Data Depth Inference for Difficult Data

by

Kelly Ramsay

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2022

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor: Shoja'eddin Chenouri
Professor, Department of Statistics and Actuarial Science
University of Waterloo

External Examiner: Peter Rousseeuw
Professor, Department of Statistics
KU Leuven

Internal-External Member: Gautam Kamath
Professor, Department of Computer Science
University of Waterloo

Internal Member: Aukosh Jagannath
Professor, Department of Statistics and Actuarial Science
University of Waterloo

Internal Member: Christopher Small
Professor Emeritus, Dept. of Statistics and Actuarial Science
University of Waterloo

Internal Member: Gregory Rice
Professor, Department of Statistics and Actuarial Science
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

We explore various ways in which a robust, nonparametric statistical tool, the *data depth function* can be used to conduct inference on data which could be described as difficult. This can include data which are difficult in structure, such as multivariate, functional, or multivariate functional data. It can also include data which are difficult in the sense that published statistics must satisfy privacy constraints.

We begin with multivariate data. In Chapter 2, we develop two robust, nonparametric methods for multiple change-point detection in the covariance matrix of a multivariate sequence of observations. We demonstrate that changes in ranks generated from data depth functions can be used to detect certain types of changes in the covariance matrix of a sequence of observations. In order to catch more than one change, the first algorithm uses methods similar to that of wild-binary segmentation (Fryzlewicz, 2014). The second algorithm estimates change-points by maximizing a penalized version of the classical Kruskal Wallis ANOVA test statistic. We show that this objective function can be maximized via the well-known pruned exact linear time algorithm. We show under mild, nonparametric assumptions that both of these algorithms are consistent for the correct number of change-points and the correct location(s) of the change-point(s). We demonstrate the efficacy of these methods with a simulation study and a data analysis. We are able to estimate changes accurately when the data are heavy tailed or skewed. We are also able to detect second order change-points in a time series of multivariate financial returns, without first imposing a time series model on the data.

In Chapter 3 we extend these methods to the setting of functional data, where we develop a group of hypothesis tests which detect differences between the covariance kernels of J samples. These tests, called functional Kruskal Wallis for covariance tests, are based on functional data depth ranks, which are combined using the classical Kruskal Wallis test statistic. These tests are very robust; we demonstrate that these tests work well when the data are very heavy tailed, both in simulation and theoretically. Specifically, in order for the test to be consistent there is no need to assume that the fourth moment of the observations is finite, which is a typical assumption of existing methods. These tests offer several other benefits: they have a simple distribution under the null hypothesis, they are computationally cheap and they posses linear invariance properties. We show via simulation that these tests have higher power than their competitors in some situations, while still maintaining a reasonable size. We characterize the behavior of these tests under the null hypothesis and show consistency of the several versions of the tests under general alternative hypotheses. We also provide a method for computing sample size and provide some analysis under local alternatives when the ranks are based on L^2-root depth.

In Chapter 4 we present methods for detecting change-points in the variability of a sequence of functional data, thus, combining the methods of Chapter 2 and Chapter 3. Our methods allow the user to test for one change-point, to test for an epidemic period, or to detect an unknown amount of change-points in the data. Since our methodology is based on depth-ranks, we have no need to estimate the covariance operator, which makes our methods computationally cheap. For example, our procedure can identify multiple change-points in $O(n \log n)$ time. Our procedure is fully non-parametric and is robust to outliers through the use of data depth ranks. We show that when n is large, our methods have simple behaviour under the null hypothesis. We also show that the functional Kruskal Wallis for covariance change-point procedures are $n^{-1/2}$-consistent. In addition to asymptotic results, we provide a finite sample accuracy result for our at-most-one change-point estimator. In simulation, we compare our methods against several other methods from the literature. We also present an application of our methods to intraday asset returns and f-MRI scans.

In Chapter 5 we investigate differentially private estimation of depth functions and their associated medians. We then present a private method for estimating depth-based medians, which is based on the exponential mechanism (McSherry and Talwar, 2007). We compute the sample complexity of these private medians as a function of the dimension, prior parameters and privacy parameter. As a by-product of our work, we present a smooth depth function, which we show has the same depth-like properties as its non-smooth counterpart. Another by-product of our work is uniform concentration for several depth functions. We also present methods and algorithms for estimating private depth values at in-sample and out-of-sample points. In addition, we extend the propose-test-release methodology of (Brunel and Avella-Medina, 2020) to be used with depth functions and the exponential mechanism. We show that when using propose-test-release to projection depth values, the probability of no reply is small, and the private depth values concentrate around their population counterparts. We also give an algorithm to approximate the "test" step in propose-test-release, since it is computationally difficult. We show that this approximation maintains the low probability of no-reply as in the original propose-test-release.

Chapter 6 presents some possible directions for future research related to network data and shape data.

Acknowledgements

I would like to acknowledge my supervisor for providing guidance and support throughout preparing this thesis. I would also like to thank the members of my committee for taking the time to read the thesis and provide feedback. I especially thank Aukosh and Greg for helping me out a lot throughout my degree. I would also like to acknowledge my partner Taylor for supporting me throughout all of the ups and downs of writing this thesis. In addition, I would like to thank my entire family for supporting me along the way! A special thank you to my father and grandma for spending so much time on the phone with me! Thank you Steven and Rennae for being so supportive and thinking of me all of the time. Thank you granddad for sending me memes and supporting me. Thank you to Devon and Shayne for being very supportive of doing what I want to do. I would also like to thank all of my friends for their support in writing this thesis. Jonalee: thanks for visiting me in Waterloo! Thanks to my gamer buddies for helping me relax and have fun! Thank you Dylan Spicker for being a very supportive friend while creating this thesis, and for helping me compute derivatives and Taylor expansions! It is fun to be potatoes together.

Dedication

I dedicate this thesis to having fun with friends and family, the best part of life! I secondly dedicate this thesis to my best friend and partner in crime Taylor Oxelgren!

Table of Contents

List of Tables xii

List of Figures xiv

1 Introduction 1

1.1 Multivariate depth functions 4

1.2 Functional depth functions 10

1.3 Inference based on depth functions 17

1.4 Contributions 18

2 Kruskal-Wallis type statistics for multivariate change-point problems 24

2.1 Introduction 24

2.2 The data model, variability changes and their relation to depth ranks . . . 28

2.3 Proposed change-point algorithms 30

2.3.1 WBS and a depth rank CUSUM statistic 30

2.3.2 KW-PELT: A Kruskal-Wallis change-point algorithm 33

2.4 Consistency of the algorithms 35

2.5 Simulation study 38

2.5.1 Choosing the algorithm parameters 39

2.5.2 Analysing and comparing the algorithm performance 42

2.6 An application to financial returns 47

3 Kruskal-Wallis type statistics covariance kernel testing problems **50**
3.1 Introduction . . . 50
3.2 Model and test statistic . . . 54
3.3 Theoretical results . . . 58
3.4 Simulation results . . . 65
3.4.1 Models and settings . . . 65
3.4.2 Results . . . 68
3.5 Applications to real data . . . 71
3.5.1 F-GARCH residual analysis of intraday stock price curves . . . 71
3.5.2 Comparing speech variability with phoneme periodograms . . . 74

4 Kruskal-Wallis type statistics for functional change-point problems **78**
4.1 Introduction . . . 78
4.2 Changepoint model and methodology . . . 80
4.3 Theoretical results . . . 84
4.4 Simulation . . . 90
4.5 Data analysis . . . 96
4.5.1 Changes in volatility of social media intraday returns . . . 96
4.5.2 Resting state f-MRI pre-processing . . . 98

5 Depth Methods for Private Data Analysis **102**
5.1 Introduction . . . 102
5.2 Differential privacy . . . 104
5.3 A smooth depth function . . . 109
5.4 Estimating private depth-medians . . . 110
5.5 Computing private depth values . . . 116
5.6 Propose-test-release and Projection Depth . . . 118
5.7 Brief discussion of future directions . . . 126

6 Future Directions **127**

6.1 Extensions for private depth-based inference . . . 127

6.2 Extensions to manifold valued data . . . 128

References **129**

APPENDICES **145**

A Chapter Appendices **146**

A.1 Proofs from Chapter 2 . . . 146

A.2 Simulation on the rank distributions . . . 161

A.3 The squared norm as a depth function . . . 163

A.4 Additional information surrounding the simulation study in Chapter 3 . . . 165

A.4.1 On the number of directions for the random projection depth . . . 165

A.4.2 If the curves have missing values at random time points . . . 165

A.5 Proofs from Chapter 3 . . . 167

A.6 Proofs from Chapter 4 . . . 175

A.7 Additional simulation results from Chapter 4 . . . 188

A.8 On simplicial depth and privacy . . . 189

A.9 Proofs from Chapter 5 . . . 189

A.9.1 Proofs related to the properties of the depth functions . . . 189

A.9.2 LDP theorem, concentration and sample complexity for the private medians . . . 193

A.9.3 Proofs related to the smoothed dual depth . . . 200

A.9.4 Propose-test-release proofs . . . 202

B Select Topics from Functional Analysis **224**

B.1 Bochner integrals . . . 224

B.2 Linear operators and functionals . . . 226

B.2.1 Operators . . . 226
B.2.2 Adjoint operator . . . 226
B.2.3 Non-negative, square-root and projection operators . . . 228
B.2.4 Operator inverses . . . 229
B.3 Compact operators and singular value decomposition . . . 229
B.3.1 Compact operators . . . 229
B.3.2 Eigenvalues of compact operators . . . 230
B.3.3 Singular value decomposition . . . 232
B.3.4 Hilbert-Schmidt operators . . . 232
B.3.5 Trace class operators . . . 233
B.3.6 Integral operators . . . 234
B.4 What are the observations in the setting of functional data? . . . 235
B.4.1 Probability on a Hilbert space . . . 235
B.4.2 Stochastic process viewpoint . . . 237
B.4.3 Combining the viewpoints . . . 238

List of Tables

3.1 Empirical power of the different tests for $J = 2$, $n = 500$ when the group sample sizes were unequal, under scale differences . . . 66

3.2 Empirical power of the different tests under the finite dimensional models when $n_1 = n_2 = 100$ . . . 68

3.3 Empirical sizes for $J = 2$ for different tests under the infinite dimensional models . . . 69

3.4 Sizes of the tests when one sample was contaminated by the different kinds of outliers . . . 71

3.5 Šidák corrected p-values of pairwise functional Steel tests performed on the centred curves . . . 76

4.1 Change-points and centered MFHD′ rank means. Notice the largest change occurs at the last change-point. . . . 98

4.2 Change-point estimates and p-values resulting from running the FKWC change-point tests on the different subjects. . . . 99

5.1 Table of α_{D} for different depth functions and underlying distributions . . . 116

A.1 Empirical power of the different tests for detecting a shape difference when the group sample sizes were unequal . . . 167

A.2 Empirical power of the different tests for $n = 200$ when the group sample sizes were unequal, under the finite dimensional models. . . . 168

A.3 Empirical powers for the epidemic FKWC test when there was an epidemic-type magnitude change . . . 222

A.4 Empirical powers at the 5% level of significance for the AMOC FKWC test under the functional autoregressive model 223

List of Figures

1.1 Heatmap of depth values of a sample of 20 points 3
1.2 Intraday log returns of SNAP stock for 50 days 10
1.3 A sample of Gaussian processes with a shape outlier 12
1.4 Financial returns for four European stocks 20

2.1 Two samples of 1000 points with different covariance matrices 29
2.2 Empirical root mean squared error of $\widehat{\ell}$ 40
2.3 Empirical root mean squared error of $\widehat{\ell}$ for different values of α under Algorithm 1 41
2.4 Boxplots of $\widehat{\ell} - \ell$ for the different algorithms 43
2.5 Boxplots of $\widehat{k}/n - \theta$ for the different algorithms 44
2.6 Boxplots of $\widehat{\ell} - \ell$ for the third simulation scenario under spatial depth . . . 45
2.7 Empirical root mean square error of $\widehat{\ell}$ as the dimension increases 46
2.8 Change-points estimated by Algorithms 1 and 2 48
2.9 Returns with estimated change-points from Algorithm 2 49

3.1 Two samples of Gaussian processes (a) and their derivatives (b) each with the same mean but a different covariance kernel 57
3.2 Five observations generated from the uncontaminated distribution compared to a drift outlier (left), wavy outlier (middle) and scale outlier (right). . . . 65
3.3 Empirical power curves of the different tests as the β parameter (left) and α parameter (right) of the second sample moves away from the null hypothesis. 67

3.4 Empirical power of the tests for detecting scale differences for $J = 2$ with $n_1 = n_2 = 100$ under the infinite dimensional Gaussian process model such that 5% each of the sample was contaminated 70

3.5 Daily log differenced intraday return curves for `fb` stock, starting on June 24th 2019 and ending March 20th 2019 73

3.6 Covariance kernels $\mathcal{K}(s,t)$ of the residuals 74

3.7 Log periodograms for the syllables 'aa' and 'dcl' 75

4.1 Power curves and accuracy boxplots when there is one change-point in the data 91

4.2 Power curves of the FKWC method compared to other methods 93

4.3 Mean absolute error $|\ell - \hat{\ell}|$ of the simulation scenarios for different values of λ_n' 94

4.4 Energy distance between the estimated and true change-point 94

4.5 Twitter differenced log returns and (b) norms of the Twitter differenced log returns over time 96

4.6 Ranks of the random projection depth values for several f-MRI scans with detected change-point means overlaid. 100

A.1 Normalised histograms of the depth ranks of sample 1 (red) and sample 2 (blue) 163

A.2 Power of the FKWC test paired with the random projection depth to detect scale and shape differences for $J = 2$ with $n_1 = n_2 = 50$, using different amounts of sampled directions 166

A.3 Power of the two sample versions of the tests when $n_1 = n_2 = 100$ when 20% of the curves were uniformly, randomly missing 169

A.4 Comparison of the FKWC methods with the derivatives to the FKWC methods without the derivatives 188

Chapter 1

Introduction

On the real line $\mathbb{R}$, there is a natural ordering of the numbers, which makes it straightforward to define quantiles, order statistics, ranks and centrality. Such concepts underlie much of nonparametric univariate inference, think boxplots, medians, rank tests et cetera. In higher dimensional Euclidean spaces $\mathbb{R}^d$ and in many general Banach spaces, there is no natural linear ordering. As a result, defining quantiles, order statistics, ranks and centrality is inherently more difficult and there is no agreed upon definition. However, extending many existing univariate, nonparametric methods to higher dimensional spaces requires a method of ordering the points in a sample; some way of relating multiple points to each other in the space. This is because nonparametric procedures must rely on the properties of the space and not those of a parametric model. This predicament has led to many different generalisations of order statistics and surrounding concepts to $\mathbb{R}^d$, and, consequentially, to general spaces.

In $\mathbb{R}^d$, data points can be represented as d-dimensional vectors, say $X_i = (x_{i1}, \ldots, x_{id})$. This representation inspires the tendency to use a component-wise approach. For example, to estimate location we could use the component-wise mean and the component-wise median. We can also define component-wise ranks by computing the univariate rank of each point for each variable, giving $R_i = (R_{i1}, \ldots, R_{id})$ and then averaging the ranks over the variables or components (Bickel, 1965). There are documented problems with the component-wise median and component-wise ranks, which can be summarised by two deficiencies. The first of which is a failure to account for the correlation structure of the data. Hypothesis tests based on component-wise averaged ranks can be problematic when several of the variables are highly correlated (Bickel, 1965). Suppose we are computing, say a rank-sum test for a difference in distribution between two groups. If there are numerous, highly correlated components which have no marginal differences then any differences

between remaining components will be washed out by noise. Further, if one component sees consistently high ranks and one low ranks for some group, then these may be washed out in the average rank. Some early works that account for this first issue are (Horrell and Lessig, 1975; Katz and McSweeney, 1980).

The second issue with many component-wise procedures is that they are not affine equivariant/invariant or even similarity equivariant/invariant. If we represent a multivariate sample $\mathbb{X}_n$ with an $n \times d$ matrix, we then say that a d-dimensional statistic $T(\mathbb{X}_n)$ is affine equivariant if

$$T(A\mathbb{X}_n + b) = AT(\mathbb{X}_n) + b,$$

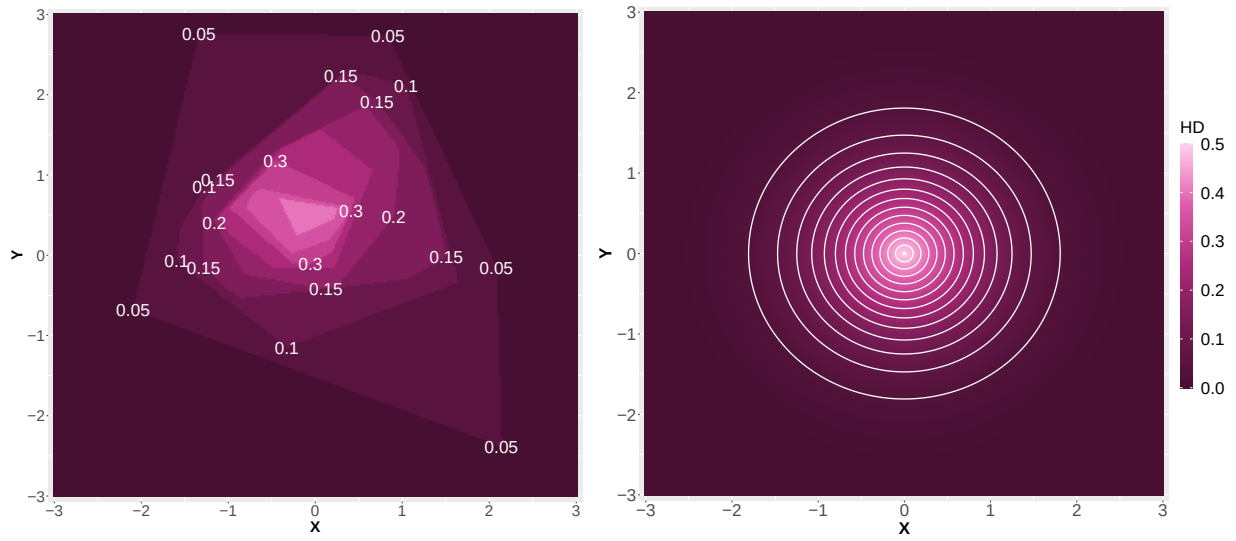
where A is an invertible matrix and b is some column vector. A statistic $T(\mathbb{X}_n)$ is affine invariant if

$$T(A\mathbb{X}_n + b) = T(\mathbb{X}_n),$$

for the same A and b. Similarly, a statistic is similarity equivariant/invariant if the above holds for orthogonal A. In a practical sense, affine equivariance implies that if the data are translated, re-scaled, rotated or reflected, the statistic is also translated, re-scaled, rotated or reflected. Similarly, affine invariance means the statistic is unchanged by such transformations. Affine invariance/equivariance enforces the idea that changing our measurement scales or coordinate system should not affect the inference procedure. Similarity invariance/equivariance is a weaker property and ensures that the statistic is immune to homogeneous scaling (scaling all variables by the same number), translations, rotations and reflections.

The component-wise median is not affine equivariant or even similarity equivariant, nor are the component-wise ranks affine invariant or similarity invariant. The component-wise median has also been shown to fall outside the convex hull of the data (Serfling, 2006). As a result of these flaws, nonparametric procedures that both take the correlation structure into account and are affine invariant/equivariant were then developed for individual testing and estimation problems (see, e.g., Oja, 1983; Randles and Peters, 1990). Such procedures were standalone, and restricted to solving one inference problem, e.g., location estimation. It was later presented that nonparametric description and inference could be performed from within a single framework if the procedures were derived from an underlying *data depth function* (Liu et al., 1999; Serfling, 2006).

A data depth function gives meaning to centrality, order and outlyingness in spaces beyond $\mathbb{R}$. Data depth functions do this by giving all points in the working space a rating based on how central the point is in the sample. For example, restricting to $\mathbb{R}^d$, multivariate depth functions can be written as $\mathrm{D}\colon \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}^+$; given a sample and a point in the domain, the depth function assigns a real valued depth to that point. Figure 1.1(a) shows



(a) (b)

Figure 1.1: (a) Sample halfspace depth values, i.e., $\mathrm{D}(X_i; F_n)$, are displayed in white text. The heatmap of the sample depth function, i.e., $\mathrm{D}(\cdot; F_n)$, is also displayed. This sample is drawn from a standard, two dimensional normal distribution. (b) Theoretical halfspace depth contours for the standard, two dimensional normal distribution.

a sample of 20 points labelled by their depth values, we can see that the points in the centre of the data cloud have larger values. Note that it is not necessary to restrict the domain of the depth function to points in the sample; we can compute depth values for each point in the sample space. The heatmap in Figure 1.1(a) gives the depth value for each point in the plot.

The concept of depth predates its extensive study by two decades, first studied by Tukey (1974), where his goal was to picture data in a way such that we can learn from it, rather than imposing what we suspect to be true on the data and creating an affirmative picture. Tukey recognized the usefulness of order statistics, and what they tell an analyst about univariate data. However, he also commented that direct generalisations of order statistics 'fail miserably' in the multivariate setting. He instead represented the information carried by order statistics in the univariate setting with a picture in the bivariate setting, using the concept of depth.

How is Figure 1.1(a) akin to univariate order statistics? First, we know that points in the lightest, triangle shaped region are very central, and can serve as medians or near

medians of the data. Further, the shape of the regions tells us that the data is spread somewhat symmetrically about the centre. This is akin to the distance of certain univariate quantiles from the centre. Lastly, we can see the observations which are outlying: those observations which have low depth. Obviously, we cannot create such a picture beyond 3-dimensions, however, we shall see that the principle ideas laid out by Tukey (1974) underlie many depth functions and their associated data visualization and inference procedures.

1.1 Multivariate depth functions

So far, the depth of a point has been taken to be with respect to some set of sample of points $\mathbb{X}_n \in \mathcal{X}$. We will now identify samples with empirical measures F_n and the depth of a point will be reinterpreted as depth with respect to a measure rather than a sample. This formulation provides a natural definition for population values of depth $\mathrm{D}(\cdot; F)$ and makes it easy to write down mathematically. If we let $\mathcal{F}$ be the set of distributions on $\mathbb{R}^d$ we can write multivariate depth functions as $\mathrm{D}\colon \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}^+$. Figure 1.1(b) shows the theoretical halfspace depth contours $\mathrm{D}(\cdot; F)$ when $F = \mathcal{N}_2(0, I)$; F is the bivariate standard normal distribution. Notice the shape is circular like that of the normal distribution, and that the depth is largest in the centre.

We have not discussed how depth functions measure centrality. To do this, we must present a list of precise mathematical properties that, when possessed, ensure a function does indeed measure centrality of a point with respect to a distribution. In their seminal paper Zuo and Serfling (2000a) give a concrete set of mathematical properties which a multivariate depth function should satisfy in order to be considered a *statistical depth function*. These properties include:

1. Affine invariance: This implies any depth based analysis is independent of the co-ordinate system. Particularly, the analysis is independent of the scale on which the data is measured.

2. Maximality at centres of symmetry: If a distribution is symmetric about a point, then surely this point should be regarded as the most central point.

3. Decreasing along rays: This property ensures that as one moves away from the deepest point, the depth decreases.

4. Vanishing at infinity: As a point moves toward infinity along some ray, the depth vanishes.

A function $\mathrm{D}\colon \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}^+$ which satisfies these four properties is known as a statistical depth function. The last three properties are all related to centrality, where the first is to ensure there is no dependence on the measurement system. Not all popular depth functions satisfy all four of these properties, but they typically satisfy most of them. We have already discussed affine invariance, let us discuss each of the remaining properties in detail. Maximality at centre means that if a distribution is symmetric about some point θ, the depth is maximal at that point. Think of the median coinciding with the mean in the univariate case. There are multiple definitions of symmetry in the multivariate setting (see Zuo and Serfling, 2000b, for more details.) Decreasing along rays means that as one moves away from the deepest point along some ray, i.e., moves away from the centre, the depth decreases. This property can be replaced by quasi-concavity. Suppose that $\alpha \in (0, 1)$ and $x, y \in \mathcal{X}$. Then, a depth function is quasi-concave if

$$\mathrm{D}(\alpha x + (1-\alpha)y; F) \geq \mathrm{D}(x; F) \wedge \mathrm{D}(y; F),$$

where $x \wedge y = \min\{x, y\}$ for two real numbers x and y. It is clear that quasi-concavity implies the decreasing along rays property; suppose that $\mathrm{D}(\theta; F) = \sup_x \mathrm{D}(x; F)$, then:

$$\mathrm{D}(\theta; F) \geq \mathrm{D}(\alpha\theta + (1-\alpha)x; F) \geq \mathrm{D}(x; F).$$

Vanishing at infinity means that as the point moves along a ray to infinity, its depth approaches zero. If all four of these properties are not satisfied, it does not necessarily mean that a depth function is invalid or not useful in data analysis. It is merely a limitation to consider.

Aside from coordinate invariance and centrality, there are other properties that are desirable for a depth function to satisfy:

- Robustness: A robust depth function implies subsequent inference will be robust. For example, robust depth functions generally lead to high breakdown multivariate medians (Chen and Tyler, 2002; Zuo, 2003).
- Consistency/Limiting Distribution: Consistency of the sample depth values for population depth values and existence of a limiting distribution of the sample depth values is useful for developing hypothesis tests and confidence intervals.
- Continuity: Continuity can be a building block for asymptotic inference and for optimizing the depth function.
- Computation: Obviously, we would want the function to be quickly computable.

From the list above, it is certainly not surprising that there are many depth functions; there are many desirable properties, making it difficult to cover them all. We now introduce several depth functions, the first of which being the halfspace depth (Tukey, 1974; Rousseeuw1 and Ruts2, 1999).

Definition 1 (Halfspace depth). *Let* $S^{d-1} = \{u \in \mathbb{R}^d : \ \|u\| = 1\}$ *be the set of unit vectors in* $\mathbb{R}^d$. *Define the halfspace depth* HD *of a point* $x \in \mathbb{R}^d$ *with respect to some distribution* $X \sim F$ *as,*

$$\mathrm{HD}(x; F) = \inf_{u \in S^{d-1}} \Pr(X^\top u \leq x^\top u) = \inf_{u \in S^{d-1}} F_u(x),$$

where F_u *is the distribution of* $X^\top u$.

Halfspace depth is the minimum of the projected mass above and below the projection of x, over all univariate projections. Halfspace depth satisfies all four properties corresponding to a statistical depth function, including the stronger version of decreasing along rays: quasi-concavity. The halfspace sample depths are uniformly consistent and, in some cases outlined in (Massé, 2004), these sample depth values each weakly converge to some non-normal random variable. Halfspace depth is an upper semi-continuous function which identifies the distribution in certain cases (but not always) (Zuo and Serfling, 2000a; Nagy, 2018). However, halfspace depth is frequently cited as being computationally complex (Serfling, 2006), especially for moderate dimensions. Recently an algorithm for computing halfspace depth in high dimensions has been proposed (Zuo, 2019) that may aid in this regard.

We can replace the infimum in Definition 1 with an average (Ramsay et al., 2019).

Definition 2 (IRW Depth). *Define integrated rank-weighted depth as*

$$\mathrm{IRW}(x; F) = \int_{S^{d-1}} \Pr(X^\top u \leq x^\top u) \wedge (1 - \Pr(X^\top u \leq x^\top u)) \ d\nu(u),$$

where ν *denotes the Haar measure.*

IRW depth vanishes at infinity and is maximal at points of symmetry of F. It is invariant under similarity transformations, which is a weaker than affine invariance. It is conjectured that this function also has the decreasing along rays property. This depth function does not have as many desirable depth-like properties as halfspace depth, but instead it can be approximately computed quickly, even in high dimensions, with statistical guarantees. This depth function also has a natural extension to functional spaces (Fraiman and Muniz, 2001), which will be covered later. IRW sample depths are also uniformly

consistent and asymptotically normal, under mild assumptions and this depth function is continuous except at atoms of the distribution (Ramsay et al., 2019). We can replace the

$$\Pr(X^\top u \leq x^\top u) \wedge (1 - \Pr(X^\top u \leq x^\top u))$$

in Definition 2 with

$$\Pr\left(X^\top u \leq x^\top u\right)\left(1 - \Pr\left(X^\top u < x^\top u\right)\right).$$

This is integrated dual depth, which we denote by IDD. Integrated dual depth which has many of the same properties of IRW depth and was introduced by Cuevas and Fraiman (2009).

After halfspace depth, simplicial depth was introduced (Liu, 1988).

Definition 3 (Simplicial Depth)**.** *Suppose that $X, Y_1, \ldots, Y_{d+1}$ are independent, $\mathbb{R}^d$-valued random variables, each having distribution F. Define simplicial depth as*

$$\mathrm{SMD}(x; F) = \Pr(X \in \mathrm{Sim}(Y_1, \ldots, Y_{d+1})),$$

where $\mathrm{Sim}(Y_1, \ldots, Y_{d+1})$ is the simplex with vertices $Y_1, \ldots, Y_{d+1}$.

The sample simplicial depth values are asymptotically normal under some conditions, which can make inference based on simplicial depth values easier to work with (Liu, 1988). Simplicial depth is, however, difficult to compute in even moderate $d > 3$ dimensions. This depth function is a statistical depth function if F is angularly symmetric (see Zuo and Serfling, 2000b), but fails to satisfy the maximality at centre and decreasing along rays for some discrete distributions. Seeing as we are typically concerned with continuous distributions in a depth-based inference context, this is not a major issue (Liu, 1988; Zuo and Serfling, 2000a).

The investigation of depth functions by Zuo and Serfling (2000a) lead to the study of a general and powerful statistical depth function based on outlyingness functions. An outlyingness function $O\colon \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}$ measures the degree of outlyingness of a point. Intuitively, we can then define a depth as

$$\mathrm{D}(x; F) = \frac{1}{1 + O(x; F)},$$

which is referred to as a Type C depth by Zuo and Serfling (2000a). A particular version of a Type C depth function is projection depth.

Definition 4 (Projection Depth)**.** *Given a univariate, translation equivariant and scale equivariant, location measure ϱ and a univariate, scale equivariant and translation invariant measure of scale σ, we can define projected outlyingness as*

$$O(x; F; \varrho, \sigma) = \sup_{u \in S^{d-1}} \frac{\left|u^\top x - \varrho(F_u)\right|}{\sigma(F_u)}$$

and thus, projection depth as,

$$\mathrm{PD}(x; F; \varrho, \sigma) = \frac{1}{1 + O(x; F; \varrho, \sigma)}.$$

Typically, ϱ and σ refer to the median and median absolute deviation, but most properties of projection depth have been investigated for general ϱ and σ. In this work, we will typically use

$$O_1(x; F_n) = \sup_{\|u\|=1} \frac{\left|u^\top x - \mathrm{MED}\left(\mathbb{X}_n^\top u\right)\right|}{\mathrm{MAD}\left(\mathbb{X}_n^\top u\right)} \qquad \text{or} \qquad O_2(x; F_n) = \sup_{\|u\|=1} \frac{\left|u^\top x - \mathrm{MED}\left(\mathbb{X}_n^\top u\right)\right|}{\mathrm{IQR}\left(\mathbb{X}_n^\top u\right)},$$

where we use a slight abuse of notation and let $\mathbb{X}_n^\top u$ refers to the sample $\{X_1^\top u, \ldots, X_n^\top u\}$.

Projection depth was discussed in various forms by Stahel (1981); Donoho (1982); Liu (1992); Zuo and Serfling (2000a) but a thorough investigation of the properties of projection depth was done in the successive papers Zuo (2003, 2004). As a result of these papers, it has been shown that projection depth is a statistical depth function, is quasi-concave and is Lipschitz continuous under very mild conditions on σ and ϱ. Computation is again difficult; there are approximation algorithms for moderate to large dimensions, but they do not have any statistical guarantees (Liu, 2017; Dyckerhoff et al., 2021; Shao et al., 2022).

Another popular depth function is spatial depth (Serfling, 2002), see also (Small, 1990) for more background on the spatial median and spatial depth. Let $u \in S^{d-1}$, where S^{d-1} is as defined in Definition 1. Spatial depth is based on spatial quantiles:

$$\mathcal{Q}(u; F) = \min_{y \in \mathbb{R}^d} \mathrm{E}\left[\|X - y\| + \langle u, X - y\rangle - \|X\| - \langle u, X\rangle\right]. \tag{1.1}$$

Spatial quantiles are extensions of univariate quantiles. Inverting this function at a point $x \in \mathbb{R}^d$ gives a measure of outlyingness: $\|\mathcal{Q}^{-1}(x; F)\|$ (Serfling, 2002). Let

$$S(x) = \begin{cases} \frac{x}{\|x\|} & x \neq 0 \\ 0 & x = 0 \end{cases}$$

and then define

$$\left\|\mathcal{Q}^{-1}(x;F)\right\| = \left\|\mathrm{E}\left[S(x-X)\right]\right\|.$$

We can now define spatial depth.

Definition 5 (Spatial Depth)**.** *Define the spatial depth of a point $x \in \mathbb{R}^d$ with respect to some distribution $X \sim F$ as,*

$$\mathrm{SD}(x;F) = 1 - \left\|\mathcal{Q}^{-1}(x;F)\right\|. \tag{1.2}$$

Spatial depth satisfies three of the above four properties required for a depth function to be classified as a statistical depth function. Spatial depth is invariant under similarity transformations but is not invariant under all affine transformations. One useful feature of spatial depth is that it has a natural extension to Banach spaces, since it's definition is based on a norm. In addition, the norm-based definition makes it easier to compute.

Related to spatial depth is Mahalanobis depth. Instead of using the Euclidean norm in Definition 5, we can obtain affine invariance by using $\|x\|_\Sigma = \sqrt{x^\top \Sigma^{-1} x}$, where Σ is the covariance matrix related to F.

Definition 6 (Mahalanobis Depth)**.** *Define the Mahalanobis depth* MH *of a point $x \in \mathbb{R}^d$ with respect to some distribution F as,*

$$\mathrm{MH}(x;F) = \frac{1}{1 + \|x - \mathrm{E}\left[X\right]\|_\Sigma^2}. \tag{1.3}$$

Mahalanobis depth satisfies all four properties required to be a statistical depth function. One criticism of Mahalanobis depth is that Σ and $\mathrm{E}\left[X\right]$ are usually replaced by estimators which are not robust, such as the sample covariance matrix and sample mean, respectively. In order for the Mahalanobis depth function to remain robust, it is necessary to use robust estimators of Σ and $\mathrm{E}\left[X\right]$. Examples of such estimators are the re-weighted MCD estimators studied by Rousseeuw and van Zomeren (1990). We denote the depth values computed using these MCD estimators by MH75, where the 75% comes from the fact that we are using the 25% breakdown version of the MCD estimators. There exists other multivariate depth functions, not covered here, but we have tried to cover ones that are popular and are useful for inference procedures proposed in this thesis.

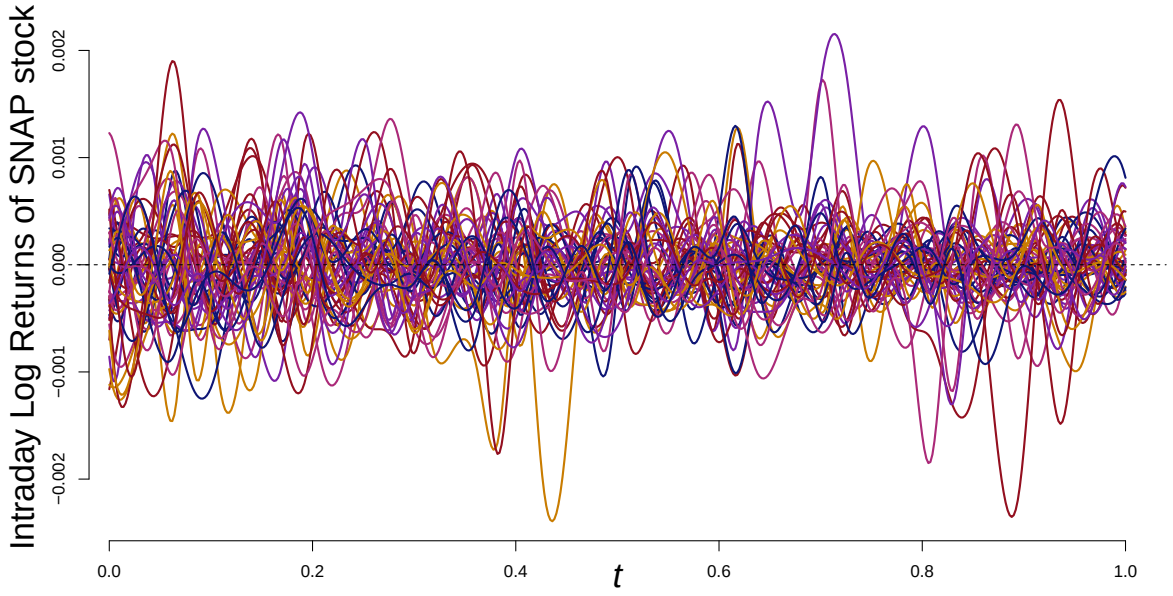


Figure 1.2: Intraday log returns of SNAP stock for 50 days. Data were smoothed to fit 50 basis functions and each edge was trimmed by 10%.

1.2 Functional depth functions

Figure 1.2 shows 50 curves, each curve is one trading day of logarithmic returns per minute for SNAP stock, these curves are a sample of *functional data*. The analysis of functional data is one of the most active research areas developed in recent decades (Ullah and Finch, 2013; Wang et al., 2016). In this setting, the observed data are not vectors but smooth functions on some domain. Generally, functional data comes in vector form or discretized form, and so it can appear multivariate at first glance. The difference from multivariate data is that there is reason to assume there is some underlying smoothness relating the columns of the vectors; there is reason to assume we have observed pieces of a function rather than separate variates. Incorporating the smoothness condition often leads to improved inference on such data. Observed functions are often assumed to have domain [0,1], we adopt that assumption as well, but the domain can be multivariate as well. For example, an image can be viewed as a function on a bivariate domain.

The analysis of functional data can be involved; the data generally have to be smoothed

(Ramsay and Silverman, 2006) and sometimes aligned (Srivastava et al., 2011) before an inference procedure is applied. There is a large amount of literature simply regarding the smoothing and aligning process, (see e.g., Ramsay and Silverman, 2006; Srivastava et al., 2011). After prepossessing, inference procedures then often require at least an elementary understanding of functional analysis and stochastic processes, which makes such procedures not as easily accessible as univariate methods. For example, functional principle component analysis (Dauxois et al., 1982), a central technique in functional data analysis, requires some understanding of the Karhunen–Loeve expansion; spectral decomposition of functions. To aid the reader, we provide a brief introduction to some of these concepts in Appendix B. In addition, there is lack of well-known parametric models for functional data outside of Gaussian process models. This has lead to the wide-spread development and application of nonparametric inference procedures for functional data (Wang et al., 2016). As a result, in the past two decades, depth functions have been extensively developed for functional data.

We can define depth in general spaces as follows: Let $\mathcal{F}$ be the set of measures on some space $\mathfrak{F}$, then, we can write depth functions generally as $\mathrm{D}\colon \mathfrak{F} \times \mathcal{F} \to \mathbb{R}^+$, where larger values of D imply that the point $x \in \mathfrak{F}$ is deep with respect to some $F \in \mathcal{F}$. In the setting of functional data, typically $\mathfrak{F}$ is a subset of one of two spaces. First, $\mathfrak{F}$ may be a subset of $\mathscr{L}^2([0,1], \mathscr{B}, \mu)$, the space of all Borel measurable, square integrable functions on the interval [0,1]. We denote the Borel sets over $[0,1]$ by $\mathscr{B}$ and the Lebesgue measure over $[0,1]$ by μ. We may write $\mathscr{L}^2([0,1], \mathscr{B}, \mu)$ as $\mathscr{L}^2$ for short when the context is clear. $\mathcal{F}$ would then be the set of measures on $\mathscr{L}^2$. The other popular choice for the working space $\mathfrak{F}$ is the space of continuous functions $\mathscr{C}([0,1])$. One may also interpret functional data as a continuous time stochastic processes indexed by $t \in [0,1]$. We typically assume conditions on the observations such that they may be interpreted equivalently as random draws from $\mathscr{L}^2$ or as stochastic processes characterized by their finite dimensional distributions $X(t)$, see Section 2.3.

The space $\mathfrak{F}$ could also be a Cartesian product of function spaces $\mathscr{L}^2$ or $\mathscr{C}$, which would mean that the observations are vectors of functions. Such data are known as multivariate functional data, we will see that many functional depths are defined for such data. To avoid confusion, we use $\mathfrak{F}^p$ to represent the space of multivariate functional observations, that is p-dimensional vectors of $\mathfrak{F}$-valued functions. We will reserve $\mathfrak{F}$ for representing a function space. Note that multivariate functional data encompasses the case where $p = 1$.

The interpretation of depth in the functional setting differs from that of the multivariate setting. In earlier work on functional depth measures, for univariate functions, a function would be 'deep' if that function was 'surrounded' by many other functions contained in the sample (Fraiman and Muniz, 2001; López-Pintado and Romo, 2009). The definition of

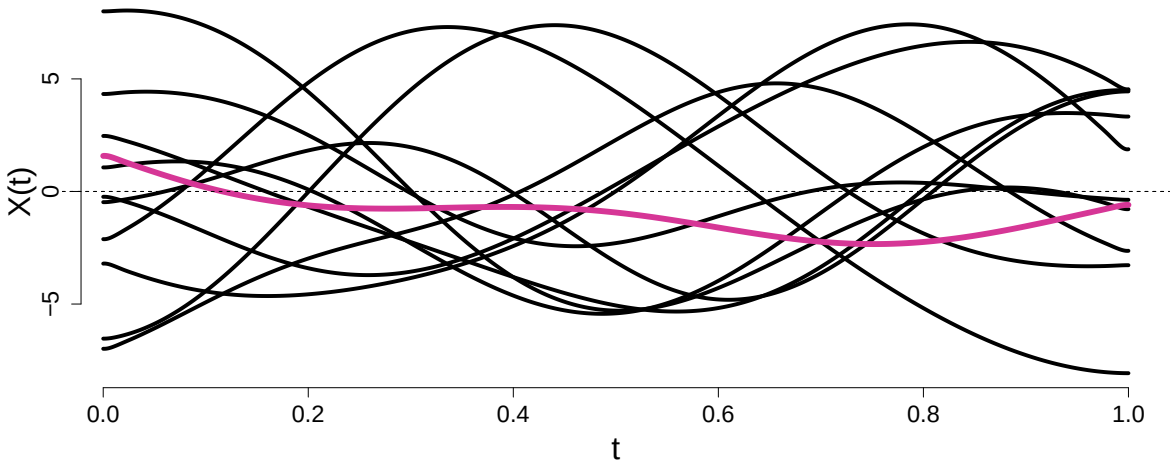


Figure 1.3: A sample of Gaussian processes with a shape outlier highlighted in pink. Notice that the outlier is deep in terms of magnitude but the shape is different from the rest of the sample.

depth in the functional context evolved to include the shape of the function as well (Hubert et al., 2015; Serfling and Wijesuriya, 2017; Dai and Genton, 2018; Harris et al., 2020). In other words, a function is considered 'deep' if it is both surrounded by many functions in the sample and similar in shape to many functions in the sample. Figure 1.3 shows an example of a shape outlier. Notice that the outlier is deep in terms of magnitude but the shape is different from the rest of the sample. Such an observation may have a low depth value due to its shape.

Beyond the meaning of deep, there are other differences between the functional setting and the multivariate setting. Functional depth functions have different transformation invariance requirements (Gijbels and Nagy, 2017; Serfling and Wijesuriya, 2017). We can characterize transformation invariance properties as follows,

$$\mathrm{D}(X; F) = \mathrm{D}(g(X); g(F)), \qquad \text{for } g \in \mathscr{G}, \tag{1.4}$$

where $\mathscr{G}$ is a class of function mappings g from $\mathfrak{F} \to \mathfrak{F}$ and $g(F)$ can be interpreted as the distribution of $g(Y)$ if $Y \sim F$. Note that $g(F_n)$ results in the empirical measure of g applied to the random sample of size n. We state two types of invariance, following the

notation of Gijbels and Nagy (2017), the first of which is scalar invariance

$$\text{P1-S: } \mathscr{G} = \{g\colon g(X) = aX + b,\ a \in \mathbb{R},\ b \in \mathfrak{F}\},$$

and the second is affine invariance in the sense that

$$\text{P1-F: } \mathscr{G} = \{g\colon g(X) = aX + b,\ a, b \in \mathfrak{F},\ a(t) \neq 0,\ t \in [0,1],\ aX \in \mathfrak{F}\}.$$

The need for translation invariance is obvious, but it is not obvious that P1-F is necessary as opposed to only P1-S. The scale invariance property P1-S has the same motivation as that of affine invariance in the multivariate setting; we would like the inference to be unaffected by changes in measurement units. It is difficult to understand the motivation for invariance under scaling all sample functions by an arbitrary function. Since the columns of the observed vectors are thought to be smoothly related, it may seem strange to have the depth function invariant with respect to some arbitrary, heterogeneous column scaling, in fact, Serfling and Wijesuriya (2017) argues this is not desirable. It could be desirable to have the depth function be unaffected by reflections and rotations of the sample, neither of which are covered by P1-F. For many existing functional depth functions, it is trivial to show that they are invariant under reflections of the data, so this is not much of a concern. Rotational invariance, however, could be added, as its own property. We can also define an equivalent property of P1-F for multivariate functional data, that is, for $X \in \mathfrak{F}^p$, affine invariance could mean

$$\text{P1-Fb: } \mathscr{G} = \{g\colon g(X) = AX + b,\ AX + b \in \mathfrak{F}^p,\ A(t) \text{ is invertible } \forall t \in [0,1]\}.$$

A popular method is to incorporate the derivatives of the observations[1] into the analysis by defining a new depth function as follows: For some differentiable function x, consider the pair $(x, x^{(1)})$. Take the depth of x to be the multivariate functional depth of $(x, x^{(1)})$ with respect to the sample $\{(X_1, X_1^{(1)}), \ldots, (X_n, X_n^{(1)})\}$. This method often provides substantially better results (Cuevas et al., 2007; Hubert et al., 2012), one can also incorporate higher orders of derivatives or other functions related to the observations. However, computing depth this way creates a transformation invariance issue; this new depth function will not satisfy P1-F, even if the multivariate functional depth used satisfies P1-Fb. Indeed, scaling each X_i by $a \in \mathfrak{F}$ results in a different transformation of the derivative and

[1]Of course, the observations are then assumed to be differentiable, which technically modifies the assumed space $\mathfrak{F}$.

produces different observations

$$(aX_i, aX_i^{(1)} + a^{(1)}X_i),$$

which would not give the same depth values. However, in order for $X_i^{(1)}$ to be affected overly by arbitrary scaling, the derivatives of a must be large relative to function X_i, suggesting the scaling function a is quite steep in some places. It is hard to see a reason to scale the data by steep functions in practice. If a is a constant function, then scaling the data corresponds to scaling the derivative and all is well; P1-S will still be satisfied.

As seen by the discussion on transformation invariance, it is not straightforward to provide an analogue of Zuo and Serfling (2000a) for the functional setting. There have been some properties, which were first outlined by Nieto-Reyes and Battey (2016) and then later expanded and refined in Gijbels and Nagy (2017). Some desirable properties have also been discussed by Serfling and Wijesuriya (2017). Summarizing all such properties is somewhat complex and so instead we can list a set of criteria we deem suitable for using functional depth measures in a hypothesis testing problem, which is conducive to the work in Chapters 3 and 4.

Namely, we would like the depth measures to admit definitions on multivariate functional data. Choosing depth measures that admit definitions for multivariate functional data allows for the use of derivatives, warping functions or other functions in the depth computation.

It is also desirable for the depth measures to be uniformly consistent with a rate of $O(n^{-1/2})$ under some general conditions, which gives that for large n. So far, this has only been shown for the depth of Fraiman and Muniz (2001) by Nagy and Ferraty (2019). We extend their result to the random projection depth and a special case of the multivariate halfspace depth in Chapter 3 and Chapter 4. Uniform consistency allows for asymptotic critical values for test statistics based on depth values and eliminates the computational burden of re-sampling methods. Lastly, we use functional data depth measures which have implementations in `R` or software that can be integrated into `R`. This last restriction allows us to readily provide simulation results for the hypothesis tests, as well as implementations that can be used in practice.

We first discuss multivariate functional halfspace depth (Hubert et al., 2012; Claeskens et al., 2014), which was introduced in the thesis by Slaets (2011). This depth is defined for multivariate functional data, so we can assume that each observation can be represented as $X_i : [0, 1] \to \mathbb{R}^p$. To compute the multivariate functional halfspace depth of $x : [0, 1] \to \mathbb{R}^p$ we need to compute the pointwise multivariate halfspace depth (Definition 1) of each $x(t)$ with respect to $F_{t,n}$, the empirical distribution of $\{X_i(t), \ldots, X_n(t)\}$. For each t,

$\mathrm{HD}(x(t); F_{t,n})$ then gives one depth value. The pointwise depth values $\mathrm{HD}(x(t); F_{t,n})$ are then combined by computing a weighted average of them, where the weight function down-weights regions where all curves are similar in amplitude.

Definition 7 (Multivariate Functional Halfspace Depth)**.** *Let $\alpha \in [0, 1/2)$, F_t be the distribution of $X(t)$ if $X \sim F$ and $\mathrm{HD}_\alpha(F_t) = \{y \in \mathbb{R}^p \colon \mathrm{HD}(y; F_t) \geq \alpha\}$. Then, multivariate functional halfspace depth is defined as*

$$\mathrm{MFHD}(x; \alpha; F) = \int w_\alpha(t)\, \mathrm{HD}(x(t); F_t) dt, \text{ where } w_\alpha(t) = \frac{\mathrm{vol}(\mathrm{HD}_\alpha(F_t))}{\int \mathrm{vol}\left(\mathrm{HD}_\alpha(F_t)\right) dt}.$$

Here, α is a parameter that controls the degree of 'downweighting'. If we choose $\alpha = 0$ and $p = 1$, multivariate functional halfspace depth reduces to the functional depth measure of Fraiman and Muniz (2001). Multivariate functional halfspace depth is invariant under transformations of the type P1-Fb. The sample depth values of this depth measure are also uniformly almost sure consistent under fairly mild conditions (see Claeskens et al., 2014). We can use the `R` package `MFHD` or the `fda.usc` package via `depth.FMp` to compute the MFHD sample depth values. We opt for the latter in our simulations in Section 3.4. We use $\alpha = 0$ throughout this thesis and thus suppress the notation to $\mathrm{MFHD}(x; F)$.

We next consider modified band depth. For some $x \in \mathfrak{F}$ and a set of functional observations $Y_1, \ldots, Y_j$ we can define

$$B\left(x; Y_1, \ldots, Y_j\right) = \left\{ t \in [0, 1] : \min_{r=1,\ldots,j} Y_{i_r}(t) \leq x(t) \leq \max_{r=1,\ldots,j} Y_{i_r}(t) \right\},$$

as the set such that x is in the j-band delimited by $Y_1, \ldots, Y_j$.

Definition 8 (Modified band depth)**.** *If $Y_1, \ldots, Y_j$ are independent and come from the distribution F, we can define*

$$\mathrm{MBD}_j^{(x)} = \mathrm{E}_F\left[\mu_{\mathbb{R}}(B\left(x; Y_1, \ldots, Y_j\right))\right]$$

where $\mu_{\mathbb{R}}$ is the standard Lebesgue measure on $\mathbb{R}$. Then the 'modified' band depth with parameter J is equal to

$$\mathrm{MBD}_J(x; F) = \sum_{j=2}^{J} \mathrm{MBD}_j^{(x)}.$$

This depth measure is invariant under transformations described by P1-F. The sample modified band depths are uniformly consistent under very mild conditions (see Gijbels

and Nagy, 2015, 2017). There are two multivariate functional extensions of this depth measure, Ieva and Paganoni (2013) and Lopez-Pintado et al. (2014). We use that of Ieva and Paganoni (2013) because of its existing implementation in `R`. This multivariate extension of modified band depth is defined as

$$\mathrm{MMBD}_J(x;F) = \sum_{k=1}^{p} w_k \,\mathrm{MBD}_J(x;F^k), \qquad \text{and} \qquad \sum_{k=1}^{p} w_k = 1,$$

where F^p is the marginal distribution pertaining to the p^{th} univariate functional argument in the vector of observations.

