NEW NONPARAMETRIC APPROACHES FOR MULTIVARIATE AND FUNCTIONAL DATA ANALYSIS IN OUTLIER DETECTION, CONSTRUCTION OF TOLERANCE TUBES, AND CLUSTERING

BY Yi Fan

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Regina Y. Liu

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2016

ABSTRACT OF THE DISSERTATION

New Nonparametric Approaches for Multivariate and Functional Data Analysis in Outlier Detection, Construction of Tolerance Tubes, and Clustering

by Yi Fan

Dissertation Director: Regina Y. Liu

Recent advances of powerful computing and data acquisition technologies have made large complex datasets ever-present, including high-dimensional or functional data. Most existing statistical approaches for multivariate or functional data rely on parametric assumptions such as normality. In reality, such assumptions are either difficult to justify or verify. The goal of this dissertation is to develop general nonparametric statistical approaches for outlier detection, tolerance tubes construction, and clustering for multivariate and functional data.

1. In Chapter 3, we propose a general approach named Antipodal Reflection Depth (ARD), to refine any existing function of depth (henceforth base depth) to form a class of new depth functions. ARD has the advantage over its base depth in capturing

the relative magnitude of deviation from all data points to the deepest one. This desirable property is key in making ARD particularly useful in many applications. Here, we focus primarily on its utility in outlier detection.

2. In Chapter 4, we introduce tolerance tubes, which can be viewed as generalizations of tolerance intervals/regions to functional settings. A tolerance tube ensures a specified portion of the functional dataset be contained within the tolerance limits with some confidence. In addition to extending the commonly accepted definitions of $\beta-$content and $\beta-$expectation, we introduce modifications by incorporating an exempt level α. The latter relaxes the definitions by allowing α portion of each functional to be exempt from the requirements and is thus particularly useful to offset allowable occasional aberrations.

3. In Chapter 5, we propose a new clustering approach named K-means on Pairwise Distance (KMPD), and show it to be effective in detecting clusters with different sizes. Moreover, KMPD has the capability of grouping anomalous sample points into a single cluster, and therefore is an effective approach for outlier detection as well.

All these approaches are completely non-parametric and data-driven, and thus can be broadly applicable. Relevant theoretical properties are investigated and justified. These approaches are also illustrated and tested using data from both simulations and a real application on a medical study of continuous glucose monitoring.

Acknowledgements

First of all, I would like to express my deepest gratitude to my advisor Professor Regina Liu. Without her encouragement and guidance, this dissertation would not have been accomplished. She gave me advice and taught me to how to think like a researcher. In the past five years, her diligence to research, patience to students, thoughtfulness to people, and many other fine characters influenced me strongly. She is more than a teacher or a successful researcher. She is a mentor and friend to me. I thank her for her unwavering support and encouragement. It was a privilege to work with her.

I would like to extend my gratitude to Professor Cun-Hui Zhang, Professor Minge Xie, and Dr. Andrew Cheng for their effort to serve as the committee members and for their generous support throughout my study. Their wisdom and profound knowledge have been invaluable on my research and they made every meeting an enjoyable one. My special thanks go to Dr. Cheng for his generosity in sharing his knowledge in aviation and risk management, which motivate several aspects of this dissertation project.

There are so many people who helped me since I arrive in the United States. I will always cherish those memories. I am truly blessed to have so many friends who had been patient to share my difficult time and offer heartwarming advices. Life may or may not bring us together in the future, but those friendships shall remain forever in my heart. Last but not the least, I would like to thank my parents for being always there for me. They are and will always be the rock and love of my life.

The years at Rutgers has been a remarkable journey for me. I thank Rutgers Statistics and Biostatictics Department for providing us students top-notch research resources and the continuous graduate support. I would also like to acknowledge the support of the following research grants, Federal Aviation Administration grant # DOT-FAA 09-G-017, National Science Foundation grant # DMS 1513483 and # DMS 1007683.

Disclaimer

The discussion on aviation safety in this dissertation reflects the views of the authors, who are solely responsible for the accuracy of the analysis results presented herein, and does NOT necessarily reflect the official view or policy of the FAA. The dataset used in this dissertation has been partially masked in order to protect confidentiality.

Dedication

To my parents, for their unconditional and persistent love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the recent advances in computing and storage technologies, complex data such as multivariate and functional data are routinely collected in many fields. There is a strong demand for effective statistical tools for the analysis of these complex data sets, especially functional data sets. Roughly speaking, a functional data point contains continuous measurements of the same object over time or other continuum, but is observed or recored only on finite discrete indices. The book Ramsay and Silverman (2005) and Ferraty and Vieu (2006) provide excellent treaties of this subject. Other more recent studies include functional PCA in Yao et al. (2005), functional regression in Muller and Stadtmuller (2005) and Delaigle and Hall (2012), just to name a few. However, most of existing approaches for functional data rely heavily on parametric assumptions such as normality. However, these assumptions are often difficult to verify or justify in practice. In this dissertation, we develop general nonparametric approaches for multivariate and functional data, which are useful to solve problems in outlier detection, tolerance tubes construction, and clustering.

Data depth, or depth for short, measures the centrality of any data point with respect to its underlying distribution or a data cloud, and thus gives rise to a natural center-outward ordering to data points in any given sample. Depth has been used to develop a class of effective nonparametric approaches to solve many problems in multivariate and functional data analysis (e.g., see Liu et al. (1999)). Nevertheless, due to the location-scale free nature,

many of them are incapable of reflecting the magnitude of deviation from each sample point to the deepest one, which can be used as an estimator of the "center". This shortcoming restricts severely the utility of depth, especially in the context of outlier detection. Thus, we proposed a general approach, referred to as Antipodal Reflection Depth (ARD), to refine any well-defined depth notion to gain the capability of capturing the relative magnitude of deviations from any data point to the center. The idea of ARD is to combine the antipodal reflection of the original sample data in the calculation of depth values but draw inferences using only the original sample with their associated depth values. This approach is completely data driven and nonparametric. It is illustrated by simulated studies and the risk management project on tracking aircraft landing performance to identify possible anomalous landings.

Tolerance intervals and tolerance regions are important tools for statistical quality control and process monitoring of univariate and multivariate data, respectively. Guttman (1970) and a recent monograph Krishnamoorthy and Mathew (2009) provide detailed discussion about this topic. Li and Liu (2008) proposed an effective approach to construct nonparametric tolerance regions using multivariate spacings derived from data depth. In this dissertation, we generalize the tolerance intervals/regions to tolerance tubes in the infinite dimensional setting for functional data. In addition to the generalizations of the commonly accepted definitions of the tolerance level of $\beta-$content or $\beta-$expectation, we introduce a modification of $\beta-$expectation tolerance tube by coupling it with an exempt level α. The latter relaxes the definition of $\beta-$expectation tolerance tube by allowing α (usually pre-set by domain experts) portion of each functional be exempt from the requirement. Specifically, a $\beta-$expectation tolerance tube with exempt level α of a sample of n functional data is expected to contain $n\beta$ functionals such that each of these functionals has at least $(1-\alpha)$ portion falling within the limits of the tube.

The proposed tolerance tubes are completely nonparametric and thus broadly applicable. We investigate their theoretical properties and justifications. We also show that this exempt tolerance tube is particularly useful in the setting where occasional short term aberrations of functional data are deemed acceptable if those aberrations do not cause substantive deviation of the norm. This desirable property is elaborated and tested further with both simulation and real applications in continuous monitoring of blood glucose levels in diabetes patients as well as of aviation risks during aircraft landing operations.

Cluster analysis plays an important role in many research areas, such as artificial intelligence, recommendation system, natural language processing, etc. In this dissertation, we propose a new clustering method for functional data, and this method is referred to as K-mean on Pairwise Distance (KMPD) method. Roughly speaking, KMPD performs clustering by evaluating the pairwise distances of each point with respect to the whole data set. It mitigates the well-known shortcoming of the popular K-means (which was introduced and studied in (Lloyd, 1957; Forgy, 1965; MacQueen, 1967; Hartigan and Manchek, 1979) that i) it tends to produce clusters with similar sizes; ii) its effectiveness relies heavily on the assumption of spherical data structure. In contrast, KMPD outperforms K-means in that i) it is able to identify small clusters if exist; ii) it recovers the true data structure when data points are not distributed in separated spheres. The method has been tested and verified using simulated data and a real aircraft landing data set.

The rest of this dissertation is organized as follows. Chapter 2 gives a brief review of data depth for both multivariate and functional data, which facilitates our discussions in the next two chapters. Chapter 3 introduces and studies theoretical properties of ARD, as well as its application in outlier detection. Chapter 4 focuses on the construction of nonparametric tolerance tubes. Chapter 5 proposes KMPD for clustering multivariate and functional data. At last, Chapter 6 summarizes the main achievement of the dissertation.

Chapter 2

Review of Data Depth

2.1 Review of Data Depth in Multivariate Settings

In this section, we begin with basic notations for multivariate data. We adhere to the convention of denoting random variables and observations in upper cases and lowercases, respectively. We also denote vectors in boldface and scalars in plain. Let $X_1, \ldots, X_n$ be a random sample from the probability distribution F in $\mathbb{R}^d$, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ their observed values. For any $\boldsymbol{x} \in \mathbb{R}^d$,

1. *Simplicial Depth (SD)* (Liu (1990)) at $\boldsymbol{x}$ w.r.t. F is defined as

$$SD(\boldsymbol{x}, F) = P_F\{\boldsymbol{x} \in S[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}]\}$$

where $S[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}]$ is a closed simplex whose vertices $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{d+1}\}$ are $(d+1)$ points randomly selected from F. Given a sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, the sample SD is defined as

$$SD_n(\boldsymbol{x}) = \frac{1}{\binom{n}{d+1}} \sum_{(\star)} I(\boldsymbol{x} \in S[\boldsymbol{X}_{i,1}, \ldots, \boldsymbol{X}_{i,d+1}])$$

where $I(\cdot)$ is an indicator function. Each $\{\boldsymbol{X}_{i,1}, \ldots, X_{i,d+1}\}$ is a subset of $(d+1)$ points from the sample, and $(\star)$ represents all possible subsets of $(d+1)$ points. The sample SD is the fraction of sample simplices which contains x.

2. *Halfspace Depth (HD)* (Tukey (1975))of $\boldsymbol{x}$ w.r.t. F is defined as

$$HD(\boldsymbol{x}, F) = \inf_{H}\{P(H) : \boldsymbol{x} \in \text{H and H is a closed half-space in } \mathbb{R}^d\}$$

Given a sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, the sample HD is defined as:

$$HD_n(\boldsymbol{x}) = \frac{1}{n} \min_{\boldsymbol{e}\in\mathbb{R}^d} \sum_{i=1}^{n} I(\boldsymbol{X}_i'\boldsymbol{e} \leq \boldsymbol{x}'\boldsymbol{e}).$$

It is the minimum fraction of data points in any closed half-space in $\mathbb{R}^d$ that contains x.

There are many other depth notions for multivariate data, see Zuo and Serfling (2000) and Liu et al. (1999). We focus mainly on SD and HD in this paper, because these two geometric depth notions are completely nonparametric and data driven.

2.2 Review of Data Depth in Functional Settings

In the functional setting, we let $\{\boldsymbol{Y}_1(t),\, t \in \mathcal{T}\}, \ldots, \{\boldsymbol{Y}_n(t),\, t \in \mathcal{T}\}$ be a random sample of d-dimensional functionals over a common support $\mathcal{T}$, and $\{\boldsymbol{y}_1(t),\, t \in \mathcal{T}\}, \ldots, \{\boldsymbol{y}_n(t),\, t \in \mathcal{T}\}$ their observed values. While $\mathcal{T}$ often refers to time, it is defined more broadly and can refer to other suitable continua such as spatial position, etc. In practice, functional data are observed discretely which are indexed by t, and the collection of indices may also vary from observation to observation. For simplicity, we assume all functional are observed at p finite indices, i.e. $\boldsymbol{t} = (t_1, \ldots, t_p)$, and the observed functional can be denoted as $\boldsymbol{y}(\boldsymbol{t})$. We stress that our new proposed approach is also suitable when the observed indices are different between functionals. We simplify the notations of functional data by dropping $t \in \mathcal{T}$ when the emphasis of $\mathcal{T}$ is not needed. Furthermore, we can simply use $\boldsymbol{Y}$ and $\boldsymbol{y}$ to represent the functionals when there is no possibility of confusion. For example, we

will use $FARD(\boldsymbol{y})$ to denote the functional antipodal reflection depth of $\{\boldsymbol{y}(t)\}$, as seen in Chapter 3.

For any univariate functional $\{y(t)\}$ over $\mathcal{T}$,

1.*Fraiman and Muniz Depth* (FM) (Fraiman and Muniz, 2001) is defined as

$$ID(y) = \int_{\mathcal{T}} D(y(t))dt,$$

where $D(\cdot)$ is a general notion of depth for univariate data. FM is the integration of pointwise depth value over $\mathcal{T}$. The sample FM is defined as

$$ID_n(y) = \int_{\mathcal{T}} D_n(y(t))dt,$$

where $D_n(\cdot)$ represents the sample version of $D(\cdot)$.

2. *Random Projection Depth (RP) and Double Random Projection Depth (RP2)* (Cuevas et al., 2007)

For any $\{y(t)\}$ in Hilbert space $L^2[0,1]$, we take a series of independently random direction $\{a_1(t)\}, \ldots, \{a_N(t)\}$ and define

$$RP(y) = \frac{1}{N}\sum_{i=1}^{N} D(< a_i, y >).$$

Here $< a_i, y >= \int_{\mathcal{T}} a_i(t)y(t)dt$ is a projection of $\{y(t)\}$ on $\{a_i(t)\}$. The sample RP is defined as

$$RP_n(y) = \frac{1}{N}\sum_{i=1}^{N} D_n(< a_i, y >).$$

In addition to considering random projections in RP, RP2 takes into account random projections simultaneously of functionals and their first-order derivatives. Specifically, it

is defined as

$$RP2(y) = \frac{1}{N}\sum_{i=1}^{N} D((< a_i, y >, < a_i, y' >)),$$

where y' is the first-order derivative of y. The sample RP2 is defined as

$$RP2_n(y) = \frac{1}{N}\sum_{i=1}^{N} D_n((< a_i, y >, < a_i, y' >)).$$

3. *Band Depth (BD) and Modified Band Depth (MBD)* (Lopez-pintado and Romo, 2009)

$$BD_J(y) = \sum_{j=2}^{J} P\{G(y) \subset B(Y_1, \ldots, Y_j)\},$$

where $G(y) = \{(t, y(t)) : t \in \mathcal{T}\}$, and $B(Y_1, \ldots, Y_j) = \{(t, y) : t \in \mathcal{T},\ \min_{r=1,\ldots,j} y_r(t) \leq y \leq \max_{r=1,\ldots,j} y_r(t)\}$ is a banded region formed by the bounds of j randomly selected curves $\{Y_1(t)\}, \ldots, \{Y_j(t)\}$ over $\mathcal{T}$. The sample BD is defined as

$$BD_{n,J}(y(t)) = \sum_{j=2}^{J} \frac{1}{\binom{n}{j}} \sum_{(\star)} I\{G(y) \subset B(y_{i_1}, \ldots, y_{i_j}), t \in \mathcal{T}\}.$$

A modified version of BD is defined as

$$MBD_J(y) = \sum_{j=2}^{J} E\lambda_r(B(y; Y_1, \ldots, Y_j)),$$

where $B(y; Y_1, \ldots, Y_j) \equiv \{t \in \mathcal{T} : \min_r Y_r(t) \leq y(t) \leq \max_r Y_r(t)\}$. Here $\lambda_r(B(y; Y_1, \ldots, Y_j))$ measures the proportion of $\mathcal{T}$ where $\{y(t)\}$ is in the banded region

$B(y; Y_1, \ldots, Y_j)$. The sample MBD is defined as

$$MBD_{n,J}(y) = \sum_{j=2}^{J} \frac{1}{\binom{n}{j}} \sum_{(\star)} \lambda_r(B(y; y_{i_1}, \ldots, y_{i_j})).$$

The choice of parameter J is suggested in Lopez-pintado and Romo (2009) to be 3 for BD and 2 for MBD.

4. *Extremal Depth (ED)*(Narisetty and Nair, 2015) is defined to be

$$ED(y, F) = 1 - P\{y \prec f\},$$

where f is a random function drawn from F, and $\prec$ indicates the *extremal ordering* of functions. Specifically, for any two functions f and g, $f \prec g$ if and only if $\text{argmin}_r\{\phi_f(r) > \phi_g(r)\} < \text{argmin}_r\{\phi_f(r) < \phi_g(r)\}$, where $\phi_f(\cdot)$ and $\phi_g(\cdot)$ are *probability mass functions*.

For any d-dimensional functional $\{\boldsymbol{y}(t)\}$,

5. *Multivariate Functional Halfspace Depth (MFHD)*(Claeskens et al., 2014) is defined as

$$MFHD(\boldsymbol{y}; F_{\boldsymbol{Y}}, \alpha) = \int_{\mathcal{T}} HD(\boldsymbol{y}(t); F_{\boldsymbol{Y}(t)}) \cdot \omega_\alpha(t, F_{\boldsymbol{Y}(t)}) dt.$$

Here, $\alpha \in (0, 1]$, $\omega_\alpha(t, F_{\boldsymbol{Y}(t)}) = [vol\{HD_\alpha(F_{\boldsymbol{Y}(t)})\}]/[\int_{\mathcal{T}} vol\{HD_\alpha(F_{\boldsymbol{Y}(s)})\} ds]$ and $HD_\alpha(F_{\boldsymbol{Y}(t)}) = \{\boldsymbol{y} \in \mathbb{R}^d : \ HD(\boldsymbol{y}, F_Y(t)) \geq \alpha\}$ is α-central region at any fixed t. $MFHD(\cdot)$ applies a weight function to pointwise halfspace depth, where the weight is proportional to the volume of α-central region at each t. The sample MFHD is defined as

$$MFHD_n(\boldsymbol{y}; \alpha) = \sum_{j=1}^{p} HD(\boldsymbol{y}(t_j); F_{\boldsymbol{Y}(t_j),n}) \cdot \omega_\alpha(t_j, F_{\boldsymbol{Y}(t_j),n}),$$

with $\omega_\alpha(t, F_{\boldsymbol{Y}(t_j),n}) = [vol\{HD_\alpha(F_{\boldsymbol{Y}(t_j),n})\}]/[\sum_{j=1}^{p} vol\{HD_\alpha(F_{\boldsymbol{Y}(t_j),n})\}]$.

Note that FM, RP, RP2, BD, and MBD only apply to univariate functionals. Although MFHD applies to multivariate functionals, and it involves the tuning parameter α which might be difficult to determine in practice, especially when the dimension of functional is large.

Spurred by the rapid development of functional data analysis, proposals for functional depth notions have grown rapidly as well. Due to the space limitation, we consider only a few that are nonparametric and data driven.

Chapter 3

Antipodal Reflection Depth and Its Application to Multivariate and Functional Data Analysis

3.1 Introduction

Data depth, or depth for short, measures the centrality of any point with respect to its underlying distribution or a data cloud, and thus gives rise to a natural center-outward ordering to the points in a given sample. Depth has been developed into powerful nonparametric approaches for multivariate and functional data analysis. There are many notions of depth (Liu et al., 1999; Zuo and Serfling, 2000). Among them, the geometric ones are particularly useful especially when the underlying distribution is unknown. Nevertheless, due to the location-scale free nature, many of them are incapable of capturing the magnitude of deviation from each sample point to the deepest one. This shortcoming can often severely restrict the utility of depth, such as in the context of outlier detection. The goal of this paper is to mitigate this shortcoming by introducing a general approach, referred to as antipodal reflection depth (ARD, henceforth), to refine any existing notion of depth (referred to as base depth henceforth) to generate a class of new notions that possess the properties expected of a well-defined notion of depth.

More specifically, the refined notion i) inherits the desirable properties from its base depth; ii) yields a center-outward ordering which simultaneously captures the centrality obtained

by the base depth and the relative magnitude of deviation. This approach is completely nonparametric and data driven. Thus, ARD not only eliminates the aforementioned shortcoming, it also substantially broadens the applicability of depth as a whole.

The key idea of ARD is to take into account an antipodal reflection sample in the calculation of depth values. More specifically, for any given sample, we first obtain its deepest point using a proper base depth. Then, we reflect each sample point against the deepest one to obtain its antipodal reflection. The collection of all the reflections will be referred to as the *antipodal reflection* sample. Finally, we obtain ARD by applying the base depth to the pooled sample, which combines the original sample and the antipodal reflection sample. Note that, for convenience, we also use ARD to denote the new depth when it does not cause any confusion.

We give a simple illustration by applying ARD to the following sample $\{-200, -40, -1, -0.5, -0.2, 0.2, 1, 1.5, 1.8, 2\}$ using simplicial depth as the base depth. We first obtain the deepest point, namely, 0 using the base depth. By reflecting against 0, we obtain the antipodal reflection sample $\{200, 40, 1, 0.5, 0.2, -0.2, -1, -1.5, -1.8, -2\}$. We then apply simplicial depth to this pooled sample $\{-200, -40, -1, -0.5, -0.2, 0.2, 1, 1.5, 1.8, 2, 200, 40, 1, 0.5, 0.2, -0.2, -1, -1.5, -1.8, -2\}$ to yield their ARD values. All points from the antipodal reflection sample are steppingstone to generate the depth values, and thus will be removed afterwards.

It is natural to expect ARD to inherit the intrinsic properties of its base depth, including the four proposed by Liu (1990) and Zuo and Serfling (2000). However, to establish those properties for ARD, we need to overcome some non-trivial difficulties, especially the dependence structure, created from the antipodal reflection, in the pooled sample. In this chapter, in addition to the four basic properties mentioned above, we justify that ARD

can be approximated by its sample version consistently. Similarly, we justify that the generalization of ARD in functional settings (referred to as FARD later) also possesses many desirable properties. These properties are recommended in Claeskens et al. (2014) and other papers for a well-defined notion of functional depth. Consistency results are also provided.

ARD has many immediate applications. For example, it can be applied to design a test of symmetry, to construct a sharper tolerance region or tube, or to detect outliers, just to name a few. In this dissertation, we mainly focus on the application of ARD in outlier detection for both multivariate and functional settings. In multivariate settings, there is a rich literature on outlier detection. However, most existing approaches require the distribution to be gaussian or at least in the exponential family. In practice, this is difficult to verify and thus casts doubt on the validity of those approaches. In functional settings, there are only limited proposals. Some of them first convert the functional data to multivariate and then solve the outlier detection problem in the multivariate setting. This scheme risks losing or distorting the important features of the original functional data. Some others resort to functional depth, which returns a "center-outward" ordering analogous to its multivariate counterpart. However, the existing functional depth notions again suffer the incapability of capturing the magnitude of deviation to the deepest point. In this paper, we use ARD to introduce a systematic nonparametric approach to detect outliers in both multivariate and functional settings. The property of ARD that it assesses the depth values of sample points together with their relative deviations from the deepest point provides a more effective scheme for outlier detection, and is key in making our approach more powerful than the usual depth approaches.

The rest of this chapter is organized as follows. In Section 3.2, we formally introduce ARD in multivariate settings, and its generalization, namely, FARD in functional settings.

Theoretical properties of ARD and FARD are studied in depth. In Section 3.3, we focus on the application of ARD in outlier detection. We compare the performance of ARD with other depth using simulated data. ARD is shown to be more effective by producing better sensitivity-specificity results. In Section 3.4, we apply ARD to an aircraft landing analysis to identify possible anomalous landings. From the comparison with other depth notions, ARD is shown to be more desirable in terms of identifying the landings which substantially deviate from the "benchmark". It is worth noting that it is this real data application which motivates the idea of ARD. We stress that ARD is not merely a theoretical generalization of the existing depth notions. It is in fact driven by the need in many practical applications to account for the relative deviations from observations to the center of the data cloud. Some concluding remarks are given in Section 3.5.

All theoretical proofs are deferred to Section 3.6.

