016b. pages 70, 73

S. Nagy, I. Gijbels, and D. Hlubinka. Depth-based recognition of shape outlying functions. *Journal of Computational and Graphical Statistics*, 26(4):883–893, 2017. pages 67, 79

S. Nagy, C. Schütt, and E. M. Werner. Halfspace depth and floating body. *Statistics Surveys*, 13:52–118, 2019. pages 51, 52

S. Nagy, R. Dyckerhoff, and P. Mozharovskyi. Uniform convergence rates for the approximated halfspace and projection depth. *Electronic Journal of Statistics*, 14(2): 3939–3975, 2020a. page 204

S. Nagy, R. Dyckerhoff, and P. Mozharovskyi. Uniform convergence rates for the approximated halfspace and projection depth. *Electronic Journal of Statistics*, 14(2): 3939–3975, 2020b. pages 61, 62, 63, 176, 181

S. Narayan, S. B. Cohen, and M. Lapata. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, 2018. page 210

N. N. Narisetty and V. N. Nair. Extremal depth for functional data and applications. *Journal of the American Statistical Association*, 111(516):1705–1714, 2016. page 79

A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons Ltd, 1983. pages 36, 144

A. Nenkova and R. J. Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152, 2004. page 210

A. Nenkova, R. Passonneau, and K. McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4–es, 2007. page 210

B. H. Neumann. On An Invariant of Plane Regions and Mass Distributions. *Journal of the London Mathematical Society*, s1-20(4):226–237, 1945. page 51

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. page 25

S. Nietert, R. Cummings, and Z. Goldfeld. Outlier-robust optimal transport: Duality, structure, and statistical analysis. *arXiv preprint arXiv:2111.01361*, 2021. page 85

A. Nieto-Reyes and H. Battey. A Topologically Valid Definition of Depth for Functional Data. *Statistical Science*, 31(1):61–79, 2016. pages 33, 68, 69, 70, 74, 78, 211, 216

A. Niinimaa and H. Oja. On the influence functions of certain bivariate medians. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):565–574, 1995. page 66

A. Niinimaa, H. Oja, and M. Tableman. The finite-sample breakdown point of the oja bivariate median and of the corresponding half-samples version. *Statistics & Probability Letters*, 10(4):325–328, 1990. page 65

D. Nolan. Asymptotics for multivariate trimming. *Stochastic Processes and their Applications*, 42(1):157–169, 1992. page 60

J. Nunez-Garcia, Z. Kutalik, K.-H. Cho, and O. Wolkenhauer. Level sets and minimum volume sets of probability density functions. *Approximate Reasoning*, 34:25–47, 2003. page 26

H. Oja. Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1(6):327–332, 1983. pages 29, 45, 54, 55

D. Paindaveine and G. V. bever. From depth to local depth: A focus on centrality. *Journal of the American Statistical Association*, 108(503):1105–1119, 2013. pages 48, 198

D. Paindaveine and G. V. Bever. Nonparametrically consistent depth-based classifiers. *Bernoulli*, 21(1):62 – 82, 2015. page 48

D. Paindaveine and M. Šiman. Computing multiple-output regression quantile regions from projection quantiles. *Computational Statistics*, 27(1):29–49, 2012. page 63

D. Paindaveine and G. Van Bever. Halfspace depths for scatter, concentration and shape matrices. *The Annals of Statistics*, 46(6B):3276–3307, 2018. pages 212, 216

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. page 208

F.-P. Paty and M. Cuturi. Subspace robust Wasserstein distances. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5072–5081, 2019. pages 84, 207

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. pages 132, 188

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. pages 35, 81, 82, 84

B. Piccoli and F. Rossi. Generalized wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014. page 85

O. Pokotylo, P. Mozharovskyi, and R. Dyckerhoff. Depth and depth-based classification with r package ddalpha. *Journal of Statistical Software*, 91(5):1–46, 2019. page 64