Functional spatial depth, described in Chakraborty and Chaudhuri (2014), is the infinite dimensional extension of multivariate spatial depth. Define spatial depth to be

$$\mathrm{SD}(x;F) = 1 - \left\|\mathrm{E}\left[s(x-X)\right]\right\|,$$

where

$$s(y) = \begin{cases} y/\left\|y\right\| & \left\|y\right\| \neq 0 \\ 0 & o.w. \end{cases}.$$

Here, $\left\|\cdot\right\|$ refers to the $\mathscr{L}^2$ norm, but this definition is valid on any Banach space. An interesting extension of spatial depth is the kernelized spatial depth (Sguera et al., 2014). Consider a mapping into a feature space $\varphi\colon \mathscr{L}^2 \to \mathbb{F}$ and some kernel function on $\mathscr{L}^2$, say $\gamma\colon \mathscr{L}^2 \times \mathscr{L}^2 \to \mathbb{R}$ defined as

$$\gamma(x,z) = \langle \varphi(x), \varphi(z) \rangle.$$

We can now define the kernel depth as

$$\mathrm{KSD}(x;F) = 1 - \left\|\mathrm{E}\left[s(\varphi(x)-\varphi(X))\right]\right\|.$$

Spatial depth and kernelized spatial depth are invariant as in P1-F if a is surjective, which is a fairly mild restriction. The sample versions of these depth measures are also uniformly consistent, (see Gijbels and Nagy, 2015, and the references therein). We can extend these depth measures to the multivariate functional setting by computing the spatial functional depth values marginally and then taking an equally weighted average of such marginal depth values; analogous to what is done above for the modified band depth (Ieva and Paganoni, 2013). One might also take a multivariate depth of the marginal functional depths, but this is more computationally expensive. Both kernelized and standard spatial functional depths are implemented in the `R` package `fda.usc`.

The last of the functional depth measures we discuss is the random projection depth

of Cuevas et al. (2007). The idea behind this depth measure is to choose many, say M, random unit functions according to some valid measure P on the unit sphere S in $\mathscr{L}^2$. Then, for each direction u_m, compute a separate depth value based on the projections onto u_m, i.e., $\mathrm{D}(\langle x, u_m\rangle; F_{u_m})$. Recall that F_{u_m} is the CDF of the random variable $\langle X, u_m\rangle$ where $X \sim F$. The depth values $\mathrm{D}(\langle x, u\rangle; F_u)$ are then averaged to give a final depth value, viz.

$$\mathrm{RP}_M(x; F) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{D}(\langle x, u_m\rangle; F_{u_m}) \approx \int_S \mathrm{D}(\langle x, u\rangle; F_u)\ dP(u). \tag{1.5}$$

We will denote the right hand term as simply RP. In this work, we take $\mathrm{D}(\langle x, u\rangle; F_u) = F_u(x)(1 - F_u(x))$, use $M = 20$ projections and $u_1, \ldots, u_M$ are Gaussian processes with exponential variogram $\gamma(s, t) = \exp(-5|s-t|)$, standardized such that they have unit norm. Cuevas et al. (2007) introduce a second version in which they calculate both the projection of the function and the projection of the function's first derivative, which provides pairs of observations. They then use a multivariate depth on the couples:

$$\mathrm{RP}'_M(x; F) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{D}((\langle x, u_m\rangle, \langle x^{(1)}, u_m\rangle); F_{u_m,(1)}) \approx \int_S \mathrm{D}((\langle x, u\rangle, \langle x^{(1)}, u\rangle); F_{u,(1)})\ dP(u),$$

where $F_{u,(1)}$ is the bivariate distribution of $(\langle X, u\rangle, \langle X^{(1)}, u\rangle)$ if $X^{(1)}$ is the first derivative of $X \sim F$. For $\mathrm{D}(\cdot)$ we take the likelihood depth (Müller, 2005). Note that this depth does not have any consistency or invariance properties presented in the original paper. However, the random projection depth is a version of integrated dual depth (Cuevas and Fraiman, 2009) for which consistency and invariance are established. The `R` function to compute the sample depth values can be found in the `fda.usc` package.

In Chapter 3 and Chapter 4 we use depth values computed on the pair $(x, x^{(1)})$. We take the depth of x to be the multivariate functional depth of $(x, x^{(1)})$ with respect to the sample $\{(X_1, X_1^{(1)}), \ldots, (X_n, X_n^{(1)})\}$. When we used the derivative of the function in the depth computation, we will denote the depth by D'. For example, when computing the multivariate functional halfspace depth with the derivative, we denote the depth by MFHD$'$.

1.3 Inference based on depth functions

Precisely how can we use these depth functions to gain insights from data? Depth functions provide definitions of order statistics because observations can be ordered by their

depth values. This idea provides immediate analogues of the aforementioned visualization, estimation and hypothesis testing procedures. Since the ordering is center outward, the definitions are modified. For example, the definition of the depth-based median is:

$$\mathrm{MED}(F) = \underset{x \in \mathfrak{F}}{\operatorname{argmax}}\, \mathrm{D}(x; F).$$

These medians are usually robust, much more so than the sample mean vector (Fraiman and Muniz, 2001; Chen and Tyler, 2002; Zuo, 2004). Furthermore, they inherent any transformation invariance properties possessed by the depth function. We can subsequently define sample depth ranks as

$$\widehat{R}_i = \#\{X_j \colon \mathrm{D}(X_j; F_n) \leq \mathrm{D}(X_i; F_n) \text{ for } j \in \{1, \ldots, n\}\}, \qquad i \in \{1, \ldots, n\}. \tag{1.6}$$

Depth-based ranks are a building block of various multivariate and functional rank tests (Liu and Singh, 1993; Serfling, 2002; López-Pintado and Romo, 2009; Chenouri et al., 2011). Ranks also give a means with which to construct multivariate, trimmed means (Fraiman and Muniz, 2001; Zuo, 2002).

The depth values can also be directly used in testing procedures (Li and Liu, 2004; López-Pintado and Wrobel, 2017; Flores et al., 2018). We have also seen depth used for functional boxplots (Serfling and Wijesuriya, 2017) and multivariate bagplots (Rousseeuw et al., 1999). In the same vein, we can visualise multivariate distributions through one dimensional curves based on depth values (Liu et al., 1999). Such curves describe scale, kurtosis, skew and more. In summary, depth functions facilitate a framework for robust, nonparametric inference in $\mathfrak{F}$-space. In fact, in the past decade this depth-based inference framework has expanded to include solutions to clustering (Jörnsten, 2004; Jeong et al., 2016; Baidari and Patil, 2019), classification (Jörnsten, 2004; Cuevas et al., 2007; Sguera et al., 2014; Hubert et al., 2017), outlier detection (Hubert et al., 2015; Sguera et al., 2016; Kuhnt and Rehage, 2016), change-point and process monitoring problems (Liu, 1995; Chenouri et al., 2020b) and discriminant analysis (Chakraborti and Graham, 2019). We further expand this framework in this thesis.

1.4 Contributions

Starting in Chapter 2, we develop a method for detecting multiple change-points in the variability of a series of multivariate observations. In change-point problems, the goal is to detect sudden changes in the underlying distribution of the data; the idea is to identify

points in time in which the model from which the data is generated changes. This can be a detection in real time (online change-point detection) or retrospectively (offline). In the offline setting, there is a distinction between methods that can detect at most one change-point, and methods that can detect multiple change-points.

Change-point detection problems originate from quality control (Page, 1954), but they have since been developed for a multitude of applications. Some of these include climate change (Reeves et al., 2007), health (Aston et al., 2017), speech recognition (Aminikhanghahi and Cook, 2017) and finance (Galeano and Wied, 2017). Despite being an old problem, there have even been some rather large innovations in the univariate setting quite recently (Killick et al., 2012; Fryzlewicz, 2014). We have also seen developments for higher dimensional data (Aston et al., 2017; Wang et al., 2020) including functional data (Aue et al., 2009a; Horváth and Kokoszka, 2012; Gromenko et al., 2017; Aue et al., 2019; Li and Ghosal, 2018; Sonmez, 2018).

Moving from univariate to multivariate data, detecting change-points becomes more difficult. This is a result of both the 'usual problems', which include visualization, computation and model complexity, as well as the fact that in the multivariate setting there exists additional types of structural changes which can occur; changes in the dependence structure of the data. In addition, we wish to consider the robustness of the procedure, especially since the lack of visualization can make detecting outliers difficult.

In fact, robustness is often ignored in the analysis of statistical techniques, including that of change-point estimation methods. It can be easy for a change-detection method to fail on an outlier, since methods are designed to set off alarms when relatively large values are observed. Despite this, many authors do not consider robustness when evaluating their change-point methodologies. In a recent comparison of popular change-point methods van den Burg and Williams (2020) consider fourteen methods, but only four (Knoblauch and Damoulas, 2018; Taylor and Letham, 2018; Fearnhead and Rigaill, 2019) of the original papers consider robustness and even less are actually robust. It is not however, that real data do not contain outliers. For example, Figure 1.4 contains the returns of four European stocks, originally analysed by Galeano and Wied (2017). Notice the two large outliers in the Siemens returns.

In addition to robustness, many multivariate change-point algorithms are not concerned with changes in the second order properties of the data. Many of the algorithms for multivariate change-point detection in the covariance or correlation matrix are very recent, e.g., (Aue et al., 2009b; Galeano and Wied, 2014; Wang et al., 2020). There does not exist one procedure which robustly and nonparametrically accounts for multiple changes in the variability of multivariate data. In Chapter 2 we present such a procedure, applying it

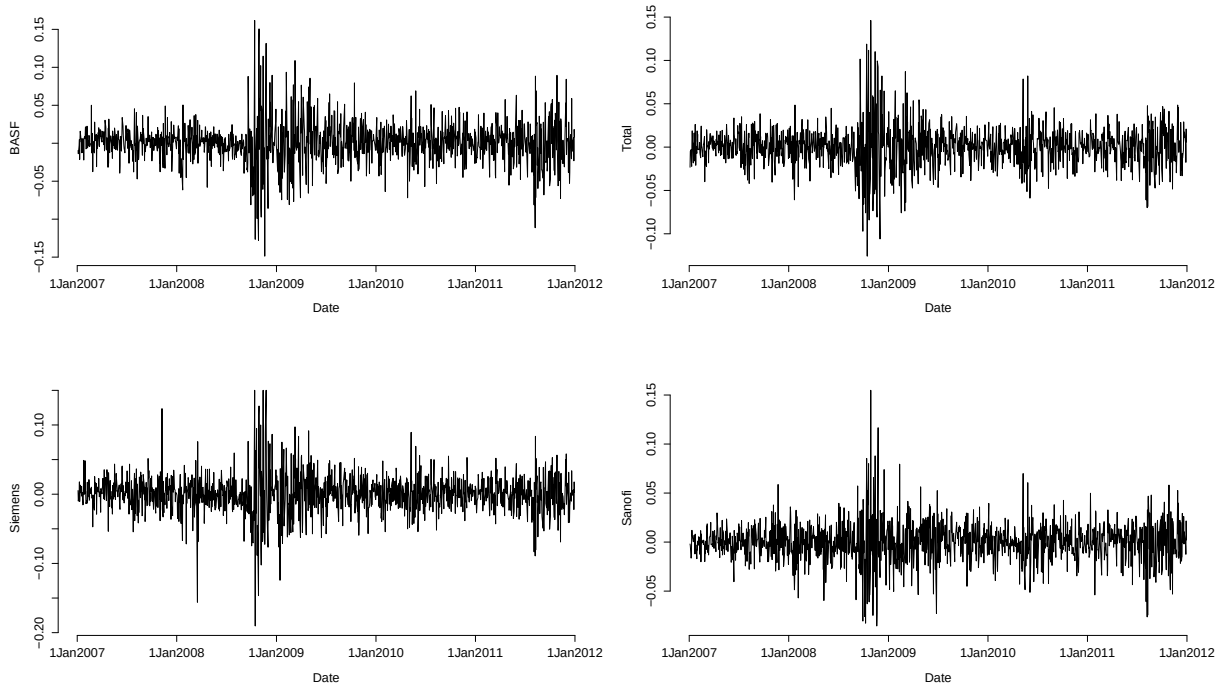


Figure 1.4: Financial returns for four European stocks. Notice the outliers in the bottom left panel.

to the data in Figure 1.4. Our method combines ideas from data depth ranks (Chenouri et al., 2020b), the PELT and wild binary segmentation algorithms (Killick et al., 2012; Fryzlewicz, 2014) and classical Kruskal Wallis k-sample testing (Kruskal, 1952). We are able to match some change-points detected by Galeano and Wied (2017) for these returns, without modelling the underlying time series.

In Chapter 3 we again use the classical Kruskal Wallis k-sample test, this time in the context of functional data. We make use of functional data depth functions to detect differences in the covariance kernels of k-samples. Equality of covariance kernels is analogous to testing for a difference in variance in the univariate set up; the *covariance kernel* of a continuous time stochastic process $X(t)$ is defined as:

$$\mathcal{K}(s,t) = \mathrm{E}\left[X(t)X(s)\right] - \mathrm{E}\left[X(t)\right]\mathrm{E}\left[X(s)\right].$$

Tests of this type have been applied to problems in engineering (Jarušková, 2013), DNA

microcircles (Panaretos et al., 2010) and biology (Fremdt et al., 2013). In this thesis, we apply our procedure to the curves in Figure 1.4, as part of applying a functional GARCH model (Cerovecki et al., 2019). There have been a few tests of this type proposed (Panaretos et al., 2010; Fremdt et al., 2013; Pigoli et al., 2014; Paparoditis and Sapatinas, 2016; Guo and Zhang, 2016; Cabassi et al., 2017; López-Pintado and Wrobel, 2017; Boente et al., 2018; Guo et al., 2018; Kashlak et al., 2019), but they typically have not been evaluated for robustness. In Chapter 3, we show that under some heavy tailed observations, many of these tests break down.

Many of these tests also involve estimation of the covariance kernel for each group, and/or resampling procedures which can be computationally expensive (Fremdt et al., 2013; Cabassi et al., 2017). Additionally, some of these tests do not possess desirable transformation invariance properties, which means they may depend on the coordinate system (Guo and Zhang, 2016). Through the use of functional data depth functions and ranks, we provide a robust nonparametric test that possesses desirable invariance properties and is computationally efficient. We show with simulation that, in some scenarios, our test is more powerful than many existing tests.

The results of the Chapters 2 and 3 can be combined to give a method for detecting change-points in the covariance kernel of a sequence of functional observations. This is the subject of Chapter 4. Detecting the presence and location of change-points in the covariance operator of a sequence of observed functions has received some recent interest in the statistics literature, see, e.g., (Aston et al., 2017; Harris et al., 2021). For example, recently change-point methods have been developed and applied to f-MRI data where the authors aim to detect many change-points, each occurring at the subject level within the data (Aston et al., 2017). This data is not only very high dimensional, but it is dependent, noisy and computationally difficult since the change-point algorithm must be run once for each subject. In Chapter 4 we introduce three procedures to perform covariance operator change-point detection: a hypothesis test for the presence of at most one change-point, a hypothesis test for the presence of an "epidemic period" and an algorithm to estimate the locations of multiple change-points when the number of change-points is not known. We may call these methods FKWC methods, or Functional Kruskal-Wallis Covariance operator methods. Since our methodology does not include estimation of the covariance operator our methods are computationally cheap. For example, our procedure can identify multiple change-points in $O(n \log n)$ time. Our procedure is fully non-parametric and is robust to outliers through the use of data depth ranks. We show that when n is large, our methods have simple behaviour under the null hypothesis. We also show that the FKWC change-point procedures are $n^{-1/2}$-consistent. In addition to asymptotic results, we provide a finite sample accuracy result for our at-most-one change-point estimator. In

simulation, we compare our methods against several others. We also present an application of our methods to intraday asset returns and f-MRI scans.

In Chapter 5 we explore ways in which depth-based methods can be used to conduct differentially private inference. Analyses that satisfy differential privacy protect study participants from harmful adversaries who wish to learn sensitive information about them. It has been shown in numerous papers, e.g., (Sweeney, 2007) that removing metadata from a database does not necessarily make a dataset private. In fact, in large data sets, given little information about the user, it is possible to identify them in the database (Narayanan and Shmatikov, 2008; Sankararaman et al., 2009; Dwork et al., 2017). For example, Narayanan and Shmatikov (2008) show that in the Netflix prize dataset, a dataset which contains the movies ratings of half of a million subscribers, knowing around 8-12 ratings is enough to identify the entire record in the database with very high certainty (>99%). They additionally argue that such a case study is valuable, in that it may be possible to learn sensitive information based on movie ratings. For example, a user may have rated movies with "predominantly gay themes" highly, which obviously may be sensitive information to some users. Further, once a released database is non-private, it is "forever non-private". By this, it is meant that any new online identity can be linked back to the database, if that new identity reveals information about their movie preferences. As a result of this work, and many related works, differential privacy was developed (Dwork et al., 2006).

Differential privacy gives the database participant the guarantee that an adversary cannot (virtually) determine whether or not they were part of the database, even if the adversary is given auxiliary information. Such a guarantee implies that there is no privacy cost (aside from leaks by say, the data curators) to participating in the database. Some other privacy definitions have been presented, however differential privacy has several advantages that have lead to its recent extensive study. First, it is resistant to many types of database attacks. Determining, or attempting to determine whether or not an individual is in a database is called a tracing attack. As a result of being resistant to tracing attacks, differentially private databases and analyses are also resistant to 'larger' privacy breaches, such as database reconstruction attacks (Dwork et al., 2017). Secondly, the definition of differential privacy is very general; it does not make rigid assumptions about what information the attacker has. The attacker can have as much information about the individual as possible; for example they may know the values of all but one of the attributes included in the dataset about the individual they are trying to trace.

To be precise, the meaning of "a study is differentially private" is that any statistic or sanitized database made public satisfies the mathematical definition of differential privacy, which we defer until Chapter 5. A non-private statistic or database is privatized through *differentially private mechanisms*, which are stochastic algorithms used to com-

pute statistics or sanitized databases. A major goal when developing both private inference procedures and private database release schemes, is to estimate and limit the information or utility loss endured as a result of privatization.

In this interest, one area of statistics that has been shown to be amenable to privatization is robust statistics. In robust statistics, the goal is to limit the influence of a small group of observations, deemed outliers. In the private setting, the goal is to limit the influence of any single observation, rather than that of a small group. Clearly, these two goals are closely related. The relationship to robust statistics has been identified and studied in a few recent works (Dwork and Lei, 2009; Lei, 2011; Sarwate and Chaudhuri, 2013; Hsu et al., 2014; Avella-Medina, 2019; Avella-Medina and Brunel, 2019; Brunel and Avella-Medina, 2020). We continue this line of work by exploring to what extent data depth functions and associated inference procedures can be privatized.

In Chapter 5 we investigate the differentially private estimation of depth functions and their associated medians. We then present a private method for estimating depth-based medians, which is based on the exponential mechanism (McSherry and Talwar, 2007). We compute the sample complexity of these private medians as a function of the dimension, prior parameters and privacy parameter. We show that, even for a decreasing privacy parameter in the dimension, the sample complexity is polynomial in the dimension. As a by-product of our work, we develop a smooth depth function, which we show has the same depth-like properties as its non-smooth counterpart. Another by-product of our work is uniform concentration inequalities for several depth functions. We also present methods and algorithms for estimating private depth values at both in-sample and out-of-sample points. We further extend the propose-test-release methodology of (Brunel and Avella-Medina, 2020) to be used with depth functions and the exponential mechanism. We show that when using propose-test-release to projection depth values, the probability of returning a 'null value' is small, and the private depth values concentrate around their population counterparts. We also give an algorithm to approximate the "test" step in propose-test-release, since it is computationally difficult. We show that this approximation maintains the small probability of returning a 'null value' mentioned above.

In the last chapter, we present some directions for future research. Recently, depth functions have been introduced on Riemennian manifolds (Fraiman et al., 2019; Harris et al., 2020). These depth functions allow for, among other things, the use of depth-based inference on shape spaces. There are also depth methods described for general Banach spaces (Cuevas and Fraiman, 2009; Chakraborty and Chaudhuri, 2014), networks Small (1997) and Hermitian positive definite matrices (Chau et al., 2019). We have yet to see serious study of inference procedures based on such concepts, and we hope to develop some of these in the future.

Chapter 2

Kruskal-Wallis type statistics for multivariate change-point problems

2.1 Introduction

The manufacturing industry motivated the development of change-point methods, that is, methods for detecting and dating distributional changes in a sequence of observations (Page, 1954). Change-point methods have since been applied to a much wider variety of research areas including climate change (Reeves et al., 2007), speech recognition (Aminikhanghahi and Cook, 2017) and finance (Wied et al., 2012), among others. With respect to a sequence of observations, the terms 'structural break' and 'change-point' refer to time points in the sequence during which there is a sudden change in the distribution from which the data is being generated. Change-point detection can be separated into two settings: 'online' and 'offline'. In the online setting, the data are being received by the analyst one datum at a time, and the goal is to detect a change as soon as possible, without too many false alarms. In the offline setting, the analyst has access to the entirety (or at least enough) of the data set, and the goal is to identify if and when changes occurred over the course of observation. In this thesis, we focus on the offline setting, for a summary of nonparametric methods in the online setting see (Chakraborti and Graham, 2019).

There are different variants of the offline change-point problem. Instead of identifying general changes in distribution, one might only be interested in identifying changes in the mean of the sequence (Chenouri et al., 2020a; Fryzlewicz, 2014), changes in the correlations of the sequence (Galeano and Wied, 2014) or changes in the covariance matrix of the sequence (Chenouri et al., 2020b; Wang et al., 2021). One may also be interested in

another type of distributional change entirely. In this chapter, we aim to detect changes in the variability of a sequence of multivariate observations.

To elaborate, suppose that the analyst suspects that there may exist time point(s) at which there are increases or decreases in the variance of one or more variates, or that there may exist time point(s) at which there is a change in the strength of the relationship between at least two variates. One can think of an analyst looking for changes in the variability and/or correlation strength of several financial asset returns. We call this type of change-point a multivariate variability change-point. For example, changes in a commonly used norm of the covariance matrix would be considered a change in variability. Another way to interpret changes in variability is through how the data cloud is affected; changes in variability produce changes in the magnitude and/or shape of the data cloud, rather than say, rotations or translations of the data cloud. A change in variability is the result of expansions and/or contractions of one ore more parameters of the covariance matrix. Note that this type of change-point differs from a change in the covariance matrix; it doesn't include cases where the covariance matrix is multiplied by an orthonormal matrix. For example, it does not include the situation where the correlation between two variates switches signs.

In order to detect changes in the variability of the data, we first compute the depth value of each of the datum in the sequence. This transforms the sequence of multivariate observations into a univariate sequence of depth values. We then compute the linear ranks of the depth values to form a sequence of ranks. We then try to detect multiple changes in the mean of this univariate sequence of ranks. We introduce two methods to complete this procedure, the first of which is a wild binary segmentation type algorithm based on rank CUSUM statistics (Fryzlewicz, 2014; Chenouri et al., 2020b). The second method is based on finding the set of change-points which maximize a penalized version of the classical Kruskal-Wallis test statistic used in nonparametric ANOVA (Kruskal, 1952). The implementation of this second method is based on the "pruned exact linear time" algorithm (Killick et al., 2012). To see the benefits of our proposed methods, we must first review existing methods.

There is a vast literature relating to the change-point problem, going back almost a century (Shewhart, 1931; Page, 1954). The literature includes a variety of approaches for both univariate, multivariate, single and multiple change-point detection methods (see the following review papers Reeves et al., 2007; Aue and Horváth, 2013; Aminikhanghahi and Cook, 2017, and the references therein). Much of the literature, especially in the multivariate setting, has focused on the detection of shifts in the mean of the process, e.g., (Truong et al., 2020).

Considerably less attention has been given to shifts in the second order behaviour of a sequence of observations. When second order change-points in the multivariate setting have been studied, the bulk of the literature has been concerned with detecting changes in the correlation structure. Galeano and Peña (2007) proposed a parametric framework for detecting changes in the correlation and variance structure of a multivariate time series, using both a likelihood ratio and a CUSUM statistic approach. Wied et al. (2012) proposed a nonparametric approach based on cumulative sums of sample correlation coefficients to detect a single change-point in the correlation structure of bivariate observations. This was later extended to multiple change-points (Galeano and Wied, 2014) and further to the multivariate setting (Galeano and Wied, 2017). Posch et al. (2019) has further extended the methods of Galeano and Wied (2017) to the high-dimensional setting by first applying dimension reduction techniques. One draw-back to the methods of Galeano and Wied (2014) is that they assume constant variances and expectations over time. Rather recently, a few alternative methods have been proposed, which include methods related to eigenvalues (Bhattacharyya and Kasa, 2018), residuals (Duan and Wied, 2018), semi-parametric CUSUM statistics (Zhao, 2017) and kernel methods (Cabrieto et al., 2018).

Literature related to estimating a change-point in the covariance matrix is quite recent, and relatively sparse. Aue et al. (2009b) take a CUSUM statistic approach similar to that of Galeano and Wied (2014). Kao et al. (2018) suggested a CUSUM statistic procedure based on eigenvalues. Chenouri et al. (2020b) considered a CUSUM based on ranks generated by data depth functions for detecting a single change-point. The high-dimensional setting has been tackled by Dette et al. (2018) and Wang et al. (2021). Dette et al. (2018) considers a two-step procedure based on dimension reduction techniques and a CUSUM statistic. Wang et al. (2021) is the only paper, to the best of our knowledge, seeking to identify multiple change-points, rather than a single change-point. They compare binary segmentation procedures (Venkatraman, E., 1992) and wild binary segmentation procedures (Fryzlewicz, 2014) based on a CUSUM statistic, under the assumption of sub-gaussian observations.

Fryzlewicz (2014) developed wild binary segmentation as an improvement on the well-known univariate multiple change-point algorithm binary segmentation (Venkatraman, E., 1992). Binary segmentation has been used to extend single change-point algorithms to multiple change-point algorithms in many settings (such as Aue and Horváth, 2013; Galeano and Wied, 2014, 2017; Duan and Wied, 2018; Wang et al., 2021; Chenouri et al., 2020a). The extension and study of wild binary segmentation in the multivariate setting, with respect to changes in the covariance structure of a time series has only been done by Wang et al. (2021).

In addition to methods where the change-point type is specified, there exists several nonparametric algorithms designed to detect general changes in the distribution of the ob-

servations. Matteson and James (2014) studied the e-divisive algorithm, which can detect the location and number of change-points in the distribution of a sequence of multivariate observations. Their method is based on distances between characteristic functions and a hierarchical clustering inspired iteration. Their methods are extended in Zhang et al. (2017), where a pruning component is added to an existing, dynamic programming-based change-point algorithm. These authors apply this pruning method to the e-divisive algorithm and the kernel change-point methods of Arlot et al. (2012). This group of methods are implemented in the `ecp` `R` package (James and Matteson, 2015). At first the rank-based multiple change-point method of Lung-Yut-Fong et al. (2011) may seem similar to our methods, but their procedure requires the number of change-points to be known. Further, their methods are based on component-wise ranks, which have several known issues, such as a lack of transformation invariance (Bickel, 1965).

The change-point literature lacks methods for specifically detecting multiple changes in the variability of multivariate data. Many of the papers discussed focus on the at most one change problem, or are not designed to detect changes in variability, or even changes in the covariance matrix of the data. The only directly comparable paper is that of Wang et al. (2021). Even this method is designed for the high dimensional setting; in our simulation study, when the dimension is low to moderate, our method outperforms this method. The other comparable methods would be those that detect multiple, general changes in the distribution of the data, such as those of Zhang et al. (2017). We demonstrate that our method is able to outperform these general methods in simulation, when the change-points are all variability change-points. This is not surprising; our method sacrifices generality for accuracy.

In addition, the aforementioned change-point methods do not consider robustness to outlying observations. For example, many of the existing methods assume that the data are sub-gaussian (Dette et al., 2018; Wang et al., 2021). Furthermore, existing papers often present no simulation results concerning a method's performance under heavy tailed data. For example, we show in simulation that the methods of Matteson and James (2014); Wang et al. (2021) do not perform well when the data are heavy tailed. By contrast, our theoretical and simulation results show that our method works well in scenarios where the data are heavy tailed.

The rest of the chapter is organized as follows, Section 2.2 introduces the data model. Section 2.3 outlines the proposed change-point detection procedures. Section 2.4 presents consistency results (with rates) for both of our presented methods. Section 2.5 presents simulation results, including a discussion of the tuning parameters. We test the proposed change-point methods in a variety of scenarios and compare the methods to one another as well as to the methods of Matteson and James (2014); Zhang et al. (2017); Wang et al.

(2021). In Section 2.6 we analyze four European daily stock returns. This is the same data set analyzed by Galeano and Wied (2017) and we compare our results to theirs.

2.2 The data model, variability changes and their relation to depth ranks

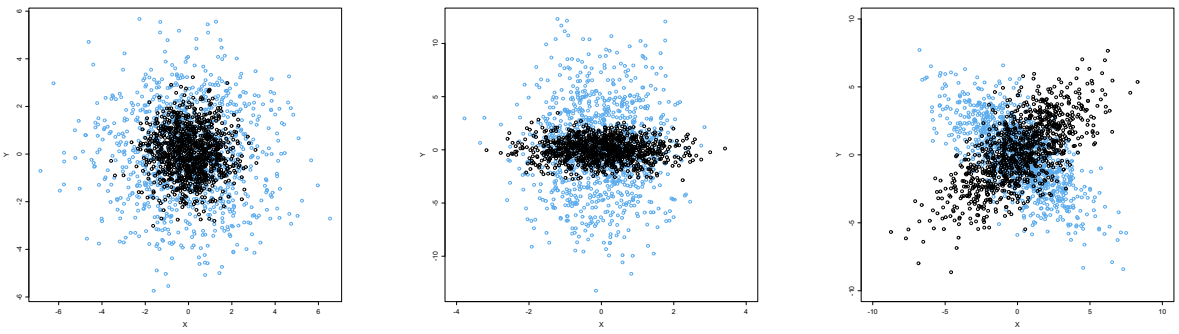
We now describe the change-point model that we will focus on. Suppose that $X_1, \ldots, X_n$ is a sequence of zero mean, independent random variables such that $X_{k_{i-1}+1}, \ldots, X_{k_i}$ have law F_i, with, $k_0 = 0 < k_1 < \cdots < k_\ell < k_{\ell+1} = n$ for some fixed, unknown ℓ. Suppose that $k_i = \lfloor n\theta_i \rfloor$ for all $i \in \{0, 1, \ldots, \ell + 1\}$. Let $\vartheta_i = \theta_i - \sum_{j=0}^{i-1} \theta_j$ be the approximate fraction of the observations coming from F_i and define

$$F_* \coloneqq \vartheta_1 F_1 + \vartheta_2 F_2 + \vartheta_3 F_3 + \cdots + \vartheta_\ell F_\ell + \vartheta_{\ell+1} F_{\ell+1}.$$

The aim is to estimate ℓ and each k_i; the correct number of change-points along with their location, given only the sample. We further suppose that for any $i = 1, \ldots, \ell$, F_i differs from F_{i+1} only in variability. That is, if we let Σ_i be the covariance matrix corresponding to F_i, recall that we can write $\Sigma_i = \mathcal{R}_i \mathcal{S}_i \mathcal{S}_i \mathcal{R}_i^\top$, where $\mathcal{R}_i$ is an orthonormal matrix and $\mathcal{S}_i$ is a scale matrix. If $\mathcal{S}_i \neq \mathcal{S}_{i+1}$, then there is a difference in variability between F_i and F_{i+1}.

We will use depth based ranks to assess the change in variability. The relationship between variability and combined sample depth values has already been explored by several other authors, e.g., (Li and Liu, 2004) and we give a more mathematical discussion in Chapter 3. For now, we give an intuitive explanation. The reader may also view a short simulation study of the distribution of the depth ranks under different covariance changes in Appendix A.2.

The fact that changes in the variability of the data produce a change in the mean of the data depth values is guaranteed from the construction of depth functions, specifically the maximality at center property combined with the quasi-concavity property. Since we assume that the pre-change and post-change data have the same location, we can also assume that the combined sample depth function will be maximized roughly at that location. Additionally, recall that a change in variability is a change in the magnitude and/or shape of the post-change data cloud. The change in the magnitude and/or shape of the data cloud will result in the post-change data being, on average, a different distance from the centre. Due to the quasi-concavity property, this change in distance will result in the post-change data having higher/lower combined sample depth values, on average.



(a) (b) (c)

Figure 2.1: Two samples of 1000 points with different covariance matrices. The differences can be characterised as (a) an expansion difference (b) a sub-matrix expansion difference (c) a sign change. Notice the magnitude or shape of the data cloud changes in panes (a) and (b), but not in pane (c).

For example, panel (a) of Figure 2.1 shows two bivariate normal samples overlaid. The blue sample has an expanded covariance matrix relative to the black sample. Notice that both data clouds have the same shape, but the magnitude differs between them. It is easily seen that the black points are more central, relative to the shared center of the samples. Therefore, when we compute the depth values with respect to the combined sample, the black points will have, on average, higher depth values. Panel (b) shows again two bivariate normal samples overlaid, but this time the expansion is only in one parameter of the covariance matrix. In this case, the shapes of the two data clouds differ, which results in the black points being more central. This results in the black points having generally higher depth values in the combined sample. Panel (c) of Figure 2.1 shows again two bivariate normal samples overlaid, except only the sign of the correlation between the two variates differs between the two samples. Notice that the data cloud does not change in size or shape, it is simply rotated. The combined sample depth values will not change in this case, since one sample is not more central relative to the other.

It is entirely possible that changes in the direction of outlyingness, e.g., a sign change in correlation, are outside of the scope of interesting, or plausible changes in a particular dataset. For example, the hypothesis that the spread of the data is increasing or decreasing in a particular direction is considerably different from the hypothesis that the relationship between two or more variates has reversed. If there is reason to believe that the only plausible changes in the data process are variability changes, rather than say, general changes,

or even general covariance changes, it is beneficial to use the depth-based procedure. This is reflected in our simulation study where we compare our method to some change-point methods which make less assumptions about the type of change in the data (see Section 2.5). This is intuitive; including more information about the data into the assumptions of the procedure should improve the results of the procedure. The downside of course would be missing other types of changes if they are not suspected to be present.

2.3 Proposed change-point algorithms

In this section we describe two multiple change-point algorithms that can be used to detect changes in the variability of multivariate data. The first algorithm, takes a local approach, in the sense that the idea is to look at small sections of the data and treat the problem as a single change problem within each small section. The second algorithm takes a global approach, such that all of the change-points are simultaneously estimated. We will compare the methods in the subsequent sections.

2.3.1 WBS and a depth rank CUSUM statistic

Wild binary segmentation, introduced by Fryzlewicz (2014) was originally developed for detecting multiple change-points in the mean of univariate data. Seeing as the problem here is essentially to detect changes in the mean of the depth-based ranks, it seems natural to use a similar approach. In fact, Chenouri et al. (2020a) combined wild binary segmentation with univariate rank statistics with quite favourable results. In the 'at most one change' setting Chenouri et al. (2020b) proposed using the following rank CUSUM statistic

$$Z_{1,n}(m/n) \coloneqq \frac{1}{\sqrt{n}} \sum_{i=1}^{m} \frac{\widehat{R}_i - (n+1)/2}{\sqrt{(n^2-1)/12}}$$

to detect scale changes in a sequence of multivariate data. Recall from Chapter 1 that $\widehat{R}_i$ are the depth-based ranks, see (1.6). Our first algorithm pairs this CUSUM statistic with wild binary segmentation. Let $[k] = \{1, \ldots, k\}$ for any integer k and let $e,\ s \in [n]$ and $s < e$. We define the following rank CUSUM statistic for the set $\{X_s, \ldots, X_e\}$ of size $n_{s,e} = e - s + 1$

$$Z_{s,e}(m/n_{s,e}) \coloneqq \frac{1}{\sqrt{n_{s,e}}} \sum_{i=1}^{m} \frac{\widehat{R}_{i,s,e} - (n_{s,e}+1)/2}{\sqrt{\left(n_{s,e}^2-1\right)/12}},$$

where $\widehat{R}_{i,s,e}$ are the linear ranks resulting from ranking the depth values of the observations in the subsample $\{X_s, \ldots, X_e\}$, with respect to only the observations in $\{X_s, \ldots, X_e\}$. More precisely, the depth values are taken with respect to the empirical distribution generated by $\{X_s, \ldots, X_e\}$. These ranks range from $1, \ldots, n_{s,e}$.

Following the lines of Fryzlewicz (2014) we can now outline our algorithm as follows. First choose many, say J, uniformly random intervals and let

$$\mathrm{INT} = \{(s_j, e_j) \colon \ j \in [J],\ e_j - s_j > \widetilde{\Delta} n,\ s_j,\ e_j \in [n]\}$$

be the set of those intervals whose length is at least $\widetilde{\Delta} n$. Note that $\widetilde{\Delta}$ is an algorithm parameter. After choosing the intervals, the algorithm runs recursively. In one recursive iteration, the algorithm starts with a supplied interval (s, e) and the aim is to detect the prominent change-point in this interval, if one exists. First, $\mathrm{INT}_{s,e} \subset \mathrm{INT}$ is computed; $\mathrm{INT}_{s,e}$ is the set of intervals (s_j, e_j) such that $e_j \leq e$ and $s_j \geq s$. This is the set of sub-intervals whose end-points are in (s, e). Then for each sub-interval $(s_j, e_j) \in \mathrm{INT}_{s,e}$, the maximal CUSUM statistic is computed:

$$\sup_{s_j \leq m < e_j} |Z_{s_j,e_j}(m/n_{s,e})|.$$

This produces $|\,\mathrm{INT}_{s,e}\,|$ change-point estimates paired with their respective CUSUM statistics, where $|A|$ denotes the cardinality of a set A. The change-point estimate which produces the maximal CUSUM statistic out of all the computed CUSUM statistics is then selected as the candidate change-point

$$(j^*, m^*)_{s,e} = \operatorname*{argmax}_{(j,m)\colon (s_j,e_j)\in \mathrm{INT}_{s,e},\ m\in\{s_j,\ldots,e_j-1\}} \left| Z_{s_j,e_j}\left(\frac{m - s_j + 1}{n_{s,e}}\right)\right|.$$

If it holds that

$$\left| Z_{s_{j^*},e_{j^*}}\left(\frac{m^* - s_{j^*} + 1}{n_{s,e}}\right)\right| > T, \tag{2.1}$$

for some threshold T, then the algorithm adds the index m^* to the list of change-points. Additionally, if (2.1) holds then the algorithm calls itself twice, once with the new supplied interval being (s, m^*) and once with the new interval being $(m^* + 1, e)$. If (2.1) does not hold then the algorithm stops and returns the set of current change-points. Pseudo-code for this algorithm is summarized in Algorithm 1.

Algorithm 1 Rank-Based Wild Binary Segmentation

procedure WBS_RANK(e, s, T, INT)
  if $e - s < 1$ **then**
    STOP
  else
    $\text{INT}_{s,e} \coloneqq$ intervals $(s_j,\ e_j) \in \text{INT}$ such that $(s_j,\ e_j) \subset (s, e)$
    $(j^*, m^*) \coloneqq \text{argmax}_{\mathcal{B}} \left| Z_{s_j,e_j}\left(\frac{m-s_j+1}{n_{s,e}}\right)\right|,$
      with $\mathcal{B} \coloneqq \{(j, m)\colon (s_j, e_j) \in \text{INT}_{s,e},\ m \in \{s_j, \ldots, e_j - 1\}\}$
    if $\left|Z_{s_{j^*},e_{j^*}}\left(\frac{k^*-s_{j^*}+1}{n_{s,e}}\right)\right| > T$ **then**
      Append m^* to the list of change-points $\widehat{\mathbf{k}}$
      WBS_Rank(s, m^*, T, INT)
      WBS_Rank$(m^* + 1, e, T, \text{INT})$
    else
      STOP
    end if
  end if
  return $\widehat{\mathbf{k}}$
end procedure

2.3.2 KW-PELT: A Kruskal-Wallis change-point algorithm

As mentioned above, Algorithm 1 takes a local approach to the problem, utilising only sections of the data to estimate each change-point. Additionally, there is the issue of subjectivity with regard to choosing the number of intervals. As an alternative, we can instead maximize a single objective function based on the whole data set. Recall from Section 2.2 that a mean change in the depth values is implied by a change in variability of the original data. Recall that the Kruskal-Wallis test statistic is used to check for median differences among multiple groups of univariate data. In other words, if there is a median difference among the groups then it is expected that the Kruskal Wallis statistic will attain a large value. It is very natural to then base the objective function on the Kruskal-Wallis ANOVA test statistic. To this end, we propose using the following as an estimator of the change-points

$$\widehat{\mathbf{k}} \coloneqq \underset{k_0=0<k_1<\cdots<k_\ell<n=k_{\ell+1}}{\operatorname{argmax}} \frac{12}{n(n+1)} \sum_{i=1}^{\ell+1} (k_i - k_{i-1}) \overline{\widehat{R}}_i^2 - 3(n+1) - \beta_n(\ell+1), \tag{2.2}$$

where β_n is a parameter for which higher values correspond to higher penalization on the number of estimated change-points and $\overline{\widehat{R}}_i$ is the mean of the sample depth ranks in group i, viz.

$$\overline{\widehat{R}}_i = \frac{1}{k_i - k_{i-1}} \sum_{j=k_{i-1}+1}^{k_i} \widehat{R}_j \ .$$

One can recall that $\widehat{R}_j$ are defined in (1.6), or, also in relation to the wild binary segmentation algorithm $\widehat{R}_j = \widehat{R}_{j,1,n}$. Note that the penalization is necessary to avoid overfitting; without it the solution to this maximization problem is simply choosing every point as a change-point. It is apparent that (2.2) is a difficult maximization problem in the sense that the number of possible solutions is 2^n. However, we can circumvent this issue by applying the pruned exact linear time algorithm (Killick et al., 2012). Indeed, rewrite the objective function, in (2.2), by which we denote $\mathbf{G}(n)$, as

$$\mathbf{G}(n) \coloneqq \sum_{i=1}^{\ell+1} -c(k_{i-1}+1 : k_i) - \beta_n \ell$$

where

$$c(s+1:e) = -\frac{12(e-s)}{n(n+1)} \left[\frac{1}{e-s} \sum_{i=s+1}^{e} \widehat{R}_i - \frac{n+1}{2} \right]^2 . \tag{2.3}$$

Letting $k_0 = 0$ and $k_{\ell+1} = e$, we can write the maximization problem in (2.2) as

$$\begin{aligned}
\max_{0<k_1<\cdots<k_\ell<e} \mathbf{G}(e) &= \min_{0<k_1<\cdots<k_\ell<e} \frac{12}{n(n+1)} \sum_{i=1}^{\ell+1} -(k_i - k_{i-1}) \left(\widehat{R}_i - \frac{n+1}{2} \right)^2 + \beta_n(\ell+1) \\
&= \min_s \left\{ \min_{k_0<k_1<\cdots<k_\ell<s} \sum_{i=1}^{\ell} (c(k_{i-1}+1 : k_i) + \beta_n) + c(s+1:e) + \beta_n \right\} \\
&= \min_s \left\{ -\mathbf{G}(s) + c(s+1:e) + \beta_n \right\}.
\end{aligned}$$

It is straightforward to show that (2.2) satisfies the assumption in (Killick et al., 2012) required for PELT to be applicable: We need to show for $0 \leq s < e < e' \leq n$ that there exists a constant C''

$$c(s+1:e) + c(e+1:e') + C'' \leq c(s+1:e').$$

Letting $\mu_n = (n+1)/2$, observe that this condition is equivalent to

$$\frac{e-s}{e'-s} \left[\frac{1}{e-s} \sum_{i=s+1}^{e} \widehat{R}_i - \mu_n \right]^2 + \frac{e'-e}{e'-s} \left[\frac{1}{e'-e} \sum_{i=e+1}^{e'} \widehat{R}_i - \mu_n \right]^2 + C'' \geq \left[\frac{1}{e'-s} \sum_{i=s+1}^{e'} \widehat{R}_i - \mu_n \right]^2.$$

Let $p \in (0,1)$ and $a, b, c \in \mathbb{R}$. The above expression can be written in the form

$$\begin{aligned}
& p(a-c)^2 + (1-p)(b-c)^2 + C'' \geq (pa + (1-p)b - c)^2 \\
\implies & pa^2 + (1-p)b^2 + C'' \geq (pa + (1-p)b)^2 \\
\implies & p(1-p)a^2 + ((1-p) - (1-p)^2)b^2 - 2pa(1-p)b + C'' \geq 0.
\end{aligned}$$

We can view the expression

$$p(1-p)a^2 + ((1-p) - (1-p)^2)b^2 - 2pa(1-p)b$$

as a quadratic form in a. The discriminant of this quadratic form is

$$(2p(1-p)b)^2 - 4p(1-p)(p(1-p))b^2 = 0$$

which means that the quadratic form has one root. Additionally, the quadratic form opens upward since $p(1-p) > 0$, implying that this function is positive for all a, b. Therefore, the condition

$$c(s+1:e) + c(e+1:e') + C'' \leq c(s+1:e')$$

is satisfied. Algorithm 2 outlines this procedure, which we call KW-PELT, in pseudo-code.

Algorithm 2 KW-PELT

procedure KW_PELT($\mathbf{R}, \beta, \widetilde{\Delta}$)
  $n \coloneqq length(\mathbf{R})$
  $\widehat{\mathbf{k}}(0) = NULL$
  $\mathcal{N}_0 \coloneqq \{0\}$
  $\mathbf{G}(0) = -\beta$
  for $k \in 1, \ldots, n$ **do**
    $\mathbf{G}(k) = \min_{s \in \mathcal{N}_k}\{\mathbf{G}(s) + c((s+1):k) + \beta\}$
    $k^1 = \text{argmin}_{s \in \mathcal{N}_k}\{\mathbf{G}(k) + c((s+1):k) + \beta\}$
    $\widehat{\mathbf{k}}(k) = (\widehat{\mathbf{k}}(k^1), k^1)$
    $\mathcal{N}_{k+1} \coloneqq \{k\} \cup \{s \in \mathcal{N}_k \colon \mathbf{G}(s) + c((s+1):k) \leq \mathbf{G}(k)\}$
  end for
  return $\widehat{\mathbf{k}}(n) \backslash \{0\}$
end procedure

We end this section with a remark about computation time. Computationally, the limiting factor for both procedures will (in general) be the computation time for the sample depths. Consequentially, we expect Algorithm 2 to be faster, due to the fact that sample depth functions need only be calculated once rather than once for every sampled interval. If $f(n; d)$ is the time it takes to compute the sample depths, then Algorithm 1 would take $O(Jn \log n + Jf(n; d))$ time as opposed to $O(n \log n + f(n; d))$ time for Algorithm 2. It is worth noting that Algorithm 2 was implemented partially in C++ whereas Algorithm 1 was implemented completely in `R` (except for possibly the depth computations, for which existing packages were used) so the empirical times in simulation are not directly comparable. This being said, both algorithms ran within minutes on a desktop computer when applied to the data set analyzed in Section 2.6.

2.4 Consistency of the algorithms

In this section we provide consistency results for both algorithms under some assumptions. For $j \in [\ell + 1]$, let $Y_j \sim F_j$, let $H_j(x) = \Pr(\mathrm{D}(Y_j; F_*) \leq x)$, let $p_{i,j} = \Pr(\mathrm{D}(Y_i; F_*) >$

$\mathrm{D}(Y_j; F_*)$) for $i, j \in [\ell]$ and let $F_{*,n}$ denote the empirical distribution invoked by the combined sample $X_1, \ldots, X_n$. The following assumptions are used in the consistency theorems that follow.

Assumption 1. *$H_j(x)$ are Lipschitz continuous with constant C, that is*

$$|H_j(x) - H_j(y)| \leq C|x - y|,$$

for $x, y \in \mathbb{R}^d$.

Assumption 2. *It holds that*

$$\mathrm{E}\left[\sup_{x \in \mathbb{R}^d} |\mathrm{D}(x; F_{*,n}) - \mathrm{D}(x; F_*)|\right] = O(n^{-1/2}).$$

Assumption 3. *The number of change-points ℓ and their locations $\lfloor n\theta_i \rfloor = k_i$ are fixed for all $i \in \{1, \ldots, \ell\}$.*

Assumption 4. *There exists $p_0 > 0$ such that for all $i \in [\ell]$, it holds that*

$$|1/2 - p_{i,i+1}| \geq p_0.$$

Assumption 5. *The threshold T is such that $T = o(\sqrt{n})$ and $T \to \infty$ as $n \to \infty$.*

Assumption 6. *There exists $p_0 > 0$ such that for any $j \in [\ell + 1]$ it holds that*

$$\left|\sum_{i=1}^{\ell+1} \vartheta_i p_{i,j} - \frac{1}{2}\right| \geq p_0.$$

Assumptions 1 and 2 are satisfied by most depth functions under absolutely continuous F, including those defined in Section 1.1 (see Liu et al., 1999, and the references therein). Assumption 3 says that the number of change-points is fixed, and their positions are fixed in n. This assumption is restrictive, but we hope to relax it in the future. Assumptions 4 and 6 are concerned with the type of changes that can be detected. In the setting of rank statistics the size of each the change-point is measured by $|p_{i,i+1} - 1/2|$. Assumption 4 says that the size of the change must be greater than some quantity fixed in n. Similarly, Assumption 6 assumes that for each segment j, the weighted average change-size between segment j and another segment is non-zero and fixed.

For example, suppose there is a single change-point and that $X_1, \ldots, X_{k_1} \sim \mathcal{N}_d(0, I)$ and that $X_{k_1+1}, \ldots, X_n \overset{\mathrm{d}}{=} \sqrt{a} X_1$ with $a > 1$. Clearly, if $X \sim F_*$ then we have that $\mathrm{E}_{F_*}[X] = \mathbf{0}$

and $\Sigma_* = (\vartheta_1 + a(1-\vartheta_1))I = \sigma_*^2 I$. Here, Σ_* is the covariance matrix corresponding to the distribution F_*. It follows that

$$\|X_1 - \mathrm{E}_{F_*}[X]\|_{\Sigma_*^{-1}} \sim \frac{1}{\sigma_*^2}\chi_d^2 \qquad \text{and} \qquad \|X_n - \mathrm{E}_{F_*}[X]\|_{\Sigma_*^{-1}} \sim \frac{a}{\sigma_*^2}\chi_d^2.$$

Now, for any $x \in \mathbb{R}^+$ we have that

$$F_{\chi_d^2}\left(\sigma_*^2 x\right) > F_{\chi_d^2}\left(\sigma_*^2 x/a\right),$$

where $F_{\chi_d^2}$ represents the cumulative distribution function of a χ_d^2 random variable. It follows immediately that $p_{1,2} = 1 - p_{2,1} \neq \frac{1}{2}$. Additionally, if $Y \sim \chi_d^2/\sigma_*^2$, then

$$\mathrm{E}_{\sigma_*^2\chi_d^2}\left[F_{\chi_d^2}\left(\sigma_*^2 Y/a\right)\right] < \mathrm{E}_{\sigma_*^2\chi_d^2}\left[F_{\chi_d^2}\left(\sigma_*^2 Y\right)\right] = \frac{1}{2};$$

both Assumption 4 and Assumption 6 are satisfied. Clearly, neither Assumption 4 or Assumption 6 hold if $a = 1$.

Theorem 1. *Suppose that* $\widetilde{\Delta} < \min_{i\in[\ell+1]}|\theta_i - \theta_{i-1}|$. *Let the estimated change-points* $\widehat{k}_1 < \widehat{k}_2 < \cdots < \widehat{k}_{\widehat{\ell}}$ *be as in Algorithm 1. Suppose that Assumptions 1-5 hold, and that the number of intervals* $J_n \to \infty$ *as* $n \to \infty$. *Then, there exists a universal constant* $C > 0$ *such that the following holds*

$$\Pr\left(\left\{\widehat{\ell} = \ell\right\} \cap \left\{\max_{i\in[\ell]}|\widehat{k}_i - k_i| \leq Cn^{1/2}\log n\right\}\right) \to 1 \text{ as } n \to \infty.$$

Theorem 1 states that for large n, it is highly probable that the change-point estimates produced by Algorithm 1 will be close to the location of the true change-points and that the number of these estimates is equal to the true number of change-points. The next theorem gives a similar, but weaker result for Algorithm 2 under a wide range of penalty terms.