3.2 Antipodal Reflection Depth (ARD)

3.2.1 ARD of Multivariate Data

Let F be a distribution in $\mathbb{R}^d$, and $\boldsymbol{\theta}$ the deepest point obtained by a suitable base depth $D(\cdot)$ mentioned in Section 2.1. For a random sample point $\boldsymbol{X}$ from F, its antipodal reflection is defined to be $\widetilde{\boldsymbol{X}} := 2\boldsymbol{\theta} - \boldsymbol{X}$. (The term "antipodal" is first mentioned in Liu et al. (1999) in the context of defining antipodal symmetry, namely, F is antipodally symmetric if $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ are identically distributed.)

Definition 3.1. [Population ARD] *Let* $\boldsymbol{Z} = \boldsymbol{\theta} + \epsilon(\boldsymbol{X} - \boldsymbol{\theta})$, *where* $P\{\epsilon = 1\} = P\{\epsilon = -1\} = 0.5$ *and* ϵ *independent of* $\boldsymbol{X}$. *We further denote the distribution of* Z *as* G. *Then,*

for any $\boldsymbol{x} \in \mathbb{R}^d$*, we define the ARD of* $\boldsymbol{x}$ *w.r.t.* F *as*

$$ARD(\boldsymbol{x}, F) = D(\boldsymbol{x}, G)$$

In principle, ARD can be defined or implemented using any well-defined depth as the base depth. To simplify the illustration, we use SD as the base depth throughout the paper. Specifically, for a random sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{d+1}$ *from* G*, ARD can be expressed as*

$$ARD(\boldsymbol{x}, F) = P_G\{\boldsymbol{x} \in S[\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{d+1}]\},$$

where $S[\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{d+1}]$ *is a closed simplex formed by* $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{d+1}\}$*.*

Definition 3.2. [Sample ARD] *Let* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ *be a random sample from* F*. Let* $\hat{\boldsymbol{\theta}}_n$ *be the sample deepest point derived from the base depth, and* $\boldsymbol{X}_1^\star, \ldots, \boldsymbol{X}_n^\star$ *the antipodal reflection of the original sample around* $\hat{\boldsymbol{\theta}}_n$*. We denote the pooled sample, consisting of the original sample and their antipodal reflection, by* $\{\boldsymbol{Y}_1^\star, \ldots, \boldsymbol{Y}_{2n}^\star\} := \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n, \boldsymbol{X}_1^\star, \ldots, \boldsymbol{X}_n^\star\}$*. Then, for any* $\boldsymbol{x} \in \mathbb{R}^d$*, the sample ARD is defined as*

$$ARD_n(\boldsymbol{x}) = D_{2n}^\star(\boldsymbol{x}),$$

where $D_{2n}^\star(\cdot)$ *is the base depth function w.r.t. the pooled sample. For example, using SD as the base depth, we obtain*

$$ARD_n(\boldsymbol{x}) = \frac{1}{\binom{2n}{d+1}} \sum_{(\star)} I(\boldsymbol{x} \in S[\boldsymbol{Y}_{i,1}^\star, \ldots, \boldsymbol{Y}_{i,d+1}^\star]),$$

where $(\boldsymbol{Y}_{i,1}^\star, \ldots, \boldsymbol{Y}_{i,d+1}^\star)$ *is any subset of* $(d+1)$ *points from the pooled sample.*

Remark 3.1. *There is an intermediate version of sample ARD, which is defined as follows. For* $1 \le i \le n$*, we let* $Z_i = \boldsymbol{\theta} + \epsilon_i(\boldsymbol{X}_i - \boldsymbol{\theta})$*. Then, for any* $\boldsymbol{x} \in \mathbb{R}^d$*, the sample ARD of* $\boldsymbol{x}$ *is*

defined as

$$ARD_n(\boldsymbol{x}) = D(\boldsymbol{x}, G_n),$$

where $D(\boldsymbol{x}, G_n)$ *is base depth function w.r.t.* G_n*. However, in reality,* $\boldsymbol{\theta}$ *is usually unknown and* $Z_i's$ *can not be observed. Thus, we use Definition 2 as an approximation of the aforementioned intermediate version. The justification is that this approximation error is negligible asymptotically. Definition 2 is not only easy to understand and implement, but also results in an asymptotically unbiased estimator of population ARD.*

In Theorem 3.1, we show that ARD satisfies four main properties introduced by Liu (1990) which subsequently were used in Zuo and Serfling (2000) as the required properties for a notion of depth. To be precise, we show that if the base depth satisfies the four properties, so will ARD.

Theorem 3.1. *Let* F *be a distribution in* $\mathbb{R}^d$*, for any* $\boldsymbol{x} \in \mathbb{R}^d$*: (a) Vanish in infinity:* $sup_{\|\boldsymbol{x}\| \geq M} ARD(\boldsymbol{x}) \to 0$*, as* $M \to \infty$

(b) Monotonicity: If F *is absolutely continuous and has the deepest point* $\boldsymbol{\theta}$*. For any* $\boldsymbol{x} \in \mathbb{R}^d$*,* $ARD(\alpha(\boldsymbol{x} - \boldsymbol{\theta}), F)$ *is monotone non-increasing in* α *for* $\alpha \geq 0$*.*

(c) Maximality at the center: If F *is absolutely continuous, ARD is maximized at the deepest point which maximizes its base depth function and retains its original properties.*

(d) Affine invariance: $ARD(A\boldsymbol{x} + \boldsymbol{b}, F_{A\boldsymbol{x}+\boldsymbol{b}}) = ARD(\boldsymbol{x}, F_{\boldsymbol{x}})$*, for any* $d \times d$ *nonsingular matrix* A *and any* $\boldsymbol{b} \in \mathbb{R}^d$*.*

Theorem 3.2. *Let* F *be an absolutely continuous distribution on* $\mathbb{R}^d$ *with bounded density* f*. Then:*

$$\sup_{\boldsymbol{x} \in \mathbb{R}^d} |ARD_n(\boldsymbol{x}) - ARD(\boldsymbol{x}, F)| \to 0 \text{ a.s. as } n \to \infty.$$

Proposition 3.1. *Let* F *be an absolutely continuous distribution on* $\mathbb{R}^d$ *with bounded density* f*, and* $\boldsymbol{\theta}$ *be the deepest point of* F*. Let* $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ *be a random sample from* F

and $\hat{\boldsymbol{\theta}}_n$ the sample deepest point. If f does not vanish in a neighborhood of $\boldsymbol{\theta}$ and the base depth is uniquely maximized at $\boldsymbol{\theta}$, it holds that $\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}$, as $n \to \infty$.

Proposition 3.2. *Sample ARD is an asymptotically unbiased estimator of population ARD, namely, for any $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\lim_{n\to\infty} E(ARD_n(\boldsymbol{x})) = ARD(\boldsymbol{x}, F)$$

Remark 3.2. *Although ARD is defined based on a new distribution G, only its values and orderings on the original sample are useful for the inference of F. The new distribution G is merely a catalyst to facilitate the modification of the base depth to ARD, but not to alter the relative ordering between any two points along the same ray from the deepest point.*

3.2.2 ARD of Functional Data

Let $F_{\mathcal{T}}$ be a functional distribution over the support $\mathcal{T}$, and F_t the distribution of the functionals at t. For each t, we obtain the deepest point w.r.t. F_t as $\boldsymbol{\theta}(t)$ from the base depth, which we call $\{\boldsymbol{\theta}(t)\}$ the *deepest functional*. For any random functional $\{\boldsymbol{Y}(t)\}$ from $F_{\mathcal{T}}$, its antipodal reflection is defined to be $\{\widetilde{\boldsymbol{Y}}_i(t) = 2\theta(t) - \boldsymbol{Y}_i(t)\}$. We define the functional ARD (henceforth FARD) as follows.

Definition 3.3. [Population FARD] *For any functional $\{\boldsymbol{y}(t)\}$, the FARD is defined as*

$$FARD(\boldsymbol{y}) = \frac{1}{\|T\|}\int_{\mathcal{T}} ARD(\boldsymbol{y}(t), F_t)dt$$

where $ARD(\boldsymbol{y}(t), F_t)$ is the multivariate ARD of $\boldsymbol{y}(t)$ w.r.t. F_t, and $\|T\|$ is the Lebesgue measure of $\mathcal{T}$.

Definition 3.4. **[sample FARD]** *Assume that all functionals are observed on discrete indices* $\boldsymbol{t} = (t_1, \ldots, t_p)$. *Given a functional sample* $\boldsymbol{Y}_1(\boldsymbol{t}), \ldots, \boldsymbol{Y}_n(\boldsymbol{t})$, *for any functional* $\boldsymbol{y}(\boldsymbol{t})$, *the sample FARD is defined as*

$$FARD_n(x) = \frac{1}{\sum_{j=1}^{p} \Delta t_j} \sum_{j=1}^{p} ARD_n(\boldsymbol{y}(t_j))\Delta t_j,$$

where $\Delta t_j = t_j - t_{j-1}$ *for* $j = 1, \ldots, p$, *and* $t_0 = \inf\{t : t \in \mathcal{T}\}$.

Note that the deepest functional obtained by FARD is a collection of pointwise ARD deepest point over $\mathcal{T}$, which may not necessarily be a functional data point in the sample. In many applications, the deepest functional can be regarded as the collective benchmark suggested by the sample functionals. Case in point is such a benchmark for aircraft landing performance used in our application in outlier detection in Section 3.4.

Claeskens et al. (2014) extended the four expected properties to functional settings. In Theorem 3.3, we show that FARD also satisfies these properties. Moreover, FARD also possesses the desirable property of non-degeneracy. This property is lacking in many notions of functional depth, such as BD and projection depth, and is thoroughly investigated in Chakraborty and Chaudhuri (2014).

Theorem 3.3. *If the base depth (or ARD) satisfies the four properties listed in Theorem 3.1, then FARD satisfies:*

(a) Vanish in infinity: For a series of functionals $\boldsymbol{y}_n$, *if there exists some set* A *with* $\lambda(A) = \lambda(\mathcal{T})$, *such that for* $\forall t \in A$, $\|y_n(t)\| \to \infty$, $FARD(\boldsymbol{y}_n, F_{\boldsymbol{Y}}) \to 0$ *as* $n \to \infty$.

(b) Monotonicity: Let $\boldsymbol{\theta} \in C(\mathcal{T}^d)$ *be the deepest functional w.r.t.* $F_{\boldsymbol{Y}}$. *For any* $\boldsymbol{y} \in C(\mathcal{T}^d)$, $FARD(\alpha(\boldsymbol{y} - \boldsymbol{\theta}), F_{\boldsymbol{Y}})$ *is monotone non-increasing in* α *for* $\alpha > 0$.

(c) Maximality at the center: Assume $\boldsymbol{\theta}(t)$ *is uniquely defined for each* $t \in \mathcal{T}$. *FARD is maximized at* $\boldsymbol{\theta}$.

(d) Affine invariance: $FARD(\boldsymbol{y}, F_{\boldsymbol{Y}}) = FARD(A\boldsymbol{y} + S, F_{A\boldsymbol{y}+S})$*, where* A *is any* $d \times d$ *nonsingular matrix,* S *is any functional over* $\mathcal{T}$*.*

(e) Non-Degeneracy: For a large class of functionals including continuous time Gaussian processes, FARD does not degenerate when the sample size goes to infinity.

Remark 3.3. *Viewing ARD as a general scheme to count for the deviation from the sample points to the deepest one to improve inferences, FARD can be viewed as a special case of integrating ARD to FM depth. In general, for any functional depth notion which is defined only using the depth value calculated at each point, we can substitute the pointwise depth by its ARD counterpart, to yield a new functional depth notion. Again, the new notion will inherit the aforementioned four properties if those properties are satisfied by the original base depth. However, it is unclear how to integrate ARD to other functional depth notions such as RP, or whether the basic properties can be inherited. We will pursue this study in the future.*

In what follows, Theorem 3.4 and Theorem 3.5 give asymptotic properties of the sample FARD and the sample deepest functional.

Theorem 3.4. *Let* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ *be a sample of d-dimensional continuous functionals from distribution* $F_{\boldsymbol{Y}}$*. They are observed at the same set of indices* $\boldsymbol{t} = (t_1, \ldots, t_p)$*. We assume that the data are observed frequently such that* $\sup_{i=1,\ldots,p-1} |t_{i+1} - t_i| = O(p^{-(1/2+\gamma)})$ *for some* $\gamma > 0$*. In addition, we assume* $H(t) := ARD(\boldsymbol{y}(t)) \in Lip(\mathcal{T})$*. Under the condition in Theorem 3.2, it holds that*

$$\sup_{\boldsymbol{y} \in C(\mathcal{T})^d} \|FARD_n(\boldsymbol{y}) - FARD(\boldsymbol{y}, F)\| \to 0, a.s.\ n \to \infty, p \to \infty,$$

where $C(\mathcal{T})^d$ *denotes d-dimensional continuous functionals.*

Theorem 3.5. *Assume there exists a Lebesgue integrable function* $g_i(t)$ *such that* $|Y^{(i)}(t)| < g_i(t)$ *over* $t \in \mathcal{T}$ *for* $i = 1, \ldots, d$. *If* $\hat{\boldsymbol{\theta}}$ *and* $\boldsymbol{\theta}$ *are both uniquely defined, we obtain* $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ *almost surely.*

3.3 Applications of ARD In Outlier Detection

The presence of outliers in a given data set may cast undue influence on the analysis results and adversely affect the inference outcome, causing, for example, estimation bias or misclassification or clustering error, etc. Outlier detection is often needed to ensure reliable analysis or inference outcomes. In this paper, we adopt the commonly used definition of outliers in Grubbs (1969), saying "an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs". In this section, we investigate the applications of ARD in outlier detection in multivariate and functional settings, respectively.

3.3.1 Outlier Detection in Multivariate Settings

Outlier detection in multivariate setting has been investigated extensively (see, e.g., Rousseeuw and Leroy (1987)). Boxplot, bagplot and their variations are useful in visualizing the data structure, but they can not be naturally extended to data with dimension higher than three. Grubbs (1969), Hardin and Rocke (2005), Riani et al. (2009) propose tests to detect the existence of one or multiple outliers. But these tests are mainly designed for samples generated from exponential family, thus not as effective outside at realm. Using data depth, Cheng et al. (2000) proposes an effective nonparametric approach to set safety thresholds for monitoring multivariate aircraft performance measures. In this section, without imposing any assumption on the underlying distribution, we propose an

effective nonparametric approach using ARD to detect outliers systematically. Note that it is difficult to determine the portion of sample points which should be labeled as outliers. In practice, the portion, denoted as α thereafter, should reflect domain knowledge or past experience. In this paper, we assume that α is given throughout. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a sample in $\mathbb{R}^d$. We obtain the reflection sample $\boldsymbol{X}_1^\star, \ldots, \boldsymbol{X}_n^\star$, and pool the original and the reflection sample together to obtain the pooled sample $Y_1^\star, \ldots, Y_{2n}^\star$. Then, we identify the $(1-\alpha)$ central region of the pooled sample based on their ARD values, and identify those outside the $(1-\alpha)$ central region as outliers. For a simple illustration, we revisit the toy example in Section 3.1, and show how ARD can detect the outliers -200 and -40 for $\alpha = 0.2$. The original sample is:

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
-200	-40	-1	-0.5	-0.2	0.2	1	1.5	1.8	2

which gives the reflection sample:

$X_1^\star$	$X_2^\star$	$X_3^\star$	$X_4^\star$	$X_5^\star$	$X_6^\star$	$X_7^\star$	$X_8^\star$	$X_9^\star$	$X_{10}^\star$
200	40	1	0.5	0.2	-0.2	-1	-1.5	-1.8	-2

Then, we get the pooled sample: $\{Y_1^\star, Y_2^\star, \ldots, Y_{20}^\star\} = \{-200, -40, \ldots, 40, 200\}$. Applying SD to $\{Y_i^\star\}_{i=1}^{20}$, we get the order statistics, $Y_{[1]}^\star, Y_{[1]}^\star, Y_{[3]}^\star, Y_{[3]}^\star, \ldots, Y_{[19]}^\star, Y_{[19]}^\star$, in descending SD values. (There are ten ties in them due to the symmetric structure of the pooled sample.) To identify $\alpha = 20\%$ outliers, we obtain first the 80% central region, $CR_{[80\%]}$, of pooled sample, which contains the following 16 points:

$Y_{[1]}^\star$	$Y_{[1]}^\star$	$Y_{[1]}^\star$	$Y_{[1]}^\star$	$\ldots$	$Y_{[15]}^\star$	$Y_{[15]}^\star$
-0.2	0.2	-0.2	0.2	$\ldots$	2	-2
X_5	$X_5^\star$	X_6	$X_6^\star$	$\ldots$	X_{10}	$X_{10}^\star$

Now, discard those in $CR_{[80\%]}$ which are not from the original sample, what remains in the central region is $\{X_3, X_4, \ldots, X_{10}\} = \{-1, -0.5, \ldots, 2\}$. The data points $X_1 = -50$ and $X_2 = -40$ are outside this region and thus labeled as outliers.

Note that ARD has the additional benefit of breaking ties in depth values. For instance, applying SD directly to the original sample, we obtain a tie in $\{-200, 2\}$, and another one in $\{-40, 1.8\}$. It is easy to see from the example above that applying ARD can break these ties and obtain a more informative ordering that also accounts for their deviations from the deepest point. Thus, along a ray stemming from the deepest point, points with higher ARD values are always closer to the deepest point. As a result, ARD usually leads to a "sharper" tolerance interval (or region) in terms of Lebesgue measure. This is further studied and extended to tolerance tubes introduced in Chapter 4.

3.3.2 Simulation Studies – Multivariate Settings

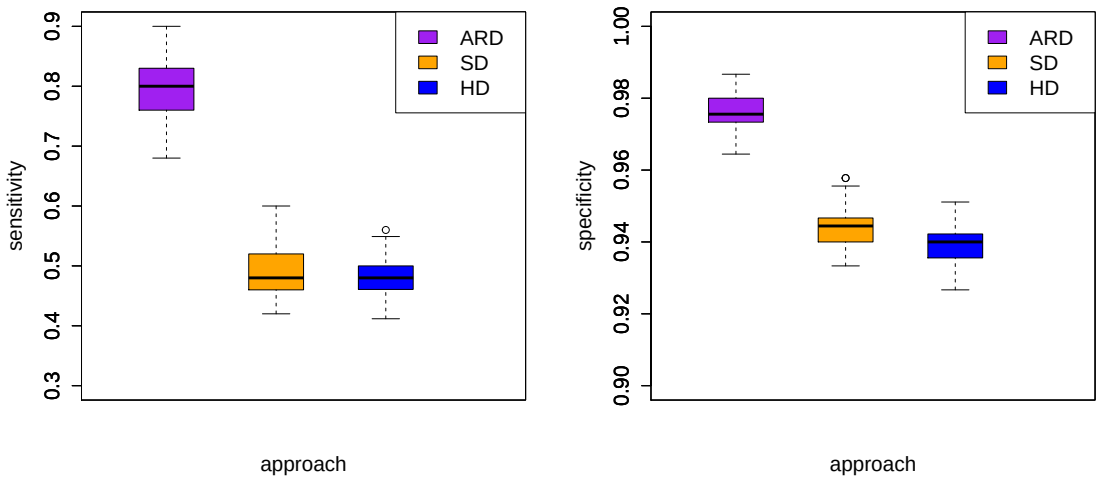
In the following, we use a simulated dataset to compare the performance of ARD approach with other depth approaches using HD and SD in outlier detection.

Simulation I. We generate 500 data points from a bivariate normal distribution $N_2(\mu, \Sigma)$ where $\mu = (0, 0)'$ and $\Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$. We contaminate 10% of data points by a different bivariate normal distribution $N_2(\mu^c, \Sigma^c)$, where $\mu^c = (3, 3)'$ and $\Sigma^c = \begin{bmatrix} 4 & -1.06 \\ -1.06 & 4 \end{bmatrix}$. By applying ARD, SD and HD to the sample, we select the 10% points with the lowest depth values as outliers, respectively. Note that, in practice, when we obtain a tie on 10% quantile of depth values, we randomly select a suitable amount of observations from the tie as outliers, such that the total number of outliers selected is exact 10% of the sample size. We

will stay with the same strategy in all the following numerical studies. Finally, we use two commonly used criteria, *sensitivity* and *specificity* to assess the accuracy and effectiveness of the outlier detection result of each approach. Here, we recall the definition of sensitivity and specificity as follows. Given any outlier detection result such as Table 3.1, the sensitivity and specificity are defined to be sensitivity $= \frac{a}{N_2}$, specificity $= \frac{d}{N_1}$. Roughly speaking, sensitivity measures the ability to identify an anomaly correctly, and specificity measures the ability to identify a regular point correctly. The two criteria together can provide a fair evaluation of outlier detection results. The sample and the result are displayed in Figure 3.3 and Figure 3.1, respectively. The simulation is replicated by 50 times.

	true outlier	true regular
detected as outlier	a	b
detected as regular	c	d
	$N_2 = a + c$	$N_1 = b + d$

Table 3.1: Sensitivity and specificity calculation.



(a) sensitivity of Simulation I (b) specificity of Simulation I

Figure 3.1: Sensitivity and specificity comparison from Simulation I: Bivariate Normal Case.

Figure 3.1 (a) and (b) clearly show that the ARD approach outperforms substantially the approaches by SD and HD. For example, the IQR of the sensitivity of ARD is $[0.76, 0.84]$, while the other two are both tightly around 0.5. Clearly, it is the capability of capturing the relative deviation to the deepest point that makes ARD more effective in detecting outliers, especially in an asymmetric data setting.

3.3.3 Outlier Detection in Functional Settings

Recent rapid development of functional data analysis has helped draw more attention to the outlier detection in functional setting. Hyndman and Shang (2010) and Yu et al. (2012) apply functional principle component analysis to the sample and analyze only the first a few principle component scores. Specifically, Hyndman and Shang (2010) constructs a bagplot

to test the outliers in the first two principle component scores, and Yu et al. (2012) designs specific tests on the selected scores. However, these results are less reliable if the selected principle component scores are not able to characterize the sample well. Moreover, the latter approach relies on the Gaussian assumption on the scores. Sun and Genton (2011) and Febrero et al. (2008) apply functional depth to detect outliers. But these approaches are only applicable to univariate functional data.

The ARD approach introduced in Section 3.3.1 can be extended naturally to the functional setting. More specifically, we summarize the procedure in the following:

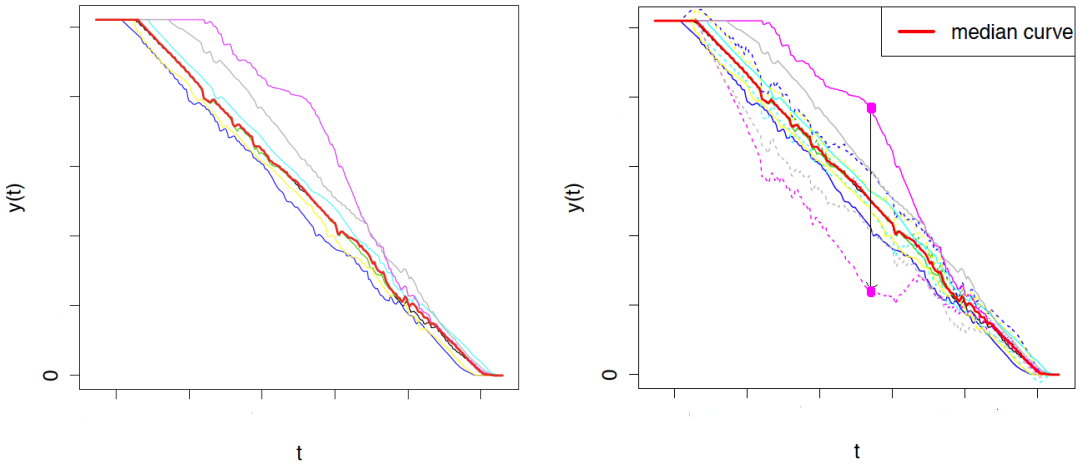
Step 1: On each $t \in \mathcal{T}$, we obtain the deepest point $\boldsymbol{\theta}_n(t)$. By connecting them throughout the interval $\mathcal{T}$, we obtain the deepest functional in the sample.