W. Polonik. Measuring Mass Concentrations and Estimating Density Contour Clusters-An Excess Mass Approach. *The Annals of Statistics*, 23(3):855 – 881, 1995. page 26

W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69(1):1–24, 1997. page 26

W. Qiao and W. Polonik. Nonparametric confidence regions for level sets: Statistical properties and geometry. *Electronic Journal of Statistics*, 13(1):985 – 1030, 2019. page 26

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer Berlin Heidelberg, 2012. pages 85, 207

S. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley Series in Probability and Statistics-Applied Probability and Statistics Section. Wiley, 1991. pages 34, 81

S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems: Volume I: Theory, volume 1*. Springer Science & Business Media, 1998. pages 36, 85

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, pages 1–67, 2020. page 210

J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis: Methods and Case Studies*. Springer-Verlag, New York, 2002. pages 30, 71, 120, 129

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer-Verlag, New York, 2005. pages 30, 31

K. Ramsay, S. Durocher, and A. Leblanc. Integrated rank-weighted depth. *Journal of Multivariate Analysis*, 173:51–69, 2019. pages 37, 53, 54, 60, 61, 170, 171, 190, 191, 199, 216

P. A. Rankel, J. Conroy, H. T. Dang, and A. Nenkova. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Association for Computational Linguistics (ACL)*, pages 131–136, 2013. page 210

P. Rigollet and X. Tong. Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12:2831–2855, 2011. page 25

P. Rigollet and R. Vert. Fast rates for plug-in estimators of density level sets. *Bernoulli*, 14(4):1154–1178, 2009. page 26

M. Romanazzi. Influence function of halfspace depth. *Journal of Multivariate Analysis*, 77(1):138–161, 2001. page 66

P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984. pages 56, 175, 182

P. J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications Vol. B*, pages 283–297, 1985. page 56

P. J. Rousseeuw and M. Hubert. Regression depth. *Journal of the American Statistical Association*, 94(446):388–402, 1999a. pages 29, 45

P. J. Rousseeuw and M. Hubert. Depth in an arrangement of hyperplanes. *Discrete & Computational Geometry*, 22(2):167–176, 1999b. page 51

P. J. Rousseeuw and M. Hubert. Computation of robust statistics: Depth, median, and related measures. In *Handbook of Discrete and Computational Geometry*, pages 1539–1552. Chapman & Hall-CRC Press, 3nd edition, 2018a. page 64

P. J. Rousseeuw and M. Hubert. Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2):e1236, 2018b. page 32

P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 1987. pages 22, 56

P. J. Rousseeuw and I. Ruts. Algorithm as 307: Bivariate location depth. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(4):516–526, 1996. page 62

P. J. Rousseeuw and I. Ruts. The depth function of a population distribution. *Metrika*, 49(3):213–244, 1999. pages 51, 198

P. J. Rousseeuw and A. Struyf. Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 8(3):193–203, 1998. pages 62, 63, 204

P. J. Rousseeuw and K. van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999. page 175

P. J. Rousseeuw, J. Raymaekers, and M. Hubert. A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, 27(2):345–359, 2018. pages 79, 129, 135, 137

I. Ruts and P. J. Rousseeuw. Computing depth contours of bivariate point clouds. *Computational Statistics & Data Analysis*, 23(1):153–168, 1996a. page 62

I. Ruts and P. J. Rousseeuw. Isodepth: A program for depth contours. In *Proceedings in Computational Statistics, COMPSTAT*, page 441–446, 1996b. page 62

A. Rényi. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961. pages 34, 35, 81, 87

F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhauser, 2015. pages 58, 82

B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems, 2016. page 85

R. Schneider. *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge University Press, 1993. pages 124, 203

R. Schneider and W. Weil. *Stochastic and Integral Geometry*. Springer-Verlag, 2008. page 124

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. pages 27, 106, 120, 129, 232

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(32), 2005. page 174

C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(8):3806–3819, 2005. page 25