Theorem 2. *For* β_n *as in (2.2), assume that* $\beta_n \to \infty$ *as* $n \to \infty$ *and that* $\beta_n = o(n)$. *Let* r *be a constant such that* $1/2 < r < 1$. *Provided Assumptions 1-3 hold and Assumption 6 holds, for* $\widehat{\mathbf{k}}$ *and* $\widehat{\ell}$ *as in Algorithm 2, there exists a constant* $C > 0$ *such that*

$$\Pr\left(\left\{\widehat{\ell} = \ell\right\} \cap \left\{\max_{i\in[\ell]}|\widehat{k}_i - k_i| \leq Cn^{r}\right\}\right) \to 1 \text{ as } n \to \infty.$$

The rate given in Theorem 1 is better than that of Theorem 2. In addition, Assumption

3 is restrictive and could be relaxed in the future, to account for an increasing number of change-points and varying locations of the change-points. Additionally, 4 and 6 could account for $p_0 \to 0$, analogous to that of Fryzlewicz (2014). We have kept simple assumptions to lay the theoretical groundwork and will improve the results in the future by adjusting the current proofs. Once we relax the assumptions, we can compare our results to the optimal rates of convergence given by Hernan Madrid Padilla et al. (2019).

2.5 Simulation study

In this section we use a simulation study to compare our methodology to existing procedures as well as to investigate different choices of the algorithm parameters: the penalty β_n and the threshold T. Specifically, we compare our method to the Binary Segmentation in Operator Norm (BSOP) algorithm and the Wild Binary Segmentation through Independent Projection (WBSIP) algorithm developed by Wang et al. (2021), as well as to the methods in the `R` package `ecp` (James and Matteson, 2015). It should be noted that the aim of Wang et al. (2021) was to detect change-points in the high dimensional setting, and not necessarily in low or moderate dimensions. The WBSIP algorithm performed much better than the BSOP algorithm, and so we only present the results of the WBSIP algorithm. As mentioned above, we also compare to the nonparametric change-point methods in the `ecp` `R` package (James and Matteson, 2015). These methods are designed to detect general types of change-points; time points where there was a change in distribution. We compared our methods with the `e.divisive`, `e.cp3o_delta` and `e.kcp3o` methods. The other methods in the package were either too slow to run with our simulation set-up or performed much worse than the chosen methods. The best of these three methods in our simulation set-up was by far the `e.divisive` method, and so we do not present the results of the other two methods.

The threshold parameter for the WBSIP algorithm was chosen to be 561, which was based on visually assessing the error $\widehat{\ell} - \ell$ so that the median was zero in most of the scenarios. We also tried choosing the threshold in order to minimize the empirical mean squared error of the estimate $\widehat{\ell}$ (which would not be known in practice) and the results were similar. For Algorithm 1 and the WBSIP algorithm we used $100\lfloor \log n \rfloor$ intervals.

The simulation study is limited to evenly spaced change-points, from distributions with independent marginals. Note that the transformation invariance properties possessed by the depth functions imply the results from similarity transformations of the data would be the same. This transformation invariance implies that the study also covers some cases

where the marginal distributions of the data are not independent. We set the mean of all distributions to be 0.

The simulation study consisted of several scenarios. The first scenario is a set of expansions and contractions controlled by the parameter σ^2. We let $\Sigma_j = \sigma_j^2 I_d$ for each F_j, $j \in [\ell + 1]$. We set

$$\sigma_1^2 = 1,\ \sigma_2^2 = 2.5,\ \sigma_3^2 = 4,\ \sigma_4^2 = 2.25,\ \sigma_5^2 = 5,\ \sigma_6^2 = 1,$$

e.g., for 2 change-points, σ^2 would vary as follows $1 \to 2.5 \to 4$. The second scenario is another set of expansions and contractions, of which the results were so similar that we do not present them here.

We simulated data from three different distribution types, normal, Cauchy and skewed normal with skewness parameter $\gamma = 0.1/d$. We ran the simulation for values of d =2, 3, 5 and 10 under 2, 3 and 5 change-points. To see results on zero change-points and one change-point (see Chenouri et al., 2020b). We used sample sizes of n =1000, n =2500, and n =5000, running each scenario 100 times. We tested our methods with halfspace depth, spatial depth and both kinds of Mahalanobis depth introduced in Section 1.1.

Lastly, we ran several simulations designed to assess the performance of our methods under sparsity and/or high dimensions. In these scenarios, d was at most 500 and/or the expansions/contractions were only applied to a submatrix of the covariance matrix.

`R` codes to replicate this simulation study, as well as implementations of Algorithm 1, Algorithm 2, the WBSIP algorithm and the BSOP algorithm are available (Ramsay, 2019b).

2.5.1 Choosing the algorithm parameters

In order to have consistency of the estimates produced by Algorithm 1, the threshold must satisfy $T = o(\sqrt{n})$. One option is to choose a fixed threshold T^*, which will produce a set of change-point estimates and their corresponding CUSUM statistics. The final set of change-points could then be chosen by testing each change-point for significance using a Bonferroni correction or Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) along with the quantiles of $\sup |B(t)|$, where $B(t)$ is the standard Brownian bridge. This would imply a threshold $T \geq T^*$. However, it might be that smaller sampled intervals are not large enough for the asymptotic approximation to work well. Additionally, one has to choose the significance level, and the threshold T^*. As a result of these considerations, we suggest a

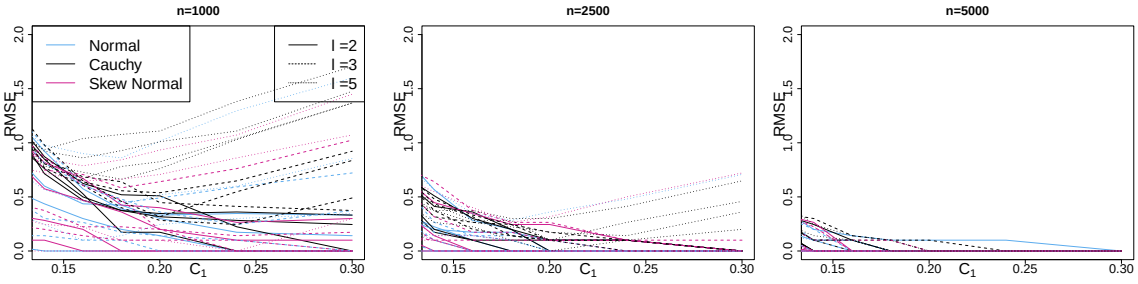


Figure 2.2: Empirical root mean squared error of $\widehat{\ell}$ for different values of C_1 under spatial depth for all the simulation parameter combinations, under Algorithm 2.

data driven thresholding approach, based on the generalized Schwartz Information Criteria, as done by Fryzlewicz (2014).

Algorithm 1 produces a nested set of models, indexed by the threshold parameter. Lowering the threshold can only add new change-points to the model; all previously estimated change-points remain. In other words, as the threshold decreases, new change-points are added to the model one at a time. It is then easier to re-index the models by the number of estimated change-points $\widehat{\ell}$. The threshold problem can then be reformulated as a model selection problem.

Suppose we have a univariate sample $Z_1, \ldots, Z_n$ and the goal is to estimate a change-point in the mean. For this problem, Fryzlewicz (2014) chooses the 'best' model by minimizing the following criteria:

$$\mathcal{G}(\widehat{\ell}) = \frac{n}{2}\log(\widehat{\varsigma}^2_{\widehat{\ell}}) + \widehat{\ell}\log^{\alpha} n, \tag{2.4}$$

with $\widehat{\varsigma}^2_{\widehat{\ell}}$ equal to the average within group squared deviation (a group is an estimated period of constant mean) and $\widehat{\ell}$ is the estimated number of change-points. Let $\bar{Z}_i$, for $i \in [n]$, be the empirical, within group mean for the group that contains univariate observation Z_i. Then we can write

$$\widehat{\varsigma}^2_{\widehat{\ell}} = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z}_i)^2.$$

Here, α is a parameter such that the larger α, the larger the penalty against choosing a model with many change-points.

The only difference for the multivariate, variability problem is that $\widehat{\varsigma}$ must be modified. As mentioned previously, each of the nested models produced by Algorithm 1 will produce

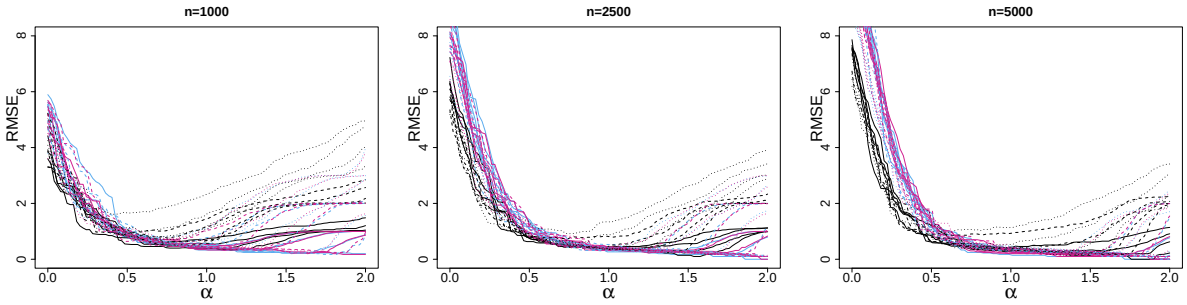


Figure 2.3: Each curve shows the empirical root mean squared error of $\widehat{\ell}$ for different values of α under Algorithm 1 paired with Mahalanobis depth, for a given simulation parameter combination, following the legend of Figure 2.2. Mahalanobis depth is used rather than spatial depth because of computational efficiency.

a candidate set of change-points $0 = w_0 < w_1 < \ldots < w_r < w_{r+1} = n$. We recall from Section 2.2 that a variability change is equivalent to a change in the mean depth-based ranks. We can treat the sample depth-based ranks produced by the depth functions as a univariate sample and thus, for a given model, define the within group deviation as

$$\widehat{\varsigma}^2_{\widehat{\ell}} = \frac{1}{n} \sum_{j=1}^{r+1} \sum_{i=w_{j-1}}^{w_j} (\widehat{R}_i - \overline{\widehat{R}}_j)^2.$$

We then choose the model with the smallest value of $\mathcal{G}(\widehat{\ell})$, out of all of the nested models produced by Algorithm 1. We remark that the use of ranks ensures $\mathcal{G}(\widehat{\ell})$ is still robust.

In order to make a practical recommendation for the parameter α, we rely on the simulation study. Figure 2.3 shows the empirical root mean squared error of $\widehat{\ell}$ under Algorithm 1 for a range of α values. Each curve is for a different combination of parameters in the first simulation scenario. The depth function used was Mahalanobis depth, which was chosen for computational ease. Figure 2.3 shows that choosing α in the range $(0.75, 1.25)$ works well.

For consistency of Algorithm 2 to hold, the penalty term should satisfy $\beta_n \to \infty$ and $\beta_n = o(n)$; this gives a wide range of choices for the penalty term. In practice what penalty term should be used? The results of the simulation study suggested using a penalty term of the form $\beta_n = C_1\sqrt{n} + C_2$. Figure 2.2 plots the empirical root mean squared errors of $\widehat{\ell}$ under spatial depth for different values of C_1 and n, with C_2 fixed at 3.74. Each curve represents one combination of the parameters in the first simulation scenario described

above. Notice that the curves are not shifting laterally as n increases, meaning that an increase in n is sufficiently captured by the $\sqrt{n}$ term in the penalty. Additionally, Figure 2.2 also shows a flattening of the RMSE curves with increased n, which is expected from the consistency theorem. Based on low root mean squared error in simulation, we recommend to fix $C_2 = 3.74$ and run Algorithm 2 for a grid of penalties defined by $C_1 \in (0.15, 0.25)$. One can then choose the set of change-points according to a model selection criteria or by visual inspection. In the simulation study, we fix C_1 at 0.18.

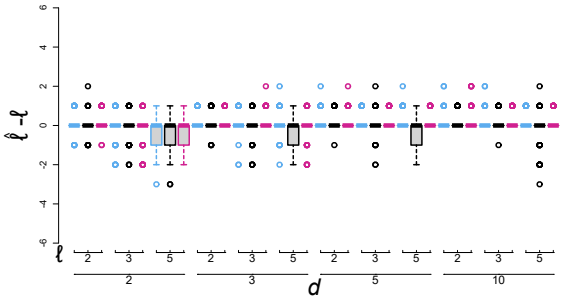
It should also be noted that a non-linear penalty could be applied, as discussed in Killick et al. (2012). Some non-linear penalties were tested in the simulation study, such as $\log \ell$, but the results were not as good as when using a linear penalty. Our investigation into non-linear penalties was fairly limited, as such, more investigation into non-linear penalties could be done in the future.

2.5.2 Analysing and comparing the algorithm performance

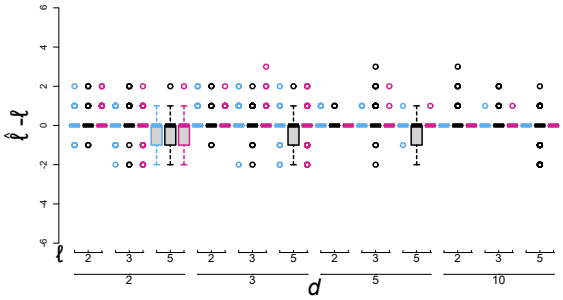
We only present the results of the simulation when the algorithms were paired with spatial depth. Spatial depth was the best performing depth function when taking into account computational speed and estimation accuracy. Halfspace depth performed similarly to spatial depth, but computationally it was much slower. Halfspace depth is affine-invariant whereas spatial depth is only similarity-invariant, therefore if affine invariance is desired in the analysis the analyst should use the halfspace depth function. Modified Mahalanobis depth and Mahalanobis depth both performed slightly worse than the other two depth functions if the data was Gaussian, but performed considerably worse when the distribution of the data was Cauchy, which could be attributed to the robustness considerations of Mahalanobis depth discussed in Section 1.1.

Figure 2.4 shows boxplots of $\widehat{\ell} - \ell$ under Algorithm 1, Algorithm 2, the WBSIP algorithm and the `e.divisive` algorithm, for the first simulation scenario with $n = 1000$. Recall that $\widehat{\ell} - \ell$ is the estimated number of change-points minus the actual number in the simulation run. Each boxplot represents a different combination of simulation parameters, e.g., the first boxplot represents the (empirical) distribution of $\widehat{\ell} - \ell$ with simulated two-dimensional Gaussian data that had 2 change-points. The empirical distributions are computed over the 100 replications of each combination of simulation parameters.

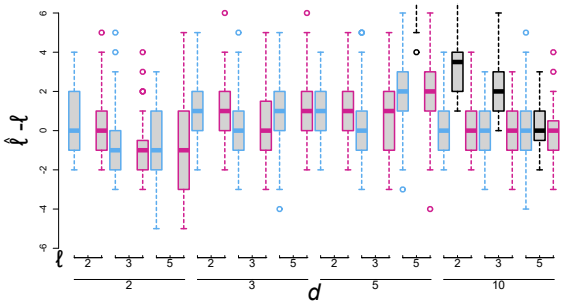
Figure 2.4 shows that both Algorithm 1 and Algorithm 2 estimate the number of change-points more accurately than the WBSIP algorithm and the e-divisive algorithm. One reason for this is that our method is less general in terms of the types of changes it can detect when compared to the other two algorithms; our method trades generality for accuracy.



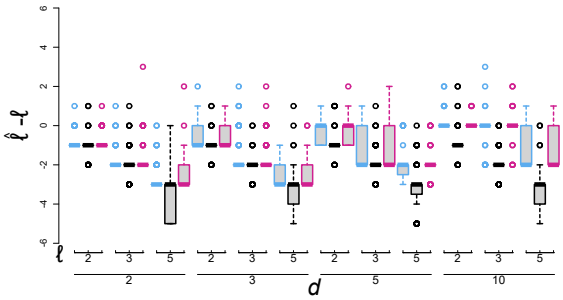
(a) WBS algorithm



(b) KW-PELT algorithm

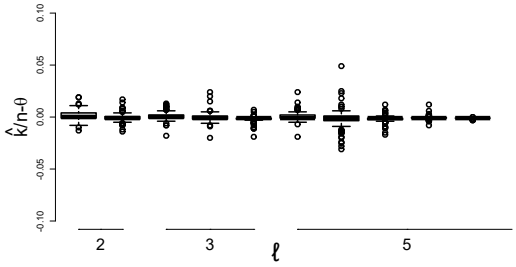


(c) WBSIP algorithm

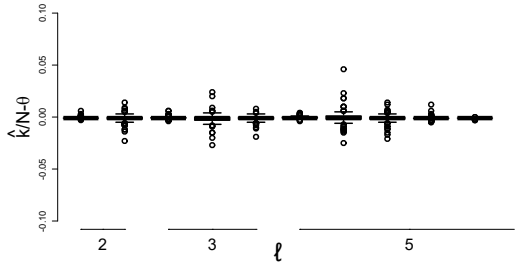


(c) e-divisive algorithm

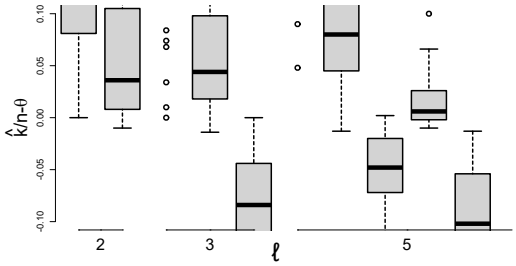
Figure 2.4: Boxplots of $\widehat{\ell} - \ell$ for the rank-basd algorithms (row 1) and the competing algorithms (row 2) with $\alpha = 0.9$, $C_1 = 0.18$ and $C_2 = 3.74$ when $n = 1000$. Each boxplot represents the values of $\widehat{\ell} - \ell$ for a particular simulation parameter combination. Here, the color of the boxplot represents the distribution. The top number on the horizontal axis represents the number of true change-points and the bottom number on the horizontal axis represents the dimension. The colors follow the legend of Figure 2.2.



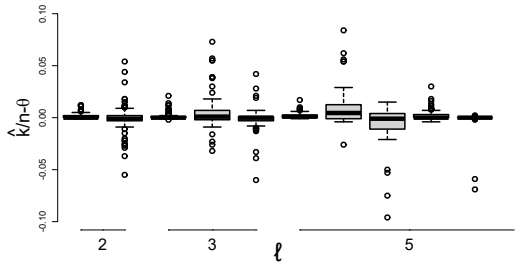
(a) WBS algorithm



(a) KW-PELT algorithm

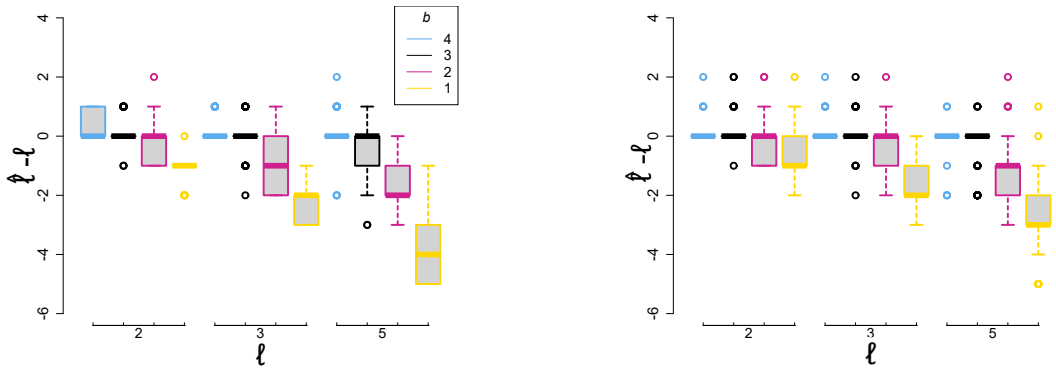


(a) WBSIP algorithm



(a) e-divisive algorithm

Figure 2.5: Boxplots of $\widehat{k}/n - \theta$ for the different algorithms with $\alpha = 0.9$, $C_1 = 0.18$ and $C_2 = 3.74$. The distribution of the data was made up of independent normal marginals with $d = 10$. Each boxplot represents the ability to estimate a particular change-point for a fixed number of true change-points. For example, the first two boxplots are the empirical distributions of $\widehat{k}_1/n - \theta_1$ and $\widehat{k}_2/n - \theta_2$ when there were two true change-points. The numbers on the horizontal axis represent the number of true change-points in that simulation parameter combination. Each boxplot represents the ability to estimate a single change-point in a run.



(a) WBS algorithm

(b) KW-PELT algorithm

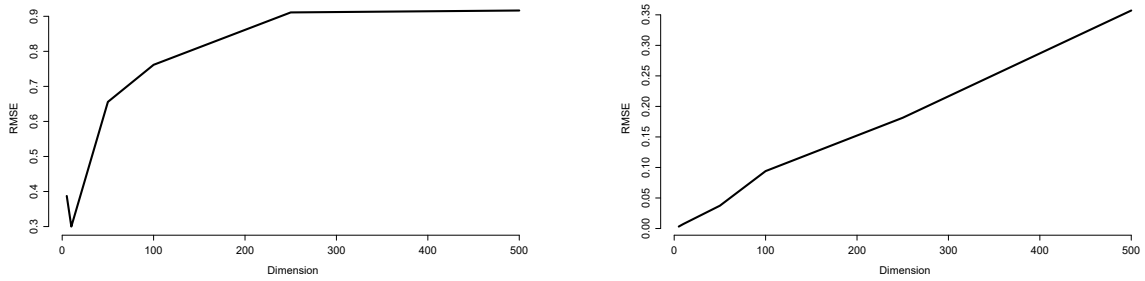
Figure 2.6: Boxplots of $\widehat{\ell} - \ell$ for the third simulation scenario under spatial depth, where the colours indicate the different values of b, the size of the submatrix to which the expansion/contraction was applied. In other words, the submatrix in which an expansion or contraction was applied had dimension $b \times b$. The numbers on the horizontal axis represent the number of true change-points in the simulation.

A second reason for these results is robustness. When the data is Cauchy, neither of the competing algorithms perform very well. Neither the WBSIP nor the e.divisive algorithm are designed to handle outliers or heavy-tailed data. For example, when the data have Cauchy marginals, the assumptions for consistency of the WBSIP procedure are violated. One should also recall that WBSIP was designed for high dimensional data which is not the main focus of this simulation study. In Figure 2.4 it is easily seen that as the dimension increases WBSIP performs better. Speaking of dimension, both Algorithms 1 and 2 were very insensitive to the dimension and the number of change-points.

In terms of accuracy when the change-point was detected, both algorithms performed very well. Figures 2.5 shows boxplots of $\widehat{k}/n - \theta$ for Algorithm 1, Algorithm 2, WBSIP and the e-divisive algorithm when d was 10 and the distribution was normally distributed. Boxplots under other simulation parameters were similar.[1] Generally, the estimates were at most about 5% off of the true break fraction, with the majority of biases being in the 1% range.

This corresponds to 10 time units away when $n = 1000$; $\widehat{k}$ was typically within 10 time

[1]They were of course worse when the data came from a Cauchy distribution.



(a) (b)

Figure 2.7: (a) Empirical root mean square error of $\widehat{\ell}$ as the dimension increases, but the size of the change remains fixed. (b) Empirical root mean square error of $\widehat{k}$ as the dimension increases, but the size of the change remains fixed. The change-points were estimated using the KW-PELT algorithm paired with spatial depth.

units of the true change-point k. Again our methods were insensitive to the dimension and the number of change-points. When compared with the WBSIP algorithm and e-divisive algorithm, we see that both Algorithm 1 and Algorithm 2 appear to estimate the location of the change-points more accurately. We can then conclude that our algorithms outperforms the WBSIP algorithm and e-divisive algorithm when the data only contains changes in variability.

In a different simulation scenario, we fixed $d = 5$ and applied the expansions and contractions of the first scenario to a $b \times b$ submatrix of the covariance matrix. Figure 2.6 shows boxplots of $\widehat{\ell} - \ell$ for both algorithms, under this simulation scenario. Figure 2.6 shows the results for spatial depth, with $n = 1000$. The colour of the boxplot in Figure 2.6 represents b, the size of the submatrix to which the expansion/contraction was applied, i.e., the submatrix in which an expansion or contraction was applied had dimension $b \times b$. The numbers on the horizontal axis represent the number of change-points in that particular simulation scenario. We see that as b decreases, the ability to detect the changes decreases. This is expected, since a smaller change should be more difficult to detect. We can also see that here, the KW-PELT algorithm performs better than the WBS algorithm. One remedy for detecting changes in relatively low dimensions might be to subsample dimensions of the data and run the procedure on each of the subsampled dimensions. We leave that for future work.

Lastly, we ran simulations designed to assess the performance of our methods in high dimensions. We use the KW-PELT algorithm with spatial depth, due to both its perfor-

mance and the fact that spatial depth can be computed quickly in high dimensions. We simulated normal data, with one change-point and with two change-points, at $n = 1000$ for $d = 50$ and $d = 500$. The KW-PELT algorithm estimated both the number of change-points and the location of the change-points were detected with 100% accuracy. This is not surprising since we might view an expansion of a very large matrix as a very large change in variability. For example, the trace of the expanded matrix is increasing as the dimension is increased, and so the signal is increasing with the dimension under an expansion-type change. The story changes if the data is high dimensional and the data is sparse, i.e., the change only occurs in a submatrix of the covariance matrix which has a fixed dimension. We ran another simulation where there was a single change-point, and the change only occurred in a 5×5 submatrix. Figure 2.7 shows that as we increase the dimension, the algorithm has a more difficult time estimating the change-point accurately. This suggests that when the data is suspected to be very sparse, we may wish to develop a depth function that accounts for sparsity.

In summary, both algorithms performed very well relative to competitors in this simulation set-up. The results also show that the rank based WBS algorithm and the KW-PELT algorithm are very comparable. The KW-PELT algorithm is computationally faster, and can be more accurate under sparsity. However, its tuning parameter requires some subjectivity. Furthermore, both algorithms have the same theoretical rate of convergence. We are tempted to recommended the KW-PELT algorithm with the understanding that the performance of the algorithms is very similar. In terms of the depth functions, halfspace depth and spatial depth performed better than the two Mahalanobis depth variants. Since halfspace depth takes longer to compute, we would ultimately recommend using spatial depth with either algorithm in practice.

2.6 An application to financial returns

In this section we apply the methodology to four daily stock returns. `R` codes for this analysis can be found on Github at (Ramsay, 2019b). We analyze the same data set analyzed by Galeano and Wied (2017) and compare our results to those produced by their method. It is expected that algorithms will produce different results, due to the fact that the aim of Galeano and Wied (2017) was to detect changes in the correlation structure of the returns; not necessarily the covariance matrix. For example, they assume constant variances over time. The results should be seen as complementary to those of Galeano and Wied (2017).

It is clear that this data has some serial dependence; it does not fit the independence

Change-point WBS	Change-point KW-PELT
Jul 18 '07	Jul 26 '07
Sep 05 '08	Sep 25 '08
Dec 08 '08	Dec 08 '08
May 01 '09	May 19 '09
Aug 25 '09	ND
ND	Jul 22 '10
Jul 25 '11	Jul 25 '11

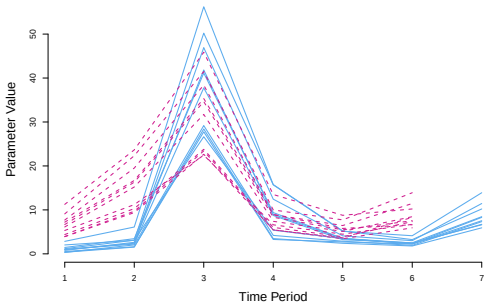


Figure 2.8: *Left*: Change-points estimated by Algorithms 1 and 2. ND stands for not detected by the Algorithm. *Right*: Covariance matrix parameters at each interval for both Algorithm 1 (pink dashed) and Algorithm 2 (blue solid) connected by lines to emphasize the change in the parameter values.

assumption. That being said, we feel that the results still provide some insight into the data. For example, the data appears to admit a weak dependence structure. As a result of the concentration inequality for rank statistics for m-dependent data (Wang et al., 2019), we only need Assumption 2 to hold under m-dependence in order for the consistency properties to hold. In fact, the consistency of many depth functions is, in part, a result of Glivenko-Cantelli type theorems. Seeing as extensions of such theorems exist for m-dependent data (Bobkov and Götze, 2010) it is likely possible to extend the results of Section 2.4. The convergence of depth functions for dependent data is an interesting topic for further research.

We applied Algorithm 1 and Algorithm 2 to the raw daily returns. We ran the WBS algorithm with 700 intervals ($100\lfloor \log n \rfloor$) using all depth functions with $\alpha = 0.9$. When running Algorithm 2, we used penalty constants $C_1 = 0.24$ and $C_2 = 3.74$, these were chosen according to the discussion in Section 2.5.1. The results did not vary at all among the different depth functions for the Algorithm 1, and were virtually the same under Algorithm 2, the only difference was that the modified Mahalanobis depth predicted the December 2008 change-point on December 9th rather than on the 8^{th}.

Table 2.8 contains the estimated change-points produced by the Algorithms. Figure 2.9 plots the estimated change-points on the data from both Algorithms. Observe that algorithms are also both unaffected by the outliers in the Siemens returns, which can be seen to the left and right of January 2008. Some of the change-points have a clear interpretation. For example, the first change-point (July 18, 2007) signifies the beginning

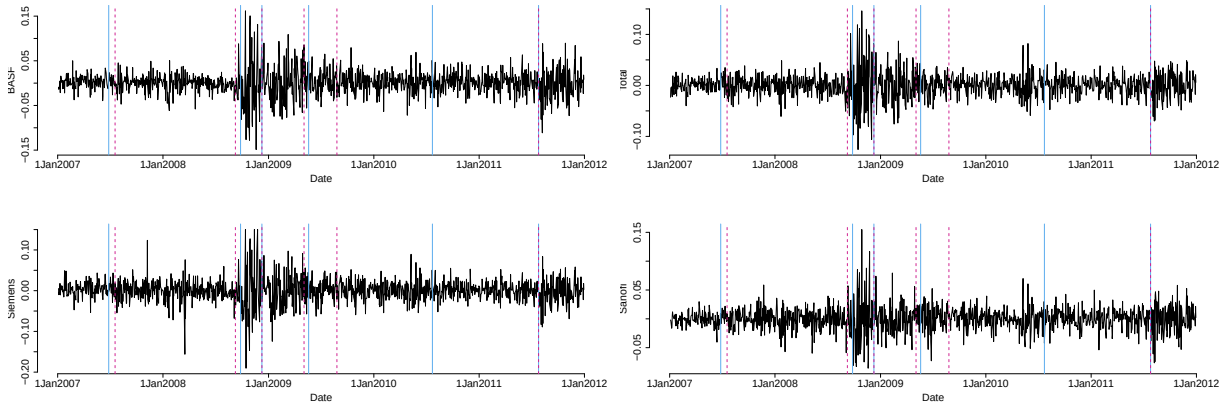


Figure 2.9: Returns with estimated change-points from Algorithm 2 marked by solid, blue lines and Algorithm 1 marked by dashed pink lines.

of the global financial crisis and the second (September 05, 2008) is associated with the collapse of Lehman brothers. In the following months, measures to stem the effects of the crisis may contribute to the next two change-points. For example, in early December 2008 the EU agreed to a 200 billion dollar stimulus package. The later change-points are associated with the Greek government debt crisis; in July 2011, the Troika approved a second bailout (of the Greek government).

The algorithms reproduced both change-points found by (Galeano and Wied, 2017) (July 18, 2007 and September 05, 2008). Changes in correlation could be accompanied by expansions or contractions in the covariance matrix of these returns. It is possible that these changes (correlation and covariance) are byproducts of a general increase/decrease in systematic volatility. Many financial returns are generally thought to have some systematic/market-wide dependence (Bodie et al., 2017). Figure 2.8 shows the estimated pairwise covariances as well as the estimated variances of each stock within each period of 'no change'. The uniform movement of the parameters indicate contractions and expansions, rather than some other type of change. Additionally, we note that all changes under Algorithm 1 were significant when the Bonferoni correction was applied to the set of test statistics at the 5% level of significance.

Chapter 3

Kruskal-Wallis type statistics covariance kernel testing problems

3.1 Introduction

Data such that the observations are each a smooth curve, deemed functional data, is being increasingly observed in a variety of fields. For example, medical images (López-Pintado and Wrobel, 2017; Aston et al., 2017), intraday financial asset returns (Cerovecki et al., 2019) and environmental "omics" data (Piña et al., 2018) can all be interpreted as functional data. As such, many functional analogues of univariate and multivariate statistical tools are needed. One such tool is the notion of common variance in the functional context; common covariance operator or covariance kernel. In this chapter, we introduce new, non-parametric functional k-sample tests for equality of covariance structures. We call this class of tests the functional Kruskal-Wallis tests for covariance structure, or for short, FKWC tests.

Before introducing the FKWC tests, we review the existing, related works. Early related works include (James and Sood, 2006), who presented a test for a difference in the shape of the mean function between two populations of curves and (Gabrys and Kokoszka, 2007; Aue et al., 2009b; Horváth et al., 2010) who all consider tests related to serial dependence, or time series characteristics.

Panaretos et al. (2010) were the first to discuss comparing the covariance structures of two populations of functional data. They present a two sample test based on the Hilbert-Schmidt norm for integral operators and restrict their attention to that of Gaussian processes. Fremdt et al. (2013) later extended the methods of Panaretos et al. (2010)

to compare two populations of non-Gaussian data. In a related work, Jarušková (2013) proposed a modification of the test of Panaretos et al. (2010), used for covariance operator change-point detection. Zhang and Shao (2015) re-normalized the test statistic of Panaretos et al. (2010) to account for dependence in the data. Gaines et al. (2011) later proposed a test for equality of two covariance operators based on univariate likelihood ratios and Roy's union intersection principle.

Up until this point, existing tests were based on the Hilbert-Schmidt metric. Pigoli et al. (2014) presented a discussion of distances between covariance operators, including criticisms of using finite dimensional distances on functional data. They argued that using a Hilbert-Schmidt metric ignores the geometry of the space of covariance kernels, and therefore is not an appropriate distance. As a result, they introduced a two sample permutation procedure, which Cabassi et al. (2017) later extended to the multi-sample case. In the same vein of resampling, Paparoditis and Sapatinas (2016) proposed a k-sample bootstrap test that can detect differences in the mean and/or the covariance structure simultaneously.

Guo and Zhang (2016) further studied a multi-sample test, which was first proposed by Zhang (2013). When the data comes from a Gaussian process, under the null hypothesis, their test statistic is a χ^2-type mixture. This distribution must be approximated in practice. They also provided a random permutation method to be used in the case of small samples and/or non-Gaussian data. Guo et al. (2018) developed a k-sample test inspired by functional ANOVA. One feature of this test is that it does not require some form of dimension reduction. Further, their method is scale invariant in the sense that re-scaling the data at any time t does not affect the test statistic. Similar to that of Guo and Zhang (2016), the distribution of the test statistic under the null hypothesis must to be estimated. The estimation of the critical value relies on parameters estimated from the data, which may pose problems if the data are contaminated.

Boente et al. (2018) studied a new type of bootstrapping method to calibrate critical values for the covariance kernel testing problem. They focused on norms between covariance operators and the resulting distribution under the null hypothesis is based on eigenvalues of fourth moments, which must be estimated. They suggested bootstrapping the eigenvalues of fourth moment operators. This can be problematic if the data is heavy tailed or contaminated.

Kashlak et al. (2019) provided a concentration inequality based analysis of covariance operators, which includes a k-sample test and a classifier. They used concentration results to develop confidence sets based on p-Schatten norms. They then used 'tuned' confidence sets to define rejection regions for k-sample tests. This test tends to underestimate the

confidence level in the case where the data is heavy tailed.

Some other related works include the following: López-Pintado and Wrobel (2017) used a version of band depth defined on images to test for a difference in dispersion between two sets of images. The measure of dispersion ignores shape or 'wigglyness' differences between the two samples. Sharipov and Wendler (2019) extended the bootstrapping procedures of Paparoditis and Sapatinas (2016) to change-point problems and dependent data. Rice and Shum (2019) introduced a test for change-points in the cross-covariance operator of two functional time series. Flores et al. (2018) presented a test for homogeneity of two distributions based on depth measures. They explicitly stated that the paper was not focused on means or covariance operators. They provided four test statistics, based on the deepest functions or absolute values of differences in the depth distributions. Aston et al. (2017) focused on testing for a condition called separability which is specific to hypersurface data such as f-MRI.

FKWC tests have several advantages over these other methods. First, FKWC tests are very robust, a feature that has not often been discussed in other works. FKWC tests are based on rank statistics, generated via functional data depth measures. Functional data depth measures are, among other things, used for outlier detection and trimmed means; data depth measures are designed to produce robust inference procedures. A test statistic based on ranks of data depth measures would thus, inherent the robustness properties of both the depth measure and of rank statistics in general. We demonstrate this robustness via simulation in Section 3.4.

Aside from being robust, many functional data depth measures are invariant under certain transformations of the data (Gijbels and Nagy, 2017). If the functional observations are all scaled by an arbitrary function, we would like the test statistic to remain unchanged. Guo et al. (2018) points out that many existing tests for equality of covariance structures are not invariant under this type of transformation, e.g., those of (Panaretos et al., 2010; Guo and Zhang, 2016). On the contrary, many data depth measures remain unchanged if the data are scaled by an arbitrary function. Such invariance properties are then inherited by the FKWC test statistic, provided derivatives are not included in the calculation of the depth, see Section 1.2. If derivatives are included, the FKWC test satisfies a weaker form of transformation invariance.

Furthermore, using data depth measures allows us to leverage existing consistency results (Nagy and Ferraty, 2019) and provide asymptotic analysis of the FKWC tests under both the null and alternative hypotheses. We show that under the null hypothesis the test statistic is a chi-squared random variable. This is a particularly nice feature, as it circumvents the need to estimate the distribution of the test statistic under the

null hypothesis using the data, which many other tests require, e.g., (Pigoli et al., 2014; Paparoditis and Sapatinas, 2016; Cabassi et al., 2017; Guo et al., 2018; Boente et al., 2018; Kashlak et al., 2019). To elaborate, for many of the existing methods, under the null hypothesis, the theoretical distribution of the test statistic contains unknown parameters that must be estimated from the data. If this is not the case, then most existing methods require data-driven bootstrap or permutation methods. Using the data to estimate the distribution of the test statistic under the null hypothesis can be complicated when the data are contaminated. This can also be computationally expensive if resampling methods are used.

Not only is there no need to use the data to estimate the distribution of the FKWC test statistic under the null hypothesis, there is also no need to estimate the sample covariance operators in the computation of the FKWC test statistic. This fact implies that one does not need to reduce the dimension of the data via truncated basis expansions of the covariance kernels, as is needed by the methods of (Fremdt et al., 2013; Pigoli et al., 2014; Paparoditis and Sapatinas, 2016; Boente et al., 2018). An additional byproduct of avoiding estimation of the covariance operators is that we do not require finite fourth moments or any fourth moment related assumptions for our theoretical analysis. Such assumptions are required for many other tests, for example, (Panaretos et al., 2010; Gaines et al., 2011; Fremdt et al., 2013; Paparoditis and Sapatinas, 2016; Guo et al., 2018; Boente et al., 2018) all require some type of fourth moment assumption on the data.

In terms of the alternative hypothesis, we show that under some mild conditions, the FKWC tests are consistent under a wide class of alternatives. We also provide a method for estimating the power and sample size under general alternatives. Some recent works have explored various local alternatives for this testing problem (Gaines et al., 2011; Guo and Zhang, 2016; Guo et al., 2018; Boente et al., 2018). We also provide a class of local alternatives under which a particular FKWC test is consistent. This FKWC test is based on a new depth measure L^2-root depth, for which we prove several elementary properties. This depth measure has a particular interpretation in this testing problem, which provides the basis for its development.

The rest of the chapter is organised as follows. Section 3.2 covers the methodology of the hypothesis tests, including the data model and the intuition behind the test statistic. Section 3.3 presents asymptotic results on the behaviour of the test statistic under both the null and alternative hypotheses. Section 3.4 presents a simulation study, in which we compare the FKWC tests to some competing tests, including those of (Guo et al., 2018; Boente et al., 2018). The last section, Section 3.5 shows an application of the FKWC test to two data sets. We first compare several samples of intraday financial asset return curves. Here, we test to see if the residuals of a functional GARCH model have similar

covariance structure. We next analyse speech recognition data, where the observations are log periodograms of five groups of recorded syllables. We perform FKWC multiple comparisons on these data to determine which pairs of syllables are similar in terms of covariance structure.

3.2 Model and test statistic

Suppose that we have observed J independent, random samples and that for each sample $j \in \{1, \ldots, J\}$ we have $X_{j1}, \ldots, X_{jn_j}$ functional observations. The combined sample size is then $n = \sum_{j=1}^{J} n_j$, where we assume that $n_j/n \to \vartheta_j$ as $n \to \infty$. We define 'functional' observations by the following assumptions: First, we assume that each X_{ji} is a mean square continuous stochastic process, meaning for each $t \in [0,1]$, $X_{ji}(t)$ is measurable with respect to some probability space $(\Omega, \mathscr{A}, P)$ and that $\lim_{t \to s} \mathrm{E}\left[|X(t) - X(s)|^2\right] = 0$. Secondly, for each $\omega \in \Omega$, $X_{ji}(\cdot, \omega)$ is a continuous function. We use $\mathfrak{F}$ to denote the space of such processes. These assumptions imply that $X_{ji}(t, \omega)$ is jointly measurable with respect to the product σ-field $\mathscr{B} \times \mathscr{A}$, where $\mathscr{B}$ denotes the Borel sets of $[0,1]$. This joint measurability implies that each X_{ji} can be interpreted as a random element which lies in $\mathscr{L}^2([0,1], \mathscr{B}, \mu)$, where μ is the Lebesgue measure on $[0,1]$. We will write $\mathscr{L}^2([0,1], \mathscr{B}, \mu)$ as $\mathscr{L}^2$ for brevity. For more details see Appendix B or Chapter 7 of Hsing and Eubank (2015). We also assume that $\mathrm{E}\left[X_{ji}\right] = \mathbf{0}$ where $\mathbf{0}$ is the zero function. If necessary, in practice, data can be centered by a robust estimator of the mean. Some variants of the proposed test involve derivatives, and for those tests we will additionally require that the derivative of X_{ji}, by which we denote $X_{ji}^{(1)}(t)$, exists on the interval $(0,1)$ and satisfies the same continuity assumptions imposed on X_{ji}. We remark that our methods extend to higher dimensional domains and range, i.e., $X_{ji} \colon [0,1]^d \to \mathbb{R}^p$ on which functional data depths are defined, but we restrict our study to the univariate domain and range setting.

The covariance kernel of a mean square continuous stochastic process X, whose mean is $\mathbf{0}$, is defined as

$$\mathcal{K}(s,t) \coloneqq \mathrm{E}\left[X(t)X(s)\right]$$

and the associated covariance operator is defined as

$$(\mathscr{K} f)(t) \coloneqq \int_{[0,1]} f(s)\mathcal{K}(s,t)ds \ .$$

The goal is to construct a test statistic for testing the following hypothesis

$$H_0 \colon \mathcal{K}_1 = \cdots = \mathcal{K}_J \qquad \text{v.s.} \qquad H_1 \colon \mathcal{K}_j \neq \mathcal{K}_k, \text{ for some } j \neq k.$$

Here $\mathcal{K}_j$ refers to the covariance kernel of group j. The assumptions on the random functions X_{ji} imply that this is equivalent to the hypothesis

$$H_0 \colon \mathscr{K}_1 = \cdots = \mathscr{K}_J \qquad \text{v.s.} \qquad H_1 \colon \mathscr{K}_j \neq \mathscr{K}_k, \text{ for some } k \neq j,$$

where $\mathscr{K}_j$ is the covariance operator of group j. We can denote the probability measure over $\mathscr{L}^2$ which describes the random behaviour of X_{ji} by F_j. Let

$$F_* \coloneqq \vartheta_1 F_1 + \vartheta_2 F_2 + \vartheta_3 F_3 + \cdots + \vartheta_\ell F_\ell + \vartheta_J F_J,$$

which is a mixture of probability measures over $\mathscr{L}^2$. Alternatively, this can be interpreted in the stochastic process sense, such that the finite dimensional distributions of an element from the combined sample X: $(X(t_1), \ldots, X(t_k))$ for $k \in \mathbb{N}$, are mixtures of the J finite dimensional distributions of each group, with weights $\{\vartheta_j\}_{j=1}^J$. Note that these finite dimensional distributions can be identified by F_* and F_j.

Typically, testing this hypothesis involves estimating $\mathcal{K}_j$. In order for estimates of $\mathcal{K}_j$ to converge weakly, it is typically required that $\mathrm{E}\left[\|X_{ji}\|^4\right] < \infty$, in some cases it is not desirable to make this assumption. Estimation of $\mathcal{K}_j$ is also high dimensional, and can be computationally intensive if repeated a number of times, such as in a bootstrap procedure. We take a different approach, and do not aim to estimate $\mathcal{K}_j$. Instead, the idea is to reduce each observation to a one dimensional rank via a data driven ranking function. The ranking function is designed such that differences in the samples' mean ranks are implied by differences in underlying covariance kernels. We can then use the classical Kruskal Wallis test statistic (Kruskal, 1952) and perform a rank test. Specifically, the test statistic proposed is

$$\widehat{\mathcal{W}}_n \coloneqq \frac{12}{n(n+1)} \sum_{j=1}^{J} n_j \left(\overline{\widehat{R}}_j - \frac{n+1}{2} \right)^2.$$

Here, $\overline{\widehat{R}}_j$ is the mean rank of the observations in group j, where the ranking mechanism will be explained in the next section. This test statistic also gives, for each sample j, a measure of how much its covariance kernel differs from the average sample kernel via

$$\left(\overline{\widehat{R}}_j - \frac{n+1}{2} \right)^2.$$

Alternatively, we can perform FKWC multiple comparisons, see Section 3.5.2.

We can further modify this test statistic using the methods of Gastwirth (1965), who presented powerful versions of several univariate rank tests. These more powerful tests were later extended to multivariate, depth-based rank tests (Chenouri et al., 2011). We further extend these methods to the functional setting. The percentile modification is predicated on the fact that it is actually the extreme rank values that allow us to detect differences between samples. The idea is to remove the middle portion of the data, and only use the outlying data or, equivalently, the low depth-based ranks. To this end, let $r \in (0,1)$ and let $n' = \lfloor rn \rfloor$. Let $\delta_j(s) = 1$ if the observation which has rank equal to s is in group j and let $\delta_j(s) = 0$ otherwise. Define the percentile modified test statistic as

$$\widehat{\mathcal{M}}_{n,r} \coloneqq \sum_{j=1}^{n} \left(1 - \frac{n_j}{n}\right) K_j \qquad \text{with} \qquad K_j = \frac{1}{\sigma_j^2} \left(\sum_{s=1}^{n'} (n' - s + 1)\delta_j(s) - \varrho_j \right)^2, \tag{3.1}$$

where

$$\varrho_j = \frac{n_j n' (n'+1)}{2n} \qquad \text{and} \qquad \sigma_j^2 = \frac{n_j (n - n_j) n' (n'+1) [2n(2n'+1) - 3n'(n'+1)]}{12n^2(n-1)}.$$

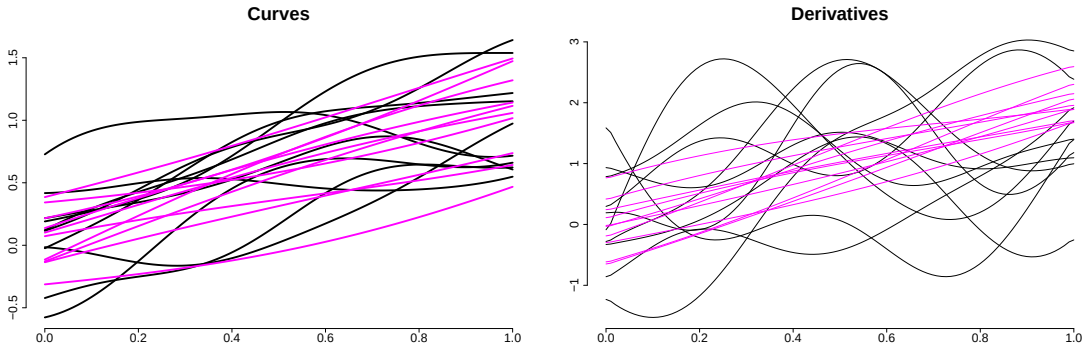
Choosing r is a matter of simulation and will be taken up in Section 3.4.

Like in Chapter 2, we use ranks based on data depth measures. In this setting, we use ranks based on the functional depth measures discussed in Chapter 1. Precisely, for some $j \in \{1, \ldots, J\}$ and some $i \in \{1, \ldots, n_j\}$, define the sample depth-based rank of X_{ji} to be

$$\widehat{R}_{ji} \coloneqq \#\{X_{\ell m} \colon \mathrm{D}(X_{\ell m}; F_{*,n}) \leq \mathrm{D}(X_{ji}; F_{*,n}),\ \ell \in \{1, \ldots, J\},\ m \in \{1, \ldots, n_\ell\}\}.$$

Note that in this chapter we rank the observations with respect to $F_{*,n}$, which places equal weight on each element of the combined sample. This means that the sample depth values describe centrality with respect to the combined sample. Recall that in the setting of functional data, central relates to both the location and the shape of the data.

Indeed, differences between the covariance kernels are often exhibited by changes in the shape and/or scale of the data, precisely the features captured by functional data depth functions. For example, Figure 3.1 shows two samples of 10 Gaussian processes and their derivatives. Each sample has the same mean but a different covariance kernel. Visually, the distinguishing factor between these two samples is the scale and shape of the curves and their derivatives. Notice that the difference is more pronounced in the derivatives. In fact, depth measures have already been shown to have good power for detecting scale changes



(a) (b)

Figure 3.1: Two samples of Gaussian processes (a) and their derivatives (b) each with the same mean but a different covariance kernel. The samples have an exponential covariance kernel with $\alpha = 0.3$ in the first sample and $\alpha = 1$ in the second sample (see Section 3.4). Notice that the difference in covariance structure is exhibited by a changed in shape of the original curves and a change in shape and scale of the derivatives.

in surface data (López-Pintado and Wrobel, 2017). In fact, it appears that ranks based on depth measures capture second order differences better in the functional setting than in the multivariate setting. For example, recall from Chapter 2 that we cannot detect changes in the rotation matrix of the covariance matrix. This is not the case in the functional setting; the FKWC test based on the random projection depth can detect a difference of the type $\mathscr{K}_1 = \mathscr{U}\mathscr{K}_2\mathscr{U}^*$ for some unitary operator $\mathscr{U}$, and $\mathscr{U}^*$ denotes the adjoint of $\mathscr{U}$.

In addition to the random projection depth, we combine our methodology with multivariate halfspace depth and the modified band depth. As discussed in Chapter 1, these depth measures that meet criteria suitable for the hypothesis testing problem. Using the ranks of the norms of the observations is another natural approach to this problem. However, the scope of differences captured by the norms of the observations is limited. We will see in Section 3.3 that some of the existing functional depth functions capture a much wider scope of differences. Nevertheless, we compare using the depth-based ranks to ranking the norms of the observations.

We can actually show that ranking the squared norms is a depth-based rank. If we define depth as follows:

$$\mathrm{LTR}(x; F) = \left(1 + \mathrm{E}_F\left[\|x - X\|^2\right]^{1/2}\right)^{-1}, \tag{3.2}$$

then the ranks from this depth function are equivalent[1] to ranking the squared norms, provided the observations have mean equal to zero. We call this depth L^2-root depth, or LTR depth for short. We demonstrate that LTR depth actually measures depth in Section A.3. Framing norm-based ranks as depth-based ranks, allows us to easily extend the norm-based ranks to include the derivative information. L^2-root depth can be extended to account for p derivatives with

$$\mathrm{LTR}_p(x;F) = \left(1 + \frac{1}{p}\sum_{k=0}^{p} \frac{1}{\mathrm{MAD}_k} \mathrm{E}_F\left[\left\|x^{(k)} - X^{(k)}\right\|^2\right]^{1/2}\right)^{-1},$$

where MAD_k is the median absolute deviation with respect to the law of the norm of the k^{th} derivative of X which has law F.