Step 2: Do antipodal reflection of all the data around the deepest point on each index t to obtain an antipodal reflection sample of the original sample. Combine the two datasets into a pooled sample.

Step 3: Within the pooled sample, we calculate depth value of each functional data. Then, we remove the antipodal sample, retaining the data from the original dataset only.

Step 4: Screen out α of the functional sample points attaining the lowest functional depth values as outliers.

We apply this approach to a functional dataset, which contains 7 functionals as in Figure 3.2 (a). We observe that the purple and the blue functionals are "outlying" w.r.t. the sample, but the blue one is much closer to the sample deepest functional. Applying ARD, we are able to discern such difference between the two and assign the lowest ARD value to the purple one. In this case, only the purple one is identified as the most likely outlier. On the contrary, if applying any non-distance based depth notion such as MBD, we would not be able to separate these two functionals.



(a) original curves (b) original curves and their reflection

Figure 3.2: An illustrative example: the left panel contains 7 curves in different colors. The right panel also includes their antipodal reflections around the median curve. The reflections are in dashed lines and preserve the same color as their corresponding original curves. The median curve is marked in red. The blue curve and purple curve are both assigned the lowest depth values in the left panel. But in the right panel, only the purple one is considered outlying.

3.3.4 Simulation Studies – Functional Settings

We now conduct three simulation studies to compare the performance of ARD approach with the others using functional depth notions reviewed in Chapter 2. In each simulation setting, we follow the same strategy as in multivariate settings that we replace 10% of the original sample with the data from a different stochastic process. By applying each depth function, the 10% points with the lowest depth values are labeled as outliers. The samples are displayed in Figure 3.3. Each simulation is repeated 50 times.

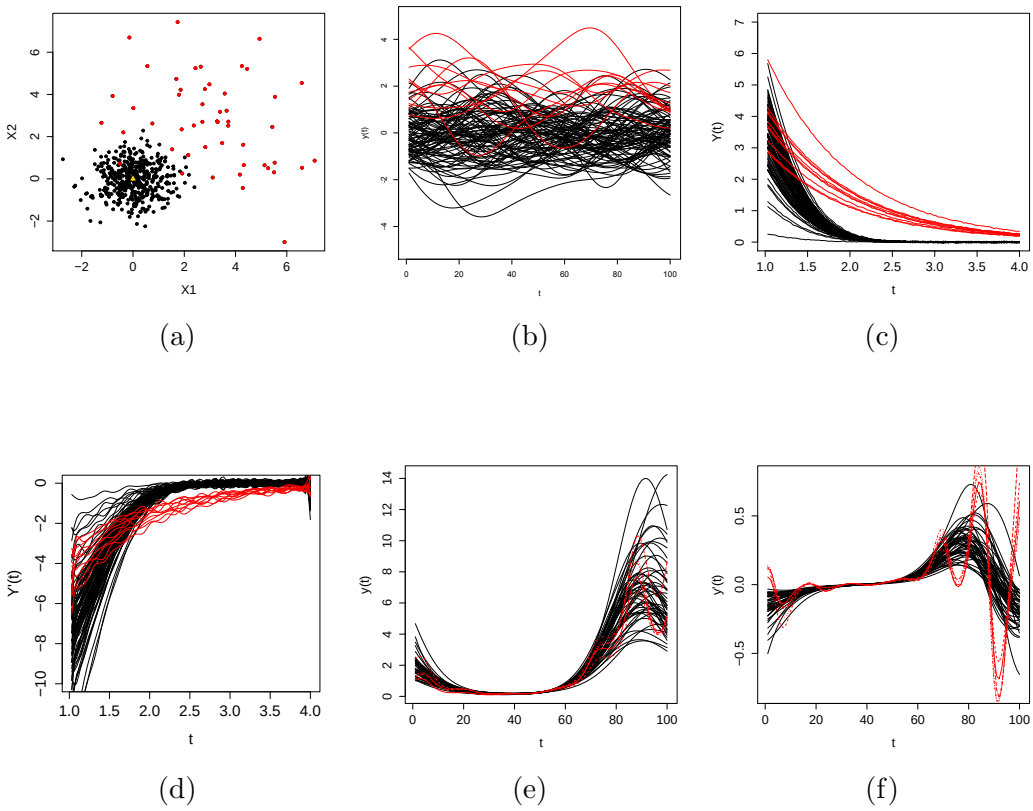
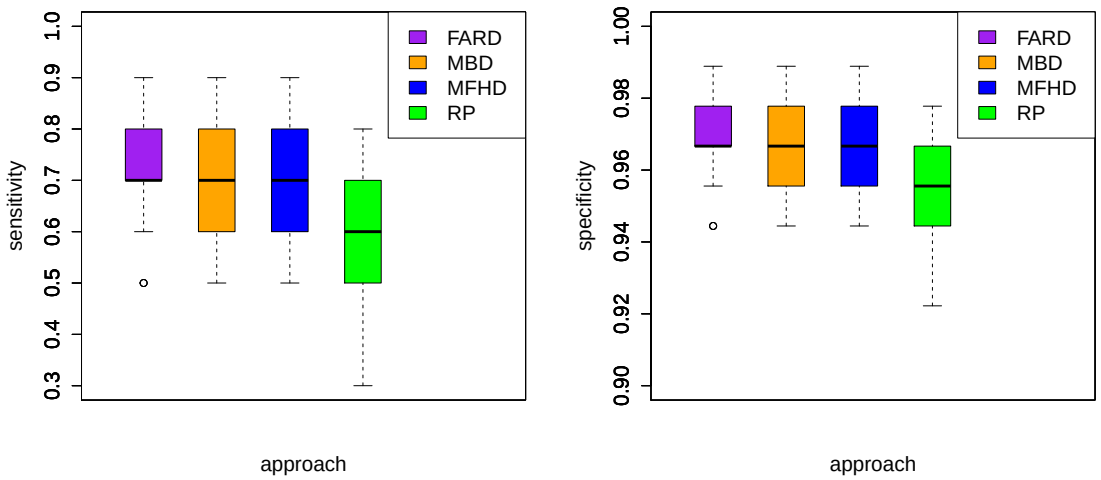


Figure 3.3: (a) contains 500 data points from the bivariate normal distribution with 50 points contaminated by a different bivariate normal distribution; (b) contains 100 curves generated from gaussian process: 90 black curves are regular curves, following $N(0, \Sigma(t))$; 10 red curves are contaminations, following $N(2, \Sigma(t))$; (c) and (d) are the original data and their derivatives from the process $Y(t) = x \cdot \exp\{-t^2\} + \epsilon$, where $x \sim N(10, 3^2)$ and $\epsilon \sim N(0, 0.01^2)$. Here, the 10 red curves are contaminations generated from $Y^c(t) = x \cdot \exp\{-t\} + 0.05$. (e) and (f) are respectively the original data and their derivatives from the log-normal process $Y(t) \sim logN(\mu(t), \Sigma(t))$ with the same covariance operator but different mean $\mu(t)$. Black curves follow the mean function $\boldsymbol{\mu}(\boldsymbol{t}) = K_{ts}(K_{ss} + D)^{-1}(a_1 \sin(\boldsymbol{s}) + a_2 \cos(\boldsymbol{s}))$ and the red contaminations follow $\boldsymbol{\mu}^\star(\boldsymbol{t}) = K^m(\sin(6\boldsymbol{s}) + \boldsymbol{\mu}(\boldsymbol{s}))$.

Simulation II. In this setting, we aim at detecting the outlier in the univariate gaussian process with shifted functional mean. We generate 100 functionals from a univariate gaussian process with mean $\mu(t) = 0$ and covariance kernel $K_y(s,t) = \exp\{-\frac{(y_{i,s}-y_{i,t})^2}{400}\}$ over

the interval $[0, 100]$. We randomly select 10% of the sample and replace them with the data from another gaussian process with the same covariance kernel but different mean, $\mu^c(t) = 2$. All the functionals are observed at equal spaced indices $t = 1, 2, \ldots, 100$.

We apply FARD, MBD, RP and MFHD to the sample, respectively. As shown in Figure 3.4 (a) and (b), FARD outperforms the other three with overall higher sensitivity and specificity. In particular, the comparison between FARD and MBD immediately shows that FARD improves the performance due to the capability of capturing the relative deviation from the deepest point without the effect of any weighting scheme. It is worth noting that despite the scale, the results of sensitivity and specificity are very similar. This is because the strategy employed to detect outliers. We fix the total number of outliers selected by each approach, which makes the sensitivity or specificity can be determined by each other. We also apply BD and FM to the samples. But their performance is no better than the current result, so we simply omit it here.



(a) sensitivity of Simulation II

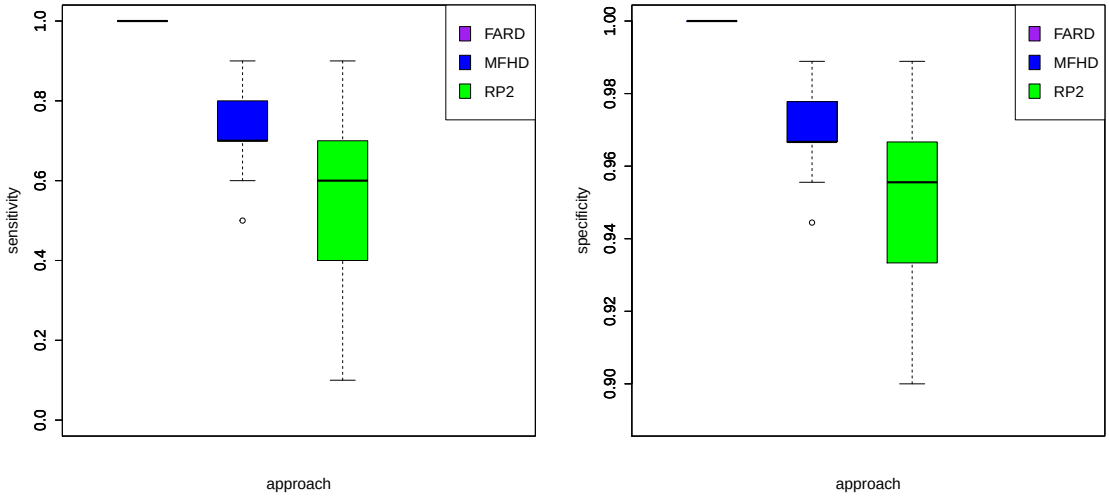
(b) specificity of Simulation II

Figure 3.4: Sensitivity and specificity comparison from Simulation II: Gaussian Process

Simulation III. In this setting, we study FARD in a multivariate functional setting. We generate 100 functionals from the process: $Y(t) = x \cdot \exp\{-t^2\} + e$ over the interval $[1, 4]$, where $x \sim N(10, 3^2)$ and $e \sim N(0, 0.01^2)$. We randomly select 10% of the sample and replace them with the data another process, $Y^c(t) = x \cdot \exp\{-t\} + 0.05 + e$, where x and e follow the same distribution as described before. All of the functionals are observed at 100 equally spaced indices, namely, $t = 1.03, \ldots, 4$. In addition, considering that derivatives may provide additional information about the shape of functionals, we create a bivariate functional sample $\{(y_i(t), y_i'(t))\}_{i=1}^{100}$, where $y_i'(t)$ is the first-order derivative of $y_i(t)$.

We apply FARD, MFHD and RP2 respectively to this sample. From Figure 3.5 (a) and (b), it is obvious that FARD outperforms the rest. It detects all the outliers accurately in every replication. The weighting scheme introduced in MFHD intrinsically underestimates the importance of the segment where the majority functionals behave

similarly. However, in this simulation, this segment is essential in distinguishing outliers from the regular functionals. In addition, it is difficult to determine the tuning parameter in MFHD, especially when the functionals are not smooth or are in high dimension. Inappropriate choices of the tuning parameter (e.g. α exceeds the maximal depth value in the sample and yields to an empty set of the $\alpha-$central region) would leave the MFHD undefined. We also observe that the performance of RP2 is less stable than others. The large variation in sensitivity and specificity may be due to the fact that they only use limited number of random projections to calculate depth values.



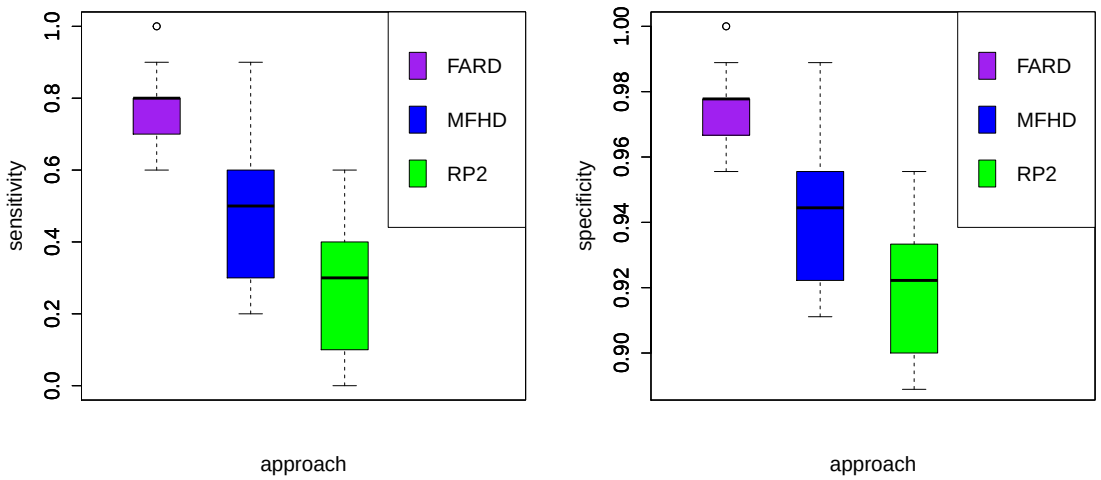
(a) sensitivity of Simulation III (b) specificity of Simulation III

Figure 3.5: Sensitivity and specificity comparison from Simulation III: Exponential Curves.

Simulation IV. In this setting, we aim at detecting outliers of log-normal process utilizing the information from its derivatives. This setting follows exactly simulation 4.2.3 in Claeskens et al. (2014). We generate 100 functionals from the log-normal process $Y(t) = \exp(X(t))$, where $X(t)$ is a Gaussian Process. Let $\boldsymbol{s} = \{s_1, \ldots, s_{20}\}$ and

$\boldsymbol{t} = \{t_1, \ldots, t_{100}\}$ be 20 and 100 equidistant indices on $[0, 2\pi]$, respectively. Then, the mean function of $X(t)$ is $\boldsymbol{\mu}(\boldsymbol{t}) = K_{ts}(K_{ss} + D)^{-1}(a_1 \sin(\boldsymbol{s}) + a_2 \cos(\boldsymbol{s}))$ and covariance function $\Sigma = K_{tt} - K_{ts}(K_{ss} + D)^{-1}K_{st}$. Here, $a_1 \sim U(-2, 2)$ and $a_2 \sim U(-1, 1)$ are two uniform variables, $D = Diag_{i=1,\ldots,20}\{min((\pi - s_i)^2, 1)\}$, $K_{ss}(i, j) = \exp(-8(s_i - s_j)^2)$, $K_{ts}(i, j) = \exp(-8(t_i - s_j)^2)$ and $K_{tt} = \exp(-8(t_i - t_j)^2)$. We randomly select 10% of the sample and replace them with data from a different process with the same covariance function but different mean $\boldsymbol{\mu}^\star(\boldsymbol{t}) = K^m(\sin(6\boldsymbol{s}) + \boldsymbol{\mu}(\boldsymbol{s}))$.

We apply FARD, MFHD ($\alpha = 0.25$) and RP2 respectively to the sample above. It is already observed in Claeskens et al. (2014) that FM, MBD can not carry out a better performance than MFHD. From Figure 3.6 (a) and (b), we observe that FARD is more effective than others in detecting the "inliers", which are anomalies but hide within the regular sample. Furthermore, the computation of RP2 is much more time consuming than the rest, and makes it infeasible to apply RP2 to a large sample of functional data with dimension higher than 2.



(a) sensitivity of Simulation IV (b) specificity of Simulation IV

Figure 3.6: Sensitivity and specificity comparison from four Simulation IV: Log-normal Process.

3.4 Application: Detecting Anomalous Aircraft Landings

With the rapid growth of air transportation, there has been a growing emphasis on proactive safety management. According to a recent study Boeing (2014), the "runway excursions, abnormal runway contact, and runway undershoot/overshoot" is the third leading cause of fatal accidents worldwide of commercial jet fleet from 2005 to 2014. The FAA provides guidance of target touchdown point to achieve safe landing, see FAA (2014).

We have been collaborating with a major domestic airline, who has provided the landing data set for the purpose of testing our proposed approaches. Each landing can be treated as a functional data point. Data are observed by snapshots frequently. We have applied the ARD approach to this aircraft landing data to detect anomalous landing traces. For the ease of visualization, we illustrate the case of using one functional variable, namely,

RALTC.

We apply respectively FARD, MFHD, MBD and RP to this dataset, and label 1% functionals with the lowest depth values as outliers. As expected, the FARD is able to capture the asymmetric structure of the landing traces. More specifically, it detects outliers from both top and bottom, but mostly from the top, where the traces have more substantial deviation from the benchmark. The comparison between FARD and MBD results immediately shows that FARD mitigates the drawback of MBD of neglecting the relative deviation from the sample points to the deepest one. Moreover, we observe that MFHD overestimates the importance of the beginning phase, and consequently misses a lot of anomalous traces which deviate a lot in the end, where should be emphasized in this application. RP is overly sensitive in declaring outliers based on only the performance at the beginning phase.

Moreover, by applying FARD, we obtain the sample deepest functional as a benchmark landing performance. Although aircrafts should follow the landing path with a glide slope of 3 degree along the center line of the runway, conditions such as adverse weather may impede carrying out landing as recommended. Instead of the path with the recommended glide slope, the benchmark selected by FARD can be viewed as the "best expected" performance that all the aircraft landings can possibly achieve. Statistical methodologies can be applied meaningfully if they are devised to withstand practical achievability and other considerations.

3.5 Discussion

In this chapter, we introduce a general approach named antipodal reflection depth (ARD), to generate a class of new depth notions for both multivariate and functional data. ARD takes into account the magnitude of deviation from any point to the deepest one, while

preserving the ordering derived by the base depth along each ray from the deepest one. It is useful in a variety of fields. In particular, this paper emphasizes its utility in detecting outliers in both multivariate and functional settings. Specifically, it mitigates the incapability of many existing depths in extracting the information from the sample with regard to their deviations to the deepest point. Both simulation studies and a real application in aircraft landing analysis show that ARD is more effective than other depth approaches in detecting outliers or identifying anomalous landing performance.

There exist some depth functions that are defined based on derivations from the deepest point, for example, L_1 depth (Vardi and Zhang, 2000) or Mahalanobis depth (Mahalanobis, 1936). Roughly speaking, using L_1 depth is implicitly imposing a spherical structure to the data, while Mahalanobis depth by an elliptical structure to the data. ARD, by adopting a geometric base depth such as SD or HD, can be applied to a more general data setting.

Finally, there are many nonparametric inference methods derived from depth. It is expected that ARD can be readily applied to all these methods as well. Similarly, the notion of depth has be shown to have a broad range of applications in many domains. For example, in addition to the application in outlier detection emphasized in this paper, ARD can be applied to estimating central regions of functional data, or testing of symmetry of a given distribution, just to name a few. We plan to pursue these ideas in separate projects. In particular, we plan to develop a "tolerance tube" for tracking aircraft landing performance to reduce the risk of landing. This tolerance tube generalizes the tolerance region in multivariate settings in Li and Liu (2008) to the functional setting.

In this paper, we consider the data setting that all functionals are observed at the same grid points. This implies that the benchmark functional selected by our approach consists of the deepest point at each grid point. This naturally gives rise to a robust estimator of the underlying mean functional. Spline smoothing or other regression approaches have

been explored to obtain a similar robust estimator. However, the latter estimator essentially utilizes the pointwise mean, which can be sensitive to potential outliers if no proper penalty is considered.

3.6 Proofs

Proof of Theorem 3.1

Following the notations in Section 3, we obtain:

$$ARD(\boldsymbol{x}, F) = D(\boldsymbol{x}, G)$$

That is, ARD of $\boldsymbol{x}$ w.r.t. F can be viewed as the base depth of $\boldsymbol{x}$ w.r.t. a corresponding antipodal symmetric distribution G. Consequently, ARD attains the four properties as long as the base depth does. □

Proof of Theorem 3.2

To show $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |ARD_n(\boldsymbol{x}) - ARD(\boldsymbol{x}, F)| \to 0$ as $n \to \infty$, following the notations in Section 3, we first observe that

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathbb{R}^d} |ARD_n(\boldsymbol{x}) - ARD(\boldsymbol{x}, F)| = \sup_{\boldsymbol{x}} |D^{\star}_{2n}(\boldsymbol{x}) - D(\boldsymbol{x}, G)| \\
&\leq \sup_{\boldsymbol{x}} |D^{\star}_{2n}(\boldsymbol{x}) - \widetilde{D}_{2n}(\boldsymbol{x})| + \sup_{\boldsymbol{x}} |\widetilde{D}_{2n}(\boldsymbol{x}) - D(\boldsymbol{x}, G)|.
\end{aligned}$$

Thus, if we can show the following two statements

$$\sup_{\boldsymbol{x}} |\widetilde{D}_{2n}(\boldsymbol{x}) - D(\boldsymbol{x}, G)| \to 0 \text{ as } n \to \infty \tag{3.1a}$$

$$\sup_{\boldsymbol{x}} |D^{\star}_{2n}(\boldsymbol{x}) - \widetilde{D}_{2n}(\boldsymbol{x})| \to 0 \text{ as } n \to \infty \tag{3.1b}$$

hold, the theorem is then proved.