C. Scott and R. Nowak. Learning minimum volume sets. *Journal of Machine Learning Resarch*, 7:665–704, 2006. page 26

A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017. page 210

P. Segaert, M. Hubert, P. Rousseeuw, and J. Raymaekers. mrfdepth: Depth measures in multivariate, regression and functional settings. *R package version 1.0.13*, 2020. pages 64, 132

V. Sehwag, A. N. Bhagoji, L. Song, C. Sitawarin, D. Cullina, M. Chiang, and P. Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 105–116, 2019. page 216

R. Serfling. *Approximation theorems of mathematical statistics*. Wiley, 1980. page 60

R. Serfling. Generalized l-, m-, and r-statistics. *The Annals of Statistics*, 12(1):76–86, 1984. page 49

R. Serfling. A depth function and a scale curve based on spatial quantiles. In *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 25–38, 2002. pages 47, 55

R. Serfling. Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference*, 123(2):259–278, 2004. page 47

R. Serfling. Depth functions in nonparametric multivariate inference. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72:1, 2006a. pages 29, 45

R. Serfling. Multivariate symmetry and asymmetry. *Encyclopedia of statistical sciences*, 8:5338–5345, 2006b. page 47

R. Serfling. Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation. *Journal of Nonparametric Statistics*, 22 (7):915–936, 2010. page 47

R. Serfling and Y. Zuo. Structural properties and convergence results for contours of sample statistical depth functions. *The Annals of Statistics*, 28(2):483–499, 2000. pages 49, 60

C. Sguera, P. Galeano, and R. Lillo. Spatial depth-based classification for functional data. *TEST*, 23(4):725–750, 2014. page 67

M. Shafique, M. Naseer, T. Theocharides, C. Kyrkou, O. Mutlu, L. Orosa, and J. Choi. Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead. *IEEE Design & Test*, 37(2):30–57, 2020. page 21

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. page 208

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 22(08):888–905, 2000. page 208

G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics*. John Wiley & Sons, 1986. pages 61, 177, 180, 181

A. Singh, C. Scott, and R. Nowak. Adaptive Hausdorff estimation of density level sets. *The Annals of Statistics*, 37(5B):2760–2782, 2009. page 26

R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35(2):876–879, 1964. page 84

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, 2006. page 208

L. Specia, D. Raj, and M. Turchi. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50, 2010. page 208

K. Sricharan, R. Raich, and A. O. Hero. k-nearest neighbor estimation of entropies with confidence. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 1205–1209, 2011. page 27

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009. page 88

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. pages 35, 88, 148, 153, 164, 165

A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1415–1428, 2011. page 79

G. Staerman, P. Mozharovskyi, S. Clémençon, and F. d'Alché Buc. Functional isolation forest. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, pages 332–347, 2019. page 232

G. Staerman, P. Mozharovskyi, and S. Clémençon. The area of the convex hull of sampled curves: a robust functional statistical depth measure. In *Proceedings of the 23nd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, volume 108, pages 570–579, 2020. page 232

G. Staerman, P. Laforgue, P. Mozharovskyi, and F. d'Alché Buc. When ot meets mom: Robust estimation of wasserstein distance. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 136–144, 2021a. page 148

G. Staerman, P. Mozharovskyi, and S. Clémençon. Affine-invariant integrated rank-weighted depth: Definition, properties and finite sample analysis. *arXiv preprint arXiv:2106.11068*, 2021b. page 232

G. Staerman, P. Mozharovskyi, P. Colombo, S. Clémençon, and F. d'Alché Buc. A pseudo-metric between probability distributions based on depth-trimmed regions. *arXiv preprint arXiv:2103.12711*, 2021c. page 232

G. Staerman, E. Adjakossa, P. Mozharovskyi, V. Hofer, J. S. Gupta, and S. Clémençon. Functional anomaly detection: a benchmark study. *arXiv preprint arXiv:2201.05115*, 2022. page 232