3.3 Theoretical results

This section is devoted to characterizing the behaviour of the FKWC tests under the null and alternative hypotheses when the sample size is large. Note all proofs of theorems presented in this section can be found in Section A.5. We also remind the reader that throughout the chapter, including in all theorem statements, we have assumed that the observations have zero mean and that $n_j/n \to \vartheta_j$ as $n \to \infty$.

Theorem 3. *Suppose that* $\Pr(\widehat{R}_{ji} = \widehat{R}_{k\ell}) = 0$ *for all* $ji \neq k\ell$*, i.e., there are no tied ranks. Then, under the null hypothesis*

$$\widehat{\mathcal{W}}_n \xrightarrow{d} \chi^2_{J-1} \text{ as } n \to \infty \qquad \text{and} \qquad \widehat{\mathcal{M}}_{n,r} \xrightarrow{d} \chi^2_{J-1} \text{ as } n \to \infty.$$

Tied ranks can be randomly broken to meet the requirements of Theorem 3. The behaviour under the null hypothesis is remarkably simple for such a complex testing problem. Therefore, the critical values can easily be obtained independently of the data and thus, this aids accuracy and computation time. The fact that the critical values are independent of the data is important for robustness, as there is no need to assess the robustness of procedures used to approximate the null distribution. Under the alternative hypothesis, we must impose additional assumptions in order to have consistency of the test.

[1] By equivalent, we mean that inferences produced by each set of these ranks are equivalent. In actuality, the ranks produced by the LTR depth function are the reverse of the ranks generated from the squared norms; an LTR rank is equal to n minus the rank generated from the norms.

Assumption 7. *For all j, it holds that $\Pr(\mathrm{D}(X_{j1}; F_*) \leq v)$, as a function of v, is a Lipschitz function.*

Assumption 8. *It holds that $\mathrm{E}\left[\sup_{x\in\mathfrak{F}} |\mathrm{D}(x; F_{n,*}) - \mathrm{D}(x; F_*)|\right] = O(n^{-1/2})$.*

Assumption 9. *Let $F_{*,u}$ be the distribution of $\langle X, u\rangle$ if $X \sim F_*$. It holds that $\mathrm{E}\left[\|X_{ji}\|^3\right] < \infty$. In addition, it holds that for any u, $F_{*,u}$ is three times differentiable, and that the first three derivatives of $F_{*,u}$ are bounded functions in u. We will denote the density of $F_{*,u}$ by $f_{*,u}$.*

Assumption 10. *Suppose that*

$$\mathrm{E}\left[\mathrm{D}(X_{j1}; F_*) - \mathrm{D}(X_{k1}; F_*)\right] \neq 0 \implies \mathrm{MED}\left(\mathrm{D}(X_{j1}; F_*) - \mathrm{D}(X_{k1}; F_*)\right) \neq 0,$$

for $j \neq k$ and $j, k \in \{1, \ldots, J\}$.

Assumption 7 is generally satisfied when the finite dimensional distributions corresponding to F_* are continuous. Assumption 8 has been shown to be satisfied for MFHD, see (Nagy and Ferraty, 2019). We can extend the results of Nagy and Ferraty (2019) to RP depth. Recall from Chapter 1 that $S = \{u \in \mathfrak{F}\colon \|u\| = 1\}$ and that F_u is the univariate distribution associated with $\langle X, u\rangle$ is $X \sim F$. Suppose that the unit vectors $u_1, \ldots, u_{M_n}$ in the definition of RP_{M_n} (see (1.5)) are drawn from some probability measure ν on S and that $M_n = O(n)$. Recall that

$$\mathrm{RP}(x; F) = \int_S F_u(\langle x, u\rangle)(1 - F_u(\langle x, u\rangle))d\nu(u). \tag{3.3}$$

Then it follows easily that[2],

$$\mathrm{E}\left[\sup_{x\in\mathfrak{F}} |\,\mathrm{RP}_{M_N}(x; F_N) - \mathrm{RP}_\infty(x; F)|\right] = O(N^{-1/2}). \tag{3.4}$$

Note that the same analysis applies when $F_u(\langle x, u\rangle)(1 - F_u(\langle x, u\rangle))$ is replaced with $1/2 - |1 - F_u(\langle x, u\rangle)|$ in (3.3). Assumption 8 need not be satisfied for LTR depth since the sample ranks are already based on $\mathrm{LTR}(\cdot; F)$. Assumption 9 is essentially a smoothness condition on the projected distributions. In addition, we require the existence of a third moment. For example, Assumption 9 is satisfied by Gaussian processes. Lastly, we impose Assumption 10. This means that the distribution of the differences of the depth values

[2]For proof see Appendix A.5

is not of the special kind where the mean is non-zero but the median is zero. This is a common assumption for a rank test; for example, for the Wilcoxon Rank-Sum test to consistently detect a difference in mean between two distributions, the distribution of the differences must have a non-zero median. Now, if there exists $k \in \{1, \ldots, J\}$ such that

$$\sum_{j=1}^{J} \vartheta_j \Pr\left(\mathrm{D}\left(X_{k1}; F_*\right) > \mathrm{D}\left(X_{j1}; F_*\right)\right) \neq \frac{1}{2}, \tag{3.5}$$

then, under Assumptions 7-8, it holds that

$$\Pr\left(\widehat{\mathcal{W}}_n > \delta\right) \to 1, \text{ as } n \to \infty, \tag{3.6}$$

for any $\delta > 0$. This result shows that the set of alternative hypotheses which induce consistency are contained completely in by (3.5). Firstly, for the FKWC test to detect a difference in covariance operator, we must have that a difference in covariance operator between groups implies that if there is a difference in the location of the depth values between groups.[3] Specifically, the median of $\mathrm{D}\left(X_{\ell 1}; F_*\right) - \mathrm{D}\left(X_{j1}; F_*\right)$ is non-zero for some $\ell \neq j$. The condition (3.5) has an additional caveat that says the differences between the groups do not perfectly 'cancel' each other out. This caveat applies to all tests based on the Kruskal-Wallis statistic, and is a minor technicality. This is not an issue if $J = 2$.

As previously mentioned, we must demonstrate that changes in the covariance kernels produce, on average, a location difference in the depth values. This can be argued qualitatively, seeing as changes in the covariance operator elicit changes in the shape or magnitude of the data. Since depth measures rate the observation on how close it is in shape and magnitude to the combined sample, then it is intuitive that observations with a different covariance kernel will have different depth values.

Beginning with the two sample case, we analyse the relationship between the covariance operator and the depth values for the RP depth. Let the sample rank of X_{ji} based on the true distribution F_* be defined as

$$R_{ji} := \#\left\{X_{\ell m} : \mathrm{D}\left(X_{\ell m}; F_*\right) \leq \mathrm{D}\left(X_{ji}; F_*\right), \ \ell \in \{1, \ldots, J\}, \ m \in \{1, \ldots, n_j\}\right\}.$$

Then we define $\mathcal{W}_n$ as the test statistic based on these ranks, which are unknown except in the special case of the L^2-root depth-based ranks. Let $\mathscr{G}$ be the set of all cumulative distribution functions on $\mathbb{R}$ that are three times differentiable. Then, define $\mathcal{H}\colon \mathscr{G} \to \mathbb{R}$ as $\mathcal{H}(F) := \frac{1}{2} f^{(1)}(0) - (F(0) f^{(1)}(0) - f^2(0))$.

[3] We only need at least one pair of groups to differ in depth location.

Theorem 4 (Random Projection Depth)**.** *Suppose that* $J = 2$*,* $u_1, \ldots, u_{M_n}$ *are drawn independently from a probability measure* ν *on* S *and that* $M_n = O(n)$*. Suppose that Assumptions 7-10 hold for* RP *and define*

$$\mathcal{R}_j = \frac{1}{6}\int_S \mathrm{E}\left[\int_0^{\langle X_{j1},u\rangle} \left(f^{(2)}_{u,*}(t)(1-2F_{u,*}(t)) - 6f_{u,*}(t)f^{(1)}_{u,*}(t)\right)\left(\langle X_{j1},u\rangle - t\right)^3 dt\right] d\nu(u).$$

Then FKWC test based on $\widehat{\mathcal{W}}_n$ *using* RP_{M_n} *with* $\mathrm{D}(z;F) = F(z)(1-F(z))$ *is consistent in the sense of* (3.6) *under alternatives of the form*

$$H_1\colon \int_S \mathcal{H}(F_{*,u})\langle \mathscr{K}_1 u, u\rangle d\nu(u) + \mathcal{R}_1 \neq \int_S \mathcal{H}(F_{*,u})\langle \mathscr{K}_2 u, u\rangle d\nu(u) + \mathcal{R}_2.$$

Remark 1. *The interpretation of* H_1 *being true is that on average (according to* ν*),* $\langle \mathscr{K}_1 u, u\rangle$ *differs from* $\langle \mathscr{K}_2 u, u\rangle$*, which can only be true if the covariance kernels are different between the groups. Indeed, the other terms* $\mathcal{H}$ *and* $\mathcal{R}_j$ *are minor nuisances. For example, we expect* $\mathcal{R}_j$ *to be small in this context because* X_{j1} *have zero mean; the expected length of the interval* $(0, \langle X_{j1}, u\rangle)$ *is 0. We discuss* $\mathcal{H}$ *below.*

The function $\mathcal{H}(F_{u,*})$ is a weighting function which has two components. First, there is a departure from symmetry term: $f^{(1)}_u(0)(\frac{1}{2} - F_u(0))$. This component will only play a role in the procedure the skewness of the projected distributions varies considerably across directions. In fact, as $F_{u,*}$ approaches symmetry, $\mathcal{H}(F_{u,*})$ approaches $f^2_{u,*}(0)$, the squared height of the projected density at zero. This is the second component of $\mathcal{H}(F_{u,*})$, which says that $\mathcal{H}(F_{u,*})$ gives more weight to directions where the height of the projected density is large at the mean, thus, directions with low spread are magnified. Therefore, the integral in H_1 can be described as a weighted average of the projections $\int\int \mathcal{K}(s,t)u(s)(t)dsdt$. Symmetry of the projected distributions implies a form of symmetry on F_*. This is a natural definition of symmetry, in the sense that it is analogous to halfspace symmetry in the multivariate setting (Zuo and Serfling, 2000a). For example, consider the case where $X_{ji} = g_j(t)Z_{ji}$ for some deterministic functions g_1 and g_2, with and $Z_{ji} \sim \mathcal{N}(0,1)$. If the groups are equal sized it follows that

$$F_{*,u}(x) = \frac{1}{2}\Phi\left(\frac{x}{\langle g_1, u\rangle}\right) + \frac{1}{2}\Phi\left(\frac{x}{\langle g_2, u\rangle}\right),$$

where Φ is the probit function. Therefore, $\mathcal{H}(F_{u,*}) \propto \left(\frac{1}{\langle g_1,u\rangle} + \frac{1}{\langle g_2,u\rangle}\right)$, and it follows that

if

$$\int_S \left(\langle g_1, u \rangle + \frac{\langle g_1, u \rangle^2}{\langle g_2, u \rangle} \right) d\nu \neq \int_S \left(\langle g_2, u \rangle + \frac{\langle g_2, u \rangle^2}{\langle g_1, u \rangle} \right) d\nu,$$

which is equivalent to

$$\int_S \left(\sqrt{\langle \mathscr{K}_1, u \rangle} + \frac{\langle \mathscr{K}_1, u \rangle}{\sqrt{\langle \mathscr{K}_2, u \rangle}} \right) d\nu \neq \int_S \left(\sqrt{\langle \mathscr{K}_2, u \rangle} + \frac{\langle \mathscr{K}_2, u \rangle}{\sqrt{\langle \mathscr{K}_1, u \rangle}} \right) d\nu,$$

then the procedure will be consistent[4].

Theorem 4 highlights an important difference between the functional data setting and the multivariate setting. Suppose that $\mathscr{K}_1$ is unitarily equivalent to $\mathscr{K}_2$, i.e., $\mathscr{K}_1 = \mathscr{U}\mathscr{K}_2\mathscr{U}$ for some unitary operator $\mathscr{U}$. Assume that our samples are multivariate: $X_i \in \mathbb{R}^b$. The covariance operators can then be represented by the covariance matrices Σ_1, Σ_2. Unitary equivalence in the multivariate setting corresponds to $\Sigma_1 = \mathcal{U}\Sigma_2\mathcal{U}^\top$ for a rotation matrix $\mathcal{U}$. Letting $\tilde{u} = u^\top \mathcal{U}$, where one notes that $\|\tilde{u}\| = 1$, it is then easy to see that

$$\int_{S^{b-1}} \mathcal{H}(F_{*,u}) u^\top \Sigma_1 u du = \int_{S^{b-1}} \mathcal{H}(F_{*,u}) \tilde{u}^\top \Sigma_2 \tilde{u} du \approx \int_{S^{b-1}} \mathcal{H}(F_{*,\tilde{u}}) \tilde{u}^\top \Sigma_2 \tilde{u} du,$$

which is why we are unable to detect changes characterized by rotations of the data via data depth functions based on projections in the multivariate setting. This equivalence does not exist in the functional setting, since the measure ν is not uniform on S. Therefore, it is not necessarily true that

$$\int_S \mathcal{H}(F_{*,u}) \langle \mathscr{K}_1 u, u \rangle d\nu(u) \approx \int_S \mathcal{H}(F_{*,u}) \langle \mathscr{K}_2 u, u \rangle d\nu(u).$$

We verify this fact in Section 3.4; in the eigenvalue simulation scenarios 1-3, the covariance operators differ between the samples, but are unitarily equivalent. We see that the FKWC test based on the RP depth detects the difference.

Theorem 5 (L^2-root Depth)**.** *Suppose Assumptions 7 and 10 hold and that $J = 2$. Then the test based on $\mathcal{W}_n$, using ranks based on the squared norms, is consistent in the sense of (3.6) under alternatives of the form H_1: $\|\mathscr{K}_1\|_{TR} \neq \|\mathscr{K}_2\|_{TR}$, where $\|\cdot\|_{TR}$ refers to the trace norm.*

Note that, by assumption, $\mathscr{K}_j$ are trace class since the observed processes are mean

[4]Provided the remainder is small.

square continuous; the kernel is continuous. Theorem 5 shows that the FKWC test using the ranks of the squared norms is consistent if the trace norm of the covariance operators differ. The alternative hypothesis $\|\mathscr{K}_1\|_{TR} \neq \|\mathscr{K}_2\|_{TR}$ is equivalent to $\sum_{k=1}^{\infty} \lambda_{k,1} \neq \sum_{k=1}^{\infty} \lambda_{k,2}$, where $\{\lambda_{k,1}\}_{k=1}^{\infty}$ and $\{\lambda_{k,2}\}_{k=1}^{\infty}$ are the decreasing sequences of singular values resulting from the singular value decomposition of $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively.

It is useful to mention a few cases where Assumption 10 is surely satisfied. If, for all j, $\|X_{ji}\|^2$ has a symmetric distribution then Assumption 10 is satisfied. If

$$\|\mathscr{K}_1\|_{TR} - \|\mathscr{K}_2\|_{TR} \geq \mathrm{V}ar\left(\|X_{11}\|^2\right)^{\frac{1}{2}} + \mathrm{V}ar\left(\|X_{21}\|^2\right)^{\frac{1}{2}}$$

then Assumption 10 is satisfied (Page and Murty, 1982). If the distribution of the squared norms is instead unimodal, Basu and DasGupta (1997) gives a sharper bound:

$$\|\mathscr{K}_1\|_{TR} - \|\mathscr{K}_2\|_{TR} \geq \left(\frac{3}{5}\left(\mathrm{V}ar\left(\|X_{11}\|^2\right) + \mathrm{V}ar\left(\|X_{21}\|^2\right)\right)\right)^{\frac{1}{2}}.$$

If X_{ji} are Gaussian processes $\mathcal{GP}(0, \mathscr{K}_j)$, then we have that

$$\|X_{11}\|^2 = \sum_{k=1}^{\infty} \lambda_{k,1} V_k \qquad \text{and} \qquad \|X_{21}\|^2 = \sum_{k=1}^{\infty} \lambda_{k,2} V_k', \quad \text{where } V_k, V_k' \overset{iid}{\sim} \chi_1^2.$$

It follows that the random variables $\|X_{11}\|^2$, $\|X_{21}\|^2$ are stochastically ordered, implying that Assumption 10 is satisfied.

To give more insight into the behaviour of the test statistic, we provide some analysis under local alternatives. Suppose Assumption 8 and (3.5) hold. A direct application of a result in Fan et al. (2011) implies that $\mathcal{W}_n$ is approximately distributed as a non-central chi-squared random variable $\chi^2_{J-1}(\tau_n)$ with non-centrality parameter

$$\tau_n = \frac{12}{n(n+1)} \sum_{j=1}^{J} n_j \left\{ n \sum_{k \neq j} \vartheta_k \left(\Pr\left(\mathrm{D}\left(X_{j1}; F_*\right) \leq \mathrm{D}\left(X_{k1}; F_*\right)\right) - \frac{1}{2} \right) \right\}^2. \tag{3.7}$$

For L^2-root depth, we are able to compute $\Pr(\mathrm{D}\left(X_{k1}; F_*\right) \leq \mathrm{D}\left(X_{j1}; F_*\right))$, which is the same quantity as $\Pr(\|X_{k1}\| \leq \|X_{j1}\|)$. Therefore one can compute the power and consequently

sample sizes for any assumed F_*, using

$$\Pr\left(\sum_{m=1}^{p}\left[\left\|X_{k1}^{(m)}\right\| - \left\|X_{j1}^{(m)}\right\|\right] \leq 0\right).$$

This could of course be done by Monte Carlo simulation for complicated models.

Theorem 6 (Local Alternative Analysis L^2-root depth)**.** *Suppose that for all i, j, k, ℓ*

$$\|X_{ji}\|^2 \stackrel{d}{=} \left[\frac{\sqrt{n}+\delta_k}{\sqrt{n}+\delta_j}\right]\|X_{k\ell}\|^2 \sim G, \tag{3.8}$$

where G has a continuously differentiable density g for some real-valued δ_js. Let $\overline{\delta} = \sum_{j=1}^{J}\vartheta_j\delta_j$ then, when based on the ranks of the squared norms,

$$\mathcal{W}_n \xrightarrow{d} \chi^2_{J-1}(\tau) \text{ with non-centrality parameter } \tau = 12\left(\int_{\mathbb{R}} zg(z)^2dz\right)^2\sum_{j=1}^{J}\vartheta_j\left(\delta_j-\overline{\delta}\right)^2.$$

Note that (3.8) holds when $\mathscr{K}_j = \mathscr{K}_0\left[1+n^{-1/2}\delta_j\right]$ and $\|X_{ji}\|^2$ form a scale family. For example, if $X_{ji} \sim \mathcal{GP}(0, \mathscr{K}_0\left[1+n^{-1/2}\delta_j\right])$, then Theorem 6 is applicable.

Due to the fact that Euclidean spaces are Hilbert spaces, the previous results provide some consequences for similar methods based on depth-based ranks in the multivariate setting. For example, Theorem 5 provides justification for assuming the hypothesis of Theorem 2 of Chenouri et al. (2020b) as well as for Assumption 4 of Chapter 2, when the ranks are based on L^2-root depth. Note that the definition of spatial depth in (Chenouri et al., 2020b) provides equivalent ranks to those of L^2-root depth. Theorem 5 then implies that the methods of Chapter 2 can detect changes in the sum of the eigenvalues of the covariance matrix. Similarly, Theorem 4 provides justification for assuming the hypothesis of Theorem 2 of (Chenouri et al., 2020a) and Assumption 4 of Chapter 2 under integrated dual depth. Liu and Singh (2006) provide a k-sample test for the covariance matrix of multivariate data. Theorems 5, 4 and 6 give analogous results for this multivariate k-sample test.

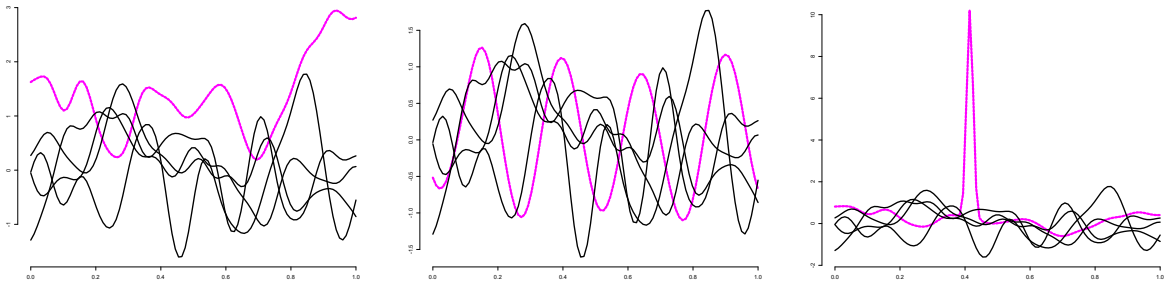


Figure 3.2: Five observations generated from the uncontaminated distribution compared to a drift outlier (left), wavy outlier (middle) and scale outlier (right).

3.4 Simulation results

3.4.1 Models and settings

In this section we evaluate the finite sample performance of the FKWC tests on both infinite and finite dimensional models. We compare the performance of the tests using the different functional depth functions, as well as the effect of the percentile modification discussed in Section 2.3. We further compare the FKWC test against seven other tests: the test of Boente et al. (2018) *Boen*, the L^2-norm tests (Guo and Zhang, 2016) *Tmax*, *L2nv*, *L2br*, *L2rp* and the ANOVA inspired tests of Guo et al. (2018) *GPFnv*, *GPFrp*, *Fmax*. For the test of Boente et al. (2018) we used 10 principal components and 5000 bootstrap samples. For the tests of Guo and Zhang (2016); Guo et al. (2018) we used 1000 permutations.

We simulated data from both infinite and finite dimensional models. For the infinite dimensional models, we simulated observations from $J = 2$ and $J = 3$ samples. The results from the simulations with three groups were the same as with two groups, and are omitted. We tested sample sizes of $n = 100$, $n = 200$ and $n = 500$ for the two sample case, where the first sample size was $n_1 = \lfloor qn \rfloor$ for $q \in \{0.2, 0.3, 0.4, 0.5\}$. For the three sample case, we used $n = 150$ and $n = 300$, with $n_1 = N_3$ and $n_2 = \lfloor qn \rfloor$ for q as above. In each infinite dimensional case, the data were sampled from either a Gaussian process $\mathcal{GP}$, a student-t process with three degrees of freedom t_3, or a skewed Gaussian process $\mathcal{SG}$. For the infinite dimensional runs we used a squared exponential covariance kernel

$$\mathcal{K}(s,t;\alpha,\beta) = \beta \ \exp\left(\frac{-(s-t)^2}{2\alpha^2}\right),$$

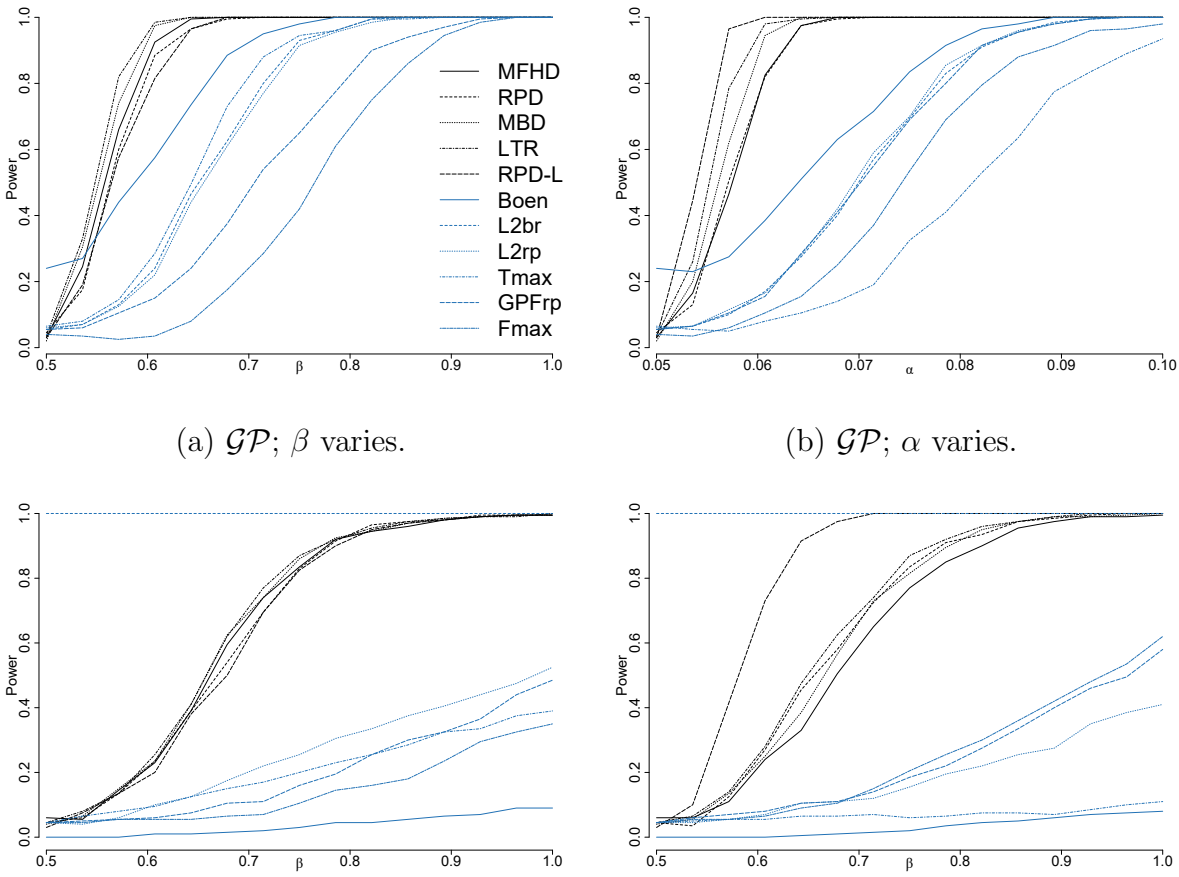
		Gaussian			Student t			Skewed Gaussian		
	n_1/n :	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
FKWC	MFHD	1.00	1.00	1.00	0.87	0.95	0.98	1.00	1.00	1.00
	RP	1.00	1.00	1.00	0.88	0.94	0.98	1.00	1.00	1.00
	MBD	1.00	1.00	1.00	0.88	0.96	0.98	1.00	1.00	1.00
	LTR	1.00	1.00	1.00	0.88	0.96	0.98	1.00	1.00	1.00
	$RP^{\dagger}$	1.00	1.00	1.00	0.87	0.94	0.96	1.00	1.00	1.00
Competing	Boen	1.00	1.00	1.00	0.17	0.15	0.04	0.99	1.00	1.00
	L2br	0.48	0.93	0.99	1.00	1.00	1.00	0.49	0.94	0.99
	L2rp	0.35	0.92	0.99	0.00	0.00	0.03	0.36	0.93	0.99
	Tmax	0.68	0.97	0.99	0.01	0.01	0.06	0.63	0.97	1.00
	GPFrp	0.03	0.56	0.91	0.01	0.00	0.04	0.01	0.54	0.87
	Fmax	0.12	0.47	0.83	0.00	0.01	0.10	0.11	0.51	0.85

Table 3.1: Empirical power of the different tests for $J = 2$, $n = 500$ when the group sample sizes were unequal, under scale differences when $\beta_1 = 0.5$ and $\beta_2 = 0.71$. Notice that when the sample sizes differ greatly, the competing tests do not perform as well. Note that $RP^{\dagger}$ is the FKWC test with the likelihood depth, rather than the simplicial depth.

and the sample differences were controlled via α (shape difference) and β (scale difference).

For the finite dimensional models, the data were simulated from a Gaussian process where we directly specified K non-zero eigenvalues of the covariance operator. A Fourier basis was used for the eigenfunctions. Here, we only tested the case of two samples and we used the same sample sizes as described above for the infinite dimensional simulation scenarios. We ran six scenarios, which resulted from combining short linear, long linear, and long exponential eigenvalue decay with either a scale difference or a unitary operator difference. See Appendix A.4 for details.

We also simulated the effects of different kinds of outliers on the different tests at different levels of contamination. The contamination level was measured as a percentage of the total sample size n and we tested levels of 0.01, 0.025, 0.05. We present the results from 2.5%, since these are the most illustrative. The three kinds of outliers were: linear drift, oscillating outliers and spike outliers, see Figure 3.2. We simulated the case where there were outliers in both samples as well as the case where there were outliers in only one sample. Lastly, we considered the affect of missing portions of the curves and the affect of the number of directions on the random projection depth. The results of these simulations can be seen in Appendix A.4.



(a) $\mathcal{GP}$; β varies.

(b) $\mathcal{GP}$; α varies.

(c) Student-t_3; β varies.

(d) Student-t_3; α varies.

Figure 3.3: Empirical power curves of the different tests as the β parameter (left) and α parameter (right) of the second sample moves away from the null hypothesis. Here, $n_j = 100$. The black curves correspond to the FKWC tests and the blue curves correspond to the existing tests.

	Simulation Scenario					
Test	1	2	3	4	5	6
MFHD	0.90	**1.00**	**1.00**	0.99	**1.00**	**1.00**
RP	0.92	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
MBD	0.80	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
LTR	0.79	**1.00**	**1.00**	**1.00**	**1.00**	0.99
$RP^{\dagger}$	0.26	0.90	0.86	**1.00**	**1.00**	**1.00**
Competing	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

Table 3.2: Empirical power of the different tests under the finite dimensional models when $n_1 = n_2 = 100$. The first row indicates the scenario number. "Competing" stands for the competing tests; all competing tests had the same empirical power. Note that $RP^{\dagger}$ is the FKWC test with the likelihood depth, rather than the simplicial depth.

We ran each simulation scenario 200 times, a grid size of 100 was used to simulate the functions. The codes can be seen on Github (Ramsay, 2019a). Note that the FKWC tests require the observations to be on the same grid in order to compute the sample depth values. If the observations are not observed on the same grid, then it will be necessary to interpolate the curves in some way, such that they can be brought to the same grid. For example, one can smooth the curves and then re-discretize them if necessary. To test the affect of missing portions of the curve, we interpolated the data with splines using the `zoo` package in `R`.

3.4.2 Results

Unanimously, the methods that incorporated the derivatives performed better than the non-derivative methods. This includes when the data contained outliers or had unequal group sizes. Also, one should note that with the exception of RP depth, the non-derivative methods do not work for shape differences. In addition, the percentile modification had little effect on the power and size in all simulation runs. Therefore, we proceed by only presenting the results from the FKWC methods based on $\mathcal{W}$ that incorporated the derivatives. In terms of which depth function was the best, they were all relatively similar in all respects. That being said, the LTR and the MBD had the most power for detecting scale differences, see Figure 3.3. Figure 3.3 also shows that for detecting shape differences, the random projection depth with the likelihood depth has the highest power, especially under heavy tails. The LTR depth and the MBD depth also performed well under shape differences. Under the very low dimensional models (three non-zero eigenvalues), the random

	Test	Gaussian			Student t			Skewed Gaussian		
		50	100	250	50	100	250	50	100	250
FKWC	MFHD	0.06	**0.03**	0.06	**0.03**	0.06	0.07	0.08	0.07	0.06
	RP	0.07	**0.04**	0.06	**0.04**	**0.04**	0.06	**0.05**	0.06	0.06
	MBD	**0.05**	**0.02**	**0.04**	**0.03**	**0.04**	0.08	0.06	0.06	0.06
	LTR	**0.05**	**0.04**	0.06	**0.04**	**0.04**	0.08	0.06	0.06	**0.04**
	RP†	0.06	**0.03**	0.07	**0.04**	**0.03**	0.07	0.07	**0.04**	0.06
Competing	Boen	0.21	0.24	0.34	**0.00**	**0.00**	**0.00**	0.16	0.21	0.32
	L2br	**0.04**	0.06	0.06	1.00	1.00	1.00	**0.04**	**0.04**	0.08
	L2rp	0.06	0.06	0.06	**0.05**	**0.04**	**0.04**	**0.04**	**0.04**	0.07
	Tmax	**0.04**	0.06	0.07	**0.04**	**0.04**	**0.04**	**0.04**	**0.03**	0.06
	GPFrp	0.06	0.06	0.06	0.06	**0.04**	**0.04**	**0.05**	**0.03**	0.06
	Fmax	**0.03**	**0.04**	**0.05**	**0.04**	**0.04**	**0.03**	0.06	**0.02**	0.06

Table 3.3: Empirical sizes for $J = 2$ for different tests under the infinite dimensional models. The first row indicates the underlying process and the second row indicates the sample size of each group. Note that RP† is the FKWC test with the likelihood depth, rather than the simplicial depth.

projection depth with the likelihood depth did not work as well as the others. The best performing depth functions in this scenario were the random projection depth with the simplicial depth and the multivariate half-space depth, see Table 3.2. In terms of when the group sizes differed or outliers were present, the depths were all comparable, see Table 3.1 and Figure 3.4. Computationally, the LTR and the modified band depth were the fastest, followed by the random projection depths and lastly the multivariate half-space depth. Overall, we recommend using the random projection depth with simplicial depth because of its theoretical interpretation, or, if computation is a concern then the modified band depth or the LTR depth can be used.

We now compare the FKWC tests to the competing tests. Note that the naive versions of the $L2$ and GPF tests are omitted, since their performance was similar to that of their biased-reduced counterparts. In addition, we only present the results from the FKWC tests based on the derivatives without the percentile modification. For the infinite dimensional models Figure 3.3 shows the power of the FKWC tests compared to the competing tests under the Gaussian and Student t processes for the two-sample case when $n_1 = n_2 = 100$. Note that the skewed Gaussian results were similar to the Gaussian results, and the results at other sample sizes were similar to Figure 3.3. Clearly, under the infinite dimensional models, the FKWC tests outperform the competing tests, especially when the distribution

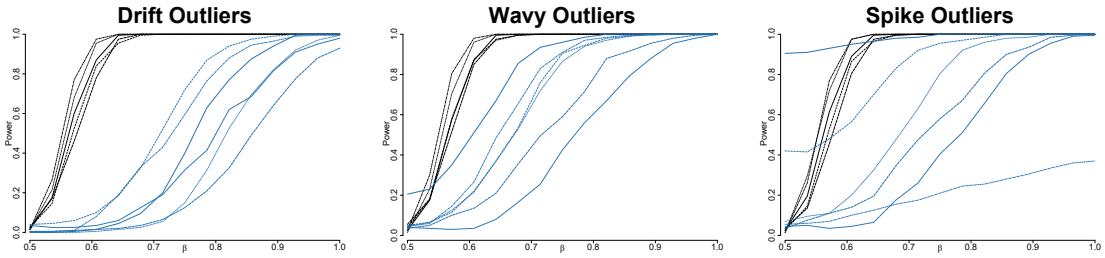


Figure 3.4: Empirical power of the tests for detecting scale differences for $J = 2$ with $n_1 = n_2 = 100$ under the infinite dimensional Gaussian process model such that 5% each of the sample was contaminated. Note that the legend follows that of Figure 3.3. The black curves correspond to the FKWC tests and the blue curves correspond to the competing tests.

is heavy tailed. In addition, the heavy tails corrupt the *L2pr* test. Table 3.1 shows the power of the tests for a fixed scale difference between the samples. Notice that many of the competing tests do not work well under very imbalanced group sizes, whereas the FKWC tests are unaffected. The tables for imbalanced group sizes with a shape change or under the finite dimensional models are given in Appendix A.4. Table 3.2 shows the power of the tests for the finite dimensional models. The competing tests all had power equal to 1, this was also generally the case for the FKWC tests except in one case. When the model had three non-zero eigenvalues, such that the covariance operators in the first and second sample were unitarily equivalent, the FKWC test paired with the likelihood depth variant of RP depth did not perform well. We can conclude that competing tests work better for very low-dimensional, light tailed models. These results could be due to the fact that the lower complexity allowed for easier approximation of the null distributions. Table 3.3 shows the size of the tests under the infinite dimensional models, note that the sizes of both the FKWC tests and the competing tests are relatively close to 0.05, with the exception of the *Boen* test and the biased-reduced L2 test.

As mentioned previously, we also simulated models where portions of the data were contaminated with outliers. Figure 3.2 shows the three different kinds of outliers the data were contaminated with. Figure 3.4 shows the power of the tests for detecting Gaussian scale differences for $J = 2$ with $n_1 = n_2 = 100$ under 5% contamination of each of the samples. For reference, this is 5 curves in in the context of Figure 3.4. Note that we expect contamination in both samples to reduce the power of the test. The drift and spike outliers have the most negative effects on the power of the competing tests. When there are drift or spike outliers in both samples, we see that the power of the competing tests is reduced,

	MFHD	RP	MBD	LTR	RP†	L2br	L2rp	Tmax	GPFrp	Fmax
Drift	0.04	0.04	0.04	0.05	0.03	0.81	0.47	0.25	0.36	0.09
Wavy	0.04	0.06	0.04	0.04	0.03	0.08	0.10	0.06	0.10	0.08
Spike	0.04	0.06	0.04	0.05	0.02	0.29	0.08	0.06	0.05	0.03

Table 3.4: Sizes of the tests when one sample was contaminated by the different kinds of outliers. Specifically, 5% of the first sample was contaminated. We see that the size of the competing tests is inflated in almost all scenarios. Note that RP† is the FKWC test with the likelihood depth, rather than the simplicial depth.

while the power of the FKWC tests is roughly the same as it was for uncontaminated data.

Consider the case where the contamination is only in one sample. We expect contamination in only one samples to increase the size of the test. Table 3.4 shows the size of the tests when one sample was contaminated by the different kinds of outliers. Notice that the size of the competing tests is inflated in almost all of the scenarios, even under the wavy outliers. The drift outlier seems to have especially negative effects on the competing tests, with the exception of the *Fmax* test. Note that we have omitted the *Boen* test in Table 3.4, as it did not perform well on the uncontaminated data. In conclusion, the FKWC tests perform very well compared to the existing tests. This is especially true in the presence of imbalanced groups, heavy tails and drift or spike type outliers. As mentioned previously, we recommend always incorporating the derivative information in the test, and for choosing a depth function we recommend using the random projection depth with simplicial depth because of its theoretical interpretation and simulation performance, or, if computation is a concern then the modified band depth or the LTR depth should be used.

3.5 Applications to real data

In this section we present an application of our methodology to two different functional datasets. One is comprised of intraday stock prices and the other is comprised of digitized speech.

3.5.1 F-GARCH residual analysis of intraday stock price curves

We analyse the daily asset price curves of $J = 3$ different stocks (`twtr`, `fb` and `snap`) starting on June 24th 2019 and ending March 20th 2020, which gives $n_1 = 207$ and $n_2 =$

$n_3 = 208$. Precisely, for each stock the price was measured over the course of the trading day in one minute intervals, for a total of 390 minutes per day. In order to account for edge effects from smoothing the curves, we trimmed 10% of the minutes from the beginning of the day and 5% of the minutes from the end of the day. This resulted in 332 minutes of stock prices. In actuality, we analysed the log returns, viz.

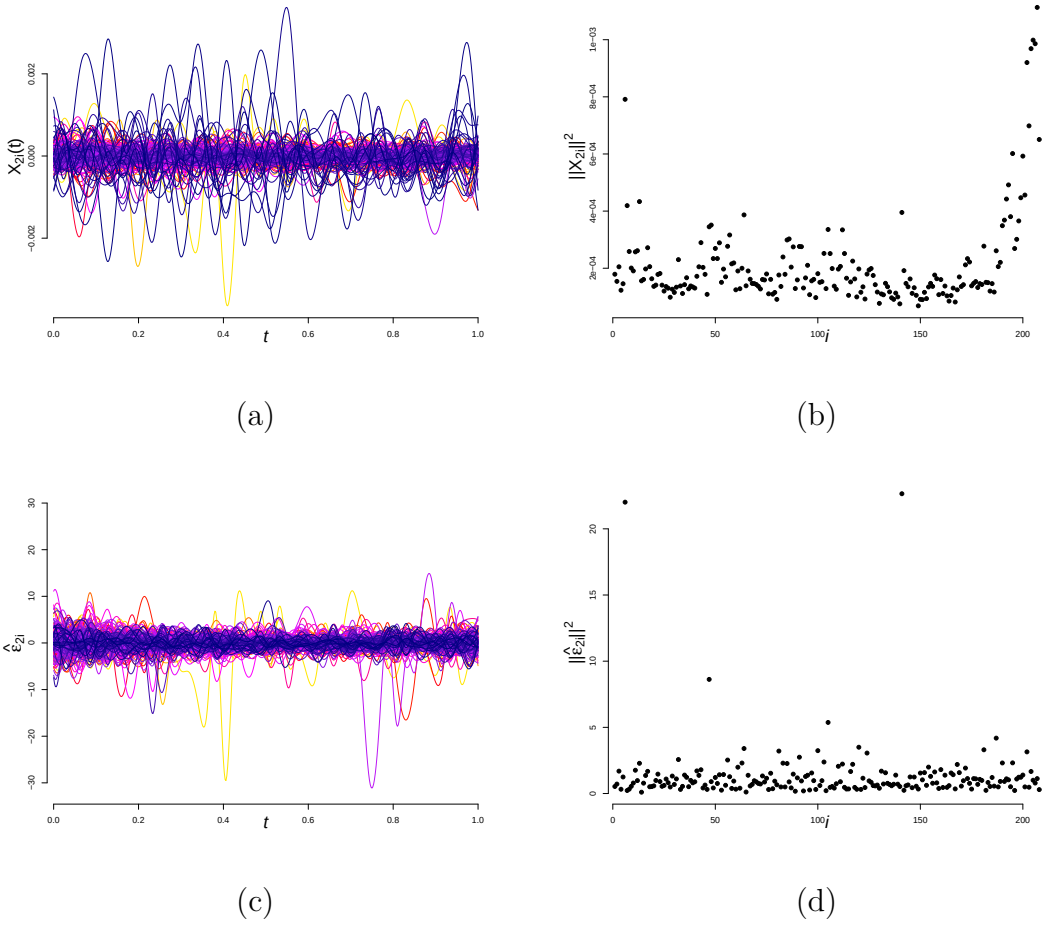
$$X_{ji}(t) = \ln(Y_{ji\lfloor 331t \rfloor + 1}) - \ln(Y_{ji\lfloor 331t \rfloor}),$$

where Y_{jik} is the j^{th} asset price on the i^{th} day at minute k. Figure 3.5(a) shows the intraday log return curves $\{X_{2i}(t)\}_{i=1}^{208}$ of Facebook (`fb`) stock. The data was fit to a B-spline basis, using 50 basis functions, see `smooth.basis` in the `fda` `R` package.

Notice that the magnitudes of the curves vary widely. Figure 3.5(b) displays the squared norms of the daily curves as a function of the day i. We can see that the magnitude of each observation is related to the day on which it was observed. For example, around March 2020 the norms are higher, likely due to the volatility which resulted from the COVID-19 pandemic.

To handle the heteroskedasticity and any serial correlation present in the data, we employ a functional GARCH(1,1) model (Aue et al., 2017; Cerovecki et al., 2019) and apply the FKWC test to the residuals. The idea is to decompose the data into the conditional volatility η_i^2 and the independent error $\epsilon_{ji}(t)$, which can be approximated by the residuals $\widehat{\epsilon}_{ji}(t)$. Unlike in the univariate GARCH model, the second order behaviour of $\epsilon_{ji}(t)$ can differ between different assets; $\mathrm{E}\left[\epsilon_{ji}^2(t)\right] = 1$ for all t is assumed for an identifiable model (Cerovecki et al., 2019) but nothing is assumed about $\mathrm{E}\left[\epsilon_{ji}(t)\epsilon_{ji}(s)\right]$ for $s \neq t$. Thus, it is also of interest to investigate the properties of $\mathrm{E}\left[\epsilon_{ji}(t)\epsilon_{ji}(s)\right]$. For example, if the errors come from the same distribution, then the residuals can be pooled and bootstrapped to provide standard errors.

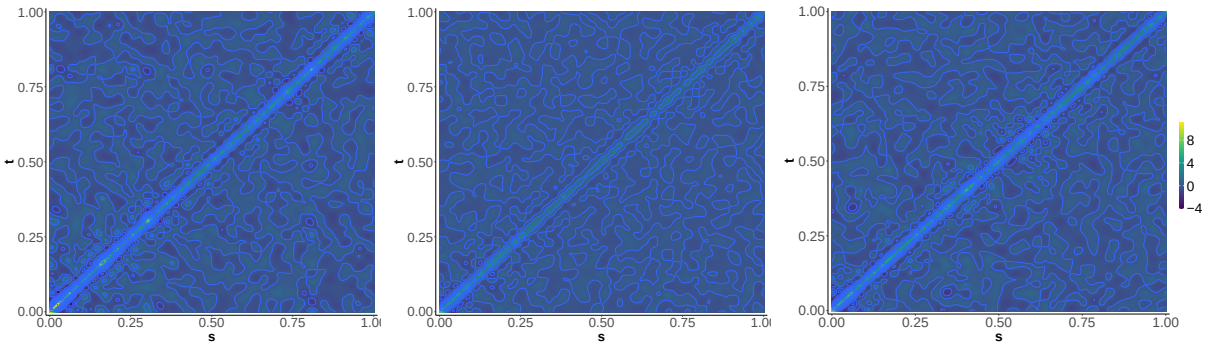
Since this type of data is typically heavy tailed, a robust test is suitable. In order to check the condition that η_{ji}^2 completely encapsulates the serial dependence in the data, we use the tests described by Rice et al. (2019). Specifically, we fit the functional GARCH(1,1) to each series of intraday returns using quasi-maximum likelihood (Cerovecki et al., 2019). We assumed that the volatility curves could be represented as linear combinations of M Bernstein basis functions. These were chosen based on a combination of the Box-Jenkins type test for the functional GARCH model (Rice et al., 2019), assessing the fit of the raw mean of the squares graphically (see Cerovecki et al. (2019)) and keeping the number of basis functions similar between assets. This resulted in choosing $M = 4$, and our results were insensitive to the number of basis functions in terms of testing the residuals for a difference in covariance. Figure 3.5(c) shows the resulting residuals of the GARCH(1,1)



(a) (b)

(c) (d)

Figure 3.5: (a) Daily log differenced intraday return curves for `fb` stock, starting on June 24th 2019 and ending March 20th 2019. (b) Daily squared norms of the intraday returns. Notice that these norms vary with the time period; the curves exhibit heteroskedastic features. For example, the most recent month of returns are much more variable. (c) Residuals for the `fb` log returns after fitting a functional GARCH(1,1) model (d) squared norms of the residuals over time.



(a) (b) (c)

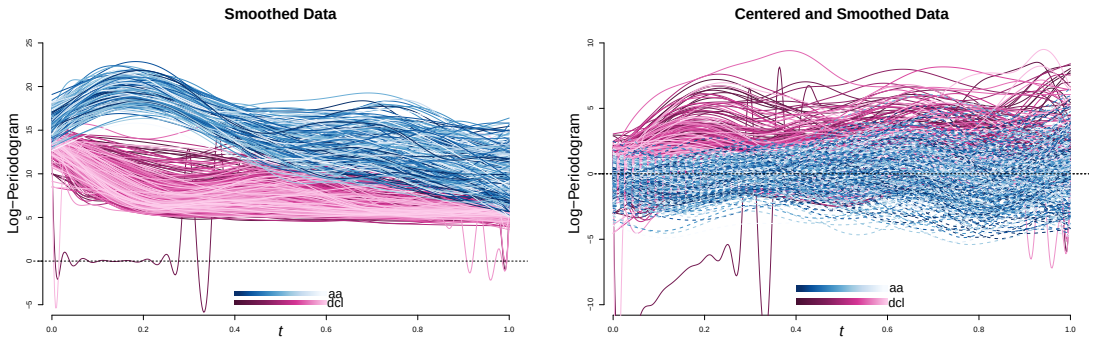
Figure 3.6: Covariance kernels $\mathcal{K}(s,t)$ of the residuals of (a) `twtr`, (b) `fb` and (c) `snap` log returns.

model fitted to the `fb` stock log returns and Figure 3.5(d) shows the norms of those residuals as a function of the day i. Notice that both the residuals and their norms are fairly uniform over time, especially when compared with the raw data. Figure 3.5(d) also shows that there are some outliers in the data.

Figure 3.6 shows contour plots of the estimated covariance kernels of the residuals of each functional time series, where 5% of the lowest random projection depth (with derivatives) observations were trimmed to account for the outliers. Notice that the estimated covariance kernels of the residuals of `fb` differs from the other two assets visually. We conducted the FKWC test at the 5% level of significance, using ranks based on the random projection depth which incorporates the derivatives. The means of the ranks are 244.4203, 424.2837, and 266.9712 for the `twtr`, `fb`, and `snap` residuals, respectively; $\widehat{\mathcal{W}}_n = 120.37$ and we reject the hypothesis that these three series have the same covariance kernels. The means of the ranks are similar for that of the `twtr` and `snap` stock, but the `fb` stock differs, which matches Figure 3.6.

3.5.2 Comparing speech variability with phoneme periodograms

In this section we analyse the Phoneme data, where the observations are log peroidograms of digitized speech. The data can be retrieved as part of the `fda R` package (Hastie et al., 1995). The data is split into five groups representing the syllables 'aa', 'ao', 'dcl', 'iy' and 'sh'. The goal is to characterize differences between the syllables' distributions in order to aid understanding of speech as well as to help improve the performance of speech



(a) (b)

Figure 3.7: Log periodograms for the syllables 'aa' and 'dcl'. After centering, we see a difference in magnitude at different times between the curves. There is also a large outlier in the 'dcl' group.

recognition models. The data has obvious location differences, for example Figure 3.7(a) shows the periodograms for two different syllables on the same plot. We centered the data by the deepest curve, as measured by random projection depth, within each group. We used a robust measure of centre to account for outliers, for example, there is a large outlier in the 'dcl' syllable group. From Figure 3.7(b) we might suspect that there are differences in the covariance kernels between the curves. To this end, we can run the FKWC test on the five groups of syllables, we run two versions of the test and compare the results. We run the FKWC test with random projection depth including the derivative information and as well as the FKWC test with L^2-root depth. These tests will differ if there are differences of the form $\mathcal{K}_j(s,t) \neq \mathcal{K}_k(s,t)$, $s \neq t$, $j \neq k$. Both tests result in incredibly small p-values, smaller than 2.2×10^{-16}. We can further examine differences between groups, by performing multiple comparisons. Suppose we would like to compare the covariance kernels of groups j and k, then there are two obvious routes for multiple comparisons. One method is to directly use the pre-calculated joint sample ranks, analogous to the univariate method of Dunn (1964). Here, for large n, one is essentially assessing the behaviour of the random variables $\mathrm{D}(X_{j1}; F_*) - \mathrm{D}(X_{k1}; F_*)$, through combined sample ranks. The other method is to extend the methods of Steel (1960) and compare the mean ranks of $\mathrm{D}(X_{ji}; F_{jk})$ and $\mathrm{D}(X_{k\ell}; F_{jk})$, where

$$F_{jk} = \frac{\vartheta_j}{\vartheta_j + \vartheta_k} F_j + \frac{\vartheta_k}{\vartheta_j + \vartheta_k} F_k.$$

	RP′					LTR				
Syllable	aa	ao	dcl	iy	sh	aa	ao	dcl	iy	sh
aa	1.00	0.85	**0.00**	0.97	**0.00**	1.00	0.13	0.40	1.00	**0.00**
ao	0.85	1.00	**0.00**	0.27	**0.00**	0.13	1.00	**0.00**	0.18	**0.00**
dcl	**0.00**	**0.00**	1.00	**0.00**	**0.00**	0.40	**0.00**	1.00	0.29	0.11
iy	0.97	0.27	**0.00**	1.00	**0.00**	1.00	0.18	0.29	1.00	**0.00**
sh	**0.00**	**0.00**	**0.00**	**0.00**	1.00	**0.00**	**0.00**	0.11	**0.00**	1.00

Table 3.5: Šidák corrected p-values of pairwise functional Steel tests performed on the centred curves.