We consider the case that simplicial depth is used as the base depth. We call two data points $\widetilde{Y}_i$, $\widetilde{Y}_j$ a reflection pair if $\widetilde{Y}_i = 2\boldsymbol{\theta} - \widetilde{Y}_j$. Let A denote the set of subsets, each of which contains $(d+1)$ points from the pooled sample $\{\widetilde{\boldsymbol{y}}_1, \ldots, \widetilde{\boldsymbol{y}}_{2n}\}$ but without any reflection pair. Let B denote the set of such $(d+1)$-point subsets containing at least one reflection pair. We can show that the depth value contributed from set B is negligible. That is to say, when $n \to \infty$, we only have to count the simplicies from set A in the calculation of $\widetilde{D}_{2n}(\boldsymbol{x})$. More formally,

$$\begin{aligned}
T &= \{(\widetilde{Y}_{i,1}, \ldots, \widetilde{Y}_{i,d+1}) : \text{ a subset of } (d+1) \text{ points from } (\widetilde{y}_1, \ldots, \widetilde{y}_{2n})\} \\
A &= \{(\widetilde{Y}_{i,1}, \ldots, \widetilde{Y}_{i,d+1}) : \text{ a subset of } (d+1) \text{ points from } (\widetilde{y}_1, \ldots, \widetilde{y}_{2n}), \\
&\qquad \text{which contains no reflection pair}\} \\
B &= T \backslash A
\end{aligned}$$

We can divide $\widetilde{D}_{2n}(\boldsymbol{x})$ into the sum of two values, contributed from set A and B, respectively. Namely,

$$\widetilde{D}_{2n}(\boldsymbol{x}) = \binom{2n}{d+1}^{-1} \{\sum_{A} I(\boldsymbol{x} \in S[\widetilde{Y}_{i,1}, \ldots, \widetilde{Y}_{i,d+1}]) + \sum_{B} I(\boldsymbol{x} \in S[\widetilde{Y}_{j,1}, \ldots, \widetilde{Y}_{j,d+1}])\}.$$

By some calculation, we obtain

$$\lim_{n\to\infty} \binom{2n}{d+1}^{-1} card(A) = 0 \tag{3.2a}$$

$$\lim_{n\to\infty} \binom{2n}{d+1}^{-1} card(B) = 1 \tag{3.2b}$$

Note that the cardinality of a set is the total number of elements in this set. We explain (3.2a) and (3.2b) by showing all the work as follows. Since set A contains all subsets without reflection pairs, we can first select corresponding indices and then select sample points from either reflection sample or original sample. That is,

$$card(A) = \binom{n}{d+1} 2^{d+1}.$$

In addition, it is easier to show $card(T) = \binom{2n}{d+1}$. Thus,

$$\begin{aligned}
\frac{card(A)}{card(T)} &= \frac{\binom{n}{d+1}2^{d+1}}{\binom{2n}{d+1}} \\
&= \frac{\frac{n!2^{d+1}}{(d+1)!(n-d-1)!}}{\frac{(2n)!}{(2n-d-1)!(d+1)!}} \\
&= \frac{n!(2n-d-1)!}{(2n)!(n-d-1)!} \cdot 2^{d+1} \\
&= \frac{n(n-1)\dots(n-d)}{2n(2n-1)\dots(2n-d)} \cdot 2^{d+1},
\end{aligned}$$

and

$$\lim_{n\to\infty} \frac{card(A)}{card(T)} = 1.$$

Similarly, we obtain

$$\lim_{n\to\infty} \frac{card(B)}{card(T)} = 0$$

Now, the problem becomes simpler since we have less simplices to count. Specifically,

$$\begin{aligned}
&\widetilde{D}_{2n}(\boldsymbol{x}) - D(\boldsymbol{x}, G) \\
&= \binom{2n}{d+1}^{-1} [\sum_{A} I(\boldsymbol{x} \in S[\widetilde{Y}_{i,1}, \dots, \widetilde{Y}_{i,d+1}]) + \sum_{B} I(\boldsymbol{x} \in S[\widetilde{Y}_{j,1}, \dots, \widetilde{Y}_{j,d+1}])] - D(\boldsymbol{x}, G)
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{n\to\infty}\{\binom{2n}{d+1}^{-1}\sum_{A} I(\boldsymbol{x}\in S[\widetilde{Y}_{i,1},\ldots,\widetilde{Y}_{i,d+1}]) - D(\boldsymbol{x},\, G)\} \\
=&\lim_{n\to\infty}\frac{card(A)}{\binom{2n}{d+1}}\frac{1}{card(A)}\sum_{A} I(\boldsymbol{x}\in S[\widetilde{Y}_{i,1},\ldots,\widetilde{Y}_{i,d+1}]) - D(\boldsymbol{x},\, G) \\
=&\lim_{n\to\infty}\{\frac{1}{card(A)}\sum_{A} I(\boldsymbol{x}\in S[\widetilde{Y}_{i,1},\ldots,\widetilde{Y}_{i,d+1}]) - D(\boldsymbol{x},\, G)\}
\end{aligned}$$

Until now, we yet can not prove the convergence directly because the terms $I(\boldsymbol{x} \in S[\widetilde{Y}_{i,1},\ldots,\widetilde{Y}_{i,d+1}])$ are not independent. Thus, in what follows, we continue to separate set A into several small sets, each of which corresponds to a unique subset of $(d+1)$ indices chosen from $\{1,2,\ldots,n\}$. Each subset, say the one corresponding to indices $\{(i,1),(i,2),\ldots,(i,d+1)\}$, contains simplices formed by $(d+1)$ points selected from $\{\boldsymbol{x}_{i,1},\ldots,\boldsymbol{x}_{i,d+1},\widetilde{\boldsymbol{x}}_{i,1},\ldots,\widetilde{\boldsymbol{x}}_{i,d+1}\}$ but without reflection pairs. Namely,

$$\begin{aligned}
&\lim_{n\to\infty}\{\frac{1}{card(A)}\sum_{A} I(\boldsymbol{x}\in S[\widetilde{Y}_{1},\ldots,\widetilde{Y}_{d+1}]) - D(\boldsymbol{x},\, G)\} \\
=&\lim_{n\to\infty}\frac{1}{card(A)}\Sigma_C\Sigma_{\Lambda_i} I(\boldsymbol{x}\in S[Z_{i,1},\ldots,Z_{i,d+1}]) - D(\boldsymbol{x},\, G)
\end{aligned}$$

where

$$C = \{((i,1),\ldots,(i,d+1)) :\ \text{a subset of } (d+1) \text{ indicies from } \{1,\ldots,\text{n}\}\}$$

$$\text{and}\quad \Lambda_i = \{(Z_{i,1},\ldots,Z_{i,d+1}) : Z_{i,j} = X_{i,j} \text{ or } Z_{i,j} = \widetilde{X}_{i,j}, j = 1,\ldots,d+1\}.$$

Thus, we only need to show $\frac{1}{card(A)}\Sigma_C\Sigma_{\Lambda_i} I\{\boldsymbol{x} \in S[Z_{i,1},\ldots,Z_{i,d+1}]\} - D(\boldsymbol{x},\, G) \to 0$ as $n \to \infty$.

We observe that for $\forall i$,

$$E\{I(\boldsymbol{x} \in S[Z_{i,1}, \ldots, Z_{i,d+1}])\} = P\{\boldsymbol{x} \in S[Z_{i,1}, \ldots, Z_{i,d+1}]\} := D(\boldsymbol{x}, G).$$

As a result, we obtain

$$E\{\frac{1}{card(\Lambda_i)} I(\boldsymbol{x} \in S[Z_{i,1}, \ldots, Z_{i,d+1}])\} = D(\boldsymbol{x}, G)$$

In fact, $\frac{1}{card(\Lambda_i)} I(\boldsymbol{x} \in S[Z_{i,1}, \ldots, Z_{i,d+1}])$ is a function of $\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,d+1}$, i.e. $h(\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,d+1})$, where $h(\cdot)$ is symmetric about its arguments. In addition, we note that

$$\begin{aligned} & card(\Lambda_i) = 2^{d+1}, \quad \forall i \\ \text{and} \quad & card(A) = \binom{n}{d+1} \cdot 2^{d+1} = card(C) \cdot card(\Lambda_i). \end{aligned}$$

It is obvious that $\frac{1}{card(A)} \Sigma_C \Sigma_{\Lambda_i} I\{\boldsymbol{x} \in S[Z_{i,1}, \ldots, Z_{i,d+1}]\}$ is a U-statistic, with first moment being $D(\boldsymbol{x}, G)$. Followed by Lemma 3 in Liu (1990), the convergence is proved.

The proof of (3.1b) can be outlined as follows. For simplicity, take $d = 2$.

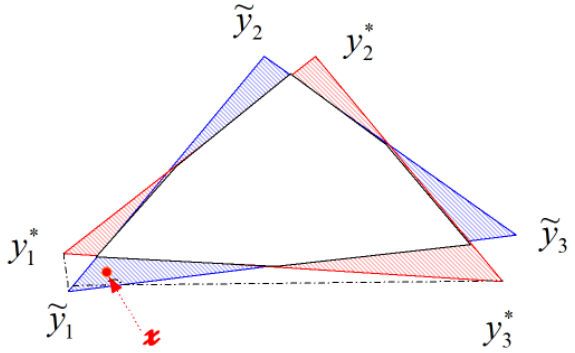


Figure 3.7: Two triangles with corresponding index. The area in shadow is the non-overlapping area of the two triangles. Only a point $\boldsymbol{x}$ within such shadowed area can make a different between $D^{\star}_{2n}(\boldsymbol{x})$ and $\widetilde{D}_{2n}(\boldsymbol{x})$.

$\forall \epsilon > 0$,

$$
\begin{aligned}
& P\{\sup_{\boldsymbol{x}\in\mathbb{R}^d} |D^{\star}_{2n}(\boldsymbol{x}) - \widetilde{D}_{2n}(\boldsymbol{x})| > \epsilon\} \\
=& P\{\sup_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{\binom{2n}{3}} \sum |I(\boldsymbol{x} \in S(\boldsymbol{y}^{\star}_{i,1}, \boldsymbol{y}^{\star}_{i,2}, \boldsymbol{y}^{\star}_{i,3})) - I(\boldsymbol{x} \in S(\widetilde{\boldsymbol{y}}_{i,1}, \widetilde{\boldsymbol{y}}_{i,2}, \widetilde{\boldsymbol{y}}_{i,3}))| > \epsilon\} \\
=& P\{\sup_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{\binom{2n}{3}} \sum I(\boldsymbol{x} \in [S(\boldsymbol{y}^{\star}_{i,1}, \boldsymbol{y}^{\star}_{i,2}, \boldsymbol{y}^{\star}_{i,3}) \Delta S(\widetilde{\boldsymbol{y}}_{i,1}, \widetilde{\boldsymbol{y}}_{i,2}, \widetilde{\boldsymbol{y}}_{i,3})]) > \epsilon\}
\end{aligned}
$$

Since $|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}| \to 0$ a.s., we obtain

$$
\begin{aligned}
& P\{\lim_{n\to\infty} \sup_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{\binom{2n}{3}} \sum I(\boldsymbol{x} \in [S(\boldsymbol{y}^{\star}_{i,1}, \boldsymbol{y}^{\star}_{i,2}, \boldsymbol{y}^{\star}_{i,3}) \Delta S(\widetilde{\boldsymbol{y}}_{i,1}, \widetilde{\boldsymbol{y}}_{i,2}, \widetilde{\boldsymbol{y}}_{i,3})]) > \epsilon\} \\
=& P\{\lim_{n\to\infty} \sup_{\boldsymbol{x}\in\mathbb{R}^d} I(\boldsymbol{x} \in \partial S(\widetilde{\boldsymbol{y}}_{i,1}, \widetilde{\boldsymbol{y}}_{i,2}, \widetilde{\boldsymbol{y}}_{i,3})) > \epsilon\} \\
=& 0.
\end{aligned}
$$

The probability is 0 because the probability of any line in $\mathbb{R}^2$ is 0.

□

Remark 3.4. *Using halfspace depth as base depth,* $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |ARD_n(\boldsymbol{x}) - ARD(\boldsymbol{x}, F)| \to 0$ *a.s. still holds.*

Proof:

We begin with the ad hoc definition of ARD using halfspace depth as base depth, which is,

$$ARD(\boldsymbol{x}, F) = \inf_H \{P_G^\star(H) : H \text{ is a closed halfspace, } \boldsymbol{x} \in H\}$$
$$ARD_n(\boldsymbol{x}) = \inf_H \{P_{2n}^\star(H) : H \text{ is a closed halfspace, } \boldsymbol{x} \in H\}$$

where $P_{2n}^\star$ is the empirical version of P obtained from the pooled sample $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_1^\star, \ldots, \boldsymbol{x}_n^\star\}$.

In general, the deepest point is not always unique. Donoho and Gasko (1992) defines the deepest point as the *centroid* of the set whose elements all attain the deepest depth value. Under this definition, we can show that $|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}| \to 0, \quad a.s.$ as $n \to \infty$ as long as the uniformly a.s. convergence of $D_n(\boldsymbol{x})$ is guaranteed. We agree to this treatment and in what follows, we assume $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}$ are uniquely defined.

We observe that if $\boldsymbol{x} \neq \hat{\boldsymbol{\theta}}_n$, any reflection pair $\boldsymbol{x}_i$ and $\boldsymbol{x}_i^\star$ can not be enclosed in the halfspace $H^\star$ which attains $ARD_n(\boldsymbol{x})$, where $H^\star = argmin_H P_{2n}^\star\{H : H \text{ is a closed halfspace, } \boldsymbol{x} \in H\}$. See Figure 3.8.

We observe that x_i and $x_i^\star$ can not be enclosed in $H^\star$ at the same time, unless x_i and $x_i^\star$ are on the boundary of $H^\star$. In this situation, the four points, $\boldsymbol{x}$, $\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_i, \boldsymbol{x}_i^\star$ form a straight line. However, the situation can not happen because we can always decrease $P_{2n}^\star(H)$ by rotating the halfspace slightly such that it excludes one point.

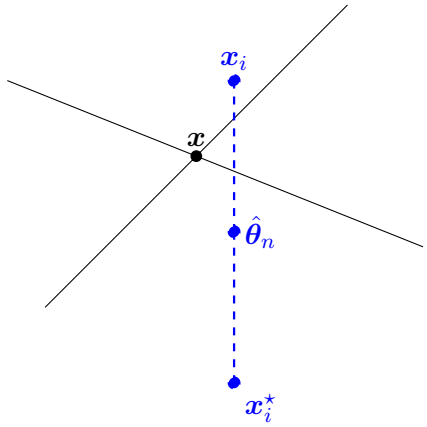


Figure 3.8: The two black lines are the boundary of any two halfspaces that crosses point $\boldsymbol{x}$. For $\boldsymbol{x} \neq \hat{\boldsymbol{\theta}}_n$, $\boldsymbol{x}_i$ and $\boldsymbol{x}_i^\star$ can not simultaneously be enclosed in the halfspace which produces halfspace depth of $\boldsymbol{x}$.

At this moment, we assume $\boldsymbol{x} \neq \hat{\boldsymbol{\theta}}_n$. The discussion of the case when $\boldsymbol{x} = \hat{\boldsymbol{\theta}}_n$ can be found afterwards. Now, we introduce a new concept "twin halfspaces". Two halfspaces H_1 and H_2 are *twin halfspaces* if they are antipodally symmetric around some point $\boldsymbol{c}$ with no intersection. Then, we can express $ARD_n(\boldsymbol{x})$ as:

$$\begin{aligned} ARD_n(\boldsymbol{x}) = \inf_{H_1,H_2} \frac{1}{2}\{&P_n(H_1 \cup H_2) : H_1 \text{ and } H_2 \text{ are twin halfspaces in } \mathbb{R}^d \\ &\text{around } \hat{\boldsymbol{\theta}}_n,\ \boldsymbol{x} \in H_1,\ \boldsymbol{x}^\star \in H_2\} \\ &=: D_{2n}^\star(\boldsymbol{x}) \end{aligned}$$

where $\boldsymbol{x}^\star = 2\hat{\boldsymbol{\theta}}_n - \boldsymbol{x}$. This is due to the following fact: as in Figure 3.9, for any halfspace $H_{\boldsymbol{x}}$ with $\boldsymbol{x}$ on its boundary, the points included in $H_{\boldsymbol{x}}$ come from two sources: (1) the original sample (e.g. $\boldsymbol{x}_1$); (2) the reflection sample (e.g. $\boldsymbol{x}_3^\star$).

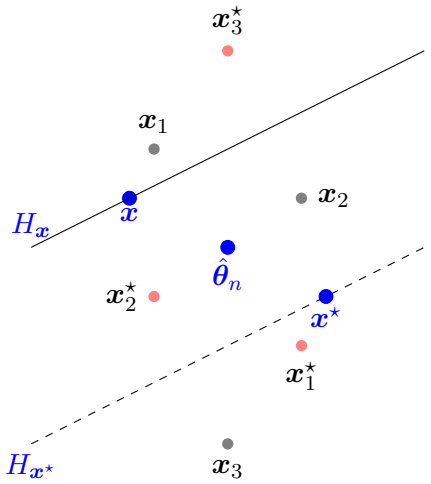


Figure 3.9: x_1, x_2, x_3 are from the original sample, and $x_1^\star, x_2^\star, x_3^\star$ are their antipodal reflections, respectively. H_x and $H_{x^\star}$ form a pair of twin halfspaces which cross x and $x^\star$ respectively.

We draw a halfspace $H_{\boldsymbol{x}^\star}$ which forms twin halfspaces together with $H_{\boldsymbol{x}}$ around $\hat{\boldsymbol{\theta}}_n$. It is not hard to prove $\boldsymbol{x}^\star$ is on the boundary of $H_{\boldsymbol{x}^\star}$. We expect that $\boldsymbol{x}_3$ would fall into $H_{\boldsymbol{x}^\star}$ as well. Consequently, we only have to count the number of points which fall into the twin halfspaces from the original sample. It implies that we can define the $ARD_n(\boldsymbol{x})$ using twin halfspaces and the original data. Similarly, we define

$$\widetilde{D}_{2n}(\boldsymbol{x}) = \inf_{H_1, H_2} \frac{1}{2}\{P_n(H_1 \cup H_2) : H_1 \text{ and } H_2 \text{ are twin halfspaces in } \mathbb{R}^d \text{ around } \boldsymbol{\theta},\ \boldsymbol{x} \in H_1,\ \widetilde{\boldsymbol{x}} \in H_2\},$$

and

$$ARD(\boldsymbol{x}, F) = D(\boldsymbol{x}, G)$$
$$= \inf_{H_1, H_2} \frac{1}{2}\{P(H_1 \cup H_2) : H_1 \text{ and } H_2 \text{ are twin halfspaces in } \mathbb{R}^d \text{ around } \boldsymbol{\theta},$$

$$\boldsymbol{x} \in H_1,\, \widetilde{\boldsymbol{x}} \in H_2\}.$$

Replacing H by $H_1 \cup H_2$ in Section 6.1 in Donoho and Gasko (1992), $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{D}_{2n}(\boldsymbol{x}) - D(\boldsymbol{x},\, G)| \to 0$ holds straightforwardly.

To prove $ARD_n(\boldsymbol{x})$ converges to $ARD(\boldsymbol{x},\, F)$ as $n \to \infty$, we still need to show $|\widetilde{D}_{2n}(\boldsymbol{x}) - D(\boldsymbol{x},\, G)| \to 0$ as $n \to \infty$.

For simplicity, take $d = 2$. Given pooled sample $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_1^\star, \ldots, \boldsymbol{x}_n^\star\}$, we obtain $\{H_1^\star,\, H_2^\star\}$ such that

$$\{H_1^\star,\, H_2^\star\} = \underset{H_1,H_2}{\operatorname{argmin}} \frac{1}{2}\{P_{2n}^\star(H_1 \cup H_2) : H_1 \text{ and } H_2 \text{ are twin halfspaces in } \mathbb{R}^d \text{ around } \hat{\boldsymbol{\theta}}_n,$$
$$\boldsymbol{x} \in H_1,\, \boldsymbol{x}^\star \in H_2\},$$

where $\boldsymbol{x}^\star = 2\hat{\boldsymbol{\theta}}_n - \boldsymbol{x}$. Analogously, we obtain $\{\widetilde{H}_1,\, \widetilde{H}_2\}$ such that

$$\{\widetilde{H}_1,\, \widetilde{H}_2\} = \underset{H_1,H_2}{\operatorname{argmin}} \frac{1}{2}\{\widetilde{P}_{2n}(H_1 \cup H_2) : H_1 \text{ and } H_2 \text{ are twin halfspaces in } \mathbb{R}^d \text{ around } \boldsymbol{\theta},$$
$$\boldsymbol{x} \in H_1,\, \widetilde{\boldsymbol{x}} \in H_2\},$$

where $\widetilde{\boldsymbol{x}} = 2\boldsymbol{\theta} - \boldsymbol{x}$.

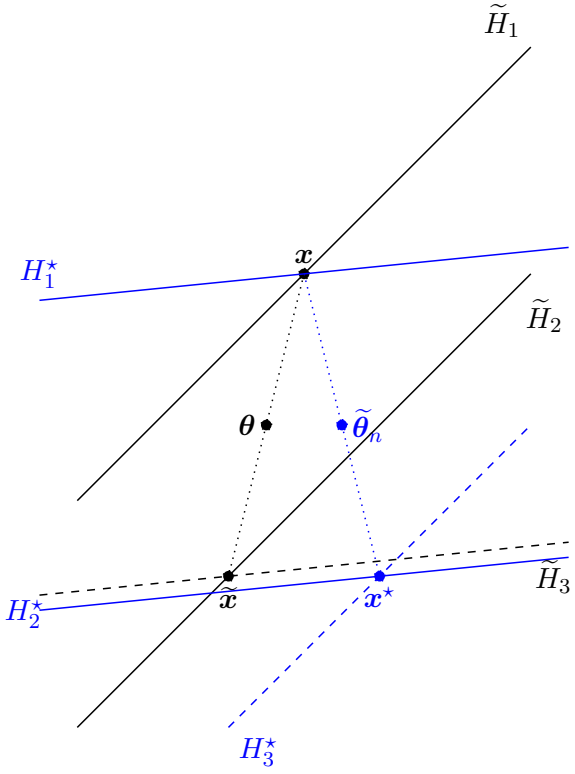


Figure 3.10: Twin halfspaces $\widetilde{H}_1$ and $\widetilde{H}_2$ around $\boldsymbol{\theta}$ contain $\boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}$ on the boundary, respectively; twin halfspaces $H_1^\star$ and $H_2^\star$ around $\hat{\boldsymbol{\theta}}_n$ contain $\boldsymbol{x}$ and $\boldsymbol{x}^\star$ on the boundary, respectively. Halfspace $\widetilde{H}_3$, which has $\widetilde{\boldsymbol{x}}$ on its boundary, has parallel boundary to $H_2^\star$; halfspace $H_3^\star$, which has $\boldsymbol{x}^\star$ on its boundary, has parallel boundary to $\widetilde{H}_2$. This figure reflects the situation described in (3.3a).

We draw a line (or hyperplane) across $\widetilde{\boldsymbol{x}}$ parallel to the boundary of $H_2^\star$ and it produces a new halfspaces, $\widetilde{H}_3$. Analogously, we draw a line (or hyperplane) across $\boldsymbol{x}^\star$ parallel to the boundary of $\widetilde{H}_2$ and it produces a new halfspace, $H_3^\star$. By the definition of halfspace depth, we know that $P_n(H_1^\star \cup \widetilde{H}_3) \geq P_n(\widetilde{H}_1 \cup \widetilde{H}_2)$. Here, $P_n(\cdot)$ is the empirical version of $P(\cdot)$. In addition, we have $P_n(\widetilde{H}_1 \cup H_3^\star) \geq P_n(H_1^\star \cup H_2^\star)$ obtained from the original sample

$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Thus, we obtain

$$P_n(H_1^\star \cup \widetilde{H}_3) \geq P_n(\widetilde{H}_1 \cup \widetilde{H}_2) \geq P_n(\widetilde{H}_1 \cup H_3^\star) \geq P_n(H_1^\star \cup H_2^\star) \tag{3.3a}$$

$$\text{or} \quad P_n(\widetilde{H}_1 \cup H_3^\star) \geq P_n(H_1^\star \cup H_2^\star) \geq P_n(H_1^\star \cup \widetilde{H}_3) \geq P_n(\widetilde{H}_1 \cup \widetilde{H}_2) \tag{3.3b}$$

depending on the data. Next, we show that for any F which has a universal bound, say $|f(\boldsymbol{x})| \leq g$,

$$\lim_{n\to\infty} P_n(H_1^\star \cup \widetilde{H}_3) - P_n(H_1^\star \cup H_2^\star) = 0.$$

For any fixed $\epsilon > 0$, we observe:

$$\begin{aligned}
&\{\sup_{\boldsymbol{x}\in\mathbb{R}^d} |D_{2n}^\star(\boldsymbol{x}) - \widetilde{D}_{2n}(\boldsymbol{x})| > \epsilon\} \\
=&\{\sup_{\boldsymbol{x}\in\mathbb{R}^d} |P_n(H_1^\star \cup H_2^\star) - P_n(\widetilde{H}_1 \cup \widetilde{H}_2)| > \epsilon\} \\
\subset&\{\sup_{\boldsymbol{x}\in\mathbb{R}^d} |P_n(H_1^\star \cup \widetilde{H}_3) - P_n(H_1^\star \cup H_2^\star)| > \epsilon\} \\
=&\{\sup_{\boldsymbol{x}\in\mathbb{R}^d} |P_n(\widetilde{H}_3 \backslash H_2^\star)| > \epsilon\}
\end{aligned}$$

Let $\rho_n = P\{\boldsymbol{x}_i \in \widetilde{H}_3 \backslash H_2^\star\} < 1$. Under some mild condition, we have $|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}| \to 0$ a.s. as $n \to \infty$. Based on that, it is easy to show that $P\{\lim_{n\to\infty} \widetilde{H}_3 \backslash H_2^\star = \partial \widetilde{H}_3\} = 1$, where $\partial \widetilde{H}_3$ is the boundary of $\widetilde{H}_3$, and $P\{\lim_{n\to\infty} \rho_n = 0\} = 1$. By Glivenko-Cantelli property, we obtain:

$$P\{\lim_{n\to\infty} \sup_{\boldsymbol{x}\in\mathbb{R}^d} |P_n(\widetilde{H}_3 \backslash H_2^\star)| > \epsilon\} = P\{\sup_{\boldsymbol{x}\in\mathbb{R}^d} |P(\partial \widetilde{H}_3)| > \epsilon\} = 0.$$

In what follows, we show that when $\boldsymbol{x} = \hat{\boldsymbol{\theta}}_n$, the convergence still holds. We observe that

$$|ARD_n(\hat{\boldsymbol{\theta}}_n) - ARD(\hat{\boldsymbol{\theta}}_n, F)| = |D_{2n}^\star(\hat{\boldsymbol{\theta}}_n) - D(\hat{\boldsymbol{\theta}}_n, G)|$$

$$
\begin{aligned}
=&|D^{\star}_{2n}(\hat{\boldsymbol{\theta}}_n) - D^{\star}_{2n}(\boldsymbol{\theta}) + D^{\star}_{2n}(\boldsymbol{\theta}) - D(\boldsymbol{\theta}, G) + D(\boldsymbol{\theta}, G) - D(\hat{\boldsymbol{\theta}}_n, G)| \\
\leq&|D^{\star}_{2n}(\hat{\boldsymbol{\theta}}_n) - D^{\star}_{2n}(\boldsymbol{\theta})| + |D^{\star}_{2n}(\boldsymbol{\theta}) - D(\boldsymbol{\theta}, G)| + |D(\boldsymbol{\theta}, G) - D(\hat{\boldsymbol{\theta}}_n, G)|.
\end{aligned}
$$

Since $|D^{\star}_{2n}(\boldsymbol{\theta}) - D(\boldsymbol{\theta}, G)| \to 0$ and we've proved that $|D(\boldsymbol{\theta}, G) - D(\hat{\boldsymbol{\theta}}_n, G)| \to 0$ as $n \to \infty$, we only need to show $|D^{\star}_{2n}(\hat{\boldsymbol{\theta}}_n) - D^{\star}_{2n}(\boldsymbol{\theta})| \to 0$ as $n \to \infty$.