W. A. Stahel. Breakdown of covariance estimators. Technical report, Fachgruppe für Statistik, ETH, Zürich, 1981. page 56

I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232, 2005. page 25

A. Struyf and P. J. Rousseeuw. Halfspace depth and regression depth characterize the empirical distribution. *Journal of Multivariate Analysis*, 69(1):135–153, 1999. page 52

A. Struyf and P. J. Rousseeuw. High-dimensional computation of the deepest location. *Computational Statistics & Data Analysis*, 34:415–426, 2000. page 63

Y. Sun and M. G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011. page 135

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014. page 216

N. Tarabelloni, A. Arribas-Gil, F. Ieva, A. M. Paganoni, and J. Romo. Roahd: Robust analysis of high dimensional data. *R package version 1.4.1*, 2018. page 132

D. M. Tax and R. P. Duin. Support vector data description. *Machine Learning*, 54(1): 45–66, 2004. page 27

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018. pages 35, 82

A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–960, 1997. pages 25, 26

J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, volume 2, pages 523–531, 1975. pages 28, 45, 50

S. Van Aelst and P. J. Rousseeuw. Minimum volume ellipsoid. *WIREs Computational Statistics*, 1(1):71–82, 2009. page 56

G. Van Bever. Contributions to nonparametric and semiparametric inference based on statistical depth. *P.h.D., qualifying paper, Université Libre de Bruxelles*, 2013. page 45

S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000. page 164

A. van der Vaart. *Asymptotic Statistics*. Cambridge university press, 2000. page 162

S. van der Walt, J. L. Schänberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. a. Yu. scikit-image: image processing in python. *PeerJ*, 2, 2014. page 107

V. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999. page 61

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2): 264–280, 1971. page 177

V. N. Vapnik and A. Y. Chervonenkis. *Theory of Pattern Recognition (in Russian)*. Nauka, 1974. page 177

Y. Vardi and C.-H. Zhang. The multivariate l1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000. pages 29, 45, 55

R. Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012. pages 178, 181

R. Vert and J.-P. Vert. Consistency and convergence rates of one-class SVM and related algorithms. *Journal of Machine Learning Research*, 17:817–854, 2006. page 27

C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003. pages 35, 58, 81

C. Villani. *Optimal Transport: old and new*. Springer, 2008. page 83

D. Wang, P. Liu, Y. Zheng, X. Qiu, and X. Huang. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*, 2020. page 210

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016. page 30

P.-Å. Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, (2):217–232, 1973. page 192

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019. pages 36, 84, 215

T.-H. Wen, M. Gašić, N. Mrkšić, P.-H. Su, D. Vandyke, and S. Young. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, 2015. page 209

H. Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen. *Mathematische Annalen*, 71(4):441–479, 1912. page 178

W. Witon, P. Colombo, A. Modi, and M. Kapadia. Disney at iest 2018: Predicting emotions using an ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 248–253, 2018. page 211

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. page 209

W. Xie, S. Kurtek, K. Bharath, and Y. Sun. A geometric approach to visualization of variability in functional data. *Journal of the American Statistical Association*, 112 (519):979–993, 2017. pages 79, 135

W. Yoon, Y. S. Yeo, M. Jeong, B.-J. Yi, and J. Kang. Learning by semantic similarity makes abstractive summarization better. *arXiv preprint arXiv:2002.07767*, 2020. page 210

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2014. page 193

J. Zhang. Some extensions of tukey depth function. *Journal of Multivariate Analysis*, 82(1):134–165, 2002. page 170

J. Zhang, Y. Zhao, M. Saleh, and P. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11328–11339, 2020. page 210

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. pages 34, 81, 209

M. Zhao and V. 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. page 27

W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, 2019a. pages 209, 210

Y. Zhao, Z. Nasrullah, and Z. Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019b. page 188

S. Zheng, Y. Song, T. Leung, and I. J. Goodfellow. Improving the robustness of deep neural networks via stability training. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4480–4488. IEEE Computer Society, 2016. page 217