Since Theorem 5 and Theorem 4 imply that differences in covariance structure will be exhibited in the pairwise ranks of the depth values, when the depth values are taken with respect to the empirical estimate of F_{jk}, it seems natural to use the methods of Steel (1960). A second argument in support of the methods of Steel (1960) is as follows. Suppose that one group, say $j' \neq k,\ j$ has a very different covariance structure when compared to the remaining groups, if the depth values are computed with respect to the combined sample, then the group j' may then 'wash away' any differences between the remaining groups. In other words, it is possible that there is a difference between the random variables $\mathrm{D}\left(X_{ji}; F_{*}\right)$ and $\mathrm{D}\left(X_{k\ell}; F_{*}\right)$, but this difference is small relative to the combined sample and therefore may not be detected. The multiple comparisons procedure is as follows. For each pair of groups, j, k, compute the combined, two-sample depth values:

$$\{\mathrm{D}\left(X_{j1}; F_{jk,n}\right), \ldots, \mathrm{D}\left(X_{jn_j}; F_{jk,n}\right), \mathrm{D}\left(X_{k1}; F_{jk,n}\right), \ldots, \mathrm{D}\left(X_{kn_k}; F_{jk,n}\right)\},$$

where $F_{jk,n}$ is the empirical distribution of $\{X_{j1}, \ldots, X_{jn_j}, X_{k1}, \ldots, X_{kn_k}\}$. Next, perform the Wilcoxon rank-sum test on the depth values for each pair. Lastly, correct the final p-values using the Šidák correction (Sidak, 1967) (or any other multiple testing correction). Again, Theorems 3, 5 and 4 justify this procedure. Table 3.5 shows Šidák corrected p-values of pairwise 'functional Steel tests' performed on the centered curves. The p-values are corrected for the tests done across both hypothesis tests, i.e., across 22 tests. We see that the results show that the syllables 'dcl' and 'sh' differ from the remaining syllables and from each other in terms of the variability of the magnitude of their log-frequencies. We can also see that under the LTR test, 'dcl' is similar to the other tests. This implies that the trace norm of the covariance operator of 'dcl' is similar to that of the other syllables, with the exception of 'ao'. Under the RP′ test, the syllable 'dcl' differs from all other syllables. We could interpret this as the log-periodograms of 'dcl' are more or less 'wiggly' when compared to the other syllables, or that the frequencies that have high variability

are different for the syllable ‘dcl’ than the other syllables.

Chapter 4

Kruskal-Wallis type statistics for functional change-point problems

4.1 Introduction

Detecting the presence and location of change-points in the covariance operator of a sequence of observed functions has received some recent interest in the statistics literature, see, e.g., (Harris et al., 2021). In this chapter, we combine the methods of Chapter 2 and Chapter 3 and use data depth ranks to detect change-points in the covariance kernel of functional data. We introduce three procedures to perform covariance operator change-point detection: a hypothesis test for the presence of at most one change-point, a hypothesis test for the presence of an "epidemic period" and an algorithm to estimate the locations of multiple change-points when the number of change-points is not known. We may call these methods FKWC methods, or Functional Kruskal-Wallis Covariance operator methods, following the naming convention in Chapter 3

In order to contextualize our procedure, we first review existing procedures for detection of change-points in the covariance structure of functional data. Jarušková (2013) is the first to discuss change-point detection for the covariance operator in the functional data setting. Under the assumption of independent observations, they introduce a test for at most one change-point. Their test is based on the first $K < \infty$ eigenvalues of the empirical covariance operator. Later, Aue et al. (2019) proposed a method to detect changes in the eigenvalues of the covariance operator, where the observations are assumed to be a dependent time series. In a similar vein, Dette and Kutta (2021) presented methodology which can detect change-points in the eigensystem of a functional time series. Sharipov

and Wendler (2019) investigated a bootstrap technique for conducting inference on the covariance operator of a functional time series, which does not require dimension reduction. They apply their bootstrapping technique to test for the presence of at most one change-point in a functional time series. They present two test statistics, both based on empirical covariance operators and the Hilbert-Schmidt norm. Stoehr et al. (2021) present CUSUM statistics combined with a bootstrap procedure in order to test for the presence of changes in the covariance operator of f-MRI data. They provide methods for detecting at most one change, as well as methods for detecting an epidemic-type change. Their methods are based on CUSUM statistics of projection scores, i.e., the CUSUM is computed from projections of the data onto the first few eigenfunctions of the estimated covariance kernel, combined with an estimate of the long-run covariance kernel. They also present a fully functional method which uses all of the eigenfunctions of the estimated covariance kernel in order to compute the test statistic. Dette and Kokot (2020) present change-point estimation and hypothesis testing methodology for detecting relevant differences in the covariance operators of functional time series. By relevant differences, it is meant that one may only wish to detect a change-point if the magnitude of the change is above a certain threshold. Their test statistic can be thought of as an adjusted estimate of the infinity norm distance between covariance kernels, which is then combined with a bootstrap procedure in order to perform the hypothesis test. These methods are generalised by Dette et al. (2020). Jiao et al. (2020) present a fully functional method for detecting changes in the covariance operator of functional data. The test statistic is an integrated univariate CUSUM-statistic, for which the null distribution must be estimated from the data. The procedure presented for estimating the null distribution relies on dimension reduction techniques. Recently, Harris et al. (2021) used a fused Lasso and a CUSUM statistic to detect multiple change-points in the mean and/or covariance operator of an observed sequence of functions. Their method can be computed in linear time and they show via simulation that their method is robust against heavy tailed data. However, their method also lacks theoretical results, and thus, the ability to test for the presence of a change-point without using resampling techniques.

Aside from Harris et al. (2021), previous works have not considered the robustness of their procedure. For example, many previous works require fourth moment assumptions (Stoehr et al., 2021; Dette and Kokot, 2020; Sharipov and Wendler, 2019), are based on CUSUM statistics and long-run covariance estimators which are not robust (Dette and Kokot, 2020; Sharipov and Wendler, 2019; Jiao et al., 2020) and/or rely on bootstrapping or other data-driven methods to estimate the null distribution of the test statistic (Sharipov and Wendler, 2019; Dette and Kokot, 2020; Jiao et al., 2020). Additionally, Stoehr et al. (2021) mentioned the need for a robust change-point procedure for the covariance operator.

The FKWC change-point procedure is robust against heavy tails and outliers, through the use of both ranking methods and functional data depth functions. Functional data depth functions lend themselves easily to robust methods, for example, they naturally define functional medians. Additionally, the use of ranks allows us to further robustify the procedure. For example, there is no need to assume finite third and fourth moments in our theoretical results, and in simulation our method performs well when the data is heavy tailed.

On top of being robust against outliers and distributional assumptions, the FKWC methods are easily computable. Many of the previous works do not consider computation, or rely on bootstrap methods (Stoehr et al., 2021; Sharipov and Wendler, 2019; Dette and Kokot, 2020) which can be computationally burdensome. There is a parallel version of the FKWC algorithm in which we can estimate the number and locations of multiple change-points in linear time. In the non-parallel version, we can compute the number and locations of multiple change-points in $O(n \log n)$ time. The FKWC methods are also easy to implement. These computational features make the FKWC methods suitable to be used with surface data; data whose domain is $\mathbb{R}^d$, $d > 1$, such as f-MRI data. We provide an implementation of our methods on Github (Ramsay, 2021).

Additionally, our FKWC methods are comprehensive, in the sense that they provide both the functionality to test for the presence of a change-point and also to detect an unknown number of change-points. Previous works typically perform only one of these tasks. We also provide a finite sample result for our at-most-one change-point estimator, where only asymptotic results have been considered thus far.

The rest of the chapter is organised as follows. Section 4.2 gives the assumed change-point model and our proposed change-point procedures. Section 4.3 presents our finite sample and asymptotic results. Section 4.4 contains a simulation study, where we compare our methods to those of Sharipov and Wendler (2019); Dette and Kokot (2020); Harris et al. (2021). Section 4.5 presents applications of our methods to intraday stock returns and resting state f-MRI scans. Technical proofs and additional simulation results can be viewed in Section A.6 and Section A.7, respectively.

4.2 Changepoint model and methodology

In this chapter we assume that each observed function X_i is a zero mean, real-valued function whose domain is $[0, 1]^d$ for fixed $d \in \mathbb{N}$. We additionally assume that each observed function is a mean square continuous stochastic process, and that we may only observe X_i

which are continuous functions. When the derivatives of the observations are involved in the inference procedure, it is assumed that each observation is differentiable on $(0,1)^d$ and that its derivative meets the same continuity conditions just described. As in Chapter 3, we denote this space of functions as $\mathfrak{F}$.

We suppose that we have observed a sequence of observations $X_1, \ldots, X_n$, such that $X_{k_{i-1}+1}, \ldots, X_{k_i}$ have a common covariance operator $\mathscr{K}_i$, with $k_0 = 0 < k_1 < \cdots < k_\ell < k_{\ell+1} = n$ for some fixed, but possibly unknown ℓ. Suppose that $k_i = \lfloor n\theta_i \rfloor$ for all $i \in \{1, \ldots, \ell\}$. Let $\vartheta_i = \theta_i - \sum_{j=0}^{i-1} \theta_j$ be the fraction of the observations with covariance kernel $\mathcal{K}_i$ and define

$$\mathcal{K}_* \coloneqq \vartheta_1\mathcal{K}_1 + \vartheta_2\mathcal{K}_2 + \vartheta_3\mathcal{K}_3 + \cdots + \vartheta_\ell\mathcal{K}_\ell + \vartheta_{\ell+1}\mathcal{K}_{\ell+1}.$$

The goal is to estimate each k_i; the location of the change-points. If we assume that $\ell = 1$, this is the at-most one-change (AMOC) model. If we assume that $\ell = 2$ and that $\mathcal{K}_1 = \mathcal{K}_3$ then this is the epidemic model. If ℓ is unknown, then this is the multiple change-point detection model. We focus on these three models in this chapter. Like in Chapter 2 and Chapter 3, our test statistic is based on the Kruskal Wallis ANOVA test statistic, where the ranks are the depth-based ranks.

We choose three depth functions, based on the results of Chapter 3. We choose the the best performing depth function in Chapter 3, the random projection depth (Cuevas et al., 2007). Due to its nice theoretical properties, we also use our procedure with the multivariate halfspace depth (Slaets, 2011) with $\alpha = 0$ otherwise known as the integrated functional depth (Fraiman and Muniz, 2001). Lastly, if there is a reason to believe that the changes in the covariance kernel are of the form

$$\mathcal{K}_j = \begin{cases} a_j\mathcal{K}_1 & t \in I_j \\ \mathcal{K}_1 & t \notin I_j \end{cases}$$

for some $I_j \subset [0,1]^d$ and some $a_j > 0$, then we could use a summary of the observations more directly related to the magnitude of the observations. For this reason, we also run our procedure on the ranks of the squared norms, which we may refer to as the L^2-root depth-ranks (see Chapter 3). As in Chapter 3, we incorporate the derivative information.

We now present our change-point detection methodology. Consider a candidate set of change-points $\mathbf{r} = \{r_1, \ldots, r_J\}$, which we will always assume to be ordered by their indices, i.e., $r_0 = 0 < r_1 < \ldots < r_j < r_{J+1} = n$. Let $\mu_n = (n+1)/2$, $\sigma_n^2 = (n^2-1)/12$ and

$\widetilde{\sigma}_n^2 = n(n+1)/12$. All of our test statistics are based on the Kruskal Wallis test statistic:

$$\mathcal{W}(\mathbf{r}) = \frac{1}{\widetilde{\sigma}_n} \sum_{j=1}^{J+1} (r_i - r_{i-1}) \overline{\widehat{R}}_j^2 - 3(n+1), \qquad \text{where} \qquad \overline{\widehat{R}}_j = \sum_{i=r_{j-1}}^{r_j - 1} \widehat{R}_i. \tag{4.1}$$

Recall from Chapter 2 that the bigger the differences between segments $\{\widehat{R}_1, \ldots, \widehat{R}_{r_1-1}\}$, $\{\widehat{R}_{r_1}, \ldots, \widehat{R}_{r_2-1}\}$, et cetera, the bigger (4.1) will be. Therefore, maximizing a version of (4.1) over $\mathbf{r}$ should give a set of time intervals which differ in median depth values. Due to the relationship between depth values and covariance kernels (see Section 4.3), this procedure will simultaneously give a set of time intervals in which the covariance operators differ. If the number of true change-points ℓ is known, it suffices to use

$$\widehat{\mathbf{k}} = \underset{r_1, \ldots, r_\ell}{\operatorname{argmax}} \mathcal{W}(\mathbf{r}),$$

as an estimate of the change-points. For example, if we assume the number of true change-points $\ell = 1$ this is the at most one change (AMOC) setting. In this context, we often wish to conduct a hypothesis test to determine whether there exists a single change-point or no change-point. Our proposed hypothesis test uses $\sup_{1<r_1<n} \mathcal{W}(r_1)$, as the test statistic. If the test is significant, we can then use $\hat{k}_1 = \operatorname{argmax}_{1<r_1<n} \mathcal{W}(r_1)$ as the estimated change-point. This procedure is equivalent to using a Wilcoxon rank-sum based CUSUM:

$$\sup_{t \in (0,1)} |\widehat{Z}_n(t)| \coloneqq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor tn \rfloor} \frac{\widehat{R}_i - \mu_n}{\sigma_n} \right|,$$

as the test statistic and then defining the change-point estimate as

$$\hat{k}_1 = \inf \left\{ r \colon |\widehat{Z}_n(r/n)| = \sup_t |\widehat{Z}_n(t)| \right\}, \tag{4.2}$$

if the test is significant. It follows that $|\widehat{Z}_n(\hat{k}_1/n)| \to \sup_t |B(t)|$ as $n \to \infty$, where $B(t)$ is a standard Brownian bridge (Chenouri et al., 2020b). Since it easy to obtain the quantiles of $\sup_t |B(t)|$, we suggest using the Wilcoxon rank-sum version of the test statistic. This Wilcoxon rank-sum version of the test statistic and associated change-point estimator are defined in the same manner as the multivariate change-point procedure proposed by Chenouri et al. (2020b).

Another change-point setting is the epidemic change-point model, where we conduct a

hypothesis test to determine whether there exists two change-points or no change-points. There is the additional assumption that the distribution in the first and third segment remain the same, and the middle segment is the "epidemic period", during which the data comes from a different distribution. In this case, we propose the following test statistic:

$$(\hat{k}_1, \hat{k}_2) = \underset{1<r_1<r_2<n}{\operatorname{argmax}} \frac{1}{\widetilde{\sigma}_n} \left(\left(\sum_{\substack{1 \le i < r_1 \\ r_2 \le i \le n}}^{n} \frac{\widehat{R}_i}{\sqrt{n - r_2 + r_1}} \right)^2 + \left(\sum_{i=r_1}^{r_2-1} \frac{\widehat{R}_i}{\sqrt{r_2 - r_1}} \right)^2 \right) - 3(n+1). \quad (4.3)$$

We can easily show that when there are no change-points

$$\sup_{1<r_1<r_2<n} \mathcal{W}_n(r_1, r_2) \xrightarrow{\text{d}} \sup_{t_1,t_2 \in (0,1)} \left(\frac{1}{(t_2 - t_1)(1 - t_2 + t_1)} \right) (B(t_2) - B(t_1))^2, \quad (4.4)$$

which can be used to obtain critical values for the hypothesis test.

If the number of true change-points ℓ is unknown then maximizing the objective function $\mathcal{W}$ over all possible candidate sets of change-points is a degenerate problem. Therefore, we must add a penalty term on the number of change-points:

$$\widehat{\mathbf{k}} = \underset{r_1,\ldots,r_j}{\operatorname{argmax}} \mathcal{W}(\mathbf{r}) - j\lambda_n. \quad (4.5)$$

Just like in Chapter 2, this estimate can be computed with the PELT algorithm (Killick et al., 2012). The PELT algorithm allows the change-point estimates to be computed in linear time given the sample depth-based ranks. Much of the computational speed will depend on computing and ranking the depth values. For example, if we let N represent the number of points in the grid on which the functions are discretized to, then the random projection depth values of a sample of size n can be computed in $O(nMN + Mn \log n)$ time. Ranking the functional depth values will always take $n \log n$ time, and therefore, theoretically, the best time for the FKWC change-point algorithm is $O(n \log n)$. One may be able to improve this result to linear time by using the parallel ranking algorithm developed by Anderson and Miller (1990).

Practically, it seems as though the majority of the computational burden comes in the form of computing the sample depth values. For example, 1 million observations can be ranked in `R` in 0.37 seconds[1] even if the algorithm for the `rank` function in `R` is not implemented in parallel. By contrast, computing the RP depth values of one million

[1] using an Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz microchip and 32 Gb of RAM.

observations would take considerably longer with existing implementations in `R`. Therefore, it is important to consider the functional depth function used with the FKWC change-point method if computation is of high concern, such as in the context of f-MRI data.

4.3 Theoretical results

In this section we present the theoretical results for our change-point procedures. We remind the reader that throughout the chapter we assume that data satisfy $\mathrm{E}[X_i] = 0$ and that the locations of the change-points is fixed $\theta_i = \lfloor nk_i \rfloor$. We first introduce some assumptions on the data.

Assumption 11. *The change-points are bounded away from the boundaries of the observed data;* $n\ c_0 < k_i < n(1-c_0)$ *where* $0 < c_0 << 1$ *and one recalls that* $k_1, \ldots, k_\ell$ *are the locations of the true change-points.*

Assumption 12. *For* $X \sim F$ *and* $z, y \in \mathbb{R}$*, the distribution function of* $\mathrm{D}(X, F)$ *is a Lipschitz function;*

$$|\Pr(\mathrm{D}(X,F) < z) - \Pr(\mathrm{D}(X,F) < y)| \leq K'|z-y|.$$

Assumption 13. *The functional depth function* D *satisfies*

$$\sup_{x \in \mathfrak{F}} \left| \mathrm{D}(x; \widehat{F}_n) - \mathrm{D}(x; F) \right| = O_p(n^{-1/2}).$$

Assumption 11 ensures that the change-points are bounded away from the edges of the sequence of observations. Assumption 12 is a smoothness condition on the distribution function of the depth function. Assumption 13 is satisfied by all three of the depth functions discussed in this chapter, see (Nagy and Ferraty, 2019) and Chapter 3. The type of change captured by the procedure is entirely encapsulated in the following condition: If there exists a changepoint at time k_i, then

$$p_i \coloneqq \Pr(\mathrm{D}(X_{k_i}; F) < \mathrm{D}(X_{k_i+1}; F)) \neq 1/2. \tag{4.6}$$

It is necessary to connect the distribution of the depth values $\mathrm{D}(X_{k_i}; F) - \mathrm{D}(X_{k_i+1}; F)$ to the covariance operators. For example, suppose that there is only one change-point which occurs at time k_1. If we use the ranks of the squared norms of the observations with our

change-point procedure, then condition (4.6) translates to

$$\text{MED}(\|X_{k_1}\|^2 - \|X_{k_1+1}\|^2) \neq 0.$$

If we further assume that the distributions of the norms have the same shape, then we can translate this condition to

$$\|\mathscr{K}_{k_1}\|_{tr} \neq \|\mathscr{K}_{k_1+1}\|_{tr}.$$

Therefore, using the ranks of the squared norms can only detect changes in the trace norms of the covariance operators. When we instead use ranks generated from the functional depth functions MFHD and RP, condition (4.6) translates to

$$\text{MED}(\text{D}(X_{k_1+1}; F) - \text{D}(X_{k_1}; F)) \neq 0.$$

Both the integrated functional depth with $\text{D}(x(t); F_t) = F_t(x)(1 - F_t(x))$ and the random projection depth can be written in the form of

$$\text{D}(X_i; F) = \int_A F_{*,a}(g(X_i, a))(1 - F_{*,a}(g(X_i, a)))dP_A \coloneqq \int_A \text{D}_a(X_{k_1})dP_A,$$

where A represents a compact set, P_A is the uniform measure on A, g is some function $g\colon \mathfrak{F} \to \mathbb{R}$ such that $\text{E}[g(X_i, a)] = 0$ and

$$\text{D}_a(X_{k_1}) = F_{*,a}(g(X_{k_1}, a))(1 - F_{*,a}(g(X_{k_1}, a))).$$

Suppressing the F in D, we can write

$$\text{D}(X_{k_1+1}) - \text{D}(X_{k_1}) = \int_A \text{D}_a(X_{k_1}) - \text{D}_a(X_{k_1+1})dP_A.$$

Now, define $\sigma^2_{1,a}$ and $\sigma^2_{2,a}$ as the variance of $g(X, a)$ for X observed before and after the change-point, respectively. To be clear, $\sigma^2_{1,a} = \text{E}[g(X_1, a)^2]$ and $\sigma^2_{2,a} = \text{E}[g(X_n, a)^2]$. In the case of the random projection depth, we have that A is a compact subset of S and $\sigma^2_{j,a} = \langle \mathscr{K}_j a, a\rangle$. In the case of the integrated functional depth we have that $A = [0, 1]^d$ and $\sigma^2_{j,a} = \mathcal{K}_j(a, a)$. If we assume that $F_{*,a}$ is thrice differentiable for all a and let $f_{*,a}$ be the

density corresponding to $F_{*,a}$, we can write

$$\begin{aligned}
\mathrm{E}\left[F_{*,a}(g(X_1,a))\right] &= F_{*,a}(0) + \frac{1}{2}f^{(1)}_{*,a}(0)\sigma^2_{1,a} + \mathcal{R}_{a,1,1} \\
\mathrm{E}\left[F^2_{*,a}(g(X_1,a))\right] &= F^2_{*,a}(0) + (f_{*,a}(0)f^{(1)}_{*,a}(0) + f^2_{*,a}(0))\sigma^2_{1,a} + \mathcal{R}_{a,1,2} \\
\mathrm{E}\left[F_{*,a}(g(X_n,a))\right] &= F_{*,a}(0) + \frac{1}{2}f^{(1)}_{*,a}(0)\sigma^2_{2,a} + \mathcal{R}_{a,2,1} \\
\mathrm{E}\left[F^2_{*,a}(g(X_n,a))\right] &= F^2_{*,a}(0) + (f_{*,a}(0)f^{(1)}_{*,a}(0) + f^2_{*,a}(0))\sigma^2_{2,a} + \mathcal{R}_{a,2,2}.
\end{aligned}$$

The remainders are defined as

$$\begin{aligned}
\mathcal{R}_{a,1,1} &= \mathrm{E}\left[\frac{1}{6}\int_0^{g(X_1,a)} f^{(2)}_{*,a}(t)(g(X_1,a)-t)^3 dt\right] \\
\mathcal{R}_{a,1,2} &= \mathrm{E}\left[\frac{1}{3}\int_0^{g(X_1,a)} (3f_{*,a}(t)f^{(1)}_{*,a}(t) + f_{*,a}(t)f^{(2)}_{*,a}(t))(g(X_1,a)-t)^3 dt\right] \\
\mathcal{R}_{a,2,1} &= \mathrm{E}\left[\frac{1}{6}\int_0^{g(X_n,a)} f^{(2)}_{*,a}(t)(g(X_n,a)-t)^3 dt\right] \\
\mathcal{R}_{a,2,2} &= \mathrm{E}\left[\frac{1}{3}\int_0^{g(X_n,a)} (3f_{*,a}(t)f^{(1)}_{*,a}(t) + f_{*,a}(t)f^{(2)}_{*,a}(t))(g(X_n,a)-t)^3 dt\right].
\end{aligned}$$

Note that we expect $\mathcal{R}_{a,j,i}$ to be small from the fact that the mean of $g(X_i,a)$ is 0. After some manipulation, it follows that

$$\mathrm{E}\left[\mathrm{D}\left(g(X_1,a);F_{*,a}\right) - \mathrm{D}\left(g(X_n,a);F_{*,a}\right)\right] = \mathcal{H}(F_{*,a})(\sigma^2_{1,a} - \sigma^2_{2,a}) + \mathcal{R}_{a,1,3} - \mathcal{R}_{a,2,3},$$

where, if F is a univariate CDF, then

$$\mathcal{H}(F) = \frac{1}{2}f^{(1)}(0) - (F(0)f^{(1)}(0) - f^2(0)) \qquad \text{and} \qquad \mathcal{R}_{a,j,3} = \mathcal{R}_{a,j,1} - \mathcal{R}_{a,j,2}.$$

We must now impose the following assumption:

Assumption 14. *The distribution F is such that*

$$\mathrm{E}\left[\mathrm{D}\left(X_{k_1+1};F\right) - \mathrm{D}\left(X_{k_1};F\right)\right] \neq 0 \implies \mathrm{MED}(\mathrm{D}\left(X_{k_1+1};F\right) - \mathrm{D}\left(X_{k_1};F\right))) \neq 0.$$

This assumption says that if the distributions of $\mathrm{D}\left(X_{k_1};F\right)$, $\mathrm{D}\left(X_{k_1+1}\right)$ differ in mean,

then they also differ in median. This assumption is akin to the assumption made in the univariate Kruskal-Wallis procedure, where we assume that the distributions in each group are not of the special form where the group means differ but the group medians do not. If Assumption 14 holds, then we see that changes in the covariance operator which can be described by

$$\int \mathcal{H}(F_t)\mathcal{K}_2(t,t)dt \neq \int \mathcal{H}(F_t)\mathcal{K}_1(t,t)dt, \tag{4.7}$$

will be captured, in the case of the integrated functional depth. In the case of the random projection depth, changes in the covariance operator of the type

$$\int_S \mathcal{H}(F_u)\langle \mathscr{K}_2 u, u\rangle dP \neq \int_S \mathcal{H}(F_u)\langle \mathscr{K}_1 u, u\rangle dP, \tag{4.8}$$

will be captured. The full proof of these results can be seen in the proof of Theorem 4 in Chapter 3. Looking at the case of random projection depth, if it is assumed that the observed functions can be written as

$$X = \sum_{i=1}^{b} c_i \phi_i$$

for some integer $b \in \mathbb{N}$, then theoretically, the procedure could detect any type of change in the covariance operator. This is of course provided that P is chosen to be uniform on $\{u^\top \Phi \colon u \in S^{p-1}\}$. From (4.7) it follows that the integrated functional depth-based procedure can only detect changes that occur on the diagonal of the covariance kernel. If one aims to detect changes characterised by $\mathrm{E}\left[X_n(s)X_n(t)\right] \neq \mathrm{E}\left[X_1(s)X_1(t)\right]$[2], then we recommend using the random projection depth based change-point procedure. Otherwise, using either of these depth functions is acceptable. If we want to use half-space depth for D in the integrated functional depth, i.e., the multivariate functional half-space depth, then the procedure can detect differences in a quantity which is akin to the univariate mean absolute deviation. In this case, a change of the form

$$\int_{[0,1]^d} \mathrm{E}\left[|(X_{k_1}(t) - \mathrm{MED}(F_t))|\right] dt \neq \int_{[0,1]^d} \mathrm{E}\left[|X_{k_1+1}(t) - \mathrm{MED}(F_t)|\right] dt, \tag{4.9}$$

will be detected by the procedure, provided that $f_t^{(1)}$ evaluated at $\mathrm{MED}(F_t)$ is small. For

[2]In this expression we can replace X_n with any observation that comes after the change-point, and we can replace X_1 with any observation that falls before the change-point.

example this will hold when the univariate distributions f_t are symmetric. We may produce something analogous to (4.9) for the random projection depth if we wish to use half-space depth instead of simplicial depth for the univariate depth function D. Of course, this analysis applies to the FKWC methods which do not incorporate the derivatives, though we expect the procedure that includes the derivative information to be able to detect these types of change-points as well as additional types of change-points. The motivation for incorporating the derivatives is based both on the successful results of (Claeskens et al., 2014), the results of Chapter 3 and the qualitative argument that increased oscillations in the observations should produce a change in the magnitude of the derivatives. We can now present our theoretical results for the FKWC change-point methods. We begin with the behaviour of our proposed test statistics under the null hypothesis.

Theorem 7. *If Assumption 13 holds and there are no change-points present in the data, then the estimator (4.2) satisfies*

$$|\widehat{Z}_n(\widehat{k}_1/n)| \xrightarrow{d} \sup_{t\in(0,1)} |B(t)|,$$

and the estimator (4.3) satisfies

$$\mathcal{W}_n(\widehat{k}_1, \widehat{k}_2) \xrightarrow{d} \sup_{0<t_1<t_2<1} \frac{(B(t_2) - B(t_1))^2}{(t_2 - t_1)(1 - t_2 + t_1)}.$$

The proof of this theorem can be seen in Section A.6.

Theorem 8. *Suppose that Assumptions 11-13 and (4.6) hold. For λ_n as in (4.5), assume that $\lim_{n\to\infty} \lambda_n = \infty$ and that $\lambda_n = o(n)$. Then the following two results hold:*

1. *The estimator (4.2) satisfies*

$$|\widehat{k}_1/n - \theta| = O_p(n^{-1/2}).$$

2. *For the estimates (4.3) and the multiple change-point estimates (4.5), we have that for all $r > 0$, there exists a constant $C > 0$ such that*

$$\Pr\left(\left\{\widehat{\ell} = \ell\right\} \cap \left\{\max_{i\in[\ell]} |\widehat{k}_i - k_i| \leq Cn^{1/2+r}\right\}\right) \to 1 \text{ as } n \to \infty.$$

The proof of this theorem follows directly from the proof of Theorem 2 in Chapter 2 and

the proof of Theorem 3.2 in (Chenouri et al., 2020b). Next we give a finite sample result for univariate rank-based change-point detection.

Theorem 9. *Suppose that there are n independent, univariate data points $X_1, \ldots, X_n$, with a change-point at time k_1. Additionally, assume that $p = \Pr(X_1 < X_n) \neq 1/2$. Let $\widehat{k}_1$ be the proposed AMOC change-point estimator based on the univariate, linear sample ranks. Then for all $t, \theta > c_0 > 0$, there exists constants C_1 and C_2 which depend on θ and c_0 such that*

$$\Pr(|\widehat{k}_1 - k_1| > tn) < C_1 e^{-C_2 n t^2 (p-1/2)^2}.$$

Theorem 9 applies to our functional data change-point estimator when we use the norms of the observations to rank the functions. If one used a different one dimensional summary of the functional data in our change-point procedure, rather than the depth values, Theorem 9 could be applied. In fact, Theorem 9 can also be used on other kinds of data, such as a sequence of one-dimensional summaries of multivariate or shape data. Theorem 9 does not cover the FKWC procedure when used with depth functions since depth functions produce dependent depth values, but it applies to the method based on the ranks of the squared norms. We now present a finite sample result for the procedure which accounts for the dependency among the sample depth values.

Theorem 10. *Suppose that Assumptions 11-12 hold and that there exists a change-point at k_1 in the sense of (4.6). Suppose that the data satisfy $X_i = \mathbf{c}_i^\top \Phi$ for a vector Φ of b orthonormal basis functions. Let $\widehat{k}_1$ be the proposed AMOC change-point estimator (4.2) paired with the random projection depth or the* MFHD *depth. Then for all $t, \theta > c_0 > 0$, there exists constants $C_1, C_2 > 0$ which depend on θ, c_0, p_1, K', the depth function choice, and F such that*

$$\Pr(|\widehat{k}_1 - k_1| > nt) \leq C_1 e^{-C_2 n t^2 (p_1 - 1/2)^2}.$$

In order to have concentration of the sample ranks around the 'population ranks', it is required that the empirical processes $\int_A \widehat{F}_{n,a} da$ also concentrates around $\int_A F_a da$, which at this time requires the finite dimensional assumption. Seeing as many functional data analyses make this assumption implicitly through smoothing methods, we do not see this as a particularly troublesome assumption. We remark that this finite dimensional assumption is only necessary for the finite sample result, and not for consistency of the change-point estimator. Recall that we have assumed that the observations are independent. In order to extend these results to dependent observations, we require three results. The first of which is a maximal inequality for ranks based on weakly dependent random variables, such as discussed in (Hoffmann-Jørgensen, 2016). The second is a concentration result for dependent U-statistics, which may be seen in (Han, 2018).

4.4 Simulation

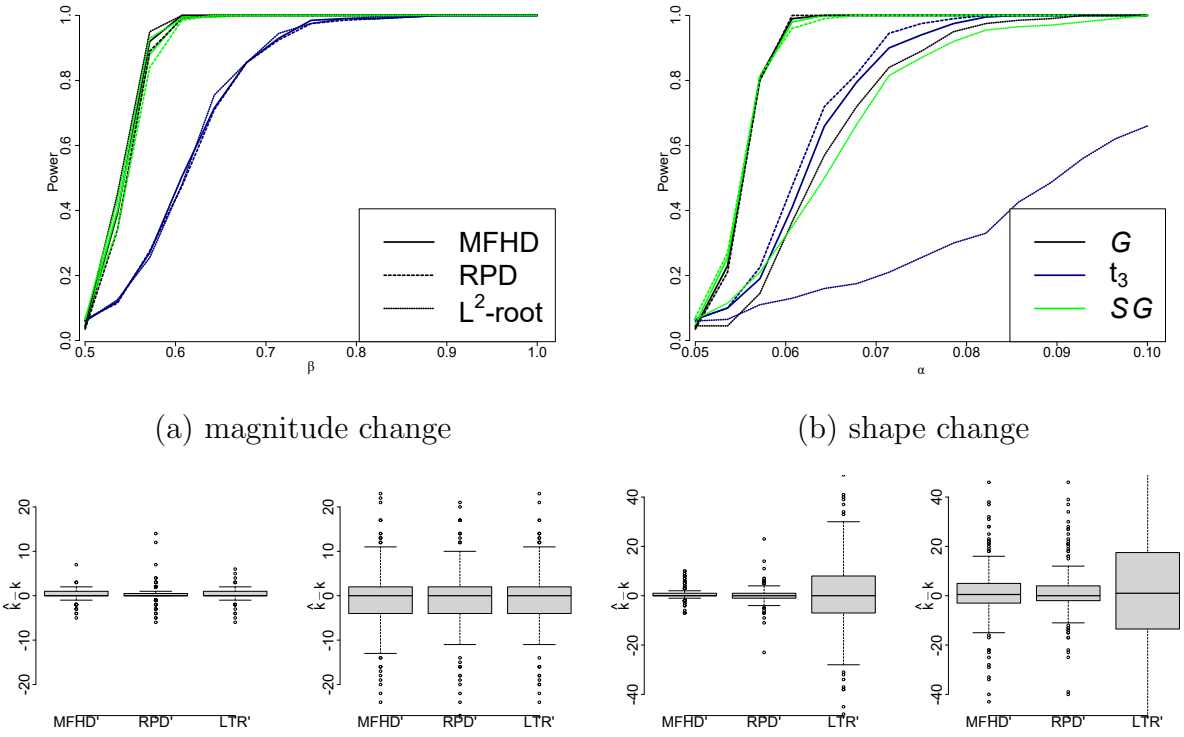
In order to test our methodology, we simulated observations from several change-point models. To test the AMOC setting, we simulated data with 0 change-points as well as with 1 change-point in the middle of the sample. To test the epidemic change-point setting, we simulated data with 2 uniformly random change-points, where we required that the change-points were at least 10% of the sample size apart. Sample sizes of 100, 200, and 500 were used for these scenarios and all tests are carried out at the 5% level of significance. Finally, to test the multiple change-point procedures, we used the simulated data for the AMOC and epidemic scenarios, as well as a simulation where there were five randomly placed change-points which also had to be at least 10% of the sample size apart. In the five change-point case, we simulated sample sizes of 200, 500, 1000 and 2500. In each case, the data were sampled from either a Gaussian process $\mathcal{GP}$, a Student-t process t_3 with degrees of freedom equal to three, or a Skewed Gaussian process $\mathcal{SG}$. We used a squared exponential covariance kernel

$$\mathcal{K}(s,t;\alpha,\beta) = \beta\ e^{\frac{-(s-t)^2}{2\alpha^2}},$$

and the sample differences were controlled by adjusting α and β. Changes in α correspond to a 'shape' difference in the data, while changes in β correspond to a magnitude difference. In the AMOC cases, we compare our methods to the methods of Sharipov and Wendler (2019); Dette and Kokot (2020). The code for these methods was kindly provided by the authors. We compare our multiple change-point algorithm to the functional multiple change-point isolation (FMCI) algorithm of Harris et al. (2021), whose code was also provided. We feel this is the most comparable algorithm, as it is computationally fast and can detect multiple change-points. We use the package `fmci` provided on Github (Harris, 2020) with the default parameters for this algorithm.

We discuss the AMOC and epidemic scenarios first, the FKWC methods which include the derivatives performed universally better than those without the derivatives, therefore we only present the results of the FKWC methods which use the derivatives. It should be noted that the MFHD and squared norm based FKWC procedures cannot detect shape changes without including the derivative information, see Section A.7. We also only present the results from the AMOC runs, since the results from the epidemic scenarios resulted in the same conclusions. The results from the epidemic scenarios can be seen in Section A.7.

Figure 4.1 shows the power curves for the AMOC FKWC test when the data had magnitude changes (panel (a)) and when the data had shape changes (panel (b)). Here, $n = 500$ and the results from other sample sizes were similar. (It was observed that a



(a) magnitude change (b) shape change

(c) (d) (e) (f)

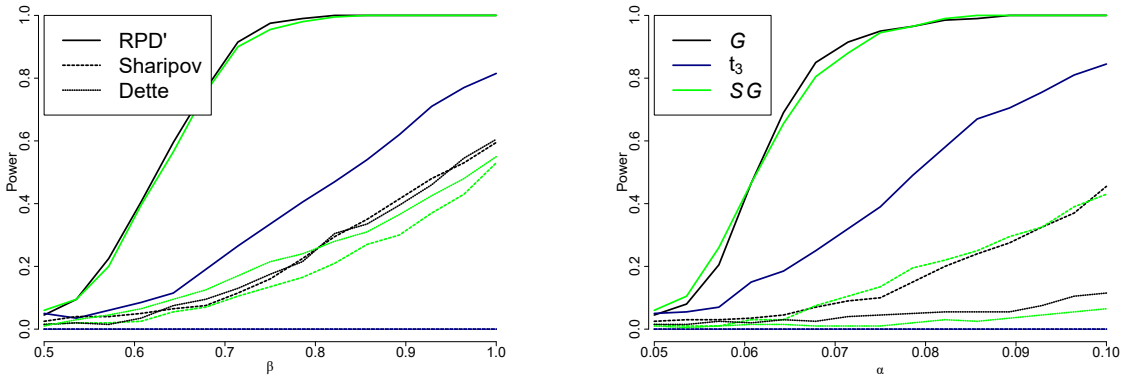
Figure 4.1: Power curves and accuracy boxplots when there is one change-point in the data and $n = 500$. Each curve represents the empirical power of the AMOC FKWC hypothesis test, for a given depth function (line type) when the data had a particular underlying distribution (line color). For example, the solid green line is the power of the procedure paired with the MFHD depth when the data came from a skewed Gaussian process. The boxplots represent the empirical distribution of $\hat{k} - k$ based on the 200 simulation runs. The data model for each of the boxplots was (c) Gaussian with a magnitude change, (d) student t with a magnitude change, (e) Gaussian with a shape change and (f) student t with a shape change. Note that "L^2-root" refers to the case where our procedure was paired with the ranks of the sum of the squared norms of the observations and their first derivative.

higher sample size indicates a higher detection power.) Figure 4.1 shows that the power of the test is lower when the underlying data comes from a heavy tailed distribution, but the method does not break down. We also see that performing the test on skewed

Gaussian data performs similarly to when performed on non-skewed Gaussian data. When the change-point altered the shape of the data, using the squared norm-based ranks did not work as well as the other two depth functions if the data was not Gaussian, even if the derivative information was incorporated. The performance of the test using ranks based on RP′ and MFHD′ was similar, with MFHD′ being slightly better for magnitude differences and RP′ being slightly better for shape differences. The empirical sizes of the test, were always within $2\% - 3\%$ of the nominal size, being within $1\% - 2\%$ most of the time. The bottom row of Figure 4.1 contains boxplots of $\hat{k}_1 - k_1$ when the change-point was detected. We observe again that when the data is heavy tailed, the estimation is less accurate, but still works. For example, the estimated change-point is often within 4 observations instead of 2 when the change-point was a magnitude change. The estimation accuracy results mirror those of the power curves; RP′ and MFHD′ perform the best with slight differences depending on whether or not the change type was shape or magnitude.

Figure 4.2 compares the power of the methods of Sharipov and Wendler (2019); Dette and Kokot (2020) to the power of the FKWC methods. We used 200 bootstrap samples with a block length of 1 for both of the competing methods. We only report the results for the integrated test of Sharipov and Wendler (2019), since it had generally higher power than the other test proposed in their paper. For the Dette method, we used 49 B-spline basis functions to smooth the data first and, note that using a Fourier basis resulted in slightly lower power. We did not smooth the data for use with the method of Sharipov and Wendler (2019). It can be seen in Figure 4.2 that the FKWC test has a higher power than its competitors for the data models in this simulation study. We also notice that the heavy tailed distribution breaks down the competing methods. We remark that though the FKWC methods have higher power than competing methods under these simulation models, the competing methods have some features that the FKWC methods do not. These methods have theoretical results for dependent data and the method by Dette and Kokot (2020) can test for "relevant" changes in the covariance operator, rather than the standard hypothesis of any change in the covariance operator.

In order to test the performance of the FKWC tests when the data are not independent, we also compared the three tests under a model where the data had some dependency. We simulated functions from the autoregressive model as discussed in the simulation section of Sharipov and Wendler (2019) and ran the FKWC test on those time series. Table A.4 in Section A.7 shows the results under this model. We see that the FKWC tests (which incorporate the derivatives) have higher power than competing methods, though they tend to have higher type one error. The RP′ test is an exception; it had an empirical size of 0.06, which is very close to the nominal level of 0.05. Overall, the performance of the FKWC test is better than its competitors under these simulation models.

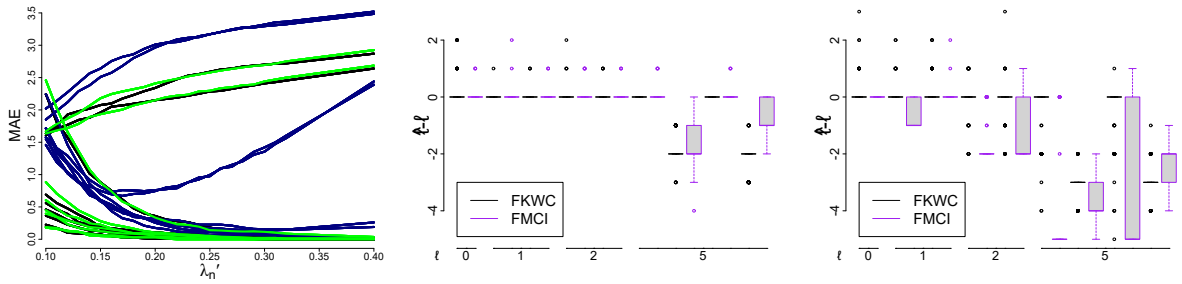


(a) (b)

Figure 4.2: Power curves of the FKWC method compared to the methods of Sharipov and Wendler (2019); Dette and Kokot (2020) when there was one change-point in the data and $n = 100$. Each curve represents the empirical power of a particular change-point procedure when the data had a particular distribution. For example, a green curve with a dashed line is the empirical power curve of the method of Sharipov and Wendler (2019) when the data was generated from a skewed Gaussian process.

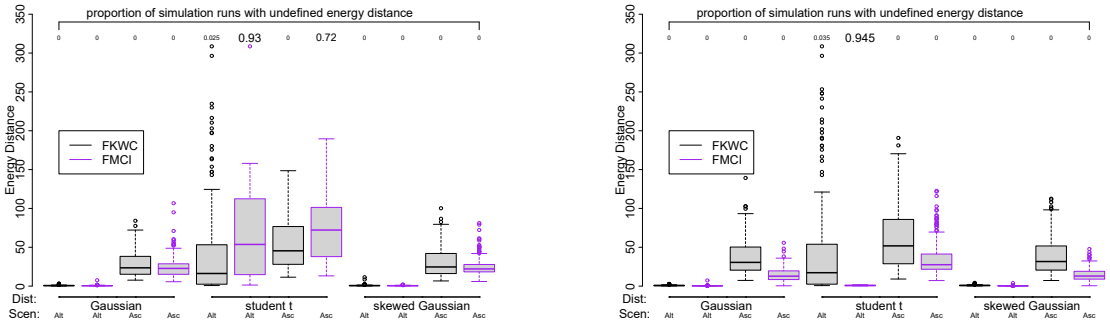
We now discuss the results from the multiple change-point algorithm. When there were five change-points, we ran four different simulation scenarios, two where the change-points were shape-type and two where the changes were magnitude-type. Within these groups, the set of change-points was either "ascending", i.e., α or β was increasing with each change or "alternating", i.e., α or β was oscillating between a high and low value with each change.

We first discuss choosing the value of λ_n. It was observed in Chapter 2 that $\lambda_n \in (3.74 + 0.15\sqrt{n}, 3.74 + 0.25\sqrt{n})$ performs well in the multivariate setting. In this study, we tested $\lambda_n = 3.74 + \lambda'_n\sqrt{n}$ for $\lambda'_n \in (0.1, 0.4)$ to see if the same parameter settings apply to the functional data setting. We ran the PELT algorithm on the simulated data for all of the scenarios, i.e., for data which had 0, 1, 2 and 5 change-points. We observed that the best choice of λ'_n was consistent across the different depth functions, and so we only present the results from RP$'$ depth. Figure 4.3 shows the mean absolute error in the estimated amount of change-points, i.e., $|\ell - \hat{\ell}|$, of the simulation scenarios for different values of λ'_n for $n = 500$. It is clear that the functional data context requires higher values of λ'_n, with the best parameters being in $0.25 - 0.4$ range. The algorithm is less sensitive to the choice of λ'_n with increased sample sizes. The group of curves presenting large errors at the top of panel (a) of Figure 4.3 are from the shape difference and/or heavy tailed scenarios with



(a) (b) (c)

Figure 4.3: (a) Mean absolute error $|\ell - \hat{\ell}|$ of the simulation scenarios for different values of λ_n' under the RP′ depth for $n = 500$. (b) Boxplots of $\hat{\ell} - \ell$ when the data contains different amounts of change-points (as labelled on the horizontal axis) under the Gaussian simulation scenario. (c) Boxplots of $\hat{\ell} - \ell$ when the data contains different amounts of change-points under the t_3 simulation scenario.



Magnitude changes Shape changes

Figure 4.4: Energy distance between the estimated and true change-points when $n = 500$ for the FKWC method and the FMCI method. The FKWC method was used with RP′ depth and $\lambda_n' = 0.3$ and the FMCI method was used with the default parameters. Only the runs in which there were five change-points are represented. The numbers at the top of the graph are the proportion of runs in which the algorithm failed to identify any change-points. The labels on the horizontal axis represent the distribution of the underlying data and the change-point scenario; either alternating or ascending.

five change-points, where there was not enough samples to detect all of the change-points.

We now compare the RP′ version of the FKWC multiple change-point method to the FMCI method of Harris et al. (2021). We chose to compare to Harris et al. (2021) because of FMCI's ability to detect multiple change-points, as well as its accessible implementation and computational speed. We use the default parameters provided in the `fmci` package. We remark that the FMCI method can detect changes in both the mean function and covariance kernel, whereas the FKWC method can only detect changes in the covariance kernel. It should be noted that we used the same simulation to evaluate the best parameter choice for the FKWC method, and so the results could be biased in favor of the FKWC method. We do not feel this plays a major in the conclusions drawn from the comparison.

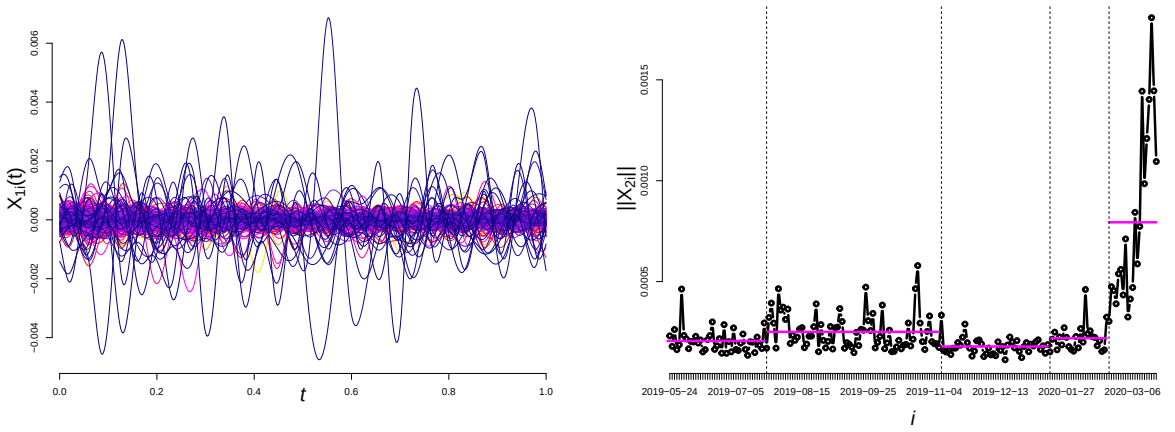
Figure 4.3 contains boxplots of $\hat{\ell} - \ell$ for different data models in the simulation. The results of skewed Gaussian and Gaussian were similar so we only present those of Gaussian and student t. We see that the results of both methods are similar under the Gaussian setting (while slightly favoring the FMCI method), but favor the FKWC method when the data is heavy tailed. Both methods tend to underestimate the number of change-points in the heavy-tailed case. We observed that both of these methods had more difficulty in the ascending scenarios, i.e., the simulation runs where either the α or β parameters were increasing at each change-point.

To evaluate the accuracy of the algorithms, we also look at the energy distance between the estimated change-point set and true change-point set for each method. The energy distance between the estimated and the true change-point set can written as

$$\frac{2}{\hat{\ell}\ell}\sum_{i=1}^{\hat{\ell}}\sum_{j=1}^{\ell}|\hat{k}_i - k_j| - \frac{1}{\hat{\ell}^2}\sum_{i=1}^{\hat{\ell}}\sum_{j=1}^{\hat{\ell}}|\hat{k}_i - \hat{k}_j| - \frac{1}{\ell^2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}|k_i - k_j|.$$

As done by Harris et al. (2021), we use the energy distance to evaluate the multiple change-point methods. We use this distance because, as discussed in Appendix B.3 of Harris et al. (2021), the energy distance measures the average error in estimating each change-point, rather than the error of the most poorly estimated change-point in the set. One criticism is that if the algorithm fails to identify any change-points, then the energy distance to a set of true change-points will not be defined.

Figure 4.4 shows boxplots of the energy distance between the estimated change-point set and true change-point set, for each method. The numbers along the top of the graph indicate the proportion of simulation runs in which the algorithm failed to identify any change-points. Figure 4.4 shows that the FMCI method performs better in the Gaussian and skewed Gaussian scenarios when the change-points are 'ascending'. However, the FMCI method can perform poorly in the heavy-tailed scenario. Notice that in the alternating



(a) (b)

Figure 4.5: (a) Twitter differenced log returns and (b) norms of the Twitter differenced log returns over time, with the FKWC detected change-points and change-point interval means overlaid.

scenario, under the heavy-tailed distribution, the FMCI method fails to identify a change-point over 90% of the time. We conclude that the FMCI method performs better when the data are not heavy tailed, but the FKWC performs better when the data are heavy tailed. In other words, the FKWC method sacrifices some of the accuracy of the FMCI method for robustness to outliers.

4.5 Data analysis

4.5.1 Changes in volatility of social media intraday returns

In this section we present an application of the multiple change-point algorithm to intraday differenced log returns of `twtr` stock. We analyse 207 daily asset price curves of `twtr` starting on June 24th 2019 and ending March 20th 2020. The price was measured in one minute intervals, over the course of the trading day, resulting in a total of 390 minutes per day. In order to account for edge effects from smoothing the curves, we trimmed 10% of the minutes from the beginning of the day and 5% of the minutes from the end of the day.