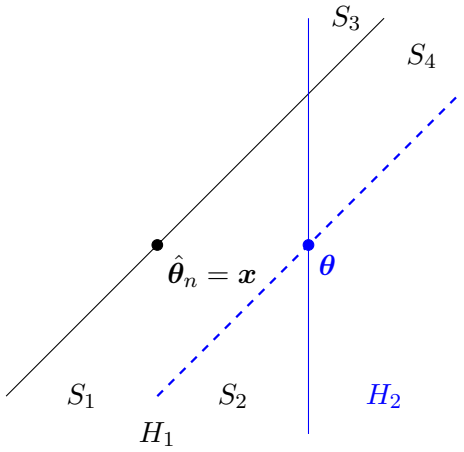


Figure 3.11: $\boldsymbol{x} = \hat{\boldsymbol{\theta}}_n$, H_1 and H_2 are halfspaces that construct $D^{\star}_{2n}(\hat{\theta}_n)$ and $D^{\star}_{2n}(\theta)$, respectively. S_1, S_2, S_3 and S_4 are subspaces produced by intersections of several halfspaces.

Suppose H_1 and H_2 are the halfspaces that construct $D^{\star}_{2n}(\hat{\boldsymbol{\theta}}_n)$ and $D^{\star}_{2n}(\boldsymbol{\theta})$, respectively. As in Figure 3.11, we denote some sub-area as S_1, S_2, S_3 and S_4, such that $H_1 \backslash H_2 = S_1 \cup S_2$ and $H_2 \backslash H_1 = S_3$. Then,

$$
\begin{aligned}
&|D^{\star}_{2n}(\hat{\boldsymbol{\theta}}_n) - D^{\star}_{2n}(\boldsymbol{\theta})| = |P^{\star}_{2n}(H_1) - P^{\star}_{2n}(H_2)| \\
=&|P^{\star}_{2n}(S_1 \cup S_2) - P_{2n}(S_3)| = |P^{\star}_{2n}(S_1 \cup S_3 \cup S_4) - P^{\star}_{2n}(S_3)| \\
=&|P^{\star}_{2n}(S_1 \cup S_4)|
\end{aligned}
$$

and for any $\epsilon > 0$, we have $P\{\lim_{n\to\infty} |P^{\star}_{2n}(S_1 \cup S_4)|\} = 0$ if F is continuous in a neighborhood of $\boldsymbol{\theta}$. □

Remark 3.5. *Changing the base depth to Mahalanobis depth, the convergence holds if* $E\|X\|^2 < \infty$.

Proof:

To recall, the definition of Mahalanobis depth as follows:

$$M_hD(\boldsymbol{x},\, G) = \frac{1}{1 + (\boldsymbol{x} - \boldsymbol{\mu}_G)'\Sigma_G^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_G)}$$

where $\boldsymbol{\mu}_G$ and Σ_G are the mean vector and covariance matrix of distribution F. Sample version $M_hD_n(\boldsymbol{x})$ is to replace $\boldsymbol{\mu}_G$ and Σ_G by their sample estimators, respectively. (Liu and Singh (1993)) For simplicity, we use $D(\cdot)$ to represent $M_hD(\cdot)$ henceforth.

Also, recall that let $X \sim F$, independent of $\epsilon \sim Bernoulli(0.5)$ and then we obtain $Z = \boldsymbol{\theta} + \epsilon(\boldsymbol{\theta} - X)$ G. Let $\boldsymbol{\theta}$ be the deepest point w.r.t. F in terms of Mahalanobis depth. In the case that the deepest point is not unique, we define a unique deepest point the same way as in Remark 2. Then, we get

$$\begin{cases} \mu_G = \boldsymbol{\theta} \\ \Sigma_G = E_G[(Z - \boldsymbol{\theta})(Z - \boldsymbol{\theta})'] = E_F[(X - \boldsymbol{\theta})(X - \boldsymbol{\theta})'] := \Sigma. \end{cases}$$

Thus, ARD is constructed as:

$$\begin{aligned} ARD(\boldsymbol{x},\, F) &= \frac{1}{1 + (\boldsymbol{x} - \boldsymbol{\mu}_G)'\Sigma_G^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_G)} \\ &= \frac{1}{1 + (\boldsymbol{x} - \boldsymbol{\theta})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\theta})}, \end{aligned}$$

Also,

$$\begin{aligned} ARD_n(\boldsymbol{x}) &= \frac{1}{1+(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_G)'\hat{\Sigma}_G^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_G)} \\ &= \frac{1}{1+(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)'c\Sigma^{\star -1}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)} \\ &:= D^{\star}_{2n}(\boldsymbol{x}), \end{aligned}$$

where $\hat{\boldsymbol{\mu}}_G = \hat{\boldsymbol{\theta}}_n$, $\hat{\Sigma}_G = c\Sigma^{\star -1}$ with $c = \frac{2(n-1)}{2n-1}$ and $\Sigma^{\star -1} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \hat{\boldsymbol{\theta}}_n)(\boldsymbol{x}_i - \hat{\boldsymbol{\theta}}_n)'$. By the same token, we define

$$\widetilde{D}_{2n}(\boldsymbol{x}) = \frac{1}{1+(\boldsymbol{x}-\boldsymbol{\theta})'a\widetilde{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})},$$

where $a = \frac{n-1}{n}$ and $\widetilde{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\theta})(\boldsymbol{x}_i - \boldsymbol{\theta})'$.

We would like to prove

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} |ARD_n(\boldsymbol{x}) - ARD(\boldsymbol{x}, F)| \to 0 \text{ as } n \to \infty.$$

Analogous to what we discussed earlier, we can prove the convergence by proving that (3.1a) and (3.1b) hold simultaneously, namely,

$$\begin{cases} \sup_{\boldsymbol{x}} |\widetilde{D}_{2n}(\boldsymbol{x}) - D(\boldsymbol{x}, G)| \to 0 & \text{as } n \to \infty \\ \sup_{\boldsymbol{x}} |D^{\star}_{2n}(\boldsymbol{x}) - \widetilde{D}_{2n}(\boldsymbol{x})| \to 0 & \text{as } n \to \infty. \end{cases}$$

To begin with, since $\widetilde{D}_{2n}(\boldsymbol{x})$ has already been expressed in a way such that no data points from the reflection sample are involved, the proof of $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{D}_{2n}(\boldsymbol{x}) - D(\boldsymbol{x}, G)| \to 0$ a.s. follows Liu and Singh (1993).

In what follows, we aim at proving $\sup_{\boldsymbol{x}} |D^{\star}_{2n}(\boldsymbol{x}) - \widetilde{D}_{2n}(\boldsymbol{x})| \to 0$ as $n \to \infty$ a.s.. By some

calculation, we obtain

$$|D^{\star}_{2n}(\boldsymbol{x}) - \widetilde{D}_{2n}(\boldsymbol{x})| = |\frac{1}{1+(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)'c\Sigma^{\star -1}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)} - \frac{1}{1+(\boldsymbol{x}-\boldsymbol{\theta})'a\widetilde{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})}|$$
$$= \frac{(\boldsymbol{x}-\boldsymbol{\theta})'a\widetilde{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta}) - (\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)'c\Sigma^{\star -1}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)}{[1+(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)'c\Sigma^{\star -1}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)][1+(\boldsymbol{x}-\boldsymbol{\theta})'a\widetilde{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})]}.$$

Consider that $\lim_{n\to\infty} a = 1$ and $\lim_{n\to\infty} c = 1$, we can prove the convergence by showing the convergence of the following term:

$$\frac{(\boldsymbol{x}-\boldsymbol{\theta})'\widetilde{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta}) - (\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)}{[1+(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)][1+(\boldsymbol{x}-\boldsymbol{\theta})'\widetilde{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})]}$$
$$= \frac{(\boldsymbol{x}-\boldsymbol{\theta})'(\widetilde{\Sigma}^{-1}-\Sigma^{\star -1})(\boldsymbol{x}-\boldsymbol{\theta}) + 2(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)}{[1+(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)][1+(\boldsymbol{x}-\boldsymbol{\theta})'\widetilde{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})]}$$

For simplicity, we denote the denominator by $[1+A^{\star}][1+\widetilde{A}]$, where $A^{\star} = (\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)$ and $\widetilde{A} = (\boldsymbol{x}-\boldsymbol{\theta})'\widetilde{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})$. Next, we show the almost sure convergence of the following three terms, namely,

$$\frac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)}{[1+A^{\star}][1+\widetilde{A}]} \tag{3.4a}$$
$$\frac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x}-\hat{\boldsymbol{\theta}}_n)}{[1+A^{\star}][1+\widetilde{A}]} \tag{3.4b}$$
$$\frac{(\boldsymbol{x}-\boldsymbol{\theta})'(\widetilde{\Sigma}^{-1}-\Sigma^{\star -1})(\boldsymbol{x}-\boldsymbol{\theta})}{[1+A^{\star}][1+\widetilde{A}]}. \tag{3.4c}$$

(a) show almost sure convergence of (3.4a).

Since the denominator is always greater or equal to 1, the ratio will converge to 0 if the numerator converges to 0. First, we show $\|\Sigma^{\star} - \Sigma\|_2 \to 0$ as $n \to \infty$, so that $\|\Sigma^{\star}\|$ can be bounded. Also, we show under some conditions, $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\| \to 0$ a.s.. Then, by

Cauchy-Schwarz Inequality, we obtain

$$\|(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)\| \leq \|(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)\|^2 \cdot \|\Sigma^{\star}\|^{-1} \to 0 \text{ a.s.}$$

Here, we consider L-2 norm, namely, $\|\cdot\| = \|\cdot\|_2$. We observe that $\|\Sigma^{\star}-\Sigma\| \leq \|\Sigma^{\star}-\widetilde{\Sigma}\| + \|\widetilde{\Sigma}-\Sigma\|$. To recall,

$$\Sigma^{\star} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\hat{\boldsymbol{\theta}}_n)(\boldsymbol{x}_i-\hat{\boldsymbol{\theta}}_n)'$$
$$\text{and} \quad \widetilde{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\theta})(\boldsymbol{x}_i-\boldsymbol{\theta})',$$

then

$$\begin{aligned}
\widetilde{\Sigma}-\Sigma^{\star} &= \frac{1}{n-1}\sum_{i=1}^{n}[(\boldsymbol{x}_i-\boldsymbol{\theta})(\boldsymbol{x}_i-\boldsymbol{\theta})' - (\boldsymbol{x}_i-\hat{\boldsymbol{\theta}}_n)(\boldsymbol{x}_i-\hat{\boldsymbol{\theta}}_n)'] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[\boldsymbol{x}_i(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)' + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)\boldsymbol{x}_i' - (\boldsymbol{\theta}\boldsymbol{\theta}'-\hat{\boldsymbol{\theta}}_n\hat{\boldsymbol{\theta}}_n')] \\
&= \frac{n}{n-1}[\bar{\boldsymbol{x}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)' + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)\bar{\boldsymbol{x}}' - (\boldsymbol{\theta}\boldsymbol{\theta}'-\hat{\boldsymbol{\theta}}_n\hat{\boldsymbol{\theta}}_n')]
\end{aligned}$$

By Cauchy-Schwarz Inequality, we obtain

$$\|\bar{\boldsymbol{x}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)'\| \leq \|\bar{\boldsymbol{x}}\| \cdot \|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n\|.$$

From Liu and Singh (1993), when $E_F\|X\|^2 < \infty$, $|D_n(\boldsymbol{x}) - D(\boldsymbol{x}, F)| \to 0$ a.s. as $n \to \infty$. It can be proved that if $\boldsymbol{\theta}$ is uniquely defined, $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\| \to 0$ a.s. as $n \to \infty$. Also, by LLN, we have $|\bar{\boldsymbol{x}} - \mu| \to 0$ a.s. as $n \to \infty$. Thus, $\|\bar{\boldsymbol{x}}\|$ is bounded almost surely. Above all, we obtain $\|\bar{\boldsymbol{x}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)'\| \to 0$ a.s..

Moreover, since $k(\boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{\theta}'$ is continuous on $\boldsymbol{\theta}$, $|k(\boldsymbol{\theta})-k(\hat{\boldsymbol{\theta}}_n)| \to 0$ as $|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}| \to 0$. We obtain

$|\widetilde{\Sigma} - \Sigma^\star| \to 0$ as $n \to \infty$. Since $h(\Sigma) = \Sigma^{-1}$ is continuous on Σ, $|\widetilde{\Sigma}^{-1} - \Sigma^{\star -1}| \to 0$ a.s. when $n \to \infty$ as well. As a result, $\|\Sigma^\star - \Sigma\|_2 \to 0$ as $n \to \infty$ and $\|(\boldsymbol{x}-\boldsymbol{\theta})'(\widetilde{\Sigma}^{-1} - \Sigma^{\star -1})(\boldsymbol{x}-\boldsymbol{\theta})\| \to 0$.

(b) show almost sure convergence of (3.4b), namely,

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} \frac{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n)}{[1 + (\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n)][1 + (\boldsymbol{x} - \boldsymbol{\theta})'\widetilde{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\theta})]} \to 0 \text{ a.s. as } n \to \infty.$$

We first consider to prove the convergence when $\boldsymbol{x}$ is in a bounded ball $B(\boldsymbol{\theta}, M)$ with a large $M > 0$. Then, we have

$$\begin{aligned} &\sup_{\boldsymbol{x}\in B(\boldsymbol{\theta},M)} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n) \\ \leq &\sup_{\boldsymbol{x}\in B(\boldsymbol{\theta},M)} \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\|\|\Sigma^{\star -1}\|\|\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n\|. \end{aligned}$$

We already showed that $\|\Sigma^{\star -1}\| \to \|\Sigma^{-1}\|$ a.s.. Thus, with $\|\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n\|$ bounded, the convergence is straightforward. Next, we prove the convergence when $\boldsymbol{x}$ is outside that ball. We observe

$$\begin{aligned} &\sup_{\boldsymbol{x}\notin B(\boldsymbol{\theta},M)} \frac{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n)}{[1 + (\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n)][1 + (\boldsymbol{x} - \boldsymbol{\theta})'\widetilde{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\theta})]} \\ \leq &\sup_{\boldsymbol{x}\notin B(\boldsymbol{\theta},M)} \frac{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n)}{1 + (\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n)'\Sigma^{\star -1}(\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n)} \end{aligned}$$

Let $\boldsymbol{y} = \Sigma^{-1/2}(\boldsymbol{x} - \hat{\boldsymbol{\theta}}_n)$, we have the above equation equals to

$$\begin{aligned} &\sup_{\boldsymbol{y}\notin B(\boldsymbol{\theta}_y,M_y)} \frac{\|\Sigma^{-1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)\boldsymbol{y}\|}{\|\boldsymbol{y}\|^2} \\ \leq &\sup_{\boldsymbol{y}\notin B(\boldsymbol{\theta}_y,M_y)} \frac{\|\Sigma^{-1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)\|\|\boldsymbol{y}\|}{\|\boldsymbol{y}\|^2} \end{aligned}$$

$$= \sup_{\boldsymbol{y} \notin B(\boldsymbol{\theta}_y, M_y)} \frac{\|\Sigma^{-1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)\|}{\|\boldsymbol{y}\|} \to 0 \text{ a.s. },$$

where $\boldsymbol{\theta}_y = \Sigma^{-1/2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)$ and $M_y = \Sigma^{-1/2}(M - \hat{\boldsymbol{\theta}}_n)$.

(c) show the almost sure convergence of (3.4c).

The proof is straightforward from the proof of (b). So far, (3.1a) holds.

Proof of Proposition 3.1: From the uniformly almost sure convergence of $D_n(\boldsymbol{x})$, we obtain

$$|D_n(\boldsymbol{\theta}) - D(\boldsymbol{\theta})| \to 0 \text{ a.s.}$$
$$|D_n(\hat{\boldsymbol{\theta}}_n) - D(\hat{\boldsymbol{\theta}}_n)| \to 0 \text{ a.s..}$$

Thus, $\forall \epsilon$, $\sigma > 0$, $\exists\, N(\epsilon, \sigma)$ s.t. we have the following two inequalities with probability 1:

$$-\epsilon \leq D_n(\hat{\boldsymbol{\theta}}_n) - D(\hat{\boldsymbol{\theta}}_n) \leq \epsilon$$
$$-\sigma \leq D_n(\boldsymbol{\theta}) - D(\boldsymbol{\theta}) \leq \sigma$$

Also, by the definition of the deepest point, we obtain

$$D(\hat{\boldsymbol{\theta}}_n) - D(\boldsymbol{\theta}) \leq 0 \quad \text{and} \quad D_n(\boldsymbol{\theta}) - D_n(\hat{\boldsymbol{\theta}}_n) \leq 0.$$

Thus, we obtain

$$D(\hat{\boldsymbol{\theta}}_n) - D(\boldsymbol{\theta}) \geq D_n(\hat{\boldsymbol{\theta}}_n) - \epsilon - D(\boldsymbol{\theta}) \geq D_n(\boldsymbol{\theta}) - \epsilon - D(\boldsymbol{\theta}) \geq D(\boldsymbol{\theta}) - \sigma - \epsilon - D(\boldsymbol{\theta}) = -(\epsilon + \sigma).$$

In a result, we have $-(\epsilon + \sigma) \leq D(\hat{\boldsymbol{\theta}}_n) - D(\boldsymbol{\theta}) \leq 0$. Consequently, $P\{\lim_{n\to\infty} |D(\hat{\boldsymbol{\theta}}_n) -$

$D(\boldsymbol{\theta})| > \gamma\} = 0$ for any $\gamma > 0$. And by the uniqueness of the deepest point, we obtain $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| \to 0$ almost surely.

□

Proof of Proposition 3.2

The proof directly follows the proof of Theorem 2.

□

Proof of Theorem 3

(a), (b), (c) and (d) hold naturally since FARD is the integration of pointwise ARD values.

(e) also holds immediately following the fact that FM depth does not degenerate.

□

Proof of Theorem 3.4

By definition,

$$FARD_n(\boldsymbol{y}) = \sum_{i=1}^{p} \frac{t_i - t_{i-1}}{\|\mathcal{T}\|} ARD_n(\boldsymbol{y}(t_i))$$
$$FARD(\boldsymbol{y}, F_{\boldsymbol{Y}}) = \int_{\mathcal{T}} \frac{1}{\|\mathcal{T}\|} ARD(\boldsymbol{y}(t), F_{\boldsymbol{Y}(t)}) dt$$

By Theorem 3.2, we obtain that for any $t \in \mathcal{T}$,

$$\sup_{\boldsymbol{y}(t) \in \mathbb{R}^d} \|ARD_n(\boldsymbol{y}(t)) - ARD(\boldsymbol{y}(t),\, F)\| \to 0,\ a.s.$$

Moreover, since $H(t) := ARD(y(t)) \in Lip(\mathcal{T})$, $\exists A > 0$ such that $\sup_{s,t} |H(t) - H(s)| \leq A|t - s|$. Thus,

$$\sup_{\boldsymbol{y} \in C(\mathcal{T})^d} \|FARD_n(\boldsymbol{y}) - FARD(\boldsymbol{y},\, F)\|$$

$$
\begin{aligned}
&= \frac{1}{\|\mathcal{T}\|} \sup_{\boldsymbol{y}\in C(\mathcal{T})^d} \int_{\mathcal{T}} |ARD_n(\boldsymbol{y}(t)) - ARD(\boldsymbol{y}(t), F_{\boldsymbol{Y}(t)})| dt \\
&\leq \frac{1}{\|\mathcal{T}\|} \sum_{i=1}^{p} \int_{t_{i-1}}^{t_i} (A(t_{i+1} - t_i) + \sup_j 2|ARD_n(\boldsymbol{y}(t_j)) - ARD(\boldsymbol{y}(t_j), F_{\boldsymbol{Y}(t_j)})|) dt \\
&= \frac{1}{\|\mathcal{T}\|} \sum_{i=1}^{p} A(t_{i+1} - t_i)^2 + \sup_j 2|ARD_n(\boldsymbol{y}(t_j)) - ARD(\boldsymbol{y}(t_j), F_{\boldsymbol{Y}(t_j)})| \\
&= O(p^{-2\gamma}) + \sup_j 2|ARD_n(\boldsymbol{y}(t_j)) - ARD(\boldsymbol{y}(t_j), F_{\boldsymbol{Y}(t_j)})|
\end{aligned}
$$

By Theorem 3.4, we know the second term converges to zero almost surely. □

Proof of Theorem 3.5

By Proposition 1, we obtain $|\hat{\boldsymbol{\theta}}_n(t) - \boldsymbol{\theta}(t)| \to 0$ a.s. point wisely over $\mathcal{T}$. From the definition of $\hat{\boldsymbol{\theta}}_n(t)$, it is easy to tell $|\hat{\boldsymbol{\theta}}_n(t)| < \max_{i\in[n]} |\boldsymbol{y}_i(t)| < \boldsymbol{g}(t)$, where $\boldsymbol{g}(t) = (g_1(t), \ldots, g_d(t))$. Then, by dominated convergence theorem, we can get the uniform convergence over $\mathcal{T}$. □

Remark 3.6. *If* $\{\boldsymbol{Y}_1(t)\}, \ldots, \{\boldsymbol{Y}_n(t)\}$ *are continuous over* $\mathcal{T}$, $\hat{\boldsymbol{\theta}}_n(t)$ *is also continuous.*

Proof:

We need to show that $\forall t \in \mathcal{T}$, $\forall \delta > 0$, we have

$$\lim_{\Delta t \to 0} |\hat{\boldsymbol{\theta}}_n(t + \Delta t) - \hat{\boldsymbol{\theta}}(t)_n| = 0$$

Given any $t \in \mathcal{T}$, without loss of generality, we assume

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_n(t) &= \boldsymbol{y}_1(t) \\
\hat{\boldsymbol{\theta}}_n(t + \Delta t) &= \boldsymbol{y}_2(t + \Delta t).
\end{aligned}
$$

Then, we would expect a switch in rank between $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ during $(t, t+\Delta t]$. Thus,

$$|\hat{\boldsymbol{\theta}}_n(t+\Delta t) - \hat{\boldsymbol{\theta}}_n(t)| \leq \max\{|y_1(t+\Delta t) - y_1(t)|,\, |y_2(t+\Delta t) - y_2(t)|\}.$$

It goes to 0 as $\Delta t \to 0$.