M. Zhong, P. Liu, D. Wang, X. Qiu, and X. Huang. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058, 2019. page 210

M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, 2020. page 210

C. Zhou and R. 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. pages 27, 188

Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, 2018. page 210

Y. Zuo. Projected based depth functions and associated medians. *The Annals of statistics*, 31(5):1460–1490, 2003. pages 56, 61, 170, 198

Y. Zuo. Robustness of weighted lp–depth and lp–median. *Allgemeines Statistisches Archiv*, 88:215–234, 2004. page 66

Y. Zuo. Multidimensional medians and uniqueness. *Computational Statistics & Data Analysis*, 66:82–88, 2013. page 56

Y. Zuo. A new approach for the computation of halfspace depth in high dimensions. *Communications in Statistics-Simulation and Computation*, 48(3):900–921, 2019. page 63

Y. Zuo and H. Cui. Depth weighted scatter estimators. *The Annals of Statistics*, 33(1): 381–413, 2005. page 61

Y. Zuo and R. Serfling. On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry. *Journal of Statistical Planning and Inference*, 84:55–79, 1999. page 47

Y. Zuo and R. Serfling. Nonparametric notions of multivariate "scatter measure" and "more scattered" based on statistical depth functions. *Journal of Multivariate Analysis*, 75:62–78, 2000a. page 50

Y. Zuo and R. Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482, 2000b. pages 29, 33, 37, 46, 47, 48, 50, 51, 54, 55, 57

Y. Zuo, H. Cui, and X. He. On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *The Annals of Statistics*, 32(1):167–188, 2004a. page 61

Y. Zuo, H. Cui, and D. Young. Influence function and maximum bias of projection depth based estimators. *The Annals of Statistics*, 32(1):189 – 218, 2004b. page 66

Titre : Détection d'anomalies fonctionnelles et estimation robuste

Mots clés : Détection d'anomalies, données fonctionnelles, profondeur de données, séries temporelles

Résumé : L'engouement pour l'apprentissage automatique s'étend à presque tous les domaines comme l'énergie, la médecine ou la finance. L'omniprésence des capteurs met à disposition de plus en plus de données avec une granularité toujours plus fine. Une abondance de nouvelles applications telles que la surveillance d'infrastructures complexes comme les avions ou les réseaux d'énergie, ainsi que la disponibilité d'échantillons de données massives, potentiellement corrompues, ont mis la pression sur la communauté scientifique pour développer de nouvelles méthodes et algorithmes d'apprentissage automatique fiables. Le travail présenté dans cette thèse s'inscrit dans cette ligne de recherche et se concentre autour de deux axes : *la détection non-supervisée d'anomalies fonctionnelles* et *l'apprentissage robuste*, tant du point de vue pratique que théorique. La première partie de cette thèse est consacrée au développement d'algorithmes efficaces de détection d'anomalies dans le cadre fonctionnel. Plus précisément, nous introduisons Functional Isolation Forest (FIF), un algorithme basé sur le partitionnement aléatoire de l'espace fonctionnel de manière flexible afin d'isoler progressivement les fonctions les unes des autres. Nous proposons également une nouvelle notion de profondeur fonctionnelle basée sur l'aire de l'enveloppe convexe des courbes échantillonnées, capturant de manière naturelle les écarts graduels de centralité. Les problèmes d'estimation et de calcul sont abordés et diverses expériences numériques fournissent des preuves empiriques de la pertinence des approches proposées. Enfin, afin de fournir des recommandations pratiques, la performance des récentes techniques de détection d'anomalies fonctionnelles est évaluée sur deux ensembles de données réelles liés à la surveillance des hélicoptères en vol et à la spectrométrie des matériaux de construction. La deuxième partie est consacrée à la conception et à l'analyse de plusieurs approches statistiques, potentiellement robustes, mêlant la profondeur de données et les estimateurs robustes de la moyenne. La distance de Wasserstein est une métrique populaire résultant d'un coût de transport entre deux distributions de probabilité et permettant de mesurer la similitude de ces dernières. Bien que cette dernière ait montré des résultats prometteurs dans de nombreuses applications d'apprentissage automatique, elle souffre d'une grande sensibilité aux valeurs aberrantes. Nous étudions donc comment tirer partie des estimateurs de la médiane des moyennes (MoM) pour renforcer l'estimation de la distance de Wasserstein avec des garanties théoriques. Par la suite, nous introduisons une nouvelle fonction de profondeur statistique dénommée Affine-Invariante Integrated Rank-Weighted (AI-IRW). Au-delà de l'analyse théorique effectuée, des résultats numériques sont présentés, confirmant la pertinence de cette profondeur. Les sur-ensembles de niveau des profondeurs statistiques donnent lieu à une extension possible des fonctions quantiles aux espaces multivariés. Nous proposons une nouvelle mesure de similarité entre deux distributions de probabilité. Elle repose sur la moyenne de la distance de Hausdorff entre les régions quantiles, induites par les profondeur de données, de chaque distribution. Nous montrons qu'elle hérite des propriétés intéressantes des profondeurs de données telles que la robustesse ou l'interprétabilité.
Tous les algorithmes développés dans cette thèse sont accessible en ligne.