This resulted in 332 minutes of stock prices. The differenced log returns are defined as

$$X_{ji}(t) = \ln(Y_{ji\lfloor 331t \rfloor + 1}) - \ln(Y_{ji\lfloor 331t \rfloor}),$$

for $t \in (0, 1]$ and where Y_{jik} is the j^{th} asset price on the i^{th} day at minute k. The data was fit to a B-spline basis, using 50 basis functions, see `smooth.basis` in the `fda R` package. These data are shown in Figure 4.5. The assumption of zero mean appears to be satisfied here. Notice the outliers, which indicates that this data may require robust inference. Obviously, these data are not independent, however, we feel that this will not overtly affect our procedure. As long as the intervals between change-points are big enough, we expect the depth values after a change to also change, even if there is say, m-dependence in the data. We would expect that m-dependence could blur the change for a short period of time and cause the change-point estimate to be biased.

We ran the PELT algorithm with $\lambda_n = 3.74 + 0.3\sqrt{n} = 8.06$, as per Section 4.4. We ran the algorithm under all of the depth functions with the derivatives. The FKWC method using the MFHD depth identified a change-point on Jan 15, 2020 which the other two depth functions did not. If we include this change-point, the algorithm identified four change-points, as given in Table 4.1. The algorithm using the ranks of the squared norms without the derivatives identified the same set of change-points[3]. We could then infer that there were change-points in the magnitude of the observations. We remark that other, additional changes in the covariance operator may have also occurred at these times. For example, the magnitude change may also have been accompanied by a change in the shape of the curves.

Figure 4.5 displays the norms of the curves over time, with the estimated change-points added as vertical lines and the means of the norms in each interval overlaid. We can see clear changes in the mean of the norms during these periods. We may also notice that our procedure was unaffected by the outlier at the beginning of the series and the one just before the last estimated change-point. Table 4.1 gives the magnitude and sign of the changes as well. We can see that the largest change-point is the last one; clearly attributed to the instability caused by the coronavirus pandemic. It is interesting to see whether or not the other change-points occurred due to market wide behaviour, or events specific to social media or even just Twitter itself. For example, running the same algorithm on `snap` stock over the same period of time reproduces the change-points on Nov 07, 2019 and Feb 21, 2020 but not the other two change-points. One possibility for the estimated change-point on July 2019 could be the Twitter earnings report released just prior, e.g., (Feiner, 2019).

[3]The estimated November change-point was said to be two days earlier.

Interval	Centered Rank Mean
Jun 24, 2019 - Jul 24, 2019	19.73
Jul 24, 2019- Nov 07, 2019	-15.64
Nov 11, 2019- Jan 15, 2020	46.04
Jan 15, 2020- Feb 21, 2020	1.08
Feb 21, 2020 - Mar 20, 2020	-90.80

Table 4.1: Change-points and centered MFHD$'$ rank means. Notice the largest change occurs at the last change-point.

Aside from determining possible causes for change-points, from a modelling perspective, one may wish to avoid using a functional GARCH model. This could be due to the fact that in order to fit a functional GARCH model at the present time, one must choose to fit the functional data to a relatively small number of basis functions in order to keep to the number of parameters in the GARCH model small. If no clear basis exists, and the principle component analysis does not work well due to outliers, one may wish for an alternate approach. Instead one can remove the heteroskedasticity in the data by re-normalizing the curves in each interval, and then proceed with alternative time series modelling from there. Of course, this would not estimate future change-points; one could model the change-point process and the return curves separately.

4.5.2 Resting state f-MRI pre-processing

Functional magnetic resonance imaging, or f-MRI, is a type of imaging for brain activity. f-MRI uses magnetic fields to determine oxygen levels of blood in the brain in order to produce 3-dimensional images of the brain. Many of these images are taken over a period of time, which results in a time series of 3-dimensional images. Note that each MRI in a given subject's f-MRI can be viewed as a function on $[0,1]^3$. Resting state f-MRI is a type of f-MRI data where no intervention is applied to the subject during the scanning process. f-MRI scans go through extensive pre-processing before being analysed.

One assumption commonly made is that, after several pre-processing steps, each subject's resulting functional time series is stationary. It is therefore important to check the scans at an individual level in order to ensure that each time series is stationary. For subjects whose time series is not stationary, we must make the necessary corrections or exclusions from the ensuing data analysis. The covariance kernel of an f-MRI time series is a 6-dimensional function. Existing methods make a separability assumption on

the covariance kernel (Stoehr et al., 2021), which we do not make here. Additionally, Stoehr et al. (2021) mentions the need for a robust method of detecting non-stationarities in f-MRI data, which leads us to apply our rank based method to this data. We analyse several scans from the Beijing dataset, which were retrieved from `www.nitrc.org`. These scans were also analysed by Stoehr et al. (2021). Following instruction provided by `https://johnmuschelli.com/`, we performed the following pre-processing steps to the data. We trimmed the first 10 seconds from the beginning of the scan, in order to have a stable signal. We then performed rigid motion correction using `antsMotionCalculation` function in the `ANTsR R` package. A 0.1 Hz high-pass Butterworth filter of order 2 was applied voxel-wise to remove drift and trend from the data. We then removed 15 observations from either end of the time series in order to remove the edge effects of the filter. The gradient of each scan was then estimated using the `numDeriv` package, which resulted in four time series of functional data, where each function is a three-dimensional image.

We then computed the RP′ sample depth values as follows. First we projected each of the four time series' onto 50 unit functions. Then, for each of the 50 projected time series, we computed the half-space depth values of each four-dimensional observation. We then averaged these depths over the 50 unit vectors. We use the half-space depth since it is faster to compute than the simplicial depth for four-dimensional data. We then applied the FKWC hypothesis tests and the FKWC multiple change-point algorithm to the resulting depth values. In addition, we restricted estimated change-points to be at least 10 observations away from either boundary. Code to run the FKWC procedure on three-dimensional functional data can be retrieved from Github (Ramsay, 2021).

	AMOC		Epidemic		
Subject	Estimate	p-value	Estimate 1	Estimate 2	p-value
sub08455	116.00	0.36	34.00	54.00	0.95
sub08992	35.00	0.00	36.00	175.00	0.00
sub08816	39.00	0.20	43.00	94.00	1.00
sub34943	159.00	0.10	21.00	173.00	0.01
sub12220	30.00	0.09	31.00	127.00	0.25
sub06880	116.00	0.00	24.00	117.00	0.00

Table 4.2: Change-point estimates and p-values resulting from running the FKWC change-point tests on the different subjects.

Figure 4.6 contains the ranks of the random projection depth values for several f-MRI scans, with the resulting change-point intervals identified by the FKWC multiple change-point algorithm overlaid. Table 4.2 contains the p-values and change-point estimates re-

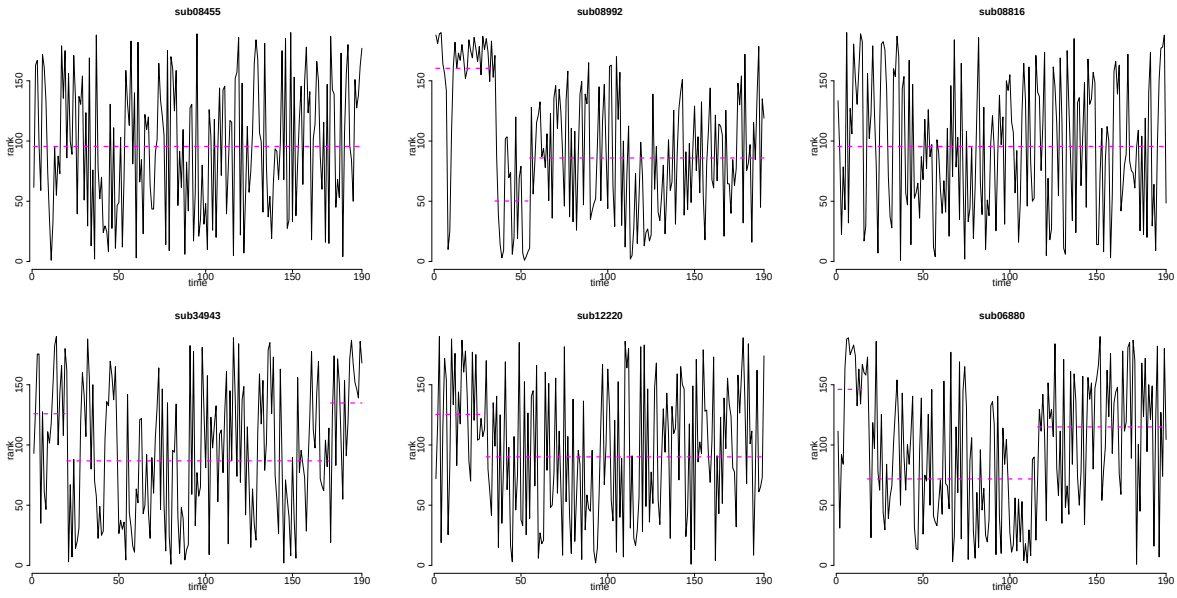


Figure 4.6: Ranks of the random projection depth values for several f-MRI scans with detected change-point means overlaid.

sulting from running the hypothesis testing procedures. We see that changes are detected in four of the six subjects analysed, two of which appear to be an epidemic type change; the ranks return to their previous means after the second interval. This is consistent with the idea that the epidemic model is more suitable for some resting state f-MRI scans (Stoehr et al., 2021). For subject `sub08455`, we do not detect any changes in the sequence, even though this subject's f-MRI has an outlier early in the sequence. This outlier can create a false positive for the AMOC alternative, as discussed by Stoehr et al. (2021). Notice that the p-values are not small for this subject when running our test. The rank sequence for subject `sub08992` was estimated to have two change-points, though the distribution of ranks in the first and third intervals are clearly different. This is why the estimates from the epidemic model and the multiple change-point procedures differ. Though, the null hypothesis is rejected by both the AMOC and epidemic model tests, even if we were to use any p-value correction procedure. In addition, the FKWC procedure ignores the outlier at the beginning of the sequence for subject `sub08992`, which showcases the robustness of the FKWC method. The FKWC methods did not detect any change-points for subject `sub08816`, whereas the methods of (Stoehr et al., 2021) detected an epidemic period. This could be due to differences in pre-processing, trimming, or the nature of the different methods' assumptions. For subject `sub34943` we see that an epidemic change is

detected by the multiple change-point procedure, and the p-value from the hypothesis is borderline significant with p-value corrections (0.01). In the case of the AMOC test, the null hypothesis is not rejected. We note that no change was detected by the functional procedure in (Stoehr et al., 2021), but a change was detected in the multivariate procedure. For subject `sub12220`, one change is detected, by the multiple change-point procedure, though the hypothesis testing yields non-significant results. We remark that the test of Stoehr et al. (2021) detects a change. The location of the change detected by Stoehr et al. (2021) occurs early in the sequence, and, as a result of the trimming we applied to the sequence, occurs very early in our time series. This makes it difficult to detect by our procedure. For subject `sub06880` we see that all three FKWC procedures agree that there are change-points, and the epidemic model and the multiple change-point procedure agree. We remark that Stoehr et al. (2021) also detected change-points in this subjects sequence of observations.

Chapter 5

Depth Methods for Private Data Analysis

5.1 Introduction

There is a large body of literature that shows simply removing the identifying information about subjects from a database is not enough to ensure data privacy (see Dwork et al., 2017, and the references therein). Even if only certain summary statistics are released, an adversary can still learn a surprising amount about individuals in a database (Dwork et al., 2017). This phenomena is largely due to auxiliary information that is known by the adversary. Given the large amount of information about individuals that is publicly available, it is not infeasible to assume that an adversary already knows some information about the individual they wish to learn about. On the contrary, if a statistic is differentially private an adversary cannot learn about the attributes of specific individuals in the original database, regardless of the amount of initial information the adversary possesses. This property, coupled with the lack of assumptions on the data itself needed to ensure privacy, accounts for the volume of recent literature on differentially private statistics.

One part of this literature represents a growing interest in the statistical community in differentially private inference, e.g., (Wasserman and Zhou, 2010; Awan et al., 2019; Cai et al., 2019; Brunel and Avella-Medina, 2020). In particular, there are connections between robust statistics and differentially private statistics, first discussed by Dwork and Lei (2009). Dwork and Lei (2009) introduced the propose-test-release algorithm for computing private estimators. This algorithm works particularly well with robust statistics. Brunel and Avella-Medina (2020) greatly expanded the propose-test-release paradigm of

Dwork and Lei (2009), using a concept from robust statistics called the finite sample breakdown point. The same authors use this idea to construct private median estimators with sub-Gaussian errors (Avella-Medina and Brunel, 2019). There are connections to other concepts from robust statistics; Chaudhuri and Hsu (2012) formalized a connection between private estimators and gross error sensitivity. The connection between private estimators and gross error sensitivity has been further exploited in order to construct differentially private statistics (Avella-Medina, 2019). Furthermore, private M-estimators were studied by several authors (Lei, 2011; Avella-Medina, 2019).

This chapter is inspired by these recent papers, where we explore the privatization of depth functions, a robust and nonparametric data analysis tool; given the recent success of robust procedures in the private setting, it is worthwhile to develop and study privatized depth functions and associated medians.

We present algorithms for computing differentially private depth values and depth-based medians. We study the cost of privacy of these methods; how the level of privacy protection affects the statistical utility of the depth function. Specifically, we consider generating a depth-based median via the exponential mechanism (McSherry and Talwar, 2007). We present the sample complexity of this estimator and show that it is polynomial in the dimension d. Furthermore, privatizing the estimator only increases the sample complexity by a factor of $\log d$ for a given error level comparable to that of the privacy parameter. Another important feature of our private median is that the range of the data can be unbounded. Many existing estimators require that the range of the data lie in a known ball or hypercube (Wasserman and Zhou, 2010; Cai et al., 2019). For many existing estimators, if the range is unknown one must sacrifice privacy budget to approximate the range of the data and the quality of the estimator depends on the size of this range. By contrast, our procedure allows for the data to have an unknown or unbounded range. In addition to private medians, we also present methods for computing private depth values and compute the associated cost of privacy. As a by-product, we extend the propose-test-release procedure of Brunel and Avella-Medina (2020) to be used with the exponential mechanism. We show that one can use propose-test-release to compute a private version of projection depth and the projection depth median. We show that the probability of returning a "null" value for the private projection depth values is small and give the cost of privacy for the projection depth values. As a by-product of this work, we present a smooth version of integrated dual depth as well as uniform concentration inequalities for several depth functions.

Note that some work has been done on the private computation of halfspace depth regions and the halfspace median (Beimel et al., 2019; Gao and Sheffet, 2020), mainly from a computational geometry point of view. Though Beimel et al. (2019) mentions that

the halfspace depth function can be used with the exponential mechanism, they do not study the estimator's properties from a statistical point of view; it is used as a method of finding a point in the convex hull of a set of points.

5.2 Differential privacy

Before getting into the fundamentals of differential privacy, it is useful to first introduce some notation. For two sets A and B, let $A \triangle B$ be the symmetric difference of A and B. In addition, recall that $|A|$ is the cardinality of A. We denote the open d-dimensional ball of radius r around x by $\mathcal{B}_r(x)$. Given $x \in \mathbb{R}^d$ we define the p-norm as $\|x\|_p = \left(\sum_{j=1}^d x_j^p\right)^{1/p}$ and given some function $\phi\colon \mathbb{R}^d \to \mathbb{R}$, we set $\|\phi\|_\infty = \sup_{y \in \mathbb{R}^d} \phi(y)$. We represent the data with $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ and assume that the data is a random sample of size n such that each observation is in $\mathbb{R}^d$. We use F_n to represent the empirical measure determined by $\mathbb{X}_n$. For a univariate distribution F, we use F^{-1} to denote the left continuous quantile function. Throughout the chapter we define the median of a continuous, univariate distribution by $\text{MED}(F) = F^{-1}(1/2)$. Both $\text{MED}(\mathbb{X}_n)$ and $\text{MED}(F_n)$ are taken to be the usual sample median. In other words, $\text{MED}(F_n)$ is the usual sample median and not $F_n^{-1}(1/2)$. We use $Q_{\mathbb{X}_n}$ to represent a measure that depends on the data set $\mathbb{X}_n$. Differentially private statistics will be denoted with the $\sim$ symbol, e.g., $\widetilde{T}$. Given a database $\mathbb{X}_n$, we let $\mathcal{D}(\mathbb{X}_n, k)$ be the set of all databases of size n which differ from $\mathbb{X}_n$ by k observations. In other words, define the space of all databases of size n containing points in $\mathbb{R}^d$ to be $\mathscr{D}_n^d$, then

$$\mathcal{D}(\mathbb{X}_n, k) = \{\mathbb{Y}_n \in \mathscr{D}_n^d\colon |\mathbb{Y}_n \triangle \mathbb{X}_n| = k\}.$$

A first essential concept when studying differential privacy is that of a mechanism. It has been shown that all differentially private statistics $\widetilde{T}(\mathbb{X}_n)$ must admit (non-degenerate) measures given the data $Q_{\mathbb{X}_n}$. This means that conditional on the observed dataset, a differentially private statistic (or database) is a random quantity (Dwork and Roth, 2014). We call the procedure that determines $Q_{\mathbb{X}_n}$ and then outputs a random draw $\widetilde{T}(\mathbb{X}_n)$ from $Q_{\mathbb{X}_n}$ a mechanism. We may also refer to the mechanism by $\widetilde{T}$ with an abuse of notation. A second essential concept for studying differential privacy is that of adjacent databases. Recall that $\mathbb{X}_n$ is a set of n d-dimensional vectors. Say $\mathbb{Y}_n$ is adjacent to $\mathbb{X}_n$ if $\mathbb{Y}_n$ is also a set of n d-dimensional vectors and the symmetric difference between $\mathbb{X}_n$ and $\mathbb{Y}_n$ contains one element. Colloquially, we say that $\mathbb{X}_n$ and $\mathbb{Y}_n$ (another random sample of size n) are adjacent if they differ by one observation; $\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, 1)$. Equipped with these concepts, we can now define differential privacy:

Definition 9. *A mechanism $\widetilde{T}$ is ϵ-differentially private for $\epsilon > 0$ if*

$$\frac{Q_{\mathbb{X}_n}(B)}{Q_{\mathbb{Y}_n}(B)} \leq e^{\epsilon} \tag{5.1}$$

holds for all measurable sets B and all pairs of adjacent datasets $\mathbb{X}_n$ and $\mathbb{Y}_n$.

The parameter ϵ should be small, implying that

$$\frac{Q_{\mathbb{X}_n}(B)}{Q_{\mathbb{Y}_n}(B)} \approx 1,$$

which gives the interpretation that the two measures $Q_{\mathbb{X}_n}$ and $Q_{\mathbb{Y}_n}$ are almost equivalent. To understand this definition, it helps to think of the problem from the adversary's point of view. Suppose that we are the adversary and that we have access to all the entries in the database except for one which we are trying to learn about. Therefore there is an unknown observation $\theta \in \mathbb{R}^d$ which we are trying to infer about. If $\widetilde{T}$ is released, how can we use it to conduct inference about θ? Suppose we want to know whether or not θ belongs to some set of observations Θ_0. For example, Θ_0 could be the set of all observations where the subject has blue eyes and has a specific rare disease. Thus, we want to test

$$H_0 \colon \theta \in \Theta_0 \text{ vs. } H_1 \colon \theta \notin \Theta_0.$$

To conduct this test, we would then ask two questions:

How likely was it to observe $\widetilde{T}$ under H_0? and *How likely was it to observe $\widetilde{T}$ under H_1?*

Differential privacy stipulates that both of these questions have practically the same answer, making it impossible to infer anything about θ from $\widetilde{T}$. Definition 9 implies that if someone in the dataset was replaced, we are just as likely to have seen $\widetilde{T}$ (or some value very close to $\widetilde{T}$ if $Q_{\mathbb{X}_n}$ is continuous). Another way to interpret the definition is to observe that differential privacy implies that $\mathrm{KL}(Q_{\mathbb{X}_n}, Q_{\mathbb{Y}_n}) < \epsilon$, where KL is the Kullback–Leibler divergence; implying that the distributions are necessarily close.

One may observe that the inequality (5.1) must hold for all pairs of adjacent databases and all possible outcomes of the estimator in order for the procedure to be differentially private. This inequality is then a worst case restriction, in the sense that (5.1) must hold for even the worst possible database and the worst possible outcome of the mechanism. Definition 9 can be difficult to satisfy because the umbrella of 'all databases and mechanism outputs' can include both some extreme databases and extreme mechanism outputs. One

may wish to relax this definition over unlikely mechanism outputs; one way to do this is if B is such that $Q_{\mathbb{X}_n}(B)$ is very small, then the bound could be allowed to fail. This is called approximate differential privacy or (ϵ, δ)-differential privacy, in which we have

$$Q_{\mathbb{X}_n}(B) \leq e^{\epsilon} Q_{\mathbb{Y}_n}(B) + \delta \tag{5.2}$$

in place of the condition (5.1). Typically, $\delta << \epsilon$, and δ can be interpreted as the probability under which the bound is allowed to fail. To see this, observe that for B such that $Q_{\mathbb{X}_n}(B) < \delta$, (5.2) holds regardless of ϵ. We mention that for remainder of the chapter, ϵ and δ are always assumed to be positive and that sometimes we may have that the privacy parameters are a function of the sample size or dimension. Taking the privacy parameter to be decreasing in n is a very safe assumption in terms of privacy; given enough samples, the database participants will surely be protected.

An important feature of differentially private statistics is that they are immune to post-processing (Dwork and Roth, 2014). Functions of differentially private estimates maintain the same level of privacy:

Proposition 1 (Dwork and Roth (2014)). *Let $\widetilde{T} \sim Q_{\mathbb{X}_n}$ be (ϵ, δ)-differentially private, then $f(\widetilde{T})$ is also (ϵ, δ)-differentially private for any data-independent map f.*

We can also compose multiple private statistics together; a vector of N differentially private procedures will also be differentially private.

Proposition 2 (Dwork and Roth (2014)). *Suppose we have $T_1, \ldots, T_N$ differentially private statistics, with privacy parameters $(\epsilon_1, \delta_1), \ldots, (\epsilon_N, \delta_N)$. Then the vector $(T_1, \ldots, T_N)$ is $\left(\sum_{j=1}^N \epsilon_j, \sum_{j=1}^N \delta_j\right)$-differentially private.*

Central to differentially private algorithms is the concept of sensitivity. Then the global sensitivity GS_n of a statistic $T \colon \mathcal{X}^n \to \mathcal{S}$ is the maximum distance between T evaluated at two adjacent databases of size n:

$$\mathrm{GS}_n(T; \|\cdot\|_{\mathcal{S}}) = \sup_{\substack{\mathbb{X}_n \in \mathscr{D}_n^d, \\ \mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, 1)}} \|T(\mathbb{X}_n) - T(\mathbb{Y}_n)\|_{\mathcal{S}} \,.$$

The norm used depends on T and the mechanism, therefore we may use the notation $\mathrm{GS}_n(T; \|\cdot\|_{\mathcal{S}})$. When the norm is clear from the context, we will simply write $\mathrm{GS}_n(T)$. Note that T does not have to be Euclidean valued; the range of $T(\mathbb{X}_n)$ can lie in any normed space.

We can now introduce some important building blocks of differentially private algorithms. Let $W_1, \ldots, W_k, \ldots$ and $Z_1, \ldots, Z_k, \ldots$ represent a sequence of independent, standard Laplace random variables and a sequence of independent, standard Gaussian random variables, respectively. The Laplace and Gaussian mechanisms are essential differentially private mechanisms; they define how much an estimator must be perturbed in order for it to be differentially private.

Mechanism 1 (Dwork et al. (2006)). *Given a statistic $T\colon \mathcal{X} \to \mathbb{R}^k$, the mechanism that outputs*

$$\widetilde{T}(\mathbb{X}_n) = T(\mathbb{X}_n) + (W_1, \ldots, W_k)\ \frac{\mathrm{GS}_n(T; \|\cdot\|_1)}{\epsilon},$$

is ϵ-differentially private.

Mechanism 2 (Dwork et al. (2006); Dwork and Roth (2014)). *Given a statistic $T\colon \mathcal{X} \to \mathbb{R}^k$, the mechanism that outputs*

$$\widetilde{T}(\mathbb{X}_n) = T(\mathbb{X}_n) + (Z_1, \ldots, Z_k)\ \frac{\sqrt{2\log(1.25/\delta)}\,\mathrm{GS}_n(T;; \|\cdot\|_2)}{\epsilon}$$

is (ϵ, δ)-differentially private.

This can be improved in strict privacy scenarios (Balle and Wang, 2018). We can also add noise based on smooth sensitivity (Nissim et al., 2007). Using smooth sensitivity allows the user to leverage improbable, worst case local sensitivities. Often in practice, statistics are computed by maximizing a data driven objective function $\phi_{\mathbb{X}_n}(\cdot)$. We can privatize such a procedure via the exponential mechanism. The exponential mechanism can be defined as follows:

Mechanism 3 (McSherry and Talwar (2007)). *Given the data, consider a function $\phi_{\mathbb{X}_n} : \mathbb{R}^k \to \mathbb{R}$. Then a random draw from the density $f(x; \phi_{\mathbb{X}_n}, \epsilon)$ that satisfies*

$$f(x; \phi_{\mathbb{X}_n}, \epsilon) \propto \exp\left(\frac{\epsilon\ \phi_{\mathbb{X}_n}(x)}{2\mathrm{GS}_n(\phi_{\mathbb{X}_n}; \|\cdot\|_\infty)}\right),$$

is an ϵ-differentially private mechanism. It is assumed that

$$\int_{\mathbb{R}^k} \exp\left(\frac{\epsilon\ \phi_{\mathbb{X}_n}(x)}{2\mathrm{GS}_n(\phi_{\mathbb{X}_n}; \|\cdot\|_\infty)}\right) dx < \infty.$$

The factor of 2 can be removed if the normalizing term is independent of the sample. All of the mechanisms discussed so far require that the non-private statistic or objective function has finite global sensitivity. This is a somewhat strict requirement; under the Gaussian model neither the sample mean nor sample median have finite global sensitivity, even when $d = 1$. Instead, we might check the local sensitivity. For a fixed database $\mathbb{X}_n$ the local sensitivity of T is defined as

$$\mathrm{LS}_n(T, \mathbb{X}_n; \|\cdot\|_{\mathcal{S}}) = \sup_{\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, 1)} \|T(\mathbb{X}_n) - T(\mathbb{Y}_n)\|_{\mathcal{S}}\,.$$

Note that LS depends on the sample. Instead of requiring that the statistic has small global sensitivity we can instead leverage the local sensitivity, provided it is small with high probability. For example, the sample median has, on average, low local sensitivity, viz.

$$\mathrm{LS}(\mathrm{MED}(\mathbb{X}_n)) \leq |F_n^{-1}(1/2 - 1/n) - F_n^{-1}(1/2 + 1/n)|,$$

when $d = 1$. Since $1/n \to 0$, we expect this value to be small (assuming the sample comes from a distribution which is continuous at its median).

The propose-test-release mechanism, or PTR, can be used to generate private versions of statistics with infinite global sensitivity but highly probable low local sensitivity. Propose-test-release was introduced by Dwork and Lei (2009) but was greatly expanded in the recent paper by Brunel and Avella-Medina (2020). The PTR algorithm of Brunel and Avella-Medina (2020) relies on the truncated breakdown point A_η, which is the minimum number of points that must be changed in order to move an estimator by η:

$$A_\eta(T; \mathbb{X}_n) = \min\left\{k \colon \sup_{\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, k)} \|T(\mathbb{X}_n) - T(\mathbb{Y}_n)\| > \eta\right\}, \tag{5.3}$$

where one recalls that $\mathcal{D}(\mathbb{X}_n, k)$ is the set of all samples that differ from $\mathbb{X}_n$ by k observations. Unlike for the traditional breakdown point, the dependence of $A_\eta(T; \mathbb{X}_n)$ on $\mathbb{X}_n$ is important. PTR works by proposing a statistic, testing if it is insensitive and then releasing it if it is, in fact, insensitive. A private version of $A_\eta(T; \mathbb{X}_n)$ is used to check the sensitivity.

Mechanism 4. *Given a statistic $T \colon \mathcal{X} \to \mathbb{R}^k$, the mechanism that outputs*

$$\widetilde{T}(\mathbb{X}_n) = \begin{cases} \perp & \text{if } A_\eta(T; \mathbb{X}_n) + \frac{1}{\epsilon} W_1 \leq 1 + \frac{\log(2/\delta)}{\epsilon} \\ T(\mathbb{X}_n) + \frac{\eta}{\epsilon} W_2 & o.w. \end{cases} \tag{5.4}$$

is $(2\epsilon, \delta)$ *differentially private and the statistic*

$$\widetilde{T}(\mathbb{X}_n) = \begin{cases} \perp & \text{if } A_\eta(T; \mathbb{X}_n) + \frac{\sqrt{2\log(1.25/\delta)}}{\epsilon} Z_1 \leq 1 + \frac{2\log(1.25/\delta)}{\epsilon} \\ T(\mathbb{X}_n) + \frac{\eta\sqrt{2\log(1.25/\delta)}}{\epsilon} Z_2 & o.w. \end{cases} \tag{5.5}$$

is $(2\epsilon, 2e^\epsilon\delta + \delta^2)$ *differentially private.*

The release of $\perp$, i.e., a "null" value, means that the dataset was too sensitive for the statistic to be released. The goal is to choose a T such that releasing $\perp$ is incredibly unlikely; $A_\eta(T; \mathbb{X}_n)$ should be large with high probability.

5.3 A smooth depth function

We will be constructing differentially private statistics from depth functions. We will focus mainly on halfspace depth, projection depth, integrated rank-weighted depth and integrated dual depth. We also discuss simplicial depth in Section A.8. One may notice that the sample versions of all of these depth functions are not smooth functions. We may wish to have a smooth sample depth function in order to use certain Markov chain Monte Carlo computational techniques to draw private estimates. For example, in the exponential mechanism, the Hessian of the objective function can be used to produce computational guarantees for Langevin dynamics. The existing depth functions are not smooth because they rely on empirical distribution functions, or EDFs, which are not smooth. One way to produce a smoothed version of the depth function is to replace the indicators in the EDF with the "Sigmoid" or "expit" function: $\mathbb{1}(Y \leq x) \approx \sigma(\beta(x - Y))$, where $\sigma(x) = (1+e^{-x})^{-1}$. The parameter β controls the smoothness of the depth function, increasing β reduces the smoothness of the depth function. The smoothed depth is then defined as

Definition 10 (β-Integrated Dual Depth)**.** *Define* β*-integrated dual depth as*

$$\mathrm{IDD}_\beta(x; F) = \int_{S^{d-1}} \mathrm{E}\left[\sigma\left(\beta(x - X)^\top u\right)\right] \cdot \left(1 - \mathrm{E}\left[\sigma\left(\beta(x - X)^\top u\right)\right]\right) d\nu(u),$$

where ν *is the uniform measure on* S^{d-1}*.*

Observe that as $\beta \to \infty$, this depth function converges to the integrated dual depth. We show that β-integrated dual depth satisfies the same "depth"-like properties as integrated dual depth. Let $AF + b$ represents the distribution of $AX + b$ if $X \sim F$.

Theorem 11. *The β-integrated dual depth satisfies the following properties:*

1. *Vanishing at infinity:* $\lim_{c\to\infty} \mathrm{IDD}_\beta(cu; F) = 0$ *for any unit vector* u.
2. *Maximality at centre: If* $X \sim F$ *and* F *is such that* $X - \theta \overset{d}{=} \theta - X$, *it holds that* $\sup_x \mathrm{IDD}_\beta(x; F) = \mathrm{IDD}_\beta(\theta; F)$.
3. *Similarity invariance: For all orthogonal matrices* A *and vectors* $b \in \mathbb{R}^d$ *it holds that* $\mathrm{IDD}_\beta(x; F) = \mathrm{IDD}_\beta(A\theta + b; AF + b)$.
4. *Decreasing along rays: If* $X \sim F$ *and* F *is such that* $X - \theta \overset{d}{=} \theta - X$, *then for all* $0 < \alpha < 1$ $\mathrm{IDD}_\beta(\alpha\theta + (1-\alpha)x; F) > \mathrm{IDD}_\beta(x; F)$.

In practice, computing the integral in Definition 10 is infeasible for even moderate dimensions. Instead, we compute a Monte Carlo estimate of the depth values:

$$\widehat{\mathrm{IDD}}_\beta(x; F_n) = \frac{1}{M}\sum_{m=1}^{M} \frac{1}{n}\sum_{i=1}^{n} \sigma\left(\beta(x - X_i)^\top u_m\right)\left(1 - \sigma\left(\beta(x - X_i)^\top u_m\right)\right).$$

We now show that the approximation $\widehat{\mathrm{IDD}}_\beta$ maintains the same uniform convergence rate of IDD when $M > n \log n$:

Theorem 12. *Suppose that for some positive integer* M, $u_1, \ldots, u_M$ *are sampled uniformly from the* d*-dimensional unit sphere. Then, for all* $t > 0$, *there exists a universal constant* K *such that*

$$\Pr\left(\sup_x \left|\widehat{\mathrm{IDD}}_\beta(x; F_n) - \mathrm{IDD}_\beta(x; F_n)\right| \geq t\right) \leq 2e^{(d+n+1)\log\left(\frac{KM}{2(d+n+1)}\right) - Mt^2/2},$$

We have now confirmed that the smoothed depth is indeed a depth function and is computable with statistical guarantees.

5.4 Estimating private depth-medians

We now discuss estimating private depth-based medians. It may seem intuitive to start with estimating private depth values, but we will see that we may use techniques from this

section in order to estimate private depth values. The goal of this section is to produce a private estimate of the depth-based median: $\operatorname{argmax} \mathrm{D}(x; F)$. There are several ways in which we could approach privatizing depth-based inference. A natural and easy way to do this is to start with a differentially private estimate of the distribution of the data $\widetilde{F}_n$ and use $\mathrm{D}(x, \widetilde{F}_n)$, which is differentially private. Computing $\widetilde{F}_n$ relies on existing methods for generating private multidimensional empirical distribution functions, which are independent of the concept of depth. Since they do not rely on the properties of the depth function, methods based on differentially private estimates of the distribution fail to take advantage of any robustness properties of depth functions.

Therefore, we opt for an algorithm based on the global sensitivity of D. Take $\mathrm{D}(x; \mathbb{X}_n)$ to be the depth of x with respect to the empirical distribution determined by $\mathbb{X}_n$. Then the global sensitivity GS_n of D is the maximum distance between the sample depth functions for two adjacent databases of size n:

$$\mathrm{GS}_n(\mathrm{D}) = \sup_{\substack{\mathbb{X}_n \in \mathscr{D}_n^d, \\ \mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, 1)}} \sup_x |\mathrm{D}(x; \mathbb{X}_n) - \mathrm{D}(x; \mathbb{Y}_n)|.$$

We observe that the global sensitivity $\mathrm{GS}_n(\mathrm{D})$ of the halfspace, simplicial, integrated dual, integrated rank-weighted and β-integrated dual depths satisfy $\mathrm{GS}_n(\mathrm{D}) = C/n$[1]. For example, halfspace depth has global sensitivity $1/n$. Since estimating depth-based medians is an optimization problem whose objective function has small global sensitivity, it is natural to use the exponential mechanism to compute a private estimate.

However, all of the depth functions we consider lie in some bounded, positive interval. This fact implies that the density given by the standard exponential mechanism is not valid:

$$\int_{\mathbb{R}^d} \exp\left(\frac{\epsilon}{2\mathrm{GS}_n(\mathrm{D})} \mathrm{D}(x; F_n)\right) dx = \infty.$$

To remedy this, we add a prior π on the median:

$$f(x; F_n) = \frac{\exp\left(\frac{\epsilon}{2\mathrm{GS}_n(\mathrm{D})} \mathrm{D}(x; F_n)\right) \pi(x)}{\int_{\mathbb{R}^d} \exp\left(\frac{\epsilon}{2\mathrm{GS}_n(\mathrm{D})} \mathrm{D}(x; F_n)\right) \pi(x) dx}. \tag{5.6}$$

Mechanism 5. *Suppose that* $\mathrm{GS}_n(\mathrm{D}) = C(\mathrm{D})/n$. *Suppose also that* $\pi(x)$ *is a density*

[1]For the proof of this, see Section A.9

chosen independently of the data and that

$$f(x; F_n) = \frac{\exp\left(\frac{n\epsilon}{2C(\mathrm{D})}\mathrm{D}(x; F_n)\right)\pi(x)}{\int_{\mathbb{R}^d}\exp\left(\frac{n\epsilon}{2C(\mathrm{D})}\mathrm{D}(x; F_n)\right)\pi(x)dx}, \tag{5.7}$$

is a valid Lebesgue density. Suppose that $\widetilde{T}(\mathbb{X}_n)$ *is a random draw from* $f(x; F_n)$*. Then* $\widetilde{T}(\mathbb{X}_n)$ *is an* ϵ*-differentially private estimate of the* D*-based median of* $\mathbb{X}_n$*.*

It is given by differential privacy of the exponential mechanism that one sample from f is ϵ-differentially private. The next step is to assess the accuracy of the private median estimate. The following is a general concentration result for one draw from the exponential mechanism, given an objective function ϕ_n. We will use this to assess the accuracy of our private median estimate.

Theorem 13. *Let* π *be a measure on* $\mathbb{R}^d$*. Suppose that* $\phi_n(\omega, x)\colon \Omega \times \mathbb{R}^d \to \mathbb{R}^+$ *is a sequence of random functions on the probability space* $(\Omega, \mathscr{A}, P)$ *and* λ_n *is a sequence of positive real numbers. Assume that*

1. $\Pr\left(\|\phi_n - \phi\|_\infty > t\right) \leq C_1(\phi, d, n)e^{-C_2(\phi)nt^2}$ *for all* $t > 0$*.*
2. ϕ *is uniquely maximized at* $x = \theta$*,* $\|\theta\| < \infty$*.*
3. ϕ *is uniformly continuous with modulus of continuity* ω_ϕ*.*
4. *Suppose that* $\min_{k\in[0,1]}[\omega_\phi(k) \wedge 1 - \lambda_n^{-1}\log\pi(\mathcal{B}_k(\theta))]$ *exists.*

Let $\widetilde{T}(\mathbb{X}_n)$ *be a draw from the measure*

$$Q_n(A) = \int_\Omega\int_A \frac{e^{\lambda_n\phi_n(\omega,x)}d\pi}{\int_{\mathbb{R}^d} e^{\lambda_n\phi_n(\omega,x)}d\pi}dP,$$

for $A \in \mathscr{B}(\mathbb{R}^d)$*. Then,*

$$\begin{aligned}\Pr\left(\left\|\widetilde{T}(\mathbb{X}_n) - \theta\right\| > t\right) \leq\ & C_1(\phi, d, n)e^{-C_2(\phi)n\left(\phi(\theta) - \sup_{\{x:\ \|x-\theta\|>t\}}\phi(x)\right)^2/16} \\ & + e^{-\lambda_n\left(\phi(\theta) - \sup_{\{x:\ \|x-\theta\|>t\}}\phi(x)\right)/2 + \lambda_n\min_{k\in[0,1]}[\omega_\phi(k)\wedge 1 - \lambda_n^{-1}\log\pi(\mathcal{B}_k(\theta))]}.\end{aligned}$$

In order to apply Theorem 13 to the depth functions, we need to check conditions 1-3 hold for the considered depth functions. We begin with condition 1. In the following we let F_u denote the law of $X^\top u$ where $X \sim F$.

Theorem 14. *Let D denote either* IRW, HD, IDD_β *or* IDD. *Then there are some universal $K, c > 0$ such that*

$$\Pr\left(\sup_{x\in\mathbb{R}^d} |\mathrm{D}(x; F_n) - \mathrm{D}(x; F)| > t\right) \leq Ke^{(d+1)\log\left(K\frac{n}{d+1}\right)-cnt^2}.$$

Now that condition 1 is settled, we need to verify condition 2 and condition 3. Let $\alpha_{\mathrm{D}}(t) = D(\theta; F) - \sup_{\mathcal{B}_t^c(\theta)} D(x; F)$. In the following if F is a law on $\mathbb{R}^d$ and we let f_u denote the density of the law of $X^\top u$ with respect to the Lebesgue measure.

Theorem 15. *Suppose that* D *is one of the four depth functions discussed above, F is such that* D *has a unique maximizer, θ, and that f_u is well-defined with $\sup_u \|f_u\|_\infty < \infty$. Let C', C be some universal constants and let $\widetilde{T}_n$ be as in (5.7). Let N_0 be such that for $n \geq N_0$, we have*

$$\min_{k>0}\left[(C' \cdot ||f||_\infty \cdot k) \wedge 1 - \frac{C}{n\epsilon}\log \pi(\mathcal{B}_k(\theta))\right] < \alpha_{\mathrm{D}}(t)/4.$$

Then there are some $K = K(F, D)$ and some universal $c, c' > 0$, such that for any $t > 0$ and $n \geq N_0$,

$$\Pr\left(\left\|\widetilde{T}_n - \theta\right\| > t\right) \leq Ke^{(d+1)\log\left(K\frac{n}{d+1}\right)-cn\alpha_D(t)^2} + Ke^{-c'n\epsilon\alpha_D(t)}.$$

Remark 2. *Halfspace depth has a unique median when F is absolutely continuous with connected support. Note that F being symmetric about a point θ is sufficient for integrated rank-weighted, β-integrated dual depth and integrated dual depth to have unique medians.*

Remark 3. *The assumption $\sup_u \|f_u\|_\infty < \infty$ adds little restriction. Note that $f_u(x) = \int_{U^\perp} f(xu + z)dz$ (see the proof of Theorem 15 in Section A.9 for details). Since f is the density of F, we know that $1 = \int_{\mathbb{R}}\int_{U^\perp} f(xu + z)dzdx$, which implies that $f_u(x)$ is bounded almost surely; the inner integral has to converge except on null sets of $\mathbb{R}$. Therefore, the assumption eliminates the case where $f_u(x)$ is unbounded on sets of measure 0.*

Remark 4. *The dependence of the bound on the prior is encapsulated in the condition*

$$\min_{k>0}\left[C' \cdot ||f||_\infty \cdot k \wedge 1 - \frac{C}{n\epsilon}\log \pi(\mathcal{B}_k(\theta))\right] < \alpha_{\mathrm{D}}(t)/4.$$

For example, if π is sub-gaussian with mean μ and coordinate variances ς^2, then we require n to be such that

$$\min_{k>0}\left[C'\cdot||f||_\infty\cdot k\wedge 1-\frac{C}{n\epsilon}C_2\log\left(1-Ke^{-\frac{k^2}{C_3\|\theta-\mu\|+C_4\varsigma(1+\sqrt{d})}}\right)\right]<\alpha_{\mathrm{D}}(t)/4.$$

Note that as $n\epsilon$ grows, $\min_{k>0}\left[C'\cdot||f||_\infty\cdot k\wedge 1-\frac{C}{n\epsilon}\log\pi(\mathcal{B}_k(\theta))\right]$ generally approaches 0; as $n\epsilon\to\infty$, the dependence on the prior becomes negligible.

We can use Theorem 15 to compute the sample complexity of these medians:

Corollary 1. *Suppose that $\epsilon>0$, $t>0$ and $0<\gamma<1$, and the conditions of Theorem 15 hold. Let N_0 be as in Theorem 15. Suppose Mechanism 5 is paired with the four depths discussed above. Then, there are constants $K_1=K_1(F,\mathrm{D}), K_2=K_2(F,\mathrm{D})$ and some universal $K_3>0$ such that the number of samples needed to estimate the private median within an error of t, with probability $1-\gamma$, is less than*

$$n(t,\gamma,d,\epsilon)=\max\left\{\left(\frac{K_1\log(1/\gamma)+K_2d}{\alpha_{\mathrm{D}}(t)^2}\right)^{\frac{1}{1-r}},\frac{K_3\log(1/\gamma)}{\epsilon\cdot\alpha_{\mathrm{D}}(t)},N_0\right\},$$

for any $r>0$.

Remark 5. *If the depth function is decreasing along rays that emanate from the deepest point, then we have that $\alpha_{\mathrm{D}}(t)=\mathrm{D}(\theta)-\sup_{\|x-\theta\|=t}\mathrm{D}(\theta)$. Therefore, $\alpha_{\mathrm{D}}(t)$ can be interpreted as the smallest difference in depth between the median and a point whose distance to the median is t. In other words, if x is a distance of t from the median, then the depth of x is at least $\alpha_{\mathrm{D}}(t)$ shallower than the median.*

Corollary 2. *Suppose that $t>0$ and $0<\gamma<1$, and the conditions of Theorem 15 hold. Let K_1,K_2 be as in Corollary 1. Suppose Mechanism 5 is paired with the four depths discussed above and that $\pi=\mathcal{N}(\mu,\varsigma^2 I)$. Then there exists universal constants K_3,K_4 such that the number of samples needed to estimate the private median within an error of t, with probability $1-\gamma$, is less than*

$$n(t,\gamma,d,\epsilon)=\max\left\{\left(\frac{K_1\log(1/\gamma)+K_2d}{\alpha_{\mathrm{D}}(t)^2}\right)^{\frac{1}{1-r}},\frac{K_3\log(1/\gamma)\vee\left[K_4\|\mu-\theta\|^2+K_4d\log\left(\frac{\varsigma}{\alpha_{\mathrm{D}}(t)}\vee d\right)\right]}{\epsilon\cdot\alpha_{\mathrm{D}}(t)}\right\},$$

for any $r > 0$. If instead $\pi \propto \mathbb{1}(x \in \mathcal{C}_k)$, where $\mathcal{C}_k$ is a cube with side lengths k, then the number of samples needed is

$$n(t, \gamma, d, \epsilon) = \max\left\{ \left(\frac{K_1 \log(1/\gamma) + K_2 d}{\alpha_{\mathrm{D}}(t)^2} \right)^{\frac{1}{1-r}}, \frac{K_3 \log(1/\gamma) \vee \left[K_4 d \log\left(\frac{1}{\alpha_{\mathrm{D}}(t)} \vee d \right) \right]}{\epsilon \cdot \alpha_{\mathrm{D}}(t)} \right\},$$

for any $r > 0$.

The Gaussian prior allows for estimation of the median when one does not want to assume the data lies in a compact set. On the other hand, the assumption that the data lies in a compact set can be incorporated by setting $\pi = \mathbb{1}(x \in \mathcal{C}_k)$. The sample complexity when using a Gaussian prior is comparable to that of trimming the data, provided we have a starting point that is within $c\sqrt{d \log d}$ of θ and variance such that $\varsigma / \alpha_{\mathrm{D}}(t)$ is polynomial in d. The first term in the sample complexity is a bound on the non-private sample complexity of the depth-based median, notice it does not contain features of the prior or the privacy parameter. On the other hand, the second term captures the cost of privacy. If the prior is either some compact set which contains θ or is Gaussian with $\varsigma / \alpha_{\mathrm{D}}(t) \propto d$ and $\|\mu - \theta\| \leq \sqrt{d \log d}$ then the private estimate adds at most a multiplicative factor of $\alpha_{\mathrm{D}}(t) \log d / \epsilon$ to the sample complexity.

For example, consider the case where $F = \mathcal{N}_d(\theta, \Sigma)$ and $\pi = \mathcal{N}(\mu, I)$. Suppose that λ_d is the smallest eigenvalue associated with Σ. Then, we have that

$$\alpha_{\mathrm{HD}}(t) = 1/2 - \sup_{\|x\| = t} \inf_{u} \Phi\left(\frac{(x - \theta)^\top u}{u^\top \Sigma u} \right) = \Phi(t/\lambda_d) - 1/2.$$

Therefore, under halfspace depth, we have that

$$n(t, \gamma, d, \epsilon) = \max\left\{ \left(\frac{K_1 \log(1/\gamma) + K_2 d}{(\Phi(t/\lambda_d) - 1/2)^2} \right)^{\frac{1}{0.99}}, \frac{K_3 \log(1/\gamma) \vee \left[C \|\mu - \theta\|^2 + C d \log\left(\frac{1}{(\Phi(t/\lambda_d) - 1/2)} \vee d \right) \right]}{\epsilon (\Phi(t/\lambda_d) - 1/2)} \right\}.$$

Suppose we would like to achieve some fixed level of error t with success rate $1 - e^{-d}$ at constant privacy level ϵ. For large d, the right hand term dominates the sample complexity. Therefore, if the prior is relatively accurate, say, $\|\mu - \theta\| \leq \sqrt{d \log d}$, we only need n to grow slightly faster than the dimension: $n \gtrsim d \log d$. There is also a trade-off

Distribution:	$N_d(\theta, \Sigma)$	d-version symmetric
$\alpha_{\text{HD}}(t)$	$1/2 - \Phi(-t/\sqrt{\lambda_d})$	$1/2 - \sup_{v \in S^{d-1}} \inf_u F_0\left(\frac{-tv^\top u}{a(u)}\right)$
$\alpha_{\text{IRW}}(t)$	$\int \lvert 1/2 - \Phi\left(\frac{tv_d^\top u}{\sqrt{u^\top \Sigma u}}\right) \rvert d\nu$	$\inf_{v \in S^{d-1}} \int \lvert \frac{1}{2} - F_0\left(\frac{tv^\top u}{a(u)}\right) \rvert d\nu$
$\alpha_{\text{IDD}}(t)$	$\frac{1}{4} - \int \Phi\left(\frac{-tv_d^\top u}{\sqrt{u^\top \Sigma u}}\right) \Phi\left(\frac{tv^\top u}{u^\top \Sigma u}\right) d\nu$	$\inf_{v \in S^{d-1}} \frac{1}{4} - \int F_0\left(\frac{tv^\top u}{a(u)}\right) F_0\left(\frac{-tv^\top u}{a(u)}\right) d\nu$

Table 5.1: Table of α_{D} for different depth functions and underlying distributions. Note that v_d is the eigenvector associated with the smallest eigenvalue of Σ. Recall that X is d-version symmetric about zero if $X^\top u \overset{\text{d}}{=} a(u)Z$ where Z has law F_0 and $Z \overset{\text{d}}{=} -Z$ and $a(u) = a(-u)$.

between the error tolerance t and the privacy parameter ϵ. For example, for very small error: $(\Phi(t/\lambda) - 1/2) \leq \epsilon/\log d$, then the privatization does not contribute to the sample complexity. Table 5.1 contains α for different depth functions and distributions. Recall that X is d-version symmetric about zero if $X^\top u \overset{\text{d}}{=} a(u)Z$ where Z has law F_0 and $Z \overset{\text{d}}{=} -Z$ and $a(u) = a(-u)$ (Eaton, 1981). For example, elliptically symmetric distributions have $a(u) = \sqrt{u'\Sigma u}$ and F constructed from independent Cauchy marginals with mean 0 and scale γ_i gives $a(u) = \sum_{j=1}^d |u_i|\gamma_i$. It is easy to see that the largest depth is then in the direction with the smallest scale. Without loss of generality, suppose the direction with the smallest scale is in the x-direction,[2] it follows that $\alpha_{\text{HD}}(t) = 1/2 - \frac{1}{\pi} \arctan\left(\frac{-t}{\gamma_1}\right)$.

5.5 Computing private depth values

There are several methods for computing private depth values. If we know a-priori at which values we would like to compute the depth value at, then we can use the Laplace/Gaussian mechanisms:

Corollary 3. *Let D denote either* IRW, HD, IDD_β *or* IDD. *For x given independently of the data, the following estimators*

$$\widetilde{\text{D}}_1(x; F_n) = \text{D}(x; F_n) + W\frac{C}{n\epsilon} \qquad \textit{and} \qquad \widetilde{\text{D}}_2(x; F_n) = \text{D}(x; F_n) + Z\frac{C\sqrt{2\log(1.25/\delta)}}{n\epsilon}$$

are ϵ-differentially private and (ϵ, δ)-differentially private, respectively. In addition, there

[2] In other words, the direction $(1, 0, \ldots, 0)$.

are some universal constants $K, c, C_1, C_2 > 0$ *such that*

1. $\Pr\left(\left|\widetilde{\mathrm{D}}_1(x; F_n) - \mathrm{D}(x; F)\right| > t\right) \leq Ke^{(d+1)\log\left(K\frac{n}{d+1}\right) - cnt(t\wedge\epsilon)}$.
2. $\Pr\left(\left|\widetilde{\mathrm{D}}_2(x; F_n) - \mathrm{D}(x; F)\right| > t\right) \leq C_1 e^{-C_2 t^2 (n\epsilon)^2/\log(1.25/\delta)} + Ke^{(d+1)\log\left(K\frac{n}{d+1}\right) - cnt^2}$.