Chapter 4

Nonparametric Tolerance Tubes for Functional Data

4.1 Introduction

Tolerance intervals and regions provide tolerance limits to univariate and multivariate data, and are deemed important tools in statistical quality control. However, the topic of tolerance limits of functional data remains relatively underdeveloped. In this chapter, we introduce tolerance tubes, as a generalization of tolerance intervals and regions, that can provide tolerance limits with a pre-specified coverage probability, say β, of functional data, with some some pre-specified level of confidence, say γ.

In $\mathbb{R}^1$, two types of tolerance intervals have been developed and commonly used, namely, $\beta-$content tolerance intervals and $\beta-$expectation tolerance intervals (Guttman, 1970). To be precise, let $X_1, \ldots, X_n$ be a sample from distribution $F \in \mathbb{R}^1$:
(1) $T(X_1, \ldots, X_n)$ is called a β-content tolerance interval at confidence level γ if

$$P\{P_F(T(X_1, \ldots, X_n)) \geq \beta\} = \gamma.$$

(2) $T(X_1, \ldots, X_n)$ is called β-expectation tolerance interval if

$$E\{P_F(T(X_1, \ldots, X_n))\} = \beta.$$

These two definitions have been generalized to multivariate settings to define tolerance regions. If the underlying distribution of the sample is known, one can easily establish a tolerance interval/region using density contours or other means. If it is unknown, there exist distribution-free tolerance intervals introduced in (Wilks, 1941; Wald, 1943) by using univariate order statistics or spacings; and there exist nonparametric tolerance regions proposed in (Chatterjee and Patra, 1980; Lei et al., 2013) by using estimated density, in (Bucchianico et al., 2001) by using index sets, and in Li and Liu (2008)by using the multivariate spacings induced by the center-outward order statistics devised from data depth. As observed in Li and Liu (2008), the approaches using estimated density or the index set have two drawbacks: i) they require that the shapes of tolerance regions be specified a priori, which is generally a difficult task without the knowledge of the underlying distribution, and also a mis-specified shape is unlikely to yield the desirable properties that one would expect from tolerance regions, and ii) if the underlying distribution is multi-modal, the resulting tolerance regions may consist of disjoint regions, which render them useless in practice, since it is difficult to provide a coherent interpretation of disjoint regions in the context of tolerance. Not surprisingly, straightforward generalizations of these two approaches to the functional setting would continue to have the same drawbacks.

In the literature, Bowden and Steinhorst (1973) and Rathnayake and Choudhary (2015) have proposed tolerance bands which provided tolerance limits for functional data. However, their approaches are valid only for univariate functionals and also only under Gaussian assumptions. In this paper, we aim to establish a general framework for tolerance tubes which can be applied broadly to functional data (including multivariate), where only continuity for each functional is assumed. Specifically, we generalize the aforementioned Definitions (1) and (2) to functional settings to define $\beta-$content and $\beta-$expectation tolerance tubes, and propose to construct such tolerance tubes by extending the idea in Li and

Liu (2008) of joining "spacings" suitably defined by the notion of data depth. To broaden further the utility of tolerance tubes, we also propose a useful modification of Definition (2) by inserting the notion of exempt level. This is elaborated below.

In many real applications, the tolerance tube does not have to be as stringent as the usual 100% compliance required in most production lines. For example, in an application of monitoring blood glucose levels of diabetes patients (which will be elaborated and illustrated in Section 4.5), it is reasonable and necessary to tolerate some temporary spikes of blood glucose levels due to normal factors such as meal-intake. Motivated by this consideration, we modify $\beta-$expectation tolerance tubes by introducing an exempt level α. The modified tolerance tube relaxes the requirement by allowing at most α portion of each functional outside of the tolerance limit. This modification is especially useful in the setting where short term aberrations in functional data are not necessarily viewed as substantive alteration of the functionals from their expected acceptable pattern.

The rest of the paper is organized as follows. In Section 4.2, we introduce formal definitions of tolerance tubes for functional data, by extending the $\beta-$content and $\beta-$expectation tolerance intervals. We further introduce the notion of an exempt level α to modify the $\beta-$expectation tolerance tube. Theoretical justifications and properties of those definitions are investigated. In Section 4.3, we propose an approach to construct the proposed tolerance tubes using the the idea of "spacings" derived by suitably defined notions of data depth in the functional setting. We also study the properties of those sample tolerance tubes. Section 4.4 provides three simulation studies to compare different tolerance tubes in terms of probability content and stability. Section 4.5 and 4.6 discuss two real data applications on glucose continuous monitoring and aircraft safe landing tracking. Some concluding remarks are presented in Section 4.7.

4.2 Nonparametric Tolerance Tubes

4.2.1 Definitions of Tolerance Tubes For Functional Data

Let $\{\boldsymbol{Y}_1(t) : t \in \mathcal{T}\}, \ldots, \{\boldsymbol{Y}_n(t) : t \in \mathcal{T}\}$ be a random sample of d-dimensional functional generated from the process F over $\mathcal{T}$, and $\{\boldsymbol{y}_1(t), t \in \mathcal{T}\}, \ldots, \{\boldsymbol{y}_n(t), t \in \mathcal{T}\}$ be their observed values. For any $t \in \mathcal{T}$, $Y(t)$ follows distribution F_t. The notation shall be simplified without $t \in \mathcal{T}$ in the bracket when the emphasis of $\mathcal{T}$ is not needed. Also, $\boldsymbol{Y}$ and $\boldsymbol{y}$ will be used to represent the functionals when there is no possibility of confusion with multivariate variables.

In general, a tolerance tube provides tolerance limits for a specified percentage of functional data with some pre-fixed level of confidence. More specifically, let $T(Y_1, \ldots, Y_n) = \{C_n(t) : t \in \mathcal{T}\}$ be a tolerance tube, where $C_n(t)$ is a set at a fixed t. A functional data y is covered within the tube $T(Y_1, \ldots, Y_n)$ if $(t, y(t)) \in T(Y_1, \ldots, Y_n)$ for any $t \in \mathcal{T}$. For simplicity, we denote the tube by T_n and the coverage of y by $y \sim T_n$. In what follows, we generalize the two definitions of tolerance intervals in $\mathbb{R}^1$ in Section 4.1 to yield the following three definitions of tolerance tubes in the functional setting.

Definition 4.1. *$T(Y_1, \ldots, Y_n)$ is called a $\beta-$content tolerance tube at confidence level γ if*

$$P\{P_F(T(Y_1, \ldots, Y_n)) \geq \beta\} = \gamma.$$

Definition 4.2. *$T(Y_1, \ldots, Y_n)$ is called $\beta-$expectation tolerance tube if*

$$E\{P_F(T(Y_1, \ldots, Y_n))\} = \beta.$$

In practice, the criterion of $Y \sim T(Y_1, \ldots, Y_n)$ can be relaxed such that only a portion of

Y inside the tube is required. For example, in a project of glucose continuous monitoring (see Section 4.5), the blood glucose level of diabetes patients surged around meal times and dropped back to the normal level afterward. In addition, we observed that such spikes appeared at different times among different patients. In such a case, the tolerance tube defined using Definition 4.1 or Definition 4.2 hardly exists in general. To define a meaningful tolerance tube that can accommodate the aforementioned allowable occasional exceptions, we introduce an exempt level in the $\beta-$expectation tolerance tube and yield Definition 4.2a below.

Definition 4.2a. *$T(Y_1, \ldots, Y_n)$ is called $\beta-$expectation tolerance tube with exempt level α if*

$$E\{P_F^{\alpha}(T(Y_1, \ldots, Y_n))\} = \beta,$$

where $0 \leq \alpha \leq 1$, $P_F^{\alpha}(T_n) := P_F(\lambda\{t : y(t) \in T_n\} \geq (1-\alpha)\lambda(\mathcal{T}))$, and $\lambda(\cdot)$ is a lebesgue measure on $\mathcal{T}$. When $\alpha = 0$, this is equal to Definition 4.2.

Roughly speaking, the tuning parameter α reflects the degree of strictness of enforcing the requirement of the tube. As α increases, the criterion for any functional falling within the tube is relaxed since it only requires at least $(1-\alpha)$ portion of the functional to be inside the tube. As a result, a larger α will usually yield a narrower tube. Ideally, α should be determined using domain knowledge. Usually, too large an α may render the tube meaningless or useless from the practical perspective. If the domain knowledge is unavailable, one may resort to cross-validation, which, however, might be computationally costly. If this cost can not be overcome, α can be determined by visual decision instead, which obviously will reduce the reliability of claimed accuracy of the tube.

4.2.2 Desirable Properties of Tolerance Tubes

For any function $\boldsymbol{Y}(t) = (Y_1(t), \ldots, Y_d(t))$, we assume all its components are Lipschitz functions. That is, for $s, t \in \mathcal{T}$, $\|Y_j(s) - Y_j(t)\| \leq \|s - t\|$, for $j \in \{1, \ldots, d\}$. A desirable tolerance tube is expected to satisfy the following properties:

P1. A tolerance tube is connected throughout the whole index domain.

P2. A tube evaluated at each index is connected. (any requirement for tolerance region?)

P3. The tolerance tube expands as the tolerance level goes up. Specifically, if $\beta_1 \leq \beta_2$, T^{β_1} is nested within T^{β_2}, namely, $T^{\beta_1} \subseteq T^{\beta_2}$.

Thus, the tolerance tube is different from the functional boxplot, bagplot (Hyndman and Shang, 2010) and other tools which are based on some statistics of the functional sample. Some other examples of invalid tolerance tubes are shown in Figure 4.1. In panel (a), boundaries are disconnected and incomplete; in panel (b) and (c), there are hollow spaces inside the tube; in panel (d), the tube degenerates to a single line.

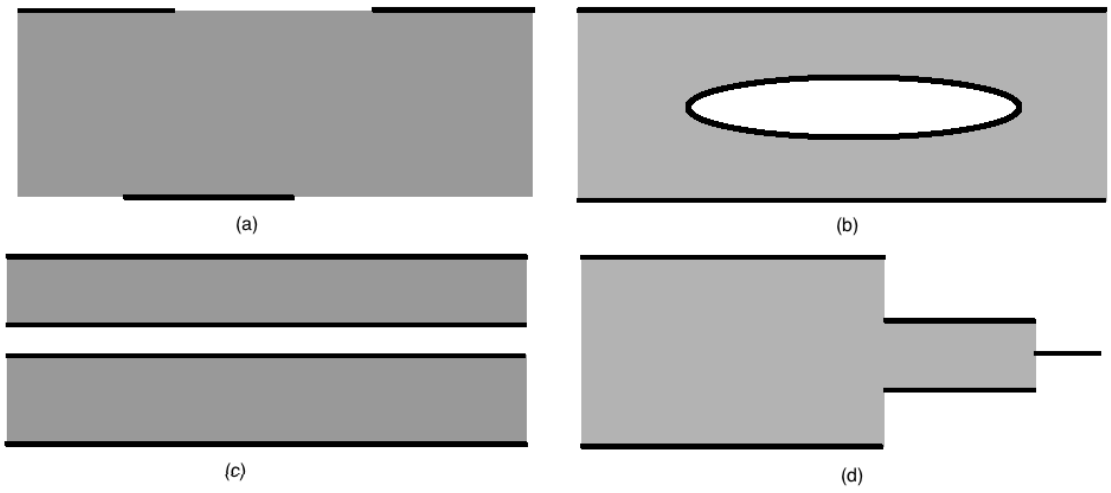


Figure 4.1: Examples of invalid tolerance tubes: (a) boundaries are disconnected and incomplete; (b) and (c) there is a hollow space inside the tube; (d)the tube degenerates. Shadow area represents the spread of functional data.

4.3 Constructing Nonparametric Tolerance Tubes Using Data Depth

Data depth has been developed to quantify the "centrality" of functional data. The central region derived from depth provides a potential formation of tolerance intervals/regions. This formation is nonparametric and data driven, thus save the effort of dealing with complex distributions of functional data. However, it can not guarantee the pre-set tolerance level, and is sensitive to aberrations in the sample. In addition, there is not a natural way of incorporating the exempt level α into the central region. Thus, we propose an effective approach to construct $\beta-$expectation tolerance tube with exempt level α using the quantile information of point-wise depth values. In what follows, we first give a brief review of existing notions of data depth and the corresponding central region in both multivariate and functional data settings.

4.3.1 Central Region Derived from Data Depth

Without loss of generality, let $FD(\cdot)$ be a general notation of functional depth. Given a sample of functionals $Y_1, \ldots, Y_n$, we obtain its depth order statistics $Y_{[1]}, \ldots, Y_{[n]}$. For any $\beta \in (0, 1]$, the $\beta-$central region is the set

$$CR_\beta(Y_1, \ldots, Y_n) = \{C_{n,\beta}(t) : t \in \mathcal{T}\},$$

where $C_{n,\beta}(t)$ is the smallest convex hull that includes $\{Y_{[1]}(t), \ldots, Y_{[r_n]}(t)\}$ with $r_n = (n+1)\beta$.

Example 1: Central Region For Univariate Functionals

For univariate functional data, the central region can be expressed as:

$$CR_\beta = \{y : y_l(t) \leq y(t) \leq y_u(t), \forall t \in \mathcal{T}\}.$$

Here, $y_l(t) = \min_y\{FD(y) \geq FD_{[r_n]}\}$, $y_u(t) = \max_y\{FD(y) \geq FD_{[r_n]}\}$, and $FD_{[r_n]}$ is the depth value of $Y_{[r_n]}$, $r_n = (n+1)\beta$.

Theorem 4.1. *Given a random sample $Y_1, Y_2, \ldots, Y_n$, let CR_β be the β central region derived via some FD, which has valid functional depth values. Then:*

(1) CR_β **P1** *to* **P3** *in Section 4.2.2;*

(2) CR_β is a β-expectation tolerance tube if $\forall y$, $y \in CR_\beta$ implies $FD(y) \geq FD_{[(n+1)\beta]}$.

Among all central regions derived by depth reviewed in Chapter 2, only ED is useful in constructing $\beta-$expectation tolerance tubes. We further investigate its performance in Section 4.4 to Section 4.6. On the other hand, there is no existing approach to construct the tube with exempt levels. Thus, in the next section, we propose an effective approach

to solve the problem.

4.3.2 Constructing $\beta-$expectation Tolerance Tubes with α Exempt Level

In the following, given a random sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, we propose a useful approach to construct the $\beta-$expectation tolerance tube on exempt level α.

1. For each functional $\boldsymbol{Y}_i$ at each point t, calculate its pointwise depth value $D(\boldsymbol{Y}_i(t), F_t)$;
2. Within each functional $\boldsymbol{Y}_i$,
   (a) identify the $\alpha-$quantile, say q_i^α, of its pointwise depth value;
   (b) identify the segment $T_i := \{t : D(\boldsymbol{Y}_i(t), F_t) \geq q_i^\alpha\}$;
3. Identify $(n+1)\beta$ curves with the highest q_i^α values. Without loss of generality, we assume they are $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{(n+1)\beta}$;
4. The $\beta-$expectation tolerance tube on α exempt level is the smallest convex hull that contains the set $\{\boldsymbol{Y}_i(t), t \in T_i\}_{i=1}^{(n+1)\beta}$.

Lemma 4.1. *Assume $D(x)$ is continuous on x, and F has Glivenko-Cantelli property uniformly over convex sets. In addition, we assume F_t is continuous for $t \in \mathcal{T}$. Then, for any Lipschitz continuous function y, $D(y,t)$ is continuous on t, namely, $|D(y,t) - D(y,t+\delta)| \to 0$ as $\delta \to 0$.*

Theorem 4.2. *The tolerance tube constructed from above procedure:*
(1) is a $\beta-$expectation tolerance tube;
(2) satisfies **P1** *to* **P3** *in Section 4.2.2 under mild conditions.*

4.4 Simulation studies

In this section, we conduct three simulation studies to investigate the performance of tolerance tubes with and without exempt levels. In particular, we are interested in studying the benefit of introducing the exempt level on the improvement of stability and coverage level under different functional data settings.

4.4.1 Simulation settings

Simulation setting I: Gaussian Processes. We generate 500 curves from gaussian process with mean $\mu(t) = 0$ and covariance function $K_y(s,t) = exp\{-\frac{|y_i(s)-y_i(t)|}{100}\}$. (All curves are observed at equal spaced grid $t = 1, 2, \ldots, 100$.)

Simulation setting II: Sinusoid Curves. (with slight location shift) We generate 500 curves from $y(t) = \sin(2\pi\theta(t+s)) + \cos(2\pi\theta(t+s))$, where $\theta = 0.05$ and $s \sim U[-10, 10]$.

Simulation setting III: Sinusoid with partial contaminations. We generate 500 curves from $y(t) = \alpha_1 \sin(2\pi\theta(t+s)) + \alpha_2 \cos(2\pi\theta(t+s))$, where $\alpha_1 \sim U[0.05, 0.1], \alpha_2 \sim U[0.05, 0.1]$, $s \sim U[-10, 10]$ and $\theta = 0.01$. All the curves are contaminated with a wave with height $h \sim U[0.05, 0.1]$ at a random place over a short period $\Delta t = 10$, that is,

$$y^\star(t) = \begin{cases} y(t) + \beta \sin(2\pi\phi(t - t_0)), & \text{if } t \in [t_0, t_0 + \Delta t) \\ y(t), & \text{otherwise,} \end{cases}$$

where $\beta \sim U[0.05, 0.1]$ and $\phi = 0.05$.

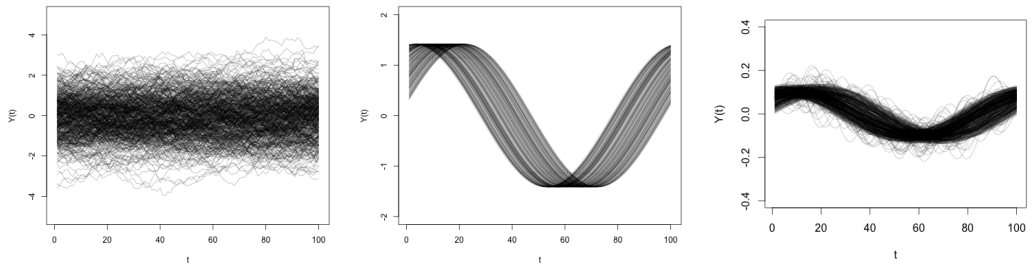


Figure 4.2: Simulation setting I: Gaussian Processes; Simulation setting II: Sinusoid Curves; Simulation III: Sinusoid with partial contaminations.

4.4.2 Simulation results

We construct $\beta-$expectation tolerance tubes with exempt level α by implementing the procedure proposed in Section 4.3.2. In addition, we use β central regions derived by ED to represent $\beta-$expectation tolerance tubes. Here, β is set to be 0.8 and α may vary from different studies. To assess their capability of achieving the target tolerance level, we randomly split the sample into two halves, one as the training set and the other as the test set. We construct β-expectation tolerance tubes using the training set, and report their coverage levels in the test set in Table 4.1. Each simulation is repeated 50 times.

	TT(ED central region)	TT with exempt level α
Simulation I	0.65(0.05)	0.74(0.03)
Simulation II	0.71(0.04)	0.79(0.03)
Simulation III	0.59(0.04)	0.74(0.04)

Table 4.1: Achieved averaged coverage levels (standard deviations) of the $\beta-$expectation tolerance tubes with and without exempt levels in the test sets. $\beta = 0.8$.

We observe that the coverage of the central region is notably less than the nominal level in all three simulated studies. Especially in Simulation III, where functionals have occasional

spikes, the averaged coverage level is even less than 0.6. In general, if any functional is likely to have many crisscrosses with other functionals, the central region often fails to meet the nominal level in the test set. On the other hand, we observe that tolerance tubes with exempt level α improve the coverage level even in Simulation III. The yield coverage is much closer to the nominal level 0.8 yet with less variance. Here, we use cross-validation to choose α to be $0.2, 0.2$ and 0.25 in three studies, respectively.

It is noteworthy that the shapes of tolerance tubes without exempt levels are much more sensitive to the randomness in the sample. This drawback is acute particularly in Simulation III. The shapes of the tubes can vary a lot between different training sets. This makes it inappropriate in practice. This drawback is largely mitigated by incorporating an exempt level. As in Figure 4.3c, the exempt tube does not involve any of the spikes and is more stable against the change of training sets.

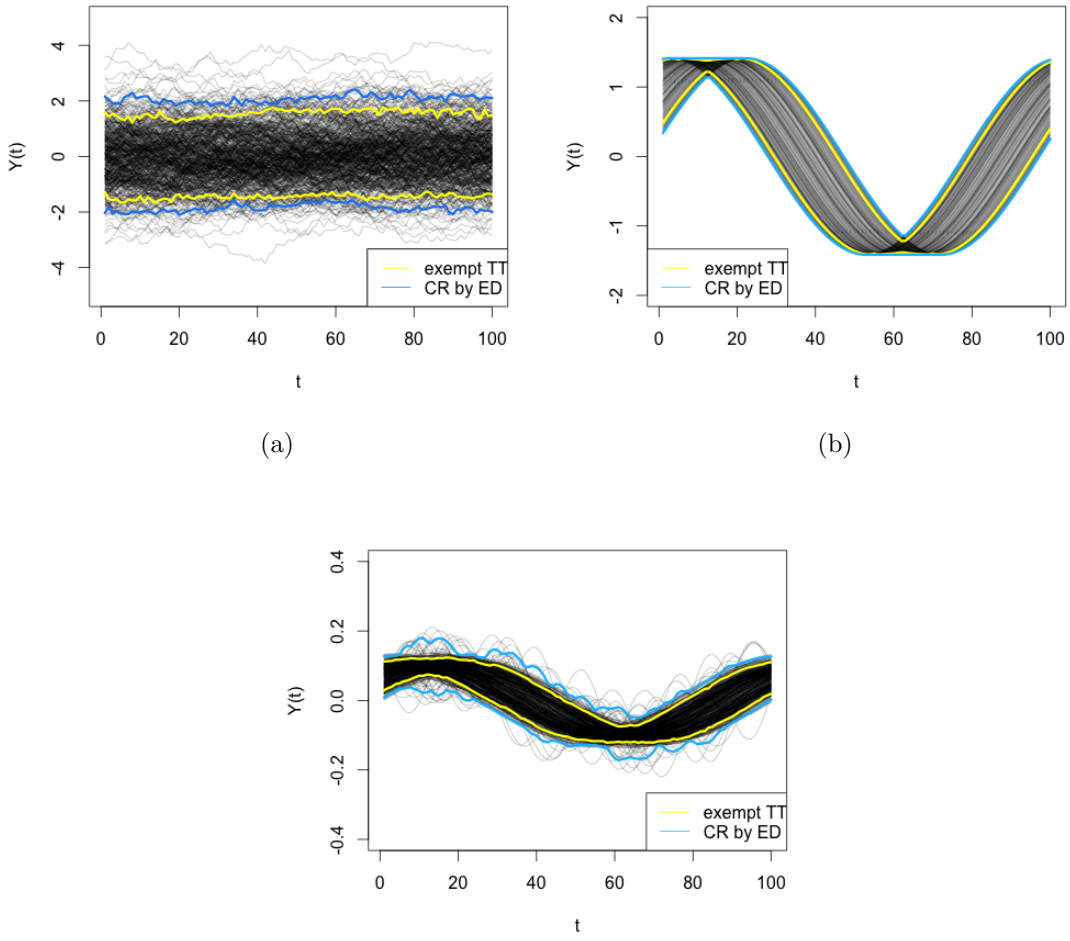


Figure 4.3: Comparison of $\beta-$expectation tolerance tubes with and without exempt levels. Simulation setting I: Gaussian Processes; Simulation setting II: Sinusoid Curves; Simulation III: Sinusoid with partial contaminations.