Title : Functional Anomaly Detection and Robust Estimation

Keywords : Anomaly detection, functional data, data depth, robustness

Abstract : Enthusiasm for Machine Learning is spreading to nearly all fields such as transportation, energy, medicine, banking or insurance as the ubiquity of sensors through IoT makes more and more data at disposal with an ever finer granularity. The abundance of new applications for monitoring complex infrastructures (e.g. aircrafts, energy networks) together with the availability of massive data samples has put pressure on the scientific community to develop new reliable Machine-Learning methods and algorithms. The work presented in this thesis focuses around two axes: *unsupervised functional anomaly detection* and *robust learning*, both from practical and theoretical perspectives. The first part of this dissertation is dedicated to the development of efficient functional anomaly detection approaches. More precisely, we introduce Functional Isolation Forest (FIF), an algorithm based on randomly splitting the functional space in a flexible manner in order to progressively isolate specific function types. Also, we propose the novel notion of functional depth based on the area of the convex hull of sampled curves, capturing gradual departures from centrality, even beyond the envelope of the data, in a natural fashion. Estimation and computational issues are addressed and various numerical experiments provide empirical evidence of the relevance of the approaches proposed. In order to provide recommendation guidance for practitioners, the performance of recent functional anomaly detection techniques is evaluated using two real-world data sets related to the monitoring of helicopters in flight and to the spectrometry of construction materials. The second part describes the design and analysis of several robust statistical approaches relying on robust mean estimation and statistical data depth. The Wasserstein distance is a popular metric between probability distributions based on optimal transport. Although the latter has shown promising results in many Machine Learning applications, it suffers from a high sensitivity to outliers. To that end, we investigate how to leverage Medians-of-Means (MoM) estimators to robustify the estimation of Wasserstein distance with provable guarantees. Thereafter, a new statistical depth function, the Affine-Invariant Integrated Rank-Weighted (AI-IRW) depth is introduced. Beyond the theoretical analysis carried out, numerical results are presented, providing strong empirical confirmation of the relevance of the depth function proposed. The upper-level sets of statistical depths—the depth-trimmed regions—give rise to a definition of multivariate quantiles. We propose a new discrepancy measure between probability distributions that relies on the average of the Hausdorff distance between the depth-based quantile regions w.r.t. each distribution and demonstrate that it benefits from attractive properties of data depths such as robustness or interpretability.
All algorithms developed in this thesis are open-sourced and available online.

Institut Polytechnique de Paris
91120 Palaiseau, France