Note that Corollary 3 follows from Theorem 14. However, many inference procedures require that we compute the depth values at the sample points. How we can estimate the vector of depth values at the sample points, i.e.,

$$\widehat{\mathbf{D}}(F_n) = (\mathrm{D}(X_1; F_n), \mathrm{D}(X_2; F_n), \ldots, \mathrm{D}(X_n; F_n)) \tag{5.8}$$

privately? The sample values now appear in both arguments of D and so we must do a bit more work to compute the global sensitivity. It works out that the global sensitivities of the vector of sample depth values for the considered depth functions with finite GS_n are all close to 1. This implies that for the full vector of sample depth values we do not get privacy for free in the limit. For the considered depth functions we have that

$$\sup_x |\mathrm{D}(x; F_n) - \mathrm{D}(x; F)| = O_p(n^{-1/2}),$$

which gives that

$$\|\mathbf{D}(F_n) - \mathbf{D}(F)\| \leq \sqrt{n\left(\sup_x |\mathrm{D}(x; F_n) - \mathrm{D}(x; F)|\right)^2} = O_p(1).$$

Let $\mathbf{W} = (W_1, \ldots, W_n)$. Then, if $\widetilde{\mathbf{D}}$ is Mechanism 1 applied to (5.8), we have that

$$\begin{aligned}\left\|\widetilde{\mathbf{D}}(F_n) - \mathbf{D}(F)\right\| &\leq \left\|\widetilde{\mathbf{D}}(F_n) - \mathbf{D}(F_n)\right\| + \|\mathbf{D}(F_n) - \mathbf{D}(F)\| \\ &= \|\mathbf{W}\mathrm{GS}_n(\mathbf{D})/\epsilon\| + O_p(1) = O_p(n^{1/2}).\end{aligned}$$

Applying the Gaussian mechanism with $\delta \propto n^{-k}$ instead gives that

$$\left\|\widetilde{\mathbf{D}}(F_n) - \mathbf{D}(F)\right\| \leq O_p(n^{1/2}\log^{1/2} n).$$

The level of noise is greater than that of the sampling error for both of these private estimates of the vector of depth values at the sample points. This result is somewhat intuitive; these vectors reveal more information about the population as n grows, which

differs markedly from the single depth value case, where the amount of information received is fixed in n. In fact, for large n the vector of depth values at the sample points contains a significant amount of information about F; the population depth function can, under certain conditions, characterize the distribution of F or F_n (see Struyf and Rousseeuw, 1999; Nagy, 2018, and the references therein). To release so much information about the population privately, we need to inject greater than negligible noise.

We cannot then, simply plug in the n private sample depth values into an inference procedure and proceed, since even large n will produce a lot of noise. If there is an obvious prior for the population, say from a previous study, then we can sample n values from such a prior and use Corollary 3 to generate sample depth values. If there is no obvious prior, then we can instead sample N values from the Mechanism 5 using a very small ϵ and then apply Corollary 3. We can then use those N values are the sample depth values, and compute ranks, perform inference and data visualization.

5.6 Propose-test-release and Projection Depth

We may wish to use another depth function in our inference procedures, such as projection depth (see Definition 4). For example, other depth functions and their associated estimators may possess more desirable properties in certain contexts. For example, projection depth is affine invariant and its median has a higher breakdown point that that of the halfspace or simplicial median[3] (Zuo, 2003). For notation simplicity, in this section we let $\hat{\xi}_{1/2,u} = \text{MED}\left(\mathbb{X}_n^\top u\right)$, $\hat{\gamma}_u = \text{MAD}\left(\mathbb{X}_n^\top u\right)$ and $\tau_u = \text{IQR}\left(\mathbb{X}_n^\top u\right)$. Similarly, we let $\xi_{r,u} = F_u^{-1}(r)$, $\gamma_u = \text{MAD}(F_u)$ and $\tau_u = \text{IQR}(F_u)$.

Recall that to approximate projection depth in practice, we use an algorithm which relies on the outlyingess computed on a finite set of unit vectors $\mathbb{U}_M$ (Liu, 2017; Dyckerhoff et al., 2021). Therefore, we work under the assumption that we have a fixed set of M unit vectors $\mathbb{U}_M$ and define outlyingness as

$$O_1(x; F_n) = \sup_{u \in \mathbb{U}_M} \frac{\left|x^\top u - \hat{\xi}_{1/2,u}\right|}{\hat{\gamma}_u} \qquad \text{or} \qquad O_2(x; F_n) = \sup_{u \in \mathbb{U}_M} \frac{\left|x^\top u - \hat{\xi}_{1/2,u}\right|}{\hat{\tau}_u}.$$

This assumption is restrictive, in the sense that we should consider the generating mechanism for $\mathbb{U}_M$. However, to extend our current proofs beyond a fixed set of unit vectors we

[3] with appropriately chosen scale and location estimators

only need to derive uniform concentration inequalities for $\hat{\xi}_{1/2,u}$, $\hat{\gamma}_u$ and $\hat{\tau}_u$ over u. This is left for future work.

Recall that $\mathrm{PD}(x;F) = (1 + O(x;F))^{-1}$, and so we only need to compute a private version of outlyingness in order to compute a private version of projection depth. It is immediately obvious that the global sensitivity of the outlyingness function is not finite. Therefore, we cannot use the exponential mechanism without injecting a significant level of noise into the estimator.

However, the MED, IQR and MAD are all robust statistics, in the sense that they are not perturbed by extreme data points. This implies that for most samples, the outlyingness function will have low local sensitivity. This fact makes projection depth a good candidate for the propose-test-release framework (Dwork and Lei, 2009; Brunel and Avella-Medina, 2020). For this section, we stick to the Laplace version of PTR introduced by Brunel and Avella-Medina (2020), but our results easily extend to the Gaussian version. With this in mind, we first consider computing the projection depth of fixed point x. We refer to the private version of O_ℓ produced by the Laplace version of PTR in Mechanism 4 by $\widetilde{O}_\ell$ for $\ell = 1$ or $\ell = 2$. The following two theorems show that the probability of returning a "null" value is unlikely, and that the private estimates of outlyingness concentrate around their population values:

Theorem 16. *Suppose that for all $u \in S^{d-1}$ it holds that γ_u is unique, $0 < k_1 < \gamma_u < k_2 < \infty$ and F_u is continuous and increasing in a neighborhood of $\xi_{1/2,u}$. Define*

- $\rho'_n = \frac{\lfloor 2\frac{\log(2/\delta_n)}{\epsilon_n} \rfloor + 1}{n}$
- $\kappa_u(t,r,c) = \left(\lfloor n(r+c) \rfloor / n - F_u(\xi_r - t)\right)^+ \wedge \left(F_u(\xi_r + t) - \lfloor n(r+c) \rfloor / n\right)^+$
- $\Delta_u(t,c) = \kappa_u(\frac{t}{2}, \frac{1}{2}, c) \wedge \Delta_{1,u}(\frac{t}{2}, c) \wedge \Delta_{2,u}(\frac{t}{2}, c)$
- $\Delta_{1,u}(t,c) = \left(\frac{1}{2} + 2c - F_u(\xi_{\frac{1}{2},u} + \gamma_u + t) + F_u(\xi_{\frac{1}{2},u} - \gamma_u - t)\right)^+$
- $\Delta_{2,u}(t,c) = \left(\frac{1}{2} - 2c - F_u(\xi_{\frac{1}{2},u} + \gamma_u - t) + F_u(\xi_{\frac{1}{2},u} - \gamma_u + t)\right)^+$.

Then, there exists constants $k_3, c_1, \ldots, c_7 > 0$ that depend on F such that for all $\delta, \epsilon, \eta, t > 0$ it holds that

$$\Pr(\widetilde{O}_1(x;F_n) = \perp) \leq \delta + Mc_1 e^{-c_2 n} + Mc_3 e^{-n \inf_u \left[\kappa_u\left(\frac{k_1^2\eta}{2k_2}, 1/2, \rho'_n\right) \wedge \Delta_u\left(\frac{k_1^2\eta}{2\|x\|^2 + 4k_3}, \rho'_n\right)\right]^2}$$

and

$$\Pr(|\widetilde{O}_1(x;F_n)-O_1(x;F)|>t)\leq Mc_4e^{-c_5n}+Mc_6e^{-c_7n\inf_u\left[\kappa_u\left(\frac{2k_1^2t}{k_2}.1/2,0\right)\wedge\Delta_u\left(\frac{2k_1^2t}{\|x\|^2+2k_3},0\right)\right]^2}+e^{-\frac{t\epsilon}{2\eta}}$$

Theorem 17. *Define the quantities* $h_1(t)=\inf_u\left[\kappa_u(\frac{t}{2},1/4,0)\wedge\kappa_u(\frac{t}{2},3/4,0)\right]$ *and*

$$h_2\left(t\right)=\inf_u\left[\kappa_u\left(\frac{t}{4},\frac{3}{4},-\rho_n'\right)\wedge\kappa_u\left(\frac{t}{4},\frac{3}{4},\rho_n'\right)\wedge\kappa_u\left(\frac{t}{4},\frac{1}{4},-\rho_n'\right)\wedge\kappa_u\left(\frac{t}{4},\frac{1}{4},\rho_n'\right)\right].$$

Suppose that for all $u\in S^{d-1}$ *it holds that* $0<k_1<\tau_u<k_2<\infty$*. In addition, suppose for all* $u\in S^{d-1}$ *that* F_u *is continuous and increasing in some neighborhood of each of* $\xi_{1/2,u}$*,* $\xi_{1/4,u}$ *and* $\xi_{3/4,u}$*. Then there exists constants* $k_3,c_1,\ldots,c_6>0$ *that depend on* F *such that for all* $\delta,\epsilon,\eta,t>0$ *it holds that*

$$\Pr(\widetilde{O}_2(x;F_n)=\perp)\leq\delta+Mc_1e^{-c_2n}+Mc_3e^{-n\left[\inf_u\kappa_u\left(\frac{k_1^2\eta}{2k_2},1/2,\rho_n'\right)\wedge h_2\left(\frac{2k_1^2\eta}{\|x\|^2+2k_3}\right)\right]^2},$$

and that

$$\Pr(|\widetilde{O}_2(x;F_n)-O_2(x;F)|>t)\leq Mc_4e^{-nc_5}+Mc_6e^{-n\left(\inf_u\kappa_u\left(\frac{k_1^2t}{k_2},\frac{1}{2},0\right)\wedge h_1\left(\frac{2k_1^2t}{\|x\|^2+2k_3}\right)\right)^2}+e^{-\frac{t\epsilon}{2\eta}}.$$

Computing the truncated breakdown point is difficult, so, when one uses the IQR with the PTR-projection depth mechanism, we propose the following procedure: Suppose that $\mathbb{Y}_n\in\mathcal{D}(\mathbb{X}_n,k^*)$, where $k^*=1+\frac{b_\delta}{\epsilon}-\frac{a_\delta}{\epsilon}V_1$. In addition, let $\tilde{\xi}_{u,1/2}=\text{MED}(\mathbb{Y}_n^\top u)$. We want to check if $A_\eta\left(\widehat{O}_2^u(x);\mathbb{X}_n\right)>k^*$. It is easy to see that that

$$|x^\top u-\tilde{\xi}_{u,1/2}|\leq|x^\top u-\xi_{1/2+k^*/n,u}|\vee|x-\xi_{1/2-k^*/n,u}|$$

and that

$$\begin{aligned}|x^\top u-\tilde{\xi}_{u,1/2}|\geq\min\big\{&|x^\top u-\xi_{1/2+k^*/n,u}|,|x^\top u-\xi_{1/2-k^*/n,u}|,\\&|x^\top u-m_1(u,k^*)|,|x^\top u-m_2(u,k^*)|\big\},\end{aligned}$$

with $m_1(u,k^*)$ being the median of a dataset the same as $\mathbb{X}_n^\top u$, except that the smallest k^* observations of $\mathbb{X}_n^\top u$ are replaced with $x^\top u$ and $m_2(u,k^*)$ being the same as $m_1(u,k^*)$,

except instead the largest k^* observations of $\mathbb{X}_n^\top u$ are replaced. Define

$$\mathcal{E}(u,k^*) = \{F_{n,u}^{-1}(3/4+k_1/n) - F_{n,u}^{-1}(1/4+k_2/n); -k^* \le k_1,k_2 \le k^*,\ |k_1|+|k_2| = k^*\},$$

with

$$\min \mathcal{E}(u,k^*) \le \mathrm{IQR}(\mathbb{Y}_n^\top u) \le \max \mathcal{E}(u,k^*).$$

We can summarise these bounds by letting

$$\begin{aligned}
\mathrm{up}(\mathrm{MED},u,k^*) &= \max\{|x-\xi_{1/2+k^*/n,u}|, |x-\xi_{1/2-k^*/n,u}|\},\\
\mathrm{lo}(\mathrm{MED},u,k^*) &= \min\left\{|x-\xi_{1/2+k^*/n,u}|, |x-\xi_{1/2-k^*/n,u}|, |x-m_1(u,k^*)|, |x-m_2(u,k^*)|\right\},\\
\mathrm{lo}(\mathrm{IQR},u,k^*) &= \min \mathcal{E}(u,k^*),\\
\mathrm{up}(\mathrm{IQR},u,k^*) &= \max \mathcal{E}(u,k^*).
\end{aligned}$$

Using this notation, we can write

$$\widehat{O}_2^u(x) \in \left[\frac{\mathrm{lo}(\mathrm{MED},u,k^*)}{\mathrm{up}(\mathrm{IQR},u,k^*)}, \frac{\mathrm{up}(\mathrm{MED},u,k^*)}{\mathrm{lo}(\mathrm{IQR},u,k^*)}\right] = \left[\mathrm{lo}(\widehat{O}_\ell^u(x,k^*)), \mathrm{up}(\widehat{O}_\ell^u(x,k^*))\right]$$

and we can check if

$$\sup_{u\in\mathbb{U}_M}\left[\left(\widehat{O}_2^u(x) - \mathrm{lo}(\widehat{O}_2^u(x))\right) \vee \left(\mathrm{up}(\widehat{O}_2^u(x)) - \widehat{O}_2^u(x)\right)\right] < \eta. \tag{5.9}$$

Then, if (5.9) holds, we must have that $A_\eta(\widehat{O}_2(x;F_n),\mathbb{X}_n) > k^*$, which gives a lower bound on the truncated breakdown point.

Remark 6. *Note that we are releasing $\perp$ if (5.9) fails to hold instead of the original condition in Mechanism 4. It is important to point out that privacy levels of the procedures remain the same. This is due to two facts, the first of which is that*

$$\mathbb{1}\left(\sup_{u\in\mathbb{U}_M}\left[\left(\widehat{O}_2^u(x) - \mathrm{lo}(\widehat{O}_2^u(x))\right) \vee \left(\mathrm{up}(\widehat{O}_2^u(x)) - \widehat{O}_2^u(x)\right)\right] < \eta\right) + \frac{a_\delta}{\epsilon}V$$

is still a differentially privacy estimator. The second fact is that $A_\eta(\widehat{O}_2(x;F_n),\mathbb{X}_n) = 1$ implies that

$$\mathbb{1}\left(\sup_{u\in\mathbb{U}_M}\left[\left(\widehat{O}_2^u(x) - \mathrm{lo}(\widehat{O}_2^u(x))\right) \vee \left(\mathrm{up}(\widehat{O}_2^u(x)) - \widehat{O}_2^u(x)\right)\right] > \eta\right) = 1.$$

Remark 7. *Let $\mathbb{U}_M$ be a set of M of unit vectors and let the concentration functions be as in Theorem 16 and Theorem 17. Using the proof of Theorem 17, we can show that this approximation of the truncated breakdown point still gives a low probability of no-reply. Let*

$$W_1 = \sup_{u\in\mathbb{U}_M}\left(\widehat{O}_2^u(x) - \mathrm{lo}(\widehat{O}_2^u(x))\right) \qquad \text{and} \qquad W_2 = \sup_{u\in\mathbb{U}_M}\left(\mathrm{up}(\widehat{O}_2^u(x)) - \widehat{O}_2^u(x)\right).$$

Then there exists constants k_3, c_1, c_2, K_1, K_2 which depend on the distribution F such that for all $\eta, \delta > 0$ it holds that

$$\Pr\left(W_1 \vee W_2 \geq \eta\right) \leq \delta_n + M\cdot c_1 e^{-c_2 n} + M\cdot K_1 e^{-K_2 n\left[h_2\left(\frac{2k_1^2\eta}{\|x\|+|2k_3|}\right)\wedge\kappa_u\left(\frac{k_1^2\eta}{2k_2},\rho_n'\right)\right]^2}.$$

Now that we know how to compute a private depth value, we can use the same ideas combined with the exponential mechanism to compute a private version of the projection depth median. We first extend the propose-test-release framework to be used with the exponential mechanism. Suppose $\phi_{\mathbb{X}_n}\colon \mathbb{R}^d \to \mathbb{R}^+$ is some objective function which we would like to maximize. We can define the truncated breakdown point of the objective function:

$$A_\eta(\phi_{\mathbb{X}_n};\mathbb{X}_n) = \min\left\{k\in\mathbb{N}\colon \sup_{\mathbb{Y}_n\in\mathcal{D}(\mathbb{X}_n,k)}\sup_x |\phi_{\mathbb{X}_n}(x) - \phi_{\mathbb{Y}_n}(x)| > \eta\right\}.$$

This is a direct extension of the truncated breakdown point of Brunel and Avella-Medina (2020) to be used with an objective function. The following mechanism extends PTR to be used with the exponential mechanism:

Mechanism 6. *Suppose that $\int_{\mathbb{R}^d}\exp(\phi_{\mathbb{X}_n}(x)\frac{\epsilon}{2\eta})dx < \infty$ and let a_δ, b_δ and V be as in Mechanism 4. Let $Q_{\mathbb{X}_n}$ be the measure defined by the density*

$$\frac{dQ_{\mathbb{X}_n}}{dx} = \frac{\exp(\phi_{\mathbb{X}_n}(x)\frac{\epsilon}{2\eta})}{\int_{\mathbb{R}^d}\exp(\phi_{\mathbb{X}_n}(x)\frac{\epsilon}{2\eta})dx}.$$

Then the estimator

$$\widetilde{T}(\mathbb{X}_n) = \begin{cases} \perp & \text{if } A_\eta(\phi_{\mathbb{X}_n};\mathbb{X}_n) + \frac{a_\delta}{\epsilon}V \leq 1 + \frac{b_\delta}{\epsilon} \\ \widehat{T}(\mathbb{X}_n) \sim Q_{\mathbb{X}_n} & o.w. \end{cases},$$

is differentially private. Under the Laplace version, the estimator is $(2\epsilon,\delta)$-differentially private and under the Gaussian version, the estimator is $(2\epsilon, 2\delta)$-differentially private.

For the proof of privacy, see the Section A.9. Mechanism 6 shows that we can still use the exponential mechanism when the objective function is highly likely to have low local sensitivity, rather than finite global sensitivity. In fact, the the Gaussian version of Mechanism 6 uses slightly less of the privacy budget than that of the original propose-test-release mechanism, which is due to the pure differential privacy of the exponential mechanism. In order to extend Theorem 13 to Mechanism 6 one simply needs to show that

$$\Pr\left(A_{\eta_n}(\phi_{\mathbb{X}_n};\mathbb{X}_n)+V\frac{a_{\delta_n}}{\epsilon_n}\leq\frac{b_{\delta_n}}{\epsilon_n}+1\right)$$

is small, either asymptotically or in a finite sample sense.

We now use Mechanism 6 with ϕ equal to the outlyingness function in order to privatize the projection depth-based median. Let ς be either IQR or MAD. A first question is whether or not the following probability density function

$$f(x)=\frac{\exp\left(-\frac{\epsilon\sup_u|x^\top u-\hat{\xi}_{1/2,u}|/\varsigma(\mathbb{X}_n^\top u)}{2\eta}\right)}{\int_{\mathbb{R}^d}\exp\left(-\frac{\epsilon\sup_u|x^\top u-\hat{\xi}_{1/2,u}|/\varsigma(\mathbb{X}_n^\top u)}{2\eta}\right)dx}$$

even exists. Zuo (2003) gives that

$$\sup_u\frac{|x^\top u-\hat{\xi}_{1/2,u}|}{\varsigma(\mathbb{X}_n^\top u)}\geq\frac{\|x\|-\sup_u\hat{\xi}_{1/2,u}}{\sup_u\varsigma(\mathbb{X}_n^\top u)}.$$

It follows that if $\sup_u\hat{\xi}_{1/2,u}<\infty$, then

$$\int_{\mathbb{R}^d}\exp\left(-\frac{\epsilon\sup_u|x^\top u-\hat{\xi}_{1/2,u}|/\varsigma(\mathbb{X}_n^\top u)}{2\eta}\right)dx\leq C_1\int_{\mathbb{R}^d}\exp\left(-\frac{C_2\|x\|}{2}\right)dx<\infty,$$

where the last inequality follows from the fact that $\exp\left(-C_3\|x\|/2\right)$ is proportional to a Gaussian density function. Unfortunately, immediately using PTR with the exponential mechanism gives no gains in estimating the projection median over using the global sensitivity of projection depth (which is 1). This is because, if the points in $\mathbb{X}_n$ are distinct, we have that $A_\eta(O_\ell(\cdot;\mathbb{X}_n);\mathbb{X}_n)=1$ for any η. To see this, suppose that $\mathbb{Y}_n=\{X_1',X_2,\ldots,X_n\}$, with X_1' being some observation such that $\varsigma(\mathbb{Y}_n^\top u)\neq\varsigma(\mathbb{X}_n^\top u)$ for $u\in\mathbb{U}_M$. Then, it holds

for any $u \in \mathbb{U}_M$, that

$$\sup_x |O_\ell^u(x; \mathbb{X}_n) - O_\ell^u(x; \mathbb{Y}_n)| \approx \sup_x \left| x^\top u \frac{\varsigma(\mathbb{X}_n^\top u) - \varsigma(\mathbb{Y}_n^\top u)}{\varsigma(\mathbb{X}_n^\top u)\varsigma(\mathbb{Y}_n^\top u)} \right| = \infty.$$

In order to estimate the projection depth-based median privately, we must then truncate the outlyingness function O in the following manner

$$\mathcal{O}_\ell(x; \mathbb{X}_n, \psi_n) = \begin{cases} O_\ell(x; \mathbb{X}_n) & \|x\| < \psi \\ \infty & \|x\| \geq \psi \end{cases}.$$

One might recognize that this is the same as putting a prior on the median proportional to $\mathbb{1}(x \in \mathcal{B}_\psi(0))$. We can then apply Mechanism 6 to $\mathcal{O}$ in order to privately estimate the projection depth-based median:

Theorem 18. *Let ς_u represent either γ_u or τ_u. Suppose that for all $u \in S^{d-1}$ it holds that $\varsigma(F_u)$ is unique, $0 < k_1 < \varsigma(F_u) < k_2 < \infty$ and that F_u is continuous and increasing in a neighborhood of $\xi_{1/2,u}$. Suppose that $\theta \in \mathcal{B}_\psi(0)$ is the unique minimizer of $O_\ell(x; F)$, and define $\alpha_{\mathcal{O}_\ell}(t) = \sup_{\|x\|=t} \mathcal{O}_\ell(x; F) - \mathcal{O}_\ell(\theta; F)$. Suppose*

$$n \geq \left(\frac{d \log\left(\frac{1}{\alpha_{\mathcal{O}_\ell}(t)} \vee d\right)}{\epsilon \cdot \alpha_{\mathcal{O}_\ell}(t)} \right)^{\frac{1}{1-r}},$$

for some $r > 0$. Let $\widetilde{T}_n$ be the private estimate of the projection depth median, with $\eta = \log n / n$. Then, there exists constants $c_1, \ldots, c_8 > 0$ that depend on F such that for all $\epsilon, \delta, t > 0$ it holds that

$$\Pr(\widetilde{T}_n = \perp) \leq \delta + M c_1 e^{-c_2 n} + M c_3 e^{-n h_3(\log n/n, \psi)},$$

and that

$$\Pr\left(\left\| \widetilde{T}_n - \theta \right\| > t \right) \leq M c_4 e^{-c_5 n} + M c_6 e^{-c_7 n h_4(\alpha_{\mathcal{O}_\ell}(t), \psi)} + e^{-c_8 \frac{n\epsilon}{\log n} \alpha_{\mathcal{O}_\ell}(t)},$$

where h_3 and h_4 are the corresponding bounds given in Theorem 16 for $\ell = 1$ or Theorem 17 for $\ell = 2$ with $x = \psi$.

Suppose that $\delta = n^{-k}$, in order to maintain a probability of returning a "null" value

proportional to δ, we must choose η such that

$$h_3(\eta, \psi) \geq k \log(nM/c)/n.$$

For example, in the case of the standard Gaussian, we have that

$$h_3(\eta, \psi) = 1/2 - \Phi\left(\Phi^{-1}(1/2) - c\frac{\eta}{\psi^2 + ck_3}\right) - \rho'_n.$$

Therefore,

$$\eta \geq (\psi^2 + ck_3)(\Phi^{-1}(1/2) - \Phi^{-1}(1/2 - k \log n/n - \rho'_n))/c.$$

Since the probit function is roughly linear around the median, we should choose $\eta = O(\log n/n)$ to ensure that the probability of no-reply is bounded about by δ.

Corollary 4. *Suppose the conditions of Theorem 18 hold and $\eta = C \log(nM)/n$ for some $C > 0$. Then, given we do not output $\perp$, there exists constants $K_1, K_2, K_3 > 0$ that depend on F such that the number of samples needed to estimate the private projection depth median within an error of $t > 0$, with probability $1 - \gamma > 0$, is less than*

$$n(t, \gamma, d, \epsilon) = \max\left\{\frac{K_1 \log(M/\gamma)}{h_4(\alpha_{\mathcal{O}_\ell}(t), \psi) \wedge K_3}, \left(K_2 \frac{\log(M/\gamma) \vee d \log\left(\frac{1}{\alpha_{\mathcal{O}_\ell}(t)} \vee d\right)}{\epsilon \cdot \alpha_{\mathcal{O}_\ell}(t)}\right)^{\frac{1}{1-r}}\right\},$$

for any $r > 0$.

The sample complexity depends on the dimension through M, for example, if we choose exponentially many unit vectors in the dimension, then $M = e^d$ and we have that

$$n(t, \gamma, d, \epsilon) = \max\left\{\frac{K_1 \log(1/\gamma) + K_1 d}{h_4(\alpha_{\mathcal{O}_\ell}(t), \psi) \wedge K_3}, \left(d \frac{K_3 \log(1/\gamma) \vee \log\left(\frac{1}{\alpha_{\mathcal{O}_\ell}(t)} \vee d\right)}{\epsilon \cdot \alpha_{\mathcal{O}_\ell}(t)}\right)^{\frac{1}{1-r}}\right\}.$$

Once again, we see that if the error level is fixed and the success rate is $1 - e^{-d}$, then we have the sample complexity scales morally like $d \log d$.

5.7 Brief discussion of future directions

We have presented several algorithms for computing differentially private depth-based medians and depth values, along with their respective private and non-private sample complexities. Under mild assumptions, for most of the presented mechanisms the cost of privacy is a multiplicative factor of $\log d$ on the sample complexity. There still remains many unanswered questions. For example, concerning the interpretation of the privacy parameter, how can we determine the appropriate choice of ϵ in practice? In addition, the construction of traditional depth functions implies that depth functions cannot be concave, which makes computation of both the non-private and private depth-based estimators challenging. Even in the case where the depth values are computable, such as for the integrated depths, sampling from the exponential mechanism remains challenging. Therefore, the next steps are to come up with objective functions that produce robust, accurate medians but are also computationally compatible with the exponential mechanism. In this direction, it may be helpful to construct depth functions that are more suited to high dimensional analysis. In fact, high dimensional datasets are more susceptible to privacy breaches because the dataset contains more external information to exploit.

Chapter 6

Future Directions

6.1 Extensions for private depth-based inference

After deriving the sample complexity of the private projection depth values and projection depth median, the next step is to develop strategies to compute these private estimators. We have started this with the algorithm to compute a lower bound on the truncated breakdown point for the outlyingness function. However, we would like to develop an algorithm to compute the estimates that involve the exponential mechanism. Algorithms for private estimators are non-trivial. Suppose we wish to sample from an approximate version of $Q_{\mathbb{X}_n}$, denoted $\widehat{Q}_{\mathbb{X}_n}$. Then we must have a means of measuring

$$d(Q_{\mathbb{X}_n}, \widehat{Q}_{\mathbb{X}_n}),$$

since this distance will affect the privacy of the procedure. For example, suppose we wish to implement Mechanism 5. Let d be the total variation distance. If we use, say, Markov chain Monte Carlo to draw from the exponential mechanism, then we should ensure that

$$d(Q_{\mathbb{X}_n}, \widehat{Q}_{\mathbb{X}_n}) \leq \delta,$$

for some δ. This would ensure that the estimator satisfies (ϵ, δ) differential privacy.

6.2 Extensions to manifold valued data

It is interesting to generalize depth functions and related inference to data which lie on a manifold. Network data, matrix data and shape data can all be modeled as objects on a manifold. Such observations are amenable to nonparametric models, since it is difficult to represent them parametrically. Some notions of depth have been introduced for manifold-valued data (Small, 1997; Fraiman et al., 2019; Chau et al., 2019; Harris et al., 2020). However, this area is still very new and there are a number of fundamental questions that are not well established. For example what does it mean to be outlying in these spaces. Just as outlyingness differs between functional and multivariate data, it will be different for manifold valued random variables. For example, outlying shapes could be abnormally shaped curves rather than high magnitude observations. It could also be a discoloured shape if the shape is 'filled in'. We need to define distances and thus, depth, between the shapes such that they are large when the shape satisfies the corresponding definition of outlyingness. Similar concepts can be discussed for network and matrix valued data. We must also incorporate the appropriate transformation invariance; if the shape is rotated, or angled to the side, the inference should not change. Some of these ideas have been explored extensively for shape data (Srivastava and Klassen, 2016). Here, each shape is part of an equivalence class that includes all valid transformations of that shape. Those transformations can include rotations, scaling, and reparameterization. This equivalence class is then considered the observation, rather than the orientation of the shape itself. The inference is then completed in the quotient space. Some of the ideas in Srivastava and Klassen (2016) can be translated to the depth-based framework, which has been started by Harris et al. (2020), but there are still things to investigate. For example, we could use depth functions to describe the variability of shape data, like we have done for multivariate data in Chapter 2 and functional data in Chapter 3.

For example, in additive manufacturing, or 3D printing, it is of interest to monitor the printing process (Tapia and Elwany, 2014). The data is of course, the object being printed, which can be represented by closed contours or a 3D surface. Given the large variety of parts that may be printed, it is useful to have a nonparametric model that can be applied to many parts. We could potentially use the methods of Chapter 2 or those of Liu (1995) in combination with those of Srivastava and Klassen (2016) to monitor the objects for changes as they are being printed. We could use hypothesis tests for variability to compare printers as well.

References

Aminikhanghahi, S. and Cook, D. (2017). A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367.

Anderson, R. J. and Miller, G. L. (1990). A simple randomized parallel algorithm for list-ranking. *Information Processing Letters*, 33(5):269–273.

Arlot, S., Celisse, A., and Harchaoui, Z. (2012). A Kernel Multiple Change-point Algorithm via Model Selection. *arXiv e-prints*, page arXiv:1202.3878.

Aston, J. A. D., Pigoli, D., and Tavakoli, S. (2017). Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, 45(4):1431–1461.

Aue, A., Gabrys, R., Horváth, L., and Kokoszka, P. (2009a). Estimation of a change-point in the mean function of functional data. *Journal of Multivariate Analysis*, 100(10):2254–2269.

Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009b). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, 37(6B):4046–4087.

Aue, A. and Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis*, 34(1):1–16.

Aue, A., Horváth, L., and F. Pellatt, D. (2017). Functional generalized autoregressive conditional heteroskedasticity. *Journal of Time Series Analysis*, 38(1):3–21.

Aue, A., Rice, G., and Sönmez, O. (2019). Structural break analysis for spectrum and trace of covariance operators. *Environmetrics*, 31(1).

Avella-Medina, M. (2019). Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, pages 1–45.

Avella-Medina, M. and Brunel, V.-E. (2019). Differentially private sub-Gaussian location estimators. *arXiv e-prints*, pages 1–16.

Awan, J., Kenney, A., Reimherr, M., and Slavković, A. (2019). Benefits and pitfalls of the exponential mechanism with applications to hilbert spaces and functional pca. *arXiv e-prints*, page arXiv:1901.10864.

Baidari, I. and Patil, C. (2019). K-data depth based clustering algorithm. In *Computational Intelligence: Theories, Applications and Future Directions*, volume 1 of *Advances in Intelligent Systems and Computing*, pages 13–24. Springer, Singapore.

Balle, B. and Wang, Y.-X. (2018). Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. *arXiv e-prints*, page arXiv:1805.06530.

Basu, S. and DasGupta, A. (1997). The mean, median, and mode of unimodal distributions:a characterization. *Theory of Probability & Its Applications*, 41(2):210–223.

Beimel, A., Moran, S., Nissim, K., and Stemmer, U. (2019). Private center points and learning of halfspaces. *arXiv e-prints*, page arXiv:1902.10731.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bhattacharyya, M. and Kasa, S. R. (2018). A test for detecting structural breakdowns in markets using eigenvalue decompositions. *arXiv e-prints*, page arXiv:1809.07114.

Bickel, P. J. (1965). On some asymptotically nonparametric competitors of hotelling's t^{21}. *Annals of Mathematical Statistics*, 36(1):160–173.

Billingsley, P. (1968). *Convergence of probability measures*. Wiley, New York, first edition.

Bobkov, S. and Götze, F. (2010). Concentration of empirical distribution functions with applications to non-i.i.d. models. *Bernoulli*, 16(4):1385–1414.

Bodie, Z., Kane, A., Marcus, A., Perrakis, S., and Ryan, P. (2017). *Investments*. McGraw-Hill Education.

Boente, G., Rodriguez, D., and Sued, M. (2018). Testing equality between several populations covariance operators. *Annals of the Institute of Statistical Mathematics*, 70(4):919–950.

Brunel, V.-E. and Avella-Medina, M. (2020). Propose, test, release: Differentially private estimation with high probability. *arXiv e-prints*, page arXiv:2002.08774.

Cabassi, A., Pigoli, D., Secchi, P., and Carter, P. A. (2017). Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology. *Electronic Journal of Statistics*, 11(2):3815–3840.

Cabrieto, J., Tuerlinckx, F., Kuppens, P., Hunyadi, B., and Ceulemans, E. (2018). Testing for the presence of correlation changes in a multivariate time series: A permutation based approach. *Scientific Reports*, 8(1):769.

Cai, T. T., Wang, Y., and Zhang, L. (2019). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv e-prints*, page arXiv:1902.04495.

Cerovecki, C., Francq, C., Hörmann, S., and Zakoïan, J.-M. (2019). Functional garch models: The quasi-likelihood approach and its applications. *Journal of Econometrics*, 209(2):353 – 375.

Chakraborti, S. and Graham, M. A. (2019). Nonparametric (distribution-free) control charts: An updated overview and some results. *Quality Engineering*, pages 1–22.

Chakraborty, A. and Chaudhuri, P. (2014). On data depth in infinite dimensional spaces. *Annals of the Institute of Statistical Mathematics*, 66(2):303–324.

Chau, J., Ombao, H., and von Sachs, R. (2019). Intrinsic Data Depth for Hermitian Positive Definite Matrices. *Journal of Computational and Graphical Statistics*, 28(2):427–439.

Chaudhuri, K. and Hsu, D. (2012). Convergence rates for differentially private statistical estimation. *arXiv e-prints*, page arXiv:1206.6395.

Chen, Z. and Tyler, D. E. (2002). The influence function and maximum bias of tukey's median. *Annals of Statistics*, 30(6):1737–1759.

Chenouri, S., Mozaffari, A., and Rice, G. (2020a). Multiple change point detection based on standard and wild rank-cusum binary segmentation. Forthcoming.

Chenouri, S., Mozaffari, A., and Rice, G. (2020b). Robust multivariate change point analysis based on data depth. *Canadian Journal of Statistics*, 48(3):417–446.

Chenouri, S., Small, C. G., and Farrar, T. J. (2011). Data depth-based nonparametric scale tests. *Canadian Journal of Statistics*, 39(2):356–369.

Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, 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.

Cuevas, A. and Fraiman, R. (2009). On depth measures and dual statistics. A methodology for dealing with general data. *Journal of Multivariate Analysis*, 100(4):753–766.

Dai, W. and Genton, M. G. (2018). Multivariate functional data visualization and outlier detection. *Journal of Computational and Graphical Statistics*, 27(4):923–934.

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136 – 154.

Dette, H. and Kokot, K. (2020). Detecting relevant differences in the covariance operators of functional time series – a sup-norm approach.

Dette, H., Kokot, K., and Volgushev, S. (2020). Testing relevant hypotheses in functional time series via self-normalization. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):629–660.

Dette, H. and Kutta, T. (2021). Detecting structural breaks in eigensystems of functional time series. *Electronic Journal of Statistics*, 15(1).

Dette, H., Pan, G. M., and Yang, Q. (2018). Estimating a change point in a sequence of very high-dimensional covariance matrices. *arXiv e-prints*, page arXiv:1807.10797.

Donoho, D. L. (1982). Breakdown properties of multivariate location estimators.

Duan, F. and Wied, D. (2018). A residual-based multivariate constant correlation test. *Metrika*, 81(6):653–687.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252.

Dwork, C. and Lei, J. (2009). Differential privacy and robust statistics. *Proceedings of the 41st annual ACM symposium on Symposium on theory of computing - STOC '09*, page 371.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407.

Dwork, C., Smith, A., Steinke, T., and Ullman, J. (2017). Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(1):61–84.

Dyckerhoff, R., Mozharovskyi, P., and Nagy, S. (2021). Approximate computation of projection depths. *Computational Statistics and Data Analysis*, 157:107166.

Dümbgen, L. (1992). Limit theorems for the simplicial depth. *Statistics & Probability Letters*, 14(2):119 – 128.

Eaton, M. L. (1981). On the Projections of Isotropic Distributions. *Annals of Statistics*, 9(2):391–400.

Fan, C., Zhang, D., and Zhang, C.-H. (2011). On sample size of the kruskal-wallis test with application to a mouse peritoneal cavity study. *Biometrics*, 67(1):213–224.

Fearnhead, P. and Rigaill, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, 114(525):169–183.

Feiner, L. (2019). Twitter shares surge after earnings report shows growth in daily users. *www.cnbc.com*.

Flores, R., Lillo, R., and Romo, J. (2018). Homogeneity test for functional data. *Journal of Applied Statistics*, 45(5):868–883.

Fraiman, R., Gamboa, F., and Moreno, L. (2019). Connecting pairwise geodesic spheres by depth: DCOPS. *Journal of Multivariate Analysis*, 169:81–94.

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

Fremdt, S., Steinebach, J. G., Horváth, L., and Kokoszka, P. (2013). Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics*, 40(1):138–152.

Fryzlewicz, P. (2014). Wild Binary Segmentation for Multiple Changepoint Detection. *The Annals of Statistics*, 42(6):2243–2281.

Gabrys, R. and Kokoszka, P. (2007). Portmanteau test of independence for functional observations. *Journal of the American Statistical Association*, 102(480):1338–1348.

Gaines, G., Kaphle, K., and Ruymgaart, F. (2011). Application of a delta-method for random operators to testing equality of two covariance operators. *Mathematical Methods of Statistics*, 20(3):232.

Galeano, P. and Peña, D. (2007). Covariance changes detection in multivariate time series. *Journal of Statistical Planning and Inference*, 137(1):194–211.

Galeano, P. and Wied, D. (2014). Multiple break detection in the correlation structure of random variables. *Computational Statistics & Data Analysis*, 76:262–282.

Galeano, P. and Wied, D. (2017). Dating multiple change points in the correlation matrix. *TEST*, 26:331–352.

Gao, Y. and Sheffet, O. (2020). Private approximations of a convex hull in low dimensions. *arXiv e-prints*, page arXiv:2007.08110.

Gastwirth, J. L. (1965). Percentile modifications of two sample rank tests. *Journal of the American Statistical Association*, 60(312):1127–1141.

Gijbels, I. and Nagy, S. (2015). Consistency of non-integrated depths for functional data. *Journal of Multivariate Analysis*, 140:259–282.

Gijbels, I. and Nagy, S. (2017). On a general definition of depth for functional data. *Statistical Science*, 32(4):630–639.

Gromenko, O., Kokoszka, P., and Reimherr, M. (2017). Detection of change in the spatiotemporal mean function. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 79(1):29–50.

Guo, J. and Zhang, J.-T. (2016). A further study of an l^2-norm based test for the equality of several covariance functions. *arXiv e-prints*, page arXiv:1609.04231.

Guo, J., Zhou, B., and Zhang, J.-T. (2018). New tests for equality of several covariance functions for functional data. *Journal of the American Statistical Association*, 0(NO. 0):1–13.

Han, F. (2018). An Exponential Inequality for U-Statistics Under Mixing Conditions. *Journal of Theoretical Probability*, 31(1):556–578.

Harris, T. (2020). fmci. https://github.com/trevor-harris/fmci.

Harris, T., Li, B., and Tucker, J. D. (2021). Scalable Multiple Changepoint Detection for Functional Data Sequences.

Harris, T., Tucker, J. D., Li, B., and Shand, L. (2020). Elastic depths for detecting shape anomalies in functional data. *Technometrics*, 0(0):1–11.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102.

Hernan Madrid Padilla, O., Yu, Y., Wang, D., and Rinaldo, A. (2019). Optimal nonparametric change point detection and localization. *arXiv e-prints*, page arXiv:1905.10019.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Hoffmann-Jørgensen, J. (2016). Maximal Inequalities for Dependent Random Variables. In *High Dimensional Probability VII*, pages 61–104. Birkhäuser, Cham.

Horrell, J. F. and Lessig, V. P. (1975). A note on a multivariate generalization of the kruskal-wallis test. *Decision Sciences*, 6(1):135–141.

Horváth, L., Hušková, M., and Kokoszka, P. (2010). Testing the stability of the functional autoregressive process. *Journal of Multivariate Analysis*.

Horváth, L. and Kokoszka, P. (2012). *Change point detection in the functional autoregressive process*, pages 253–276. Springer New York, New York, NY.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley Series in Probability and Statistics. Wiley.

Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C., and Roth, A. (2014). Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410.

Hubert, M., Claeskens, G., Ketelaere, B. D., and Vakili, K. (2012). A new depth-based approach for detecting outlying curves. In *Proceedings of COMPSTAT 2012*, pages 329–340.

Hubert, M., Rousseeuw, P., and Segaert, P. (2017). Multivariate and functional classification using depth and distance. *Advances in Data Analysis and Classification*.

Hubert, M., Rousseeuw, P. J., and Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods and Applications*.

Ieva, F. and Paganoni, A. M. (2013). Depth measures for multivariate functional data. *Communications in Statistics-Theory and Methods*, 42(7):1265–1276.

James, G. M. and Sood, A. (2006). Performing hypothesis tests on the shape of functional data. *Computational Statistics & Data Analysis*, 50:1774–1792.

James, N. A. and Matteson, D. S. (2015). ecp: An r package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, 62(7):1–25.

Jarušková, D. (2013). Testing for a change in covariance operator. *Journal of Statistical Planning and Inference*, 143(9):1500 – 1511.

Jeong, M.-H., Cai, Y., Sullivan, C. J., and Wang, S. (2016). Data depth based clustering analysis. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '16*, pages 1–10, New York, New York, USA. ACM Press.

Jiao, S., Frostig, R. D., and Ombao, H. (2020). Break point detection for functional covariance.

Jörnsten, R. (2004). Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis*, 90(1):67–89.

Kao, C., Trapani, L., and Urga, G. (2018). Testing for instability in covariance structures. *Bernoulli*, 24(1):740–771.

Kashlak, A. B., Aston, J. A. D., and Nickl, R. (2019). Inference on covariance operators via concentration inequalities: k-sample tests, classification, and clustering via rademacher complexities. *Sankhya A*, 81(1):214–243.

Katz, B. M. and McSweeney, M. (1980). A multivariate kruskal-wallis test with post hoc procedures. *Multivariate Behavioral Research*, 15(3):281–297. PMID: 26794183.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Knoblauch, J. and Damoulas, T. (2018). Spatio-temporal Bayesian On-line Changepoint Detection with Model Selection. *arXiv e-prints*, page arXiv:1805.05383.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, 23(4):525–540.

Kuhnt, S. and Rehage, A. (2016). An angle-based multivariate functional pseudo-depth for shape outlier detection. *Journal of Multivariate Analysis*, 146:325–340.

Lei, J. (2011). Differentially private m-estimators. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24, pages 361–369. Curran Associates, Inc.

Li, J. and Liu, R. Y. (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statist. Sci.*, 19(4):686–696.

Li, X. and Ghosal, S. (2018). Bayesian Change Point Detection for Functional Data. *arXiv e-prints*, page arXiv:1808.01236.

Liu, R. and Singh, K. (2006). *Rank tests for multivariate scale difference based on data depth*, volume 72 of *DIMACS: Series in Discrete Mathematics and Theoretical Computer Science*, pages 17–35. American Mathematical Society.

Liu, R. Y. (1988). On a notion of simplicial depth. *Proceedings of the National Academy of Sciences*, 85(6):1732–1734.

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

Liu, R. Y. (1992). *Data depth and multivariate rank tests*, page 279–294. Y. Dodge, ed., North-Holland, Amsterdam.

Liu, R. Y. (1995). Control charts for multivariate processes. *Journal of the American Statistical Association*, 90(432):1380–1387.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27(3):783–840.

Liu, R. Y. and Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260.

Liu, X. (2017). Approximating the projection depth median of dimensions $p \geq 3$. *Communications in Statistics - Simulation and Computation*, 46(5):3756–3768.

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

Lopez-Pintado, S., Sun, Y., K. Lin, J., and Genton, M. (2014). Simplicial band depth for multivariate functional data. *Advances in Data Analysis and Classification*, 8:321–338.

López-Pintado, S. and Wrobel, J. (2017). Robust non-parametric tests for imaging data based on data depth. *Stat*, 6(1):405–419.

Lung-Yut-Fong, A., Lévy-Leduc, C., and Cappé, O. (2011). Homogeneity and change-point detection tests for multivariate data using rank statistics. *arXiv e-prints*, page arXiv:1107.1971.

Massé, J.-C. (2004). Asymptotics for the tukey depth process, with an application to a multivariate trimmed mean. *Bernoulli*, 10(3):397–419.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Müller, C. H. (2005). Depth estimators and tests based on the likelihood principle with application to regression. *Journal of Multivariate Analysis*, 95(1):153 – 181.

Nagy, S. (2018). Halfspace depth does not characterize probability distributions. *Mathematics Subject Classification*.

Nagy, S. and Ferraty, F. (2019). Data depth for measurable noisy random functions. *Journal of Multivariate Analysis*, 170:95–114.

Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125.

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

Nissim, K., Raskhodnikova, S., and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '07, page 75–84, New York, NY, USA. Association for Computing Machinery.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1(6):327 – 332.

Page, E. S. (1954). Continuous Inspection Schemes. *Biometrika*, 41(1-2):100–115.

Page, W. and Murty, V. N. (1982). Nearness relations among measures of central tendency and dispersion: Part 1. *The Two-Year College Mathematics Journal*, 13(5):315–327.

Panaretos, V. M., Kraus, D., and Maddocks, J. H. (2010). Second-order comparison of gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*.

Paparoditis, E. and Sapatinas, T. (2016). Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data. *Biometrika*, 103(3):727–733.

Pigoli, D., Aston, J. A. D., Dryden, I. L., and Secchi, P. (2014). Distances and inference for covariance operators. *Biometrika*, 101(2):409–422.

Pitcan, Y. (2017). A Note on Concentration Inequalities for U-Statistics. *arXiv e-prints*, page arXiv:1712.06160.

Piña, B., Raldúa, D., Barata, C., Portugal, J., Navarro-Martín, L., Martínez, R., Fuertes, I., and Casado, M. (2018). Chapter twenty - functional data analysis: Omics for environmental risk assessment. In Jaumot, J., Bedia, C., and Tauler, R., editors, *Data Analysis for Omic Sciences: Methods and Applications*, volume 82 of *Comprehensive Analytical Chemistry*, pages 583–611. Elsevier.

Posch, P. N., Ullmann, D., and Wied, D. (2019). Detecting structural changes in large portfolios. *Empirical Economics*, 56(4):1341–1357.

Pruss, A. R. (1998). A maximal inequality for partial sums of finite exchangeable sequences of random variables. *Proceedings of the American Mathematical Society*, 126(6):1811–1819.

Ramsay, J. and Silverman, B. (2006). *Functional Data Analysis*. Springer Series in Statistics. Springer New York.

Ramsay, K. (2019a). FKWC. https://github.com/12ramsake/FKWC.

Ramsay, K. (2019b). MVT-WBS-RankCUSUM. https://github.com/12ramsake/MVT-WBS-RankCUSUM.

Ramsay, K. (2021). FKWC_Changepoint. https://github.com/12ramsake/FKWC_Changepoint.

Ramsay, K., Durocher, S., and Leblanc, A. (2019). Integrated rank-weighted depth. *Journal of Multivariate Analysis*, 173:51 – 69.

Randles, R. H. and Peters, D. (1990). Multivariate rank tests for the two-sample location problem. *Communications in Statistics - Theory and Methods*, 19(11):4225–4238.

Reeves, J., Chen, J., Wang, X. L., Lund, R., Lu, Q. Q., Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915.

Rice, G. and Shum, M. (2019). Inference for the lagged cross-covariance operator between functional time series. *Journal of Time Series Analysis*, 40(5):665–692.

Rice, G., Wirjanto, T., and Zhao, Y. (2019). Tests for conditional heteroscedasticity with functional data and goodness-of-fit tests for FGARCH models. MPRA Paper 93048, University Library of Munich, Germany.

Rousseeuw, P. J., Ruts, I., and Tukey, J. W. (1999). The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–387.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.

Rousseeuw1, P. J. and Ruts2, I. (1999). The depth function of a population distribution. *Metrika*, 49.

Sankararaman, S., Obozinski, G., Jordan, M. I., and Halperin, E. (2009). Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41(9):965–967.

Sarwate, A. D. and Chaudhuri, K. (2013). Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Processing Magazine*, 30(5):86–94.

Serfling, R. (2002). A depth function and a scale curve based on spatial quantiles. In *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, pages 25–38. Birkhäuser, Basel.

Serfling, R. and Mazumder, S. (2009). Exponential probability inequality and convergence results for the median absolute deviation and its modifications. *Statistics & Probability Letters*, 79:1767–1773.

Serfling, R. and Wijesuriya, U. (2017). Depth-based nonparametric description of functional data, with emphasis on use of spatial depth. *Computational Statistics & Data Analysis*, 105:24–45.

Serfling, R. J. (2006). Depth functions in nonparametric multivariate inference. *Data Depth: Robust Multivariate Analysis, Computational Geometry, and Applications*, pages 1–16.

Sguera, C., Galeano, P., and Lillo, R. (2014). Spatial depth-based classification for functional data. *TEST*, 23:725–750.

Sguera, C., Galeano, P., and Lillo, R. E. (2016). Functional outlier detection by a local depth with application to NOx levels. *Stochastic Environmental Research and Risk Assessment*, 30(4):1115–1130.

Shao, W., Zuo, Y., and Luo, J. (2022). Employing the mcmc technique to compute the projection depth in high dimensions. *Journal of Computational and Applied Mathematics*, 411:114278.

Sharipov, O. S. and Wendler, M. (2019). Bootstrapping covariance operators of functional time series. *arXiv e-prints*, page arXiv:1904.06721.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product.* Van Nostrand, Oxford, England.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Slaets, L. (2011). *Analyzing Phase and Amplitude Variation of Functional Data*. PhD thesis, KU Leuven, Faculty of Business and Economics.