4.5 Real Example: Blood Glucose Monitoring

An effective closed-loop artificial pancreas system is essential in glycemic management of diabetes patients. It monitors the blood glucose density level continuously and manages the amount of insulin injection accordingly. Insufficient insulin is unable to lower the blood glucose level when it is stimulated by taking meals or other activities. On the contrary, overdosed injection will result in low blood glucose levels (or hypoglycemia), which may cause coma or more severely, death. The dataset we used in this study contains the blood glucose measurements of 121 diabetes patients over a whole day. The measurement was taken every 5 minutes from 00:00 to 23:59. Generally, we expect that the tolerance tube i) signals abnormal blood glucose levels; ii) allows occasional short-term spikes in the curve which can be due to normal reasons such as taking meals. The latter is critical because the blood glucose level of patients is never constant over time. For example, it rises after taking meals and drops back if it is under proper control. Thus, temporary spikes due to normal factors should not rise severe health concerns. Satisfying ii) can reduce the false alarm rate and avoid unnecessary panic in patients.

As a simple illustration, we present the $80\%-$expectation tolerance tubes with and without exempt levels in Figure 4.4. Here, the exempt level is set to be 0.15. We observe that the tolerance tube with the exempt level (on the right panel) possesses the following desirable properties: i) it achieves the nominal level of coverage in the sample; ii) the exempt level α can incorporate additional information from domain knowledge; iii) the tolerance limit across time is more robust to individual turbulence which is not representative of the panel. In particular, the latter is generally lacked in tolerance tubes without exempt levels. As a comparison, we observe that the shape of the tolerance tube varies dramatically across time on the left panel of Figure 4.4.

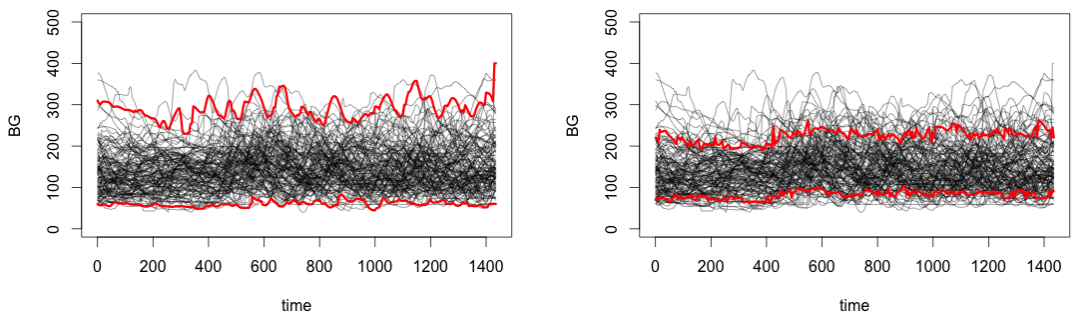


Figure 4.4: 80%−expectation tolerance tubes with (left) and without (right) exempt levels for blood glucose levels of 121 diabetes patients.

4.6 Real Example: Aircraft Landing Monitoring

The proposed tolerance tube with exempt levels can be applied to the aircraft landing data set discussed in Section 3.4 to continuous monitor and alert potential anomalies in aircraft performance during the approach phase of flight. Since in practice, the transient deviation from the recommended criteria is often inevitable, the tolerance tube with exempt levels would be particularly suitable for spotting anomalies. Ideally, the exempt level (α) and tolerance level (β) should incorporate domain knowledge.

4.7 Discussion

In this chapter, we formally define tolerance tubes of functional data, as a generalization of tolerance intervals and regions in finite dimensional settings. Tolerance tubes provide tolerance limits to ensure that a specified portion of a functional data set to fall within the tube with some desired level of confidence. Although the usual definitions of β−content and β−expectation tolerance intervals can be generalized in a straightforward fashion, such

defined tolerance tubes, well defined as they are, may be too stringent to have broad utility for tracking functional data in real applications over the increasing volatility in functional data over the continuum. Thus, we further introduce an exempt level α in the definition of tolerance tubes, which allows the "worst" α portion of each functional to be exempt from the requirement. Under the modified definition, a functional is considered contained inside the tube as long as at least $(1-\alpha)$ portion of it is within the tube. Such tolerance tube allows us to incorporate possible practical considerations from domain experts in order to draw more meaningful and practical inferences. Tolerance tubes with a suitable exempt level are shown to be more effective in the real applications of continuous tracking of blood glucose and aircraft landing performance.

In this chapter, our approach to construct the tolerance tubes is based on data depth. In principle, using any well-defined notion of functional depth should be able to produce $\beta - expectation$ tolerance tubes using its β central region. However, when exempt levels are considered, the functional central region are no longer valid, and it is difficult to see how this approach can be modified. On the other hand, our proposed approach to utilize the quantile information of depth values evaluated at each domain index and construct the tube by excluding the "worst" part from each functional, and it is thus particularly suited for incorporating exempt levels.

The idea of exempt levels behind the proposed tolerance tube in Definition 4.2a is to accommodate allowable occasional aberrations or random oscillation, and yet remains effective in detecting any continuous long stretch outside the tolerance tube. The latter could be indicative of pattern of behavior, and requires further consideration or analysis. For example, blood glucose stays higher above for a sustained period of time may indicate insufficient intake of insulin. In such a scenario, the input from domain experts may be utilized to devise an additional dynamic reactionary procedure into our construction of

tolerance tubes.

Finally, it is worth noting that our proposed tolerance tubes can be implemented with suitable weighting schemes if needed. Case in point is the example of aircraft tracking project above. Depending on the domains or contexts of the applications, different weighting schemes may be employed to further refine or sharpen the tolerance tube. Obviously, a prolonged deviation from the tolerance tube closer to the runway threshold would potentially incur higher risk in safe landing, as in contrast with the occurrence of such a prolonged deviation that is still far off the runway threshold. Therefore, it is useful to work with domain experts to devise a suitable weighting scheme with more stringent weight as the aircraft approaches the runway. Note that Chapter 3 has pursued a similar goal in this aircraft landing performance project by developing an effective outlier detection approach using ARD. It would seem that our approach using tolerance tubes coupled with ARD and suitable weighting schemes may be a more practical approach, because such tolerance tubes would be able to account for more forcefully the specific pattern of the functions.

4.8 Proofs

Proof of Theorem 4.1:

(1) CR_β is constructed using the smallest convex hull which contains the deepest $(n+1)\beta$ functions throughout the domain. Thus, it is straightforward to show that CR_β is connected throughout the domain and convex at each index.

Given a sample of functional data, we obtain its depth order statistics, say, $Y_{[1]}, \ldots, Y_{[n]}$ with decreasing depth values. Then CR_{β_1} and CR_{β_2} are the smallest convex hull which contains $Y_{[1]}, \ldots, Y_{[(n+1)\beta_1]}$, and $Y_{[1]}, \ldots, Y_{[(n+1)\beta_2]}$, respectively. Since $\beta_1 \leq \beta_2$, it is clear

that $CR_{\beta_1} \subseteq CR_{\beta_2}$.

(2) Let $S_r = \{y : FD(y) \geq FD_{[r]}$, for $r = 1, \ldots, n$. It can be shown that $P_F(s_r) \sim beta(r, n-r+1)$. To achieve $\beta-$expectation, we have to solve for $\frac{r}{n+1} = \beta$, which implies $r = (n+1)\beta$. By the construction of central region, it is easy to obtain that $CR_\beta \supseteq S_r$. Thus, to achieve β expected coverage, it should satisfy $CR_\beta \subseteq S_r$. In other words, for any $y^\star \in CR_\beta$, $FD(y^\star) \geq FD_{[r]}$. □

Proof of Lemma 4.1:

$\forall t \in \mathcal{T}$,

$$\begin{aligned} |D(y, t+\delta) - D(y,t)| &= |D(y_{t+\delta}, F_{t+\delta}) - D(y_t, F_t)| \\ &= |D(y_{t+\delta}, F_{t+\delta}) - D(y_t, F_{t+\delta}) + D(y_t, F_{t+\delta}) - D(y_t, F_t)| \\ &\leq |D(y_{t+\delta}, F_{t+\delta}) - D(y_t, F_{t+\delta})| + |D(y_t, F_{t+\delta}) - D(y_t, F_t)| \\ &\triangleq A \quad + \quad B \end{aligned}$$

Since y is a Lipschitz function, $|y_{t+\delta} - y_t| \to 0$ when $\delta \to 0$. By the continuity of the depth function $D(\cdot)$, we obtain $A \to 0$. Next, we show that $B \to 0$ as well.

First of all, we observe that

$$|F_{t+\delta}(x) - F_t(x)| \leq |F_{t+\delta}(x) - F^n_{t+\delta}(x)| + |F^n_t(x) - F_t(x)| + |F^n_{t+\delta}(x) - F^n_t(x)|.$$

By the Glivenko-Cantelli property, the first two terms on the right hand side go to zero uniformly almost sure. Then, we only have to show $\sup_x |F^n_{t+\delta}(x) - F^n_t(x)| \to 0$ almost

surely, namely, for any $\gamma > 0$, and any x, t,

$$P\{\lim_{n\to\infty}\lim_{\delta\to 0}|F^n_{t+\delta}(x) - F^n_t(x)| > \gamma\} = 0.$$

Given any sample of functional data, there are two scenarios:

(1) for any function f_i in the sample, there is no one reaching x at t, i.e., $f_i(t) \neq x$ for $i = 1, \ldots, n$. Then, let $\nu = \min_i |f_i(t) - x|$. Since all functions are Lipchitz functions, $\exists M > 0$, s.t. $|f(t+\delta) - f(t)| \leq M|t + \delta - t| = M\delta$, $\forall f \sim F$. We select δ s.t. $M\delta < \nu$, namely, $\delta < \nu/M$. Then, we obtain $P\{|F^n_{t+\delta}(x) - F^n_t(x)| > \nu\} = 0$.

(2) there are some functions reaching x at t. From the discussion above, to satisfy $|F^n_{t+\delta}(x) - F^n_t(x)| > \gamma$, there have to be at least $n\gamma + 1$ functions reaching x at t. As $n \to \infty$, we obtain $P_{F_t}(x) > \gamma$. Since F_t is continuous, $P_{F_t}(x) = 0$ for all x. Contradiction. From the discussion above, we conclude that $\sup_x |F_{t+\delta}(x) - F_t(x)| \to 0$ almost surely. As a result, it would follow that $\sup_x |D(x, F_{t+\delta}) - D(x, F_t)| \to 0$. □

Proof of Theorem 4.2:

(1) Let $Y_1, \ldots, Y_n$ be a random sample of functional points following distribution F. For $i \in 1, \ldots, n$, let $z_i = FD(Y_i)$, $T_i = P_F(y : FD(Y) > z_i)$. By Theorem 4.1, we only need to show $y \in TT^\alpha_\beta \iff FD(y) \geq FD_{[r_n]}$, with $r_n = (n+1)\beta$.

Let $y^\star$ be a new functional with no less than α portion is inside the tube. Let $q^\alpha_\star$ be the $\alpha-$quantile of $\{D(\boldsymbol{y}^\star(t)), t \in \mathcal{T}\}$. For t_0 such that $y^\star(t_0) \in TT^\alpha_\beta$, $\exists j \in \{1, \ldots, (n+1)\beta\}$, such that $D_{t_0}(Y_{[j]}) \leq D_{t_0}(y^\star)$. Since $D_{t_0}(Y_{[j]}) \geq q^\alpha_j \geq q^\alpha_{[(n+1)\beta]}$, we obtain $D_{t_0}(y^\star) \geq q^\alpha_{[(n+1)\beta]}$. Thus, $q^\alpha_\star \geq q^\alpha_{[(n+1)\beta]}$. Contradiction.

(2)(P1). Without loss of generality, we assume $G = \{Y_1, Y_2, \ldots, Y_{(n+1)\beta}\}$ are selected to

construct the tolerance tube TT_β. Suppose TT_β breaks right after $t_0 \in \mathcal{T}$, namely, $\exists \delta_0 > 0$, $\forall 0 < \delta < \delta_0$ and $t = t_0 + \delta$, $D(Y_i, t) < D^\alpha(Y_i)$ for $i \in \{1, 2, \ldots, n\beta\}$. Here, $D^\alpha(Y_i)$ is the $\alpha-$quantile of the set $\{D(Y_i, t) : t \in \mathcal{T}\}$. Since the tube is not broken at t_0, $\exists G_0 \in G$, such that $y(t_0)$ are selected $\forall y \in G_0$. By Lemma 4.1, we obtain $\forall y \in G_0$, $D(y, t) = D^\alpha(y)$, and $\exists y^\star \notin G$, such that $y(t_0) = y^\star(t_0)$. We denote these two requirements as $(\star)$. By the randomness of functions, the probability of satisfying $(\star)$ is very small if not zero. Thus, unless under weird heteroscedastic sample structures, the tube would not break over the whole interval. □

(P2). We show that for any $t \in \mathcal{T}$, $\exists \delta_0 > 0$, such that at least one function y selected over the interval $[t, t + \delta]$.

(1) If $D(y, t) \geq D^\alpha(y)$, say, $\Delta = D(y, t) - D^\alpha(y) > 0$, by Lemma 4.1, $\forall 0 < \delta < \delta_0$,

$$D(y, t + \delta) - D^\alpha(y) = D(y, t + \delta) - D(y, t) + D(y, t) - D^\alpha(y)$$

By continuity, $\exists \delta > 0$, such that $|D(y, t + \delta) - D(y, t)| < 1/2\Delta$. Thus,

$$D(y, t + \delta) - D^\alpha(y) > 1/2\Delta > 0.$$

Thus, y is also selected over $[t, t + \delta]$.

(2) If $D(y, t) = D^\alpha(y)$ for every function y that is selected at t_0, let $G_1 = \{y : y(t_0) \text{ is selected}\}$. By P1, $\exists \delta_0 > 0$, such that $\forall 0 < \delta < \delta_0$, denote $G_2 = \{y : y(t_0 + \delta) \text{ is selected}\}$. If the scenario in Figure 4.5 happens, $(G_1 \cap G_2) = \varnothing$. From discussions in (1), it is clear that $\forall y \in G_2$, $y(t_0) = D^\alpha(y)$ holds as well. Thus, it should be selected at t_0, namely, $(G_1 \cap G_2) \supseteq y$, contradiction! □

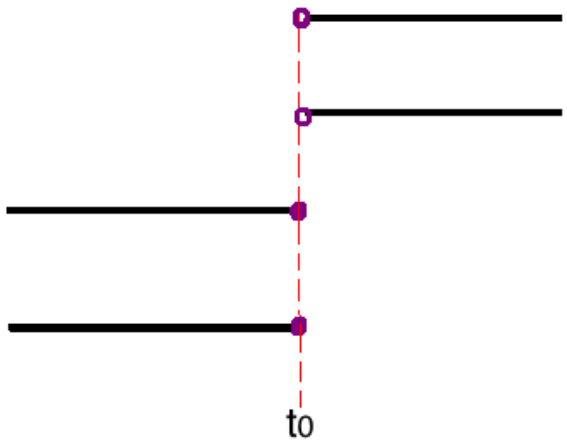


Figure 4.5: The showcase of a disconnected tube.

(P3). Tolerance tubes are nested. Let $0 < \beta_1 < \beta_2 \leq 1$. We show that $TT_{\beta_1(t)} \in TT_{\beta_2(t)}, \forall t \in \mathcal{T}$. For any $\boldsymbol{y}_i(t) \in TT_{\beta_1}(t)$ for some t, we obtain $q_i^\alpha \geq q_{[(n+1)\beta_1]}^\alpha \geq q_{[(n+1)\beta_2]}^\alpha$. Thus, $\boldsymbol{y}_i(t) \in TT_{\beta_2}(t)$. $\square$

Chapter 5

KMPD (K-means on Pairwise Distance): A New Clustering Approach and Its Application to Aircraft Landing Pattern Recognition

5.1 Introduction

Cluster analysis aims generally at grouping a collection of objects into clusters such that members within the same cluster are more similar to each other than those in different clusters. While this holds for most clustering approaches, the interpretation of "similar" can vary greatly in different approaches, and notions of dissimilarities can take a variety of forms. Popular choices include Euclidean distance or L_2 norm, L_1 norm, L_∞ norm, cosine, Mahalanobis distance, or a mixture of these. Naturally, the choice of dissimilarity notions would strongly impact the clustering result, and thus should be, and often so, chosen to reflect the data structure or subject matter condition. Among the existing clustering methods, K-means introduced and studied in (Lloyd, 1957; Forgy, 1965; MacQueen, 1967; Hartigan and Manchek, 1979), is arguably the most popular one due to its straightforward interpretation and simplicity in computation. However, it has been observed by Garcia-Escudero et al. (2008) that "this method is designed for clustering spherical groups of roughly equal sizes and, thus the method is not reliable for analyzing constellations of groups that depart strongly from this assumption". The spherical condition can sometimes

be achieved by applying suitable transformations, but, in reality, the condition of equal-size generally does not hold or is impossible to verify. Thus, the latter "equal-size" clustering result has been a well-known shortcoming for the K-means method.

Another important consideration in cluster analysis is the detection and handling outliers. Outliers can present as the noise in a sample or abnormal behaviors of legitimate observations, namely, observations being discordant with the vast majority in the sample. Desirable clustering approaches should be able to separate outliers from normal observations, ideally as a separate cluster. However, some approaches, K-means included, are sensitive to the presence of outliers, and thus likely to distort the clustering result. Although there are several approaches on robust cluster analysis such as Garcia-Escudero et al. (2008), Gallegos and Ritter (2005), most of them are parametric in nature and require specified underlying distributions, which is not easy to verify in practice. Thus, a broadly applicable nonparametric approach is strongly desired.

The goal of this paper is to introduce a new clustering approach, referred to as *K-means on Pairwise Distance* (KMPD), which can i) achieve the two aforementioned desirable properties; ii) detect and cluster separately outliers. The idea of KMPD is simple and intuitive. Roughly speaking, KMPD conducts similar algorithms as in K-means, but on the *proximity matrix* of the data set. Here, the proximity matrix is a symmetric and positive definite matrix which contains the pairwise dissimilarity of all points in the data set. In different contexts, proximity or dissimilarity may be defined differently, for example, Jaccard coefficient for binary vectors, cosine measure for string vectors. Since this paper focuses on mainly data sets of continuous observations, for convenience, we simply use "distance" to represent the proximity metric throughout. The KMPD method delivers such a result that the sample points assigned to the same cluster would have similar distance with respect to the overall data set. For example, if A and B are in the same cluster, the

distance of A with respect to the other points are similar to B would have. That is to say, if A is far away from the majority of points, B would also be far away. Thus, in the context of outlier detection, it is effective to detect anomalies and gather them into one cluster, separated from the regular sample points.

KMPD is also applicable to functional data as long as the distance between functionals is well defined. Theoretically, functional data are infinite dimensional. This nature would present some difficulty to many existing clustering methods which are often based on density estimation or some assumed model. However, KMPD works on the proximity or distance matrix instead of the original data, it applies to any dimensional settings with well-defined distance measures.

The rest of this chapter is organized as follows. In Section 5.2, we introduce the procedure of KMPD, followed by a discussion of the advantages of KMPD over K-means. Comparison studies of KMPD and K-means are presented in Section 5.3 using simulated examples. It is shown that KMPD outperforms K-means with much higher accuracy in discovering the underlying clusters in both multivariate and functional examples. In Section 5.4, we apply KMPD to a real data application of aviation risk management to detect possible risky landings. More concluding remarks are provided in Section 5.5.

5.2 Methodology: KMPD

In this section, we introduce the procedure of KMPD method in detail. Given a sample $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ or in functional spaces,

1. transform the sample to $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ where $\boldsymbol{p}_i = (p_{i,1}, \ldots, p_{i,n})$. Here, $p_{i,j} = d(x_i, x_j)$ is the dissimilarity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where $d(\cdot, \cdot)$ is a distance measure;

2. for a specified K, we solve

$$C^* = \underset{C}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{\substack{C(i)=k \\ i=1,\ldots,n}} \|\boldsymbol{p}_i - \boldsymbol{m}_k^{(p)}\|^2. \tag{5.1}$$

Here C is a cluster assignment and $C(i) = k$ means to assign object i to cluster k; m_k is the mean of cluster k, namely, $\boldsymbol{m}_k^{(p)} = \sum_{\substack{C(i)=k \\ i=1,\ldots,n}} \boldsymbol{p}_i / \sum_{i=1}^{n} I_{C(i)=k}$.

Note that, for $i = 1, \ldots, n$, $\boldsymbol{p}_i$ represents the "relationship" between $\boldsymbol{x}_i$ and the whole data set. The second step essentially uses the same iterative algorithm in K-means to obtain the centers of the transformed data set $\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n\}$. Thus, the sample points in the same cluster are similar regarding to their relationship with the whole sample. For instance, if A and B are assigned to the same cluster and A is far away from the majority of sample points, B would be far away as well.

It is known that K-means is ineffective in detecting clusters with different sizes. More specifically, when there exist several clusters with different sizes, K-means tends to separate the large clusters and absorb small clusters into the big ones. This is because the objective function of K-means implicitly penalizes both the size and the dispersion of the clusters, which tends to force the large clusters to split. More formally, given a sample $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \in \mathbb{R}^d$, K-means minimizes the objective function:

$$\sum_{k=1}^{K} \sum_{\substack{C(i)=k \\ i=1,\ldots,n}} \|\boldsymbol{x}_i - m_k\|^2,$$

where $m_k = \frac{\sum_{C(i)=k} p_i}{\sum_{i=1}^{n} I_{C(i)=k}}$. Here, $\sum_{\substack{C(i)=k \\ i=1,\ldots,n}} \|\boldsymbol{x}_i - m_k\|^2$ evaluates the dispersion of the k^{th} cluster. Thus, this objective function can be further expressed as

$$\sum_{k=1}^{K}(N_k - 1)\hat{\sigma_k^2},$$

where $\hat{\sigma_k^2}$ is an estimator of the variance of the k^{th} cluster. Consequently, this algorithm implicitly penalizes the size of large clusters twice. (First of all, it multiplies $\hat{\sigma_k^2}$ by N_k; secondly, large clusters tend to have greater dispersions.) As a result, this algorithm encourage large cluster to split and merge with small clusters.

KMPD, on the contrary, encourages the data points which are originally from the same large cluster to stay. If there exists two small clusters scatters around a big cluster, KMPD tends to merge the two small ones into one cluster, separated from the big one.

Determining the number of clusters has always been a critical issue for all partitioning types of clustering methods including KMPD and K-means. In literature, there are many proposals aiming to provide solutions to this issue, such as silhouette (Rousseeuw, 1986), gap statistic (Tibshirani et al., 2001). In practice, the choice should also depend on domain knowledge or experts inputs.

5.3 Simulation Studies

5.3.1 KMPD for Univariate Data

We use samples generated from mixture Gaussian distributions to illustrate the benefit of KMPD. Assume X is generated from the mixture of two Gaussian distributions, namely, $X \sim \delta Y_1 + (1-\delta)Y_2$, where $Y_1 \sim N(\mu_1, \sigma_1^2)$, $Y_2 \sim N(\mu_2, \sigma_2^2)$, $\delta \sim Bernoulli(p)$, and Y_1, Y_2, δ

are mutually independent. Here, the mixture proportion δ is rooted in p. In other words, if p is close to 1, most sample points would be generated from the first cluster. Without loss of generality, we assume $\mu_1 = -1$ and $\mu_2 = 1$. Figure 5.1 gives a simple illustration of the mixture distribution.

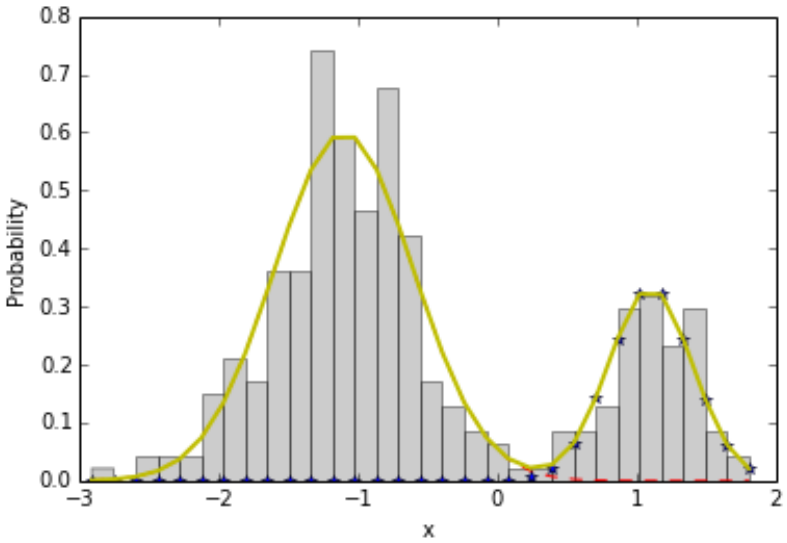


Figure 5.1: Histogram and pdf of the data setting: $p = .75, \sigma_1 = 0.5, \sigma_2 = 0.3, \mu_1 = -1.1, \mu_2 = 1.1$.

When $\sigma_1 = \sigma_2$ and $p = 0.5$, the data structure is spherical with equal-size. Both KMPD and K-means produce the same clustering result as that contained by the Bayes rule. Outside the special structured and equal-sized setting, the validity of K-means clustering result is often in doubt. Next, we provide a comparison study between KMPD and K-means on clustering univariate mixture Guassian samples under different parameter settings. Again, we use the Bayes rule as benchmark for evaluation. Specifically, we conduct the comparisons in the following experiments:

(1) Fix $\sigma_1 = 0.5$, $\sigma_2 = 0.5$, and vary p.

(2) Fix $p = 0.8$, $\sigma_1 = 0.5$, and vary σ_2.

Figure 5.2 and Figure 5.3 compare the decision boundaries produced by the two methods against the Bayes rule. Here, the decision boundary can be simply viewed as the boundary point between clusters.

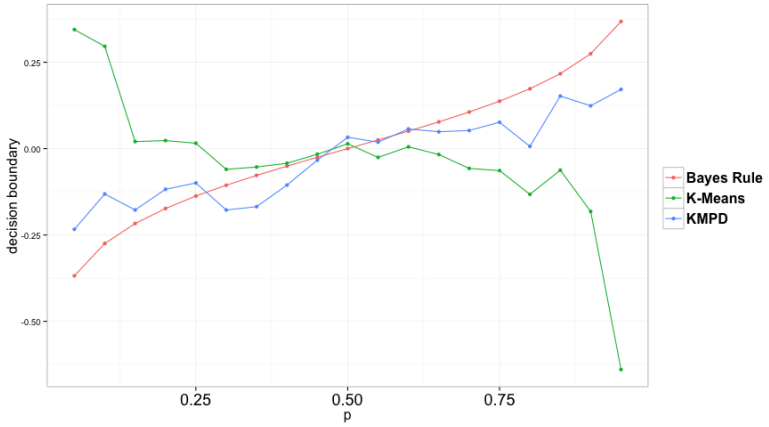


Figure 5.2: Decision boundaries of KMPD v.s. K-means when $\sigma_1 = \sigma_2 = 0.5$ and p varies over $(0, 1)$

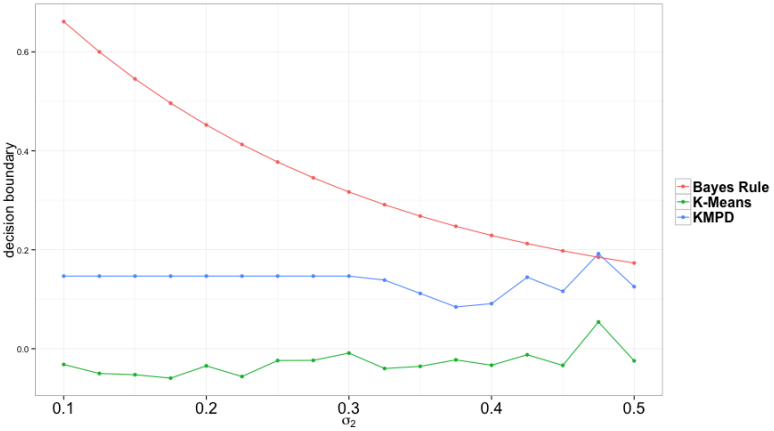


Figure 5.3: Decision boundaries of KMPD v.s. K-means when $p = 0.8$, $\sigma_1 = 0.5$ and σ_2 varies over $(0, 0.5)$

We observe that when the two oracle clusters are contrasting in size, KMPD outperforms K-means and obtains results closer to the result achieved by the Bayes rule. In experiment (1), when the two groups have equal variance, the boundary produced by KMPD is around the one produced by Bayes rule as p varies over $(0, 1)$. On the contrary, K-means only performs well when p is around 0.5. As p deviated further from 0.5, K-means result deteriorates more the Bayes rule result, and is substantively inferior as p is close to 0 or 1.

In many model-based clustering problem settings, such as the one described in our experiment (2), using Bayes rule would lead to two decision boundaries. Here, we only focus on the boundary between two cluster centers, because the other one can not always be quantified by all clustering methods, including K-means method. Thus, in experiment (2), only one boundary from Bayes rule result was plotted here. We observe in Figure 5.3, although neither KMPD nor K-means result follows closely the Bayes rule result, KMPD is consistently closer. With the gap seemingly approaches zero as σ_2 increases to the same value as σ_1 (e.g., it coincides with the Bayes rule when $\sigma_2 = 0.475$).

It is worth noting that when the two clusters are of similar sizes and dispersions, the improvement of KMPD over K-mean would not be pronounced. Instead, the two methods lead to similar clustering results.

We conduct three simulation studies to compare the performance of K-Means and KMPD in multivariate and functional data settings.

5.3.2 KMPD for Multivariate and Functional Data

Simulation Setting I: Mixture Gaussian in $\mathbb{R}^2$ We generate 300 points from a mixture bivariate Gaussian distribution as follows:

$$\begin{cases} N(\boldsymbol{\mu}_1, \Sigma_1), & \text{with probability } p \\ N(\boldsymbol{\mu}_2, \Sigma_2), & \text{with probability } 1-p, \end{cases}$$

where $\boldsymbol{\mu}_1 = (0,0)$, $\boldsymbol{\mu}_2 = (2,0)$,$\Sigma_1 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0.5 \end{smallmatrix}\right)$, $\Sigma_2 = \left(\begin{smallmatrix} 0.5 & 0 \\ 0 & 0.25 \end{smallmatrix}\right)$, and $p = 0.9$.

Simulation Setting II: Ring-type of outliers We generate 300 two-dimensional points from a bivariate normal distribution, namely, $\{(x,y) : (x,y) \sim N(\mu, \Sigma)\}$, where $\boldsymbol{\mu} = c(0,0)$ and $\Sigma = \begin{pmatrix} 0.25^2 & 0 \\ 0 & 0.25^2 \end{pmatrix}$, with 10% of them contaminated by points on the ring centered at the origin with radius 2.

Simulation Setting III: Brownian Motion We generate two clusters of 300 paths of brownian motion over the interval $[0, 2000]$. One cluster is more stable, which follows the generating process: $X(0) \sim N(0, 0.1^2)$, $X(t+1) = X(t) + N(0, 0.1^2)$, for $t = 1, \ldots, 1999$. It consists of 90% of the sample that one colored in black in the right panel of Figure 5.6. The other is more volatile, which follows the generating process: $X(0) \sim N(0,1)$, $X(t+1) = X(t) + N(0,1)$, for $t = 1, \ldots, 1999$, colored in red in the same figure.

Figures 5.4 to 5.6 compare the clustering results of KMPD and K-Means to the true labeling in each simulated setting. Different clusters are marked in different colors, black or red. Overall, as seen in Figure 5.7, KMPD outperforms K-means by yielding a lower misclassification rate in all three examples. K-means method shows a general phenomenon in yielding cross sectional clusters by some form of linear devision. For example, in the

middle panel of Figure 5.5, it performs a linear "cut" which produces two clusters of similar sizes, both of which contain sample points from the central normal distribution as well as from the ring. On the contrary, KMPD successfully recovers the data structure by grouping the points in the ring as a single cluster. Similarly, in Simulation III, KMPD separated the vast majority of the volatile cluster from upper and lower side of the stable cluster. However, the K-Means approach merged all the volatile trajectories into the stable group and caused a high misclassification rate. We replicate the simulation each by 50 times, and summarize the misclassification rates of the two methods using boxplots.

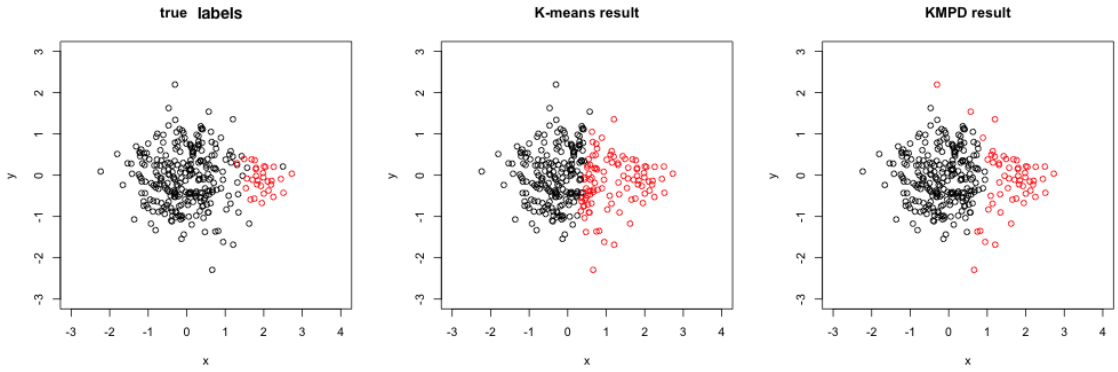


Figure 5.4: Clustering results of Simulation I: multivariate gaussian example.

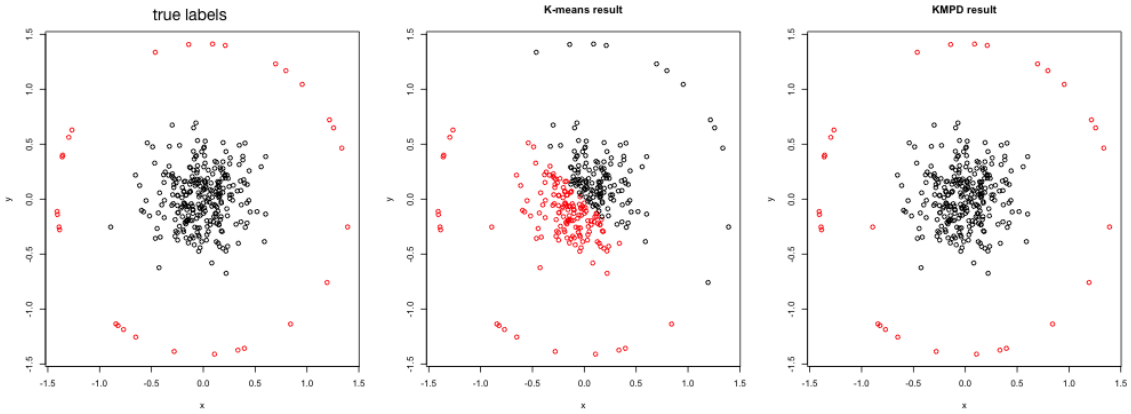


Figure 5.5: Clustering results of Simulation II: Ring-type outliers.

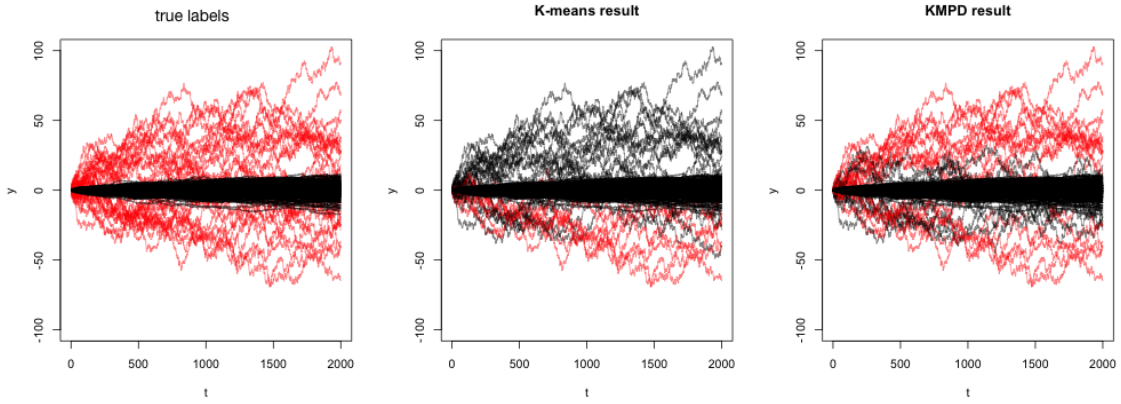


Figure 5.6: Clustering results of Simulation III: Brownian motion example.

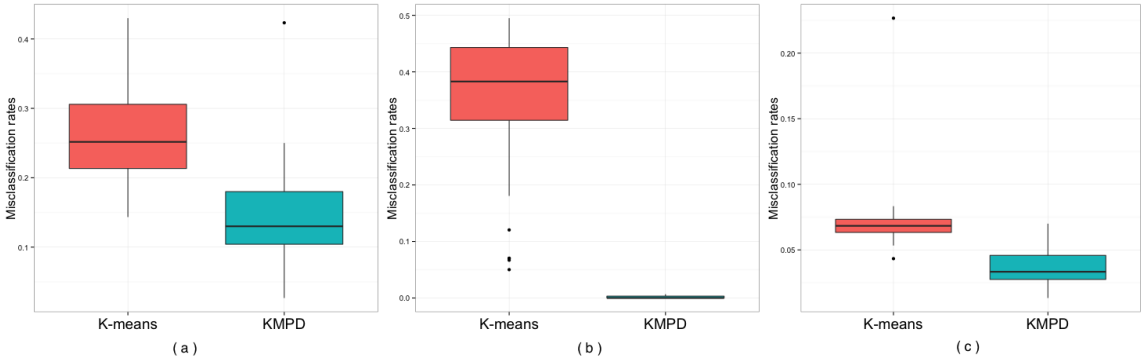


Figure 5.7: Boxplots of misclassification rates of K-means and KMPD for three simulation settings for multivariate and functional data. (a) to (c) are the results for Simulation I-Multivariate Gaussian, Simulation II-Ring-type Outliers and Simulation III-Brownian Motions, respectively.

5.4 Application on Aircraft Landing Pattern Recognition

In this section, we revisit the aircraft landing data set provided by our collaborating airline as discussed in Section 3.4. We propose to apply KMPD to this dataset to detect different landing patterns if there is any. First, we need to choose a suitable dissimilarity measure between each pair of functionals. Although there exist several dissimilarity measures for continuous functionals, such as L_1 and L_∞, we use the weighted L_2 norm to measure the dissimilarity in order to account for varying importance in different landing phases. Mathematically, we define the dissimilarity between any function $\{\boldsymbol{y}_i(t)\}$ and $\{\boldsymbol{y}_j(t)\}$ as:

$$p_{i,j} = \int_{\mathcal{T}} w(t)(y_i(t) - y_j(t))^2 \, dt, \tag{5.2}$$

where $w(t) \geq 0$ and $\int_{\mathcal{T}} w(t)dt = 1$. In practice, functional data are observed in finite discrete points, say T points in total. Thus, we obtain

$$p_{i,j} = \sum_{t=1}^{T} w_t (y_{i,t} - y_{j,t})^2 \tag{5.3}$$

with $w_t \geq 0$ and $\sum_{t=1}^{T} w_t = 1$.

To accentuate the importance of final approaching phase (i.e., close to touchdown point), we impose a weight function $w(t)$ which is monotone increasing with t. This is motivated by two practical considerations. First, without weight, the early landing phase where landing traces are much more spread-out and would dominate the overall clustering result. Thus, the landings in the final landing phase generally is more dense and would be marginalized. Second, the final landing phase is generally considered more critical than all early phases. If a aircraft landing deviates in the initial phase while it is still far away from the target touchdown point, the pilot has more leeway and time to correct the course to land within the target touchdown range. However, this correction would be difficult to achieve if an aircraft deviates substantively from the normal landing course near the target touchdown point. In this section, we consider the following weight function by incorporating the input from domain experts in aviation safety,

$$w(t) = \frac{1}{c + A(t)} \tag{5.4}$$

where $A(t)$ is the width of the cone at t, the distance to the runway threshold. The cone should accommodate the allowable deviation from the recommended gliding slope landing path, and c is a tuning parameter to be calibrated.

We applied KMPD to the data set with different choices of K, the number of clusters.

The final K is determined to be 3 by combining Silhouette and other statistics mentioned in Section 5.1 with as well as opinions from domain experts, and the need for a clear and meaningful interpretation of the clustering outcome. We use $K = 3$ to compare the performance of KMPD and K-means. It is worth noting that K-means approach fails to detect small clusters which is usually the case of the cluster of anomalous landing performance.

5.5 Discussion and Concluding Remarks

In this paper, we developed the new clustering method KMPD (K-means on Pairwise Distance) method. It utilizes the proximity matrix, and partitions data points based their dissimilarities to the entire data set. In other words, data points assigned to the same cluster are similar in terms of their overall dissimilarity with respect to the entire sample. This method is effective in discovering clusters with different sizes. We use simulated examples to show that KMPD outperforms K-means, in terms of the accuracy of clustering. KMPD is especially superior when the underlying clusters are widely different in size. Indeed, it works well regardless the difference in the relative size of the clusters.

KMPD is also useful in the context of outlier detection. It can gather the points which are "far" from the rest majority into one separate cluster, and thus label it as an outlier cluster. In the setting of functional data, we use an aircraft landing analysis to show that KMPD is effective in detecting an outlier cluster containing landing trajectories which deviate substantially from the vast majority which represent the expected normal operations. In Chapter 3, we pursue this same project using the so-called *antipodal reflection depth* approach, which is powerful in detecting outliers in both multivariate and functional data. It would be worthwhile to compare this method with KMPD.

Finally, we note that KMPD does not incur additional computational complexity, as it only requires additionally initial computation of pairwise distances.

Chapter 6

Summary

In this dissertation, we developed nonparametric approaches for several aspects in multivariate and functional data analysis. These approaches are useful for but not limited to solving problems in outlier detection, construction of tolerance tubes, and clustering. All the proposed approaches are shown to be effective using data from simulation and real applications. In Chapter 3, we proposed ARD approach to refine any existing notion of data depth to a class of new depth functions. The new depth functions gain the additional capability of capturing the relative magnitude of deviation from all data points to the deepest one. It broadens the utility of depth to research areas where the magnitude is indeed essential, for example, in the context of outlier detection. In Chapter 4, we introduced and investigated nonparametric tolerance tubes for functional data. In addition to the generalization of tolerance regions, we proposed modifications by coupling the definitions with an exempt level α. The idea of exempt levels is to mitigate the effect of occasional aberration or oscillation by randomness. Nonetheless, a continuous long stretch outside the tolerance tube could indicate pattern of behavior and needs further investigation. In Chapter 5, we proposed a new clustering approach KMPD to detect different patterns in functional data. This new approach focuses on the pairwise distance matrix of any given sample and thus brings a new and insightful interpretation to clustering results. It is particularly suitable for the scenarios when underlying clusters are substantially different in sizes.

All these approaches above are completely non-parametric and data-driven. Their utilities have been tested in both simulated studies and real applications, including an aviation risk management project involving monitoring aircraft landing performance and a medical study of blood glucose levels in diabetes patients. In the future, we plan to explore their boarder applications.

Bibliography

Boeing (1959-2014). Statistical summary of commercial jet airplane accidents: worldwide operations.

Bowden, D. and Steinhorst, R. (1973). Tolerance bands for growth curves. *Biometrics*, 29:361–371.

Bucchianico, A., Einmahl, J., and Mushkudiani, N. (2001). Smallest nonparametric tolerance regions. *Ann. Stat.*, 29:1320–1343.

Chakraborty, A. and Chaudhuri, P. (2014). On data depth in infinite dimensional spaces. *Annals of the Institute of Statistical Mathematics*, 66:303–324.

Chatterjee, S. and Patra, N. (1980). Asymptotically minimal multivariate tolerance sets. *Calcutta Statist. Assoc. Bull.*, 29:1828–1843.

Cheng, A., Liu, R., and Luxhoj, J. (2000). Monitoring multivariate aviation safety data by data depth: control charts and threshold system. *IIE Transactions*, 32:861–872.

Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. (2014). Multivariate functional halfspace depth. *J. Amer. Statist. Assoc.*, 109:411–423.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22:481–496.

Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.*, 40:322–352.

Donoho, D. and Gasko, M. (1992). breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Stat.*, 20:1803–1827.

FAA (2014). Advisory circular 91-79a - mitigating the risks of a runway overrun upon landing.

Febrero, M., Galeano, P., and Gonzalez-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics*, 19:331–345.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional data analysis*. Springer New York.

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21:768–780.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10:419–440.

Gallegos, M. and Ritter, G. (2005). A robust method for cluster analysis. *Ann. Statist.*, 33:347–380.

Garcia-Escudero, L., Gordaliza, A., Matran, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *Ann. Statist.*, 36:1324–1345.

Grubbs, F. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11:1–21.

Guttman, I. (1970). Statistical tolerance regions: classical and bayesian.

Hardin, J. and Rocke, D. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statististics*, 14:910–927.

Hartigan, J. and Manchek, A. (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C*, 28:100–108.

Hyndman, R. and Shang, H. (2010). Rainbow plots, bag plots, and box plots for functional data. *Journal of Computational and Graphical Statististics*, 19:29–45.

Krishnamoorthy, K. and Mathew, T. (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation*. John Wiley, New York.

Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free prediction sets. *J. Amer. Statist. Assoc.*, 108:278–287.

Li, J. and Liu, R. (2008). Multivariate spacings based on data depth: I. construction of nonparametric multivariate tolerance regions. *Ann. Statist.*, 36:1299–1323.

Liu, R. (1990). On a notion of data depth based on random simplices. *Ann. Statist.*, 18:405–414.

Liu, R., Parelius, J., and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. (special invited paper). *Ann. Statist.*, 27:252–260.

Liu, R. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.*, 88:405–414.

Lloyd, S. (1957). Least square quantization in pcm. *Bell Telephone Laboratories Paper*.

Lopez-pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *J. Amer. Statist. Assoc.*, 104:718–734.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability I*. University of California Press.

Mahalanobis, P. (1936). On the generalized distance in statistics. In *Proceedings of the National Institute of Sciences (Calcutta)*, pages 49–55.

Muller, H. and Stadtmuller, U. (2005). Generalized functional linear models. *Ann. Statist.*, 33:774–805.

Narisetty, N. and Nair, V. (2015). Extremal notion of depth for functional data. *J. Amer. Statist. Assoc.* (to appear).

Ramsay, J. and Silverman, B. (2005). *Functional data analysis*. Springer-Verlag New York.

Rathnayake, L. and Choudhary, P. (2015). Tolerance bands for functional data. *Biometrics*.

Riani, M., Atkinson, A., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *J. R. Statist. Soc. B*, 71:447–466.

Rousseeuw, P. (1986). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *computational and applied mathematics*, 20:53–65.

Rousseeuw, P. J. and Leroy, A. (1987). Robust regression and outlier detection.

Sun, Y. and Genton, M. (2011). Functional box plots. *Journal of Computational and Graphical Statist.*, 20:316–334.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J.R. Statistic. Soc. B*, 63:411–423.

Tukey, J. (1975). Mathematics and picturing data. In *Proceedings of the 1975 International Congress of Mathematics*, volume 2, pages 523–531.

Vardi, Y. and Zhang, C. (2000). The multivariate l1-median and associated data depth. In *Proceedings of the National Academy of Sciences*, volume 97, pages 1423–1426.

Wald, A. (1943). An extension of wilks' method for setting tolerance limits. *Ann. Math. Statist.*, 14:45–55.

Wilks, S. (1941). Determination of sample sizes for setting tolerance limits. *Ann. Math. Statist.*, 12:91–96.

Yao, F., Muller, H., and Wang, J. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.*, 100:577–590.

Yu, G., Zou, C., and Wang, Z. (2012). Outliers detection in functional observations with applications to profile monitoring. *Technometrics*, 54:308–318.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Ann. Statist.*, 28:461–482.