Small, C. G. (1990). A survey of multidimensional medians.

Small, C. G. (1997). Multidimensional medians arising from geodesics on graphs. *Ann. Statist.*, 25(2):478–494.

Sonmez, O. (2018). *Structural Breaks In Functional Time Series Data*. PhD thesis, University of California Davis.

Srivastava, A. and Klassen, E. P. (2016). *Functional and Shape Data Analysis*. Springer Series in Statistics. Springer New York, New York, NY.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. (2011). Registration of Functional Data Using Fisher-Rao Metric. *arXiv e-prints*, page arXiv:1103.3817.

Stahel, W. A. (1981). Breakdown of covariance estimators.

Steel, R. G. D. (1960). A rank sum test for comparing all pairs of treatments. *Technometrics*, 2(2):197–207.

Stoehr, C., Aston, J. A. D., and Kirch, C. (2021). Detecting changes in the covariance structure of functional time series with application to fmri data. *Econometrics and Statistics*, 18(C):44–62.

Struyf, A. and Rousseeuw, P. J. (1999). Halfspace depth and regression depth characterize the empirical distribution. *Journal of Multivariate Analysis*, 69:1355153.

Sweeney, L. (2007). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110.

Talagrand, M. (1994). Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76.

Tapia, G. and Elwany, A. (2014). A Review on Process Monitoring and Control in Metal-Based Additive Manufacturing. *Journal of Manufacturing Science and Engineering*, 136(6). 060801.

Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Tukey, J. W. (1974). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*.

Ullah, S. and Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13(1):43.

van den Burg, G. J. J. and Williams, C. K. I. (2020). An Evaluation of Change Point Detection Algorithms. *arXiv e-prints*, page arXiv:2003.06222.

Venkatraman, E. (1992). *Consistency Results in Multiple Change-Point Problems*. PhD thesis, Stanford University, Department of Statistics.

Wang, D., Yu, Y., and Rinaldo, A. (2020). Optimal covariance change point localization in high dimension. *Bernoulli*, to appear.

Wang, D., Yu, Y., and Rinaldo, A. (2021). Optimal covariance change point localization in high dimensions. *Bernoulli*, 27(1):554–575.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295.

Wang, Y., Wang, Z., and Zi, X. (2019). Rank-based multiple change-point detection. *Communications in Statistics - Theory and Methods*, 0(0):1–17.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389.

Weber, N. C. (1980). A martingale approach to central limit theorems for exchangeable random variables. *Journal of Applied Probability*, 17(3):662–673.

Wied, D., Krämer, W., and Dehling, H. (2012). Testing for a change in correlation at an unknown point in time using an extended functional delta method. *Econometric Theory*, 28(3):570–589.

Yu, M. and Chen, X. (2017). Finite sample change point inference and identification for high-dimensional mean vectors. *arXiv e-prints*, page arXiv:1711.08747.

Zhang, J.-T. (2013). Test of equality of covariance functions. In *Analysis of Variance for Functional Data*, pages 368–384. Chapman and Hall/CRC.

Zhang, W., James, N. A., and Matteson, D. S. (2017). Pruning and nonparametric multiple change point detection. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 288–295.

Zhang, X. and Shao, X. (2015). Two sample inference for the second-order property of temporally dependent functional data. *Bernoulli*, 21(2):909–929.

Zhao, Y. (2017). *An analysis of the stability in multivariate correlation structures*. PhD thesis, Birmingham Business School, Department of Economics.

Zuo, Y. (2002). Multivariate trimmed means based on data depth. In Dodge, Y., editor, *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 313–322, Basel. Birkhäuser Basel.

Zuo, Y. (2003). Projection-based depth functions and associated medians. *The Annals of Statistics*, 31(5):1460–1490.

Zuo, Y. (2004). Influence function and maximum bias of projection depth based estimators. *Annals of Statistics*, 32(1):189–218.

Zuo, Y. (2019). A new approach for the computation of halfspace depth in high dimensions. *Communications in Statistics - Simulation and Computation*, 48(3):900–921.

Zuo, Y. and Serfling, R. (2000a). General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482.

Zuo, Y. and Serfling, R. (2000b). On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry. *Journal of Statistical Planning and Inference*, 84(1):55–79.

APPENDICES

Appendix A

Chapter Appendices

A.1 Proofs from Chapter 2

Note that I call it a proof sketch, since I refer to sections of Fryzlewicz (2014) where which the details for our context are not explicitly written, but all of the conclusions should be correct.

Sketch of the proof of Theorem 1. The proof of Theorem 1 follows the format of the proof of consistency for the original WBS algorithm, specifically, the proof of Theorem 3.2 in (Fryzlewicz, 2014). However, in their setting they use properties of the Gaussian distribution, where here we must use tools which apply to the rank based setting. We also remark that we have simpler assumptions than Fryzlewicz (2014), which simplifies some steps of the proof. Let $\Delta = \min_{i,j\in[\ell]} |\theta_j - \theta_i|$. We first define the following ranks based on the population depth functions

$$R_{i,s,e} \coloneqq \#\left\{X_j \colon \mathrm{D}(X_j; F_{*,s,e}) \leq \mathrm{D}(X_i; F_{*,s,e}),\ j \in \{s,\dots,e\}\right\},\ i \in \{s,\dots,e\}.$$

The distribution $F_{*,s,e}$ is a mixture distribution with weights proportional to the number of observations coming from F_j in the subsample $\{X_s,\dots,X_e\}$. It should be noted that these weights depend on n, since they depend on the subsample. More specifically, for some interval with length that satisfies $n_{s,e} > \widetilde{\Delta}\cdot n$ we have that $F_{*,s,e} = \sum_{j=1}^{\ell+1} \widetilde{\vartheta}_{j,s,e} F_j$, for

some $\widetilde{\vartheta}_{j,s,e} \geq 0$. We also define the quantities

$$\widetilde{Z}_{s,e}(k/n_{s,e}) \coloneqq \frac{1}{\sqrt{n_{s,e}}} \sum_{i=1}^{k} \frac{R_{i,s,e} - (n_{s,e}+1)/2}{\sqrt{\left(n_{s,e}^2 - 1\right)/12}}$$

$$G_{s,e}(k/n_{s,e}) \coloneqq \widetilde{Z}_{s,e}(k/n_{s,e}) - Z_{s,e}(k/n_{s,e}) = \frac{1}{\sqrt{n_{s,e}}} \sum_{i=1}^{k} \frac{R_{i,s,e} - \widehat{R}_{i,s,e}}{\sqrt{\left(n_{s,e}^2 - 1\right)/12}}.$$

Recall that $n_{s,e} \geq \widetilde{\Delta} n$. We first show that $\max_{\substack{s<k<e \\ e-s>\widetilde{\Delta} n}} |G_{s,e}(k/n_{s,e})| \lesssim \sqrt{\log n}$:

Lemma 1. *Let*

$$A_n = \left\{ \max_{\substack{s<k<e \\ e-s>\widetilde{\Delta} n}} |G_{s,e}(k/n_{s,e})| \leq c\sqrt{\log n} \right\},$$

then it holds that $\lim_{n\to\infty} \Pr(A_n) = 1$.

Proof. We can now write

$$\begin{aligned}
\max_{\substack{s<k<e \\ e-s>\widetilde{\Delta} n}} |G_{s,e}(k/n_{s,e})| &\leq \max_{\substack{s<k<e \\ e-s>\widetilde{\Delta} n}} \frac{1}{\sqrt{n_{s,e}}} \sum_{i=1}^{k} \left| \frac{R_{i,s,e} - \widehat{R}_{i,s,e}}{\sqrt{\left(n_{s,e}^2 - 1\right)/12}} \right| \\
&\leq \frac{1}{\widetilde{\Delta}\sqrt{n}} \sum_{i=1}^{n} \left| \frac{R_{i,1,n} - \widehat{R}_{i,1,n}}{\sqrt{\widetilde{\Delta}^2 \left(n^2 - 1\right)/12}} \right| \\
&\leq \frac{1}{K\sqrt{n}} \sum_{i=1}^{n} \left| \frac{R_{i,1,n} - \widehat{R}_{i,1,n}}{\sqrt{\left(n^2 - 1\right)/12}} \right|,
\end{aligned}$$

for some absolute constant K. From (A5) on page 439 of Chenouri et al. (2020b) we have, under Assumptions 1 and 2, that

$$\mathrm{E}\left[\left| \frac{R_{i,1,n} - \widehat{R}_{i,1,n}}{\sqrt{\left(n^2 - 1\right)/12}} \right| \right] = O_p(n^{-1/2}).$$

This gives that

$$\mathrm{E}\left[\max_{\substack{s<k\leq e\\ e-s>\widetilde{\Delta}n}} |G_{s,e}(k/n_{s,e})|\right] = O(1),$$

and it follows by Markov's inequality that

$$\max_{\substack{s<k\leq e\\ e-s>\widetilde{\Delta}n}} |G_{s,e}(k/n_{s,e})| = O_p(1).$$

□

Next, we show that $\max_{\substack{1\leq s\leq k\leq e\leq n\\ e-s>\widetilde{\Delta}n}} |\widetilde{Z}_{s,e}(t) - \mathrm{E}[\widetilde{Z}_{s,e}(t)]| \lesssim \sqrt{\log n}$:

Lemma 2. *There exists $K > 0$ such that*

$$\Pr\left(A_n'\right) \coloneqq \Pr\left(\max_{\substack{1\leq s\leq k\leq e\leq n\\ e-s>\widetilde{\Delta}n}} |\widetilde{Z}_{s,e}(k/ns, e) - \mathrm{E}[\widetilde{Z}_{s,e}(k/ns, e)]| \leq K\sqrt{\log n}\right) \geq 1 - C/n.$$

Proof. Let $\lambda_n = K\sqrt{\log n}$ and $\widetilde{\sigma}_{n_{s,e}} = \sqrt{(n_{s,e}^2 - 1)/12}$. We can then write

$$\frac{1}{\widetilde{\sigma}_{n_{s,e}}\sqrt{n_{s,e}}}(R_{k,s,e} - \mathrm{E}\left[R_{k,s,e}\right]) = \frac{1}{\sqrt{n_{s,e}}}\sum_{j=1}^{\ell} Z_j,$$

where Z_j are sums of $\gamma_{n,j} \geq 0$ terms of $i.i.d.$ random variables which have mean 0, variance bounded by c and a bounded range. It follows that

$$\Pr\left(\frac{1}{\sqrt{n_{s,e}}}\left|\sum_{j=1}^{\ell} Z_i\right| > \lambda_n\right) \leq \sum_{j=1}^{\ell} \Pr\left(|Z_j| > C\lambda_n\sqrt{\widetilde{\Delta}n/\gamma_{n,j}}\right) \leq ce^{-C\frac{n\log n}{\max_j \gamma_{n,j}}} \leq ce^{-C\log n},$$

for some $c, C = C(K)$ dependent on K. We now use the Bonferroni inequality to get that

$$\begin{aligned}\Pr(A_n') &= \Pr\left(\max_{\substack{1\le s\le k\le e\le n\\ e-s>\widetilde{\Delta}n}} |\widetilde{Z}_{s,e}(k/ns, e) - \mathrm{E}[\widetilde{Z}_{s,e}(k/ns, e)]| \le K\sqrt{\log n}\right)\\ &\le Cn^3 \max_{\substack{1\le s\le k\le e\le n\\ e-s>\widetilde{\Delta}n}} \Pr\left(\frac{1}{\widetilde{\sigma}_{n_{s,e}}\sqrt{n_{s,e}}}|R_k - \mathrm{E}[R_k]| \le K\sqrt{\log n}\right)\\ &\le ce^{-C(K)\log n} \le C/n,\end{aligned}$$

where the last inequality holds with appropriately chosen K. □

Now, we can use the triangle inequality to show that $\max_t |Z_{s,e}(t) - \mathrm{E}\left[\widetilde{Z}_{s,e}(t)\right]| \lesssim c\log n$. Precisely, $\lim_{n\to\infty}\Pr(A_n'') \to 1$, where

$$A_n'' = \{\max_t |Z_{s,e}(t) - \mathrm{E}\left[\widetilde{Z}_{s,e}(t)\right]| \le c\log n\},$$

where we let $\lambda_n = c\log n$ for the remainder of the proof.

The next step is to show that with high probability "nice" intervals are sampled. Define $\mathcal{I}_i$ to be the interval $[k_{i-1} + \frac{1}{3}(k_i - k_{i-1}), k_{i-1} + \frac{2}{3}(k_i - k_{i-1})]$. It follows from (Fryzlewicz, 2014) and our assumptions that

$$D_n = \{\forall i \in [\ell]\ \exists\ m \in [M]\colon (s_m, e_m) \in \mathcal{I}_i \times \mathcal{I}_{i+1}\}. \tag{A.1}$$

In addition, Fryzlewicz (2014) gives that $\Pr(D_n) \to 1$ as $n \to \infty$. We now condition on A_n'' and D_n.

Let s, e be such that

$$k_{i_0} \le s < k_{i_0+1} < \ldots < k_{i_0+\ell'} \le k_{i_0+\ell'+1},$$

for $0 \le i_0 \le \ell - \ell'$. Observe the following conditions on the interval (s, e)

$$s < k_{i_0+y} - C\cdot n < k_{i_0+y} + C\cdot n < e \text{ for some } 1 \le y \le \ell' \tag{A.2}$$

$$\max(\min(k_{i_0+1} - s, s - k_{i_0}), \min(k_{i_0+\ell'+1} - e, e - k_{i_0+\ell'})) \le \frac{Cn^{1/2}\log n}{\Delta^2 p_0^2}. \tag{A.3}$$

First, suppose an interval $(s_m, e_m) \in \mathcal{I}_i \times \mathcal{I}_{i+1}$ contains the change-point k^*. It is easy

to see that:

$$\mathrm{E}\left[\widetilde{Z}_{s,e}(k/n_{s,e})\right] = \begin{cases} \frac{1}{a_{s,e}} k(n_{s,e} - k^*)(p - 1/2) & k \leq k^* \\ \frac{1}{a_{s,e}} k^*(n_{s,e} - k)(p - 1/2) & k > k^* \end{cases}.$$

Clearly, $|\mathrm{E}\left[\widetilde{Z}_{s,e}(k/n_{s,e})\right]|$ is maximized at the change-point k^*. It is also clear from Assumptions 3 and Assumption 4 that $|\mathrm{E}\left[\widetilde{Z}_{s,e}(k^*/n_{s,e})\right]| > Cn^{1/2}p_0$.

Suppose that (A.2) and (A.3) hold. We make the following arguments conditional on the set A_n''. Consider

$$(m_0, \hat{k}_0) = \underset{m \in \mathrm{INT}_{s,e},\ s_m \leq k < e_m}{\operatorname{argmax}} |\widetilde{Z}_{s_m,e_m}(k/n_{s_m,e_m})|.$$

Note that eventually, $\Delta \cdot n \geq 3Cn^r$. When this is true, we have that for any undetected change-points $k_i \in (s, e)$ it holds that $\mathcal{I}_i \cup \mathcal{I}_{i+1} \subset (s, e)$. Denote this set of change-point indices by $\mathcal{K}_{s,e}$. By conditioning, we know that for each $i \in \mathcal{K}_{s,e}$, there exists $m_i \in \mathrm{INT}_{s,e}$ such that $(s_{m_i}, e_{m_i}) \in \mathcal{I}_i \cup \mathcal{I}_{i+1}$. Now,

$$\begin{aligned} |\widetilde{Z}_{s_{m_0},e_{m_0}}(\hat{k}_0/n_{s_m,e_m})| &\geq \max_{s_{m_i} \leq k < e_{m_i}} |\widetilde{Z}_{s_{m_i},e_{m_i}}(k/n_{s_m,e_m})| \\ &\geq |\widetilde{Z}_{s_{m_i},e_{m_i}}(k_i/n_{s_m,e_m})| \geq |\mathrm{E}[\widetilde{Z}_{s_{m_i},e_{m_i}}(k_i/n_{s_m,e_m})]| - \lambda_n \geq Cn^{1/2}p_0, \end{aligned}$$

which holds for large n. It easily follows that

$$|\mathrm{E}[\widetilde{Z}_{s_{m_0},e_{m_0}}(\hat{k}_0/n_{s_m,e_m})]| \geq |\widetilde{Z}_{s_{m_0},e_{m_0}}(\hat{k}_0/n_{s_m,e_m})| - \lambda_n \geq C'n^{1/2}p_0.$$

Now, $|\mathrm{E}[\widetilde{Z}_{s_{m_0},e_{m_0}}(\hat{k}_0/n_{s_m,e_m})]|$ has a local maximum at $|\mathrm{E}[\widetilde{Z}_{s_{m_0},e_{m_0}}(k_{i_0+y}/n_{s_m,e_m})]|$, where k_{i_0+y} is the nearest change-point to either the left of $\hat{k}_0$ or the right of $\hat{k}_0$.

To see this, suppose that k_{i^*} is immediately to the right of $\hat{k}_0$. Then, letting $n_0 = n_{s_{m_0},e_{m_0}}$, we have that

$$\begin{aligned} \mathrm{E}\left[\widetilde{Z}_{s_{m_0},e_{m_0}}(k_{i^*}/n_0) - \widetilde{Z}_{s_{m_0},e_{m_0}}(\hat{k}_0/n_0)\right] &= \sum_{j=\hat{k}_0 - s_{m_0}+1}^{k_{i^*}-s_{m_0}} \mathrm{E}\left[R_{j,s_{m_0},e_{m_0}} - \frac{n_0+1}{2}\right] \\ &= (k_{i^*} - \hat{k}_0)K, \end{aligned}$$

which is monotonic as a function of $\hat{k}_0$ with $k_{i^*-1} < \hat{k}_0 \leq k_{i^*}$. An analogous argument

applies for k_{i^*-1} immediately to the left of $\hat{k}_0$:

$$\mathrm{E}\left[\widetilde{Z}_{s_{m_0},e_{m_0}}(k_{i^*-1}/n_0)\right] - \mathrm{E}\left[\widetilde{Z}_{s_{m_0},e_{m_0}}(\hat{k}_0/n_0)\right] = (\hat{k}_0 - k_{i^*-1})K.$$

Now it is clear that either k_{i^*-1} or k_{i^*} maximizes

$$|\mathrm{E}\left[\widetilde{Z}_{s_{m_0},e_{m_0}}(k/n_0)\right]|,$$

on the interval (k_{i^*-1}, k_{i^*}). Therefore,

$$|\mathrm{E}[\widetilde{Z}_{s_{m_0},e_{m_0}}(k_{i_0+y}/n_{s_m,e_m})]| \geq |\mathrm{E}[\widetilde{Z}_{s_{m_0},e_{m_0}}(\hat{k}_0/n_{s_m,e_m})]| \geq C' n^{1/2} p_0,$$

where k_{i_0+y} is the nearest change-point to either the left of $\hat{k}_0$ or the right of $\hat{k}_0$. Using the same argument in (Fryzlewicz, 2014) on page 2273, it must hold that (s_{m_0}, e_{m_0}) satisfies (A.2). Then, following the same argument in Lemma A.2 of Fryzlewicz (2014), we have that $|\hat{k}_0 - k_{i_0+y}| \leq C n^{1/2} \log n / p_0$.

Therefore, we have shown that A_n'' and D_n both occur with probability approaching 1. Obviously, at the start of the algorithm (A.2) and (A.3) are satisfied. Furthermore, we have that each detected change-point $\hat{k}_i$ satisfies $|\hat{k}_i - k_i| \leq C n^{1/2} \log n / p_0$. Therefore, (A.2) and (A.3) are always satisfied when there are undetected change-points. It remains to show that when there are no change-points remaining, the algorithm does not return false positives.

We now prove that there will be no false positives: Suppose some interval (s, e) contains no change-points. Let $\widetilde{\sigma}_{n_{s,e}} = \sqrt{(n_{s,e}^2 - 1)/12}$. We can then write

$$\widetilde{\sigma}_{n_{s,e}} \sqrt{n_{s,e}} \widetilde{Z}_{s,e}(k/cn) = \frac{k(k+1)}{2} + k(n_{s,e} - k) U_1,$$

where U_1 is a one sample U-statistic with $\mathrm{E}[U_1]$. It holds that $\mathrm{E}\left[\widetilde{\sigma}_{n_{s,e}} \sqrt{n_{s,e}} \widetilde{Z}_{s,e}(k/cn)\right] = \frac{k(n_{s,e}+1)}{2} - \frac{k(k+1)}{2} + \frac{k(n_{s,e}-k)}{2} = 0$. Now, note that $\widetilde{\sigma}_{n_{s,e}} \sqrt{n_{s,e}} \geq c_0 \sqrt{n}$, for some universal c_0. Now, eventually $0 < T < c_0 \sqrt{n}$. We have from (Hoeffding, 1963) that

$$\Pr\left(\widetilde{Z}_{s,e}(k/cn) > T\right) = \Pr\left(U_1 - 1/2 > T \frac{\widetilde{\sigma}_{n_{s,e}} \sqrt{n_{s,e}}}{k(n_{s,e} - k)}\right) \leq \Pr\left(U_1 - 1/2 > T \frac{K}{\sqrt{n}}\right) \leq 2e^{-KT^2},$$

for some universal K. Clearly, $\widetilde{Z}_{s,e}(k/cn)/T \xrightarrow{\mathrm{p}} 0$ for all $e - s > \widetilde{\Delta} n$. □

Proof of Theorem 2. Let $\Delta = \min_{i,j\in[\ell]} |\theta_j - \theta_i|$. Let C_i be fixed positive constants independent of n, $|A|$ represent the cardinality of the set A, and

$$\widetilde{\sigma}_n^2 = \frac{n(n+1)}{12}.$$

To complete the proof, we first prove two lemmas.

Lemma 3. *For $i \in [n]$, it holds that*

$$\mathrm{V}ar(\widehat{R}_i)/\mathrm{V}ar(R_i) = O(1) \qquad \textit{and} \qquad \mathrm{V}ar(R_i)/\widetilde{\sigma}_n^2 = O(1). \tag{A.4}$$

Proof. The right-side identity follows easily from Assumption 3; $\mathrm{V}ar(R_i) = O(n^2)$, for any $i \in [n]$. Using (A.6), we can write

$$\begin{aligned}
\mathrm{V}ar\left(\widehat{R}_i\right) &= \mathrm{V}ar\left(R_i + \mathcal{E}_i\right) \\
&= \mathrm{V}ar\left(R_i\right) + \mathrm{V}ar\left(\mathcal{E}_i\right) + 2\mathrm{C}ov\left(\mathcal{E}_i, R_i\right) \\
&\leq \mathrm{V}ar\left(R_i\right) + \mathrm{V}ar\left(\mathcal{E}_i\right) + 2\mathrm{E}\left[\left|\mathcal{E}_i - \mathrm{E}\left[\mathcal{E}_i\right]\right|\right] n \\
&= \mathrm{V}ar\left(R_i\right) + \mathrm{V}ar\left(\mathcal{E}_i\right) + O(n^{3/2}).
\end{aligned}$$

We also have that

$$\begin{aligned}
\mathrm{V}ar\left(\mathcal{E}_i\right) &= \mathrm{E}\left[\left(\sum_{m=1}^{n} \mathbb{1}\left(B_{i,m}\right) - \sum_{m=1}^{n} \mathbb{1}\left(A_{i,m}\right)\right)^2\right] + O(n) \\
&= \mathrm{E}\left[\sum_{m_1=1}^{n}\sum_{m_2=1}^{n} \left[\mathbb{1}\left(B_{i,m_1}\right) - \mathbb{1}\left(A_{i,m_1}\right)\right]\left[\mathbb{1}\left(B_{i,m_2}\right) - \mathbb{1}\left(A_{i,m_2}\right)\right]\right] + O(n) \\
&\leq \mathrm{E}\left[\sum_{m_1=1}^{n}\sum_{m_2=1}^{n} \left[\mathbb{1}\left(B_{i,m_1}\right) + \mathbb{1}\left(A_{i,m_1}\right)\right]\right] + O(n) \\
&\leq O(n^{3/2}),
\end{aligned}$$

where the first line comes from applying equation (A5) of Chenouri et al. (2020b) and the last line is from the the fact that $\mathrm{E}\left[\mathbb{1}\left(B_{i,m}\right)\right] = O(n^{-1/2})$ and $\mathrm{E}\left[\mathbb{1}\left(A_{i,m}\right)\right] = O(n^{-1/2})$ (Chenouri et al., 2020b). Now,

$$\lim_{n\to\infty} \frac{\mathrm{V}ar(\widehat{R}_i)}{\mathrm{V}ar(R_i)} = \lim_{n\to\infty} \frac{\mathrm{V}ar(\widehat{R}_i)/n^2}{\mathrm{V}ar(R_i)/n^2} = \lim_{n\to\infty} \frac{\mathrm{V}ar(R_i)/n^2 + o(1)}{\mathrm{V}ar(R_i)/n^2} = 1.$$

□

Define the set $\mathbf{X}_n = 2^{[n-1]} \times \{0\} \times \{n\}$; elements of $\mathbf{X}_n$ are sets of indices ranging from 0 to n, which represent locations of change-points. A member of $\mathbf{X}_n$ is a set $\mathbf{x}$ that contains 0 and n joined with an element of the power set of $[n-1]$. We will represent such an element with $\mathbf{x} = \{x_0, \ldots, x_{p+1}\}$ where $x_0 \coloneqq 0 < x_1 < \ldots < x_p < x_{p+1} \coloneqq n$. $\mathbf{X}_n$ forms the space of possible sets of change-points for a fixed n. We can then write the objective function based on the population depth ranks $\mathcal{T}$ and the objective function based on the sample depth ranks $\widehat{\mathcal{T}}$ as follows:

$$\widehat{\mathcal{T}}(\mathbf{x}) \coloneqq \frac{1}{\widetilde{\sigma}_n^2} \sum_{i=1}^{|\mathbf{x}|} (x_i - x_{i-1}) \overline{\widehat{R}}_i^2 - 3(n+1) - \beta_n(|\mathbf{x}| - 1) \coloneqq \widehat{\mathcal{C}}(\mathbf{x}) - \beta_n(|\mathbf{x}| - 1)$$

$$\mathcal{T}(\mathbf{x}) \coloneqq \frac{1}{\widetilde{\sigma}_n^2} \sum_{i=1}^{|\mathbf{x}|} (x_i - x_{i-1}) \overline{R}_i^2 - 3(n+1) - \beta_n(|\mathbf{x}| - 1) \coloneqq \mathcal{C}(\mathbf{x}) - \beta_n(|\mathbf{x}| - 1), \tag{A.5}$$

where $\mathbf{x} \in \mathbf{X}_n$ and

$$\overline{R}_i = \frac{1}{x_i - x_{i-1}} \sum_{j=x_{i-1}+1}^{x_i} R_j \qquad \text{and} \qquad \overline{\widehat{R}}_i = \frac{1}{x_i - x_{i-1}} \sum_{j=x_{i-1}+1}^{x_i} \widehat{R_j}.$$

Lemma 4. *Let $\mathbf{x}_n \in 2^{[n-1]} \times \{0\} \times \{n\}$ be such that for each $x_j \in \mathbf{x}_n \backslash n$, it holds that $\min_{i \in \ell, k_i > x_j} x_j + 1 - k_i > cn$. Additionally impose that $|\mathbf{x}_n| = O(1)$. Then,*

$$|\widehat{\mathcal{T}}(\mathbf{x}_n) - \mathcal{T}(\mathbf{x}_n)| = O_p(1).$$

Proof. First, we remind the reader that each $x_j \in \mathbf{x}_n$ depends on n, which we omit in the notation for brevity. Note that $R_{k_i+1}, \ldots, R_{k_{i+1}}$ is a triangular array of exchangeable random variables, for any $i \in \{0\} \cup [\ell]$. The same holds for $\widehat{R}_{k_i+1}, \ldots, \widehat{R}_{k_{i+1}}$. It follows that for any $j \in [c\ell]$ the sequences $R_{x_{j-1}+1}, \ldots, R_{x_j}$ and $\widehat{R}_{x_{j-1}+1}, \ldots, \widehat{R}_{x_j}$ are both sums of segments of triangular arrays of exchangeable random variables. Specifically, suppose without loss of generality that the segment $(x_{i-1}+1, x_i)$ contains $0 \leq M \leq \ell$ change-points,

by which we denote $k_1, \ldots, k_M$. Then, we have that

$$\sum_{j=x_{i-1}+1}^{x_i} R_j = \sum_{j=x_{i-1}+1}^{k_1} R_j + \sum_{j=k_1+1}^{k_2} R_j + \ldots + \sum_{j=k_M+1}^{x_i} R_j.$$

This form allows us to apply the central limit theorem of Weber (1980); for any interval (s, e) which contains no change-points and is such that $e - s > cn$, it holds that

$$\frac{1}{\widetilde{\sigma}_n \sqrt{e-s}} \sum_{j=s}^{e} R_j = O_p(1) + \frac{(e-s+1)\mathrm{E}[R_s]}{\widetilde{\sigma}_n \sqrt{e-s}}.$$

This result follows from Lemma 3. In addition, note that

$$\mathrm{E}\left[\overline{R}_i\right] = \sum_{j=1}^{M+1} \frac{c_j n \mathrm{E}[R_{k_j}]}{x_i - x_{i-1}},$$

where c_j is the proportion of points in segment j. It immediately follows that

$$\frac{\sqrt{x_i - x_{i-1}}}{\widetilde{\sigma}_n} \sum_{j=x_{i-1}+1}^{x_i} (\overline{R}_i - \mathrm{E}\left[\overline{R}_i\right]) = O_p(1).$$

The same argument gives that

$$\frac{\sqrt{x_i - x_{i-1}}}{\widetilde{\sigma}_n} \sum_{j=x_{i-1}+1}^{x_i} (\overline{\widehat{R}}_i - \mathrm{E}[\overline{\widehat{R}}_i]) = O_p(1).$$

We now relate these to quantities. Consider the representation of $\widehat{R}_i$

$$\widehat{R}_i = R_i + \sum_{m=1}^{n} \mathbb{1}\left(B_{i,m}\right) - \sum_{m=1}^{n} \mathbb{1}\left(A_{i,m}\right) \coloneqq R_i + \mathcal{E}_i, \tag{A.6}$$

where

$$\begin{aligned} A_{i,j} &= \left\{D\left(X_j, F_*\right) \leq D\left(X_i, F_*\right)\right\} \cap \left\{D\left(X_j, F_{*,n}\right) > D\left(X_i, F_{*,n}\right)\right\} \\ B_{i,j} &= \left\{D\left(X_j, F_*\right) > D\left(X_i, F_*\right)\right\} \cap \left\{D\left(X_j, F_{*,n}\right) \leq D\left(X_i, F_{*,n}\right)\right\}. \end{aligned}$$

We can use this representation, Assumption 1 and Assumption 2 to show that

$$\mathrm{E}[\mathcal{E}_i] = \mathrm{E}[\widehat{R}_i] - \mathrm{E}[R_i] = O(n^{1/2}).$$

For more details, see pages 436-437 of Chenouri et al. (2020b).

It then follows from Slutsky's theorem, continuous mapping theorem and the central limit theorem of Weber (1980) that

$$\begin{aligned}
\widehat{\mathcal{T}}(\mathbf{x}_n) - \mathcal{T}(\mathbf{x}_n) &= \frac{1}{\widetilde{\sigma}_n^2} \sum_{i=1}^{\ell+2} (x_i - x_{i-1}) \left(\overline{\widehat{R}}_i^2 - \overline{R}_i^2 \right) \\
&= \sum_{i=1}^{|\mathbf{x}_n|} \left(\frac{\sqrt{(x_i - x_{i-1})}\ \overline{\widehat{R}}_i}{\widetilde{\sigma}_n} \right)^2 - \left(\frac{\sqrt{(x_i - x_{i-1})}\overline{R}_i}{\widetilde{\sigma}_n} \right)^2 \\
&= O_p(1) + \frac{1}{\widetilde{\sigma}_n^2} \sum_{i=1}^{|\mathbf{x}_n|} (x_i - x_{i-1}) [\mathrm{E}[\overline{\widehat{R}}_i]^2 - \mathrm{E}[\overline{R}_i]^2 + \mathrm{E}[\overline{R}_i]\overline{R}_i - \mathrm{E}[\overline{\widehat{R}}_i]\overline{\widehat{R}}_i] \\
&= O_p(1) + \frac{1}{\widetilde{\sigma}_n^2} \sum_{i=1}^{|\mathbf{x}_n|} (x_i - x_{i-1}) [\mathrm{E}[\overline{\widehat{R}}_i](\mathrm{E}[\overline{\widehat{R}}_i] - \overline{\widehat{R}}_i) - \mathrm{E}[\overline{R}_i](\mathrm{E}[\overline{R}_i] - \overline{R}_i)] \\
&= O_p(1) + \frac{1}{\widetilde{\sigma}_n^2} \sum_{i=1}^{|\mathbf{x}_n|} (x_i - x_{i-1}) [\mathrm{E}[\overline{R}_i](\mathrm{E}[\overline{\mathcal{E}}_i] - \overline{\mathcal{E}}_i) + \mathrm{E}[\overline{\mathcal{E}}_i](\mathrm{E}[\overline{\widehat{R}}_i] - \overline{\widehat{R}}_i)] \\
&= O_p(1).
\end{aligned}$$

This analysis gives the result that

$$\widehat{\mathcal{T}}(\mathbf{x}_n) - \mathcal{T}(\mathbf{x}_n) = \widehat{\mathcal{C}}(\mathbf{x}_n) - \mathcal{C}(\mathbf{x}_n) = O_p(1), \tag{A.7}$$

when $\mathbf{x}_n$ is as described. □

Proceeding with the proof, we make an argument by contradiction, similar to that of Wang et al. (2021). However, we use the previously discussed exchangeablility results, i.e., (Weber, 1980) which were not used in their paper. Recall, $\widehat{\mathbf{k}}$ is the estimated set of change-points and $\mathbf{k}$ is the true set of change-points. We examine the events $\{\widehat{\ell} < \ell\}$, $\{\widehat{\ell} > \ell\}$ and $\left\{\max_{k \in \mathbf{k}} \min_{\widehat{k} \in \widehat{\mathbf{k}}} |\widehat{k} - k| \geq Cn^r\right\}$ separately. We start with $\{\widehat{\ell} < \ell\}$:

Lemma 5. *It holds that* $\Pr(\widehat{\ell} < \ell) \to 0$ *as* $n \to \infty$.

Proof. Assume $\widehat{\ell} < \ell$; by Assumption 3, there is at least one change-point $0 < k_{i^*} < n$ such that for any $j \in [\widehat{\ell}]$ it is true that $|k_{i^*} - \widehat{k}_j| \geq \Delta n/2$ with Δ independent of n. Now, define

$$\mathbf{w}_1 = \{k_{i^*} - \Delta n/2, k_{i^*} + \Delta n/2\} \cup \mathbf{k} \backslash k_{i^*} \qquad \text{and} \qquad \mathbf{w}_2 = \mathbf{w}_1 \cup \widehat{\mathbf{k}}.$$

Clearly, $\widehat{\mathcal{C}}(\mathbf{w}_2) \geq \widehat{\mathcal{C}}(\widehat{\mathbf{k}})$ (which is the necessary condition for PELT, recall that $\widehat{\mathcal{C}}$ is the portion of the objective function without the penalty) and so we work with $\widehat{\mathcal{C}}(\mathbf{w}_2)$. The goal is to show that following contradiction to the assumption that some $\widehat{\mathbf{k}}$ such that $\widehat{\ell} < \ell$ is the maximizer of $\widehat{\mathcal{T}}$. To see this, we have

$$\begin{aligned}
\mathcal{T}(\mathbf{k}) - \widehat{\mathcal{T}}(\widehat{\mathbf{k}}) &= \mathcal{C}(\mathbf{k}) - \widehat{\mathcal{C}}(\widehat{\mathbf{k}}) - O(\beta_n) \\
&\geq \mathcal{C}(\mathbf{k}) - \widehat{\mathcal{C}}(\mathbf{w}_2) - O(\beta_n) \\
&= \mathcal{C}(\mathbf{k}) - \widehat{\mathcal{C}}(\mathbf{w}_1) + O_p(1) - O(\beta_n) \\
&= \mathcal{C}(\mathbf{k}) - \mathcal{C}(\mathbf{w}_1) + O_p(1) - O(\beta_n) \\
&= O_p(n) - O(\beta_n) + O_p(1) \xrightarrow{\text{p}} \infty,
\end{aligned}$$

as $n \to \infty$, since $\beta_n = o(n)$.

First, we show that

$$\widehat{\mathcal{C}}(\mathbf{w}_2) = \widehat{\mathcal{C}}(\mathbf{w}_1) + O_p(1).$$

To this end, letting $w_0 = 0$, $w_{\ell+\widehat{\ell}+2} = n$ and $\mathbf{w}_2 = \{w_0, w_1, w_2, \ldots, w_{\ell+\widehat{\ell}+1}, w_{\ell+\widehat{\ell}+2}\}$ where $w_m < w_j$ for $m < j$, we can write

$$\widehat{\mathcal{C}}(\mathbf{w}_1) - \widehat{\mathcal{C}}(\mathbf{w}_2) = \frac{1}{\widetilde{\sigma}_n^2} \sum_{j=1}^{|\mathbf{w_2}|} (w_j - w_{j-1}) \left[\overline{\widehat{R}}_j(\mathbf{w}_1)^2 - \overline{\widehat{R}}_j(\mathbf{w}_2)^2 \right]$$

where

$$\overline{\widehat{R}}_j(\mathbf{x}) = \frac{1}{n_{j,2}(\mathbf{x}) - n_{j,1}(\mathbf{x})} \sum_{i=n_{j,1}(\mathbf{x})+1}^{n_{j,2}(\mathbf{x})} \widehat{R}_i,$$

and

$$n_{j,1}(\mathbf{x}) = \underset{x \in \mathbf{x}:\, x \leq w_{j-1}}{\operatorname{argmin}} |x - w_{j-1}|, \qquad n_{j,2}(\mathbf{x}) = \underset{x \in \mathbf{x}:\, x \geq w_j}{\operatorname{argmin}} |x - w_j|.$$

In this context,

$$\overline{\widehat{R}}_j(\mathbf{w}_2) = \frac{1}{(w_j - w_{j-1})} \sum_{m=w_{j-1}+1}^{w_j} \widehat{R}_m \quad \text{and} \quad \overline{\widehat{R}}_j(\mathbf{w}_1) = \frac{1}{n_{j,2}(\mathbf{w}_1) - n_{j,1}(\mathbf{w}_1)} \sum_{m=n_{j,1}(\mathbf{w}_1)+1}^{n_{j,2}(\mathbf{w}_1)} \widehat{R}_m \ .$$

To elaborate, ordering the points in $\mathbf{w}_1$ defines $\ell+2$ disjoint groups of ranks and therefore $\ell+2$ group means. The value $\overline{\widehat{R}}_j(\mathbf{w}_1)$ is the mean of such a group of ranks which also contains the ranks $\{\widehat{R}_{w_{j-1}}, \ldots, \widehat{R}_{w_j}\}$.

Let j^* represent $w_{j^*} = k_{i^*} + \Delta n/2$. Then we have that

$$\begin{aligned}
\widehat{\mathcal{C}}(\mathbf{w}_1) - \widehat{\mathcal{C}}(\mathbf{w}_2) &= \frac{1}{\widetilde{\sigma}_n^2} \sum_{j=1}^{|\mathbf{w}_2|} (w_j - w_{j-1}) \left(\overline{\widehat{R}}_j(\mathbf{w}_1)^2 - \overline{\widehat{R}}_j(\mathbf{w}_2)^2 \right) \\
&= \frac{1}{\widetilde{\sigma}_n^2} \sum_{j \in [\ell+\widehat{\ell}+1] \setminus j^*} (w_j - w_{j-1}) \left(\overline{\widehat{R}}_j(\mathbf{w}_1)^2 - \overline{\widehat{R}}_j(\mathbf{w}_2)^2 \right). \qquad \text{(A.8)}
\end{aligned}$$

For any $j \neq j^*$ Lemma 3 gives

$$\frac{(w_j - w_{j-1})}{\widetilde{\sigma}_n^2} \left(\overline{\widehat{R}}_j(\mathbf{w}_1)^2 - \overline{\widehat{R}}_j(\mathbf{w}_2)^2 \right) = O(1) \frac{(w_j - w_{j-1})}{\mathrm{Var}\left(R_{w_j}\right)} \left(\overline{\widehat{R}}_j(\mathbf{w}_1)^2 - \overline{\widehat{R}}_j(\mathbf{w}_2)^2 \right).$$

Now, it is easy to see that the central limit theorem of Weber (1980) gives that

$$\frac{(w_j - w_{j-1})}{\mathrm{Var}\left(R_{w_j}\right)} \overline{\widehat{R}}_j(\mathbf{w}_1)^2 = O_p(1) + O_p(1) \sqrt{\frac{(w_j - w_{j-1})}{\mathrm{Var}\left(R_{w_j}\right)}} \mathrm{E}[\widehat{R}_{w_j}] + \frac{(w_j - w_{j-1})}{\mathrm{Var}\left(R_{w_j}\right)} \mathrm{E}[\widehat{R}_{w_j}]^2,$$

and similarly

$$\frac{(w_j - w_{j-1})}{\mathrm{Var}\left(R_{w_j}\right)} \overline{\widehat{R}}_j(\mathbf{w}_2)^2 = O_p(1) + O_p(1) \sqrt{\frac{(w_j - w_{j-1})}{\mathrm{Var}\left(R_{w_j}\right)}} \mathrm{E}[\widehat{R}_{w_j}] + \frac{(w_j - w_{j-1})}{\mathrm{Var}\left(R_{w_j}\right)} \mathrm{E}[\widehat{R}_{w_j}]^2.$$

Now, we have that

$$\widehat{\mathcal{C}}(\mathbf{w}_1) - \widehat{\mathcal{C}}(\mathbf{w}_2) = O_p(1). \qquad \text{(A.9)}$$

It follows immediately from Lemma 4 that $\widehat{\mathcal{C}}(\mathbf{w}_1) - \mathcal{C}(\mathbf{w}_1) = O_p(1)$; $\mathbf{w}_1$ satisfies the conditions of $\mathbf{x}_n$ in Lemma 4.

Now, we want to show that

$$\lim_{n\to\infty} \mathcal{C}(\mathbf{k}) - \mathcal{C}(\mathbf{w}_1) = O_p(n).$$

Let k_{i^*-1} and k_{i^*+1} be the true change-points immediately preceding and following k_{i^*} respectively. Recall k_{i^*} is the change-point that is at least $\Delta n/2$ points away from any estimated change-point. Note that $\mathbf{k} - \mathbf{w}_1 = \{k_{i^*}\}$ and $\mathbf{w}_1 - \mathbf{k} = \{k_{i^*} \pm \Delta n/2\}$. We have

$$\begin{aligned}
\frac{n+1}{n}\left(\mathcal{C}(\mathbf{k}) - \mathcal{C}(\mathbf{w}_1)\right) = {} & \frac{12\vartheta_{i^*}n}{n^2}\left[\frac{1}{n\vartheta_{i^*}}\sum_{j=k_{i^*\text{-}1}+1}^{k_{i^*}} R_j\right]^2 + \frac{12\vartheta_{i^*+1}n}{n^2}\left[\frac{1}{n\vartheta_{i^*+1}}\sum_{j=k_{i^*}+1}^{k_{i^*+1}} R_j\right]^2 \\
& - \frac{12\Delta n}{n^2}\left[\frac{1}{n\Delta}\sum_{j=k_{i^*}\text{-}\Delta n/2}^{k_i+\Delta n/2} R_j\right]^2 \\
& - \frac{12n(\vartheta_{i^*} - \Delta/2)}{n^2}\left[\frac{1}{n(\vartheta_{i^*} - \Delta/2)}\sum_{j=k_{i^*\text{-}1}+1}^{k_{i^*}-\Delta n/2} R_j\right]^2 \\
& - \frac{12n(\vartheta_{i^*+1} - \Delta/2)}{n^2}\left[\frac{1}{n(\vartheta_{i^*+1} - \Delta/2)}\sum_{j=k_{i^*}+\Delta n/2}^{k_{i^*+1}} R_j\right]^2.
\end{aligned}$$

For arbitrary $k_m \in \mathbf{k}$ choose $j \in \{k_{m-1}+1, \ldots, k_m\}$, then

$$\begin{aligned}
\mathrm{E}\left[R_j\right] &= \sum_{j\in[\ell+1]\setminus m} n\vartheta_j p_{m,j} - \frac{n\vartheta_i - 1}{2} = n\left[\sum_{j=1}^{\ell+1}\vartheta_j p_{m,j} - \frac{1}{2}\right] \\
\mathrm{V}ar\left(R_j\right) &\le n - 1 + n(n-1)/2.
\end{aligned}$$

It follows from continuous mapping theorem and (Weber, 1980) that

$$\frac{1}{n^2}\left[\frac{1}{n\vartheta_i}\sum_{j=k_{i^*-1}+1}^{k_{i^*}} R_j\right]^2 \xrightarrow{\text{p}} \left[\sum_{j=1}^{\ell+1}\vartheta_j p_{i^*,j} - \frac{1}{2}\right]^2,$$

$$\frac{1}{n^2}\left[\frac{1}{n(\vartheta_i - \Delta/2)}\sum_{j=k_{i^*-1}+1}^{k_{i^*}-\Delta n/2} R_j\right]^2 \xrightarrow{\text{p}} \left[\sum_{j=1}^{\ell+1}\vartheta_j p_{i^*,j} - \frac{1}{2}\right]^2,$$

$$\frac{1}{n^2}\left[\frac{1}{n\vartheta_{i+1}}\sum_{j=k_{i^*-1}+1}^{k_{i^*}} R_j\right]^2 \xrightarrow{\text{p}} \left[\sum_{j=1}^{\ell+1}\vartheta_j p_{i^*+1,j} - \frac{1}{2}\right]^2,$$

$$\frac{1}{n^2}\left[\frac{1}{n(\vartheta_{i+1} - \Delta/2)}\sum_{j=k_{i^*}+\Delta n/2}^{k_{i^*+1}} R_j\right]^2 \xrightarrow{\text{p}} \left[\sum_{j=1}^{\ell+1}\vartheta_j p_{i^*+1,j} - \frac{1}{2}\right]^2,$$

$$\frac{1}{n^2}\left[\frac{1}{n\Delta}\sum_{j=k_{i^*}-\Delta n/2}^{k_{i^*}+\Delta n/2} R_j\right]^2 \xrightarrow{\text{p}} \frac{1}{4}\left[\sum_{j=1}^{\ell+1}\vartheta_j p_{i^*,j} - \frac{1}{2} + \sum_{j=1}^{\ell+1}\vartheta_j p_{i^*+1,j} - \frac{1}{2}\right]^2.$$

Slutsky's lemma and continuous mapping theorem directly imply that

$$\begin{aligned}\frac{n+1}{n^2}\left(\mathcal{C}(\mathbf{k}) - \mathcal{C}(\mathbf{w}_1)\right) \xrightarrow{\text{p}} \frac{12\Delta}{4}\Bigg[&\left(\sum_{j=1}^{\ell+1}\vartheta_j p_{i^*,j} - \frac{1}{2}\right)^2 + \left(\sum_{j=1}^{\ell+1}\vartheta_j p_{i^*+1,j} - \frac{1}{2}\right)^2 \\ &- 2\left(\sum_{j=1}^{\ell+1}\vartheta_j p_{i^*+1,j} - \frac{1}{2}\right)\left(\sum_{j=1}^{\ell+1}\vartheta_j p_{i^*,j} - \frac{1}{2}\right)\Bigg] \\ = 3\Delta\left[\sum_{j=1}^{\ell+1}\vartheta_j p_{i^*+1,j} - \frac{1}{2} - \sum_{j=1}^{\ell+1}\vartheta_j p_{i^*,j} + \frac{1}{2}\right]^2 &> 0.\end{aligned}$$

We can then conclude that $\mathcal{C}(\mathbf{k}) - \mathcal{C}(\mathbf{w}_1) \to +\infty$ in probability at a rate of $O_p(n)$. Then, we have that

$$\mathcal{T}(\mathbf{k}) - \widehat{\mathcal{T}}(\widehat{\mathbf{k}}) = O_p(n) + O_p(1) - \beta_n \to \infty,$$

providing a contradiction to the assumption that $\widehat{\ell} < \ell$. □

Lemma 6. *It holds that* $\Pr(\widehat{\ell} > \ell) \to 0$ *as* $n \to \infty$.

Proof. Now assume that $\widehat{\ell} > \ell$. It is easy to see that $\widehat{\mathcal{C}}(\widehat{\mathbf{k}}) \leq \widehat{\mathcal{C}}(\widehat{\mathbf{k}} \cup \mathbf{k})$. Using this fact and a similar analysis as to that of the event $\{\widehat{\ell} < \ell\}$, we can write that

$$\mathcal{C}(\mathbf{k}) - \widehat{\mathcal{C}}(\widehat{\mathbf{k}}) \geq \mathcal{C}(\mathbf{k}) - \widehat{\mathcal{C}}(\widehat{\mathbf{k}} \cup \mathbf{k}) = \mathcal{C}(\mathbf{k}) - \mathcal{C}(\widehat{\mathbf{k}} \cup \mathbf{k}) + O_p(1) = O_p(1).$$

We then have that

$$\mathcal{T}(\mathbf{k}) - \widehat{\mathcal{T}}(\widehat{\mathbf{k}}) = \mathcal{C}(\mathbf{k}) - \widehat{\mathcal{C}}(\widehat{\mathbf{k}}) + \beta_n(\widehat{\ell} - \ell) \geq \mathcal{C}(\mathbf{k}) - \widehat{\mathcal{C}}(\widehat{\mathbf{k}} \cup \mathbf{k}) + \beta_n(\widehat{\ell} - \ell) = O(\beta_n) + O_p(1) \to \infty,$$

as $n \to \infty$. □

Lemma 7. *It holds that* $\max_{k \in \mathbf{k}} \min_{\widehat{k} \in \widehat{\mathbf{k}}} \frac{1}{Cn^r} |\widehat{k} - k| \xrightarrow{p} 0$ *as* $n \to \infty$.

Proof. We take the contradiction approach again; consider there exists $k_{i^*} \in \mathbf{k}$ such that $\min_{k \in \mathbf{k}} |\widehat{k} - k_{i^*}| > Cn^r$. Define $\mathbf{w}_1'$ in the same way as $\mathbf{w}_1$ but replace Δ with C:

$$\mathbf{w}_1' = \{k_{i^*} - Cn^r/2, k_{i^*} + Cn^r/2\} \cup \mathbf{k} \backslash k_{i^*} \qquad \text{and} \qquad \mathbf{w}_2' = \mathbf{w}_1' \cup \widehat{\mathbf{k}}.$$

Similar to the analysis of $\{\widehat{\ell} < \ell\}$, we can write

$$\begin{aligned}
\frac{n+1}{n} \left(\mathcal{C}(\mathbf{k}) - \mathcal{C}(\mathbf{w}_1')\right) = {} & \frac{12\vartheta_{i^*} n}{n^2} \left[\frac{1}{n\vartheta_{i^*}} \sum_{j=k_{i^*\text{-}1}+1}^{k_{i^*}} R_j\right]^2 + \frac{12\vartheta_{i^*+1} n}{n^2} \left[\frac{1}{n\vartheta_{i^*+1}} \sum_{j=k_{i^*}+1}^{k_{i^*+1}} R_j\right]^2 \\
& - \frac{12Cn^r}{n^2} \left[\frac{1}{Cn^r} \sum_{j=k_{i^*}\text{-}Cn^r/2}^{k_i + Cn^r/2} R_j\right]^2 \\
& - \frac{12n(\vartheta_{i^*} - Cn^{r-1}/2)}{n^2} \left[\frac{1}{n(\vartheta_{i^*} - Cn^{r-1}/2)} \sum_{j=k_{i^*\text{-}1}+1}^{k_{i^*} - Cn^r/2} R_j\right]^2 \\
& - \frac{12n(\vartheta_{i^*+1} - Cn^{r-1}/2)}{n^2} \left[\frac{1}{n(\vartheta_{i^*+1} - Cn^{r-1}/2)} \sum_{j=k_{i^*}+Cn^r/2}^{k_{i^*+1}} R_j\right]^2.
\end{aligned}$$

We can let $\mu_{k_i^*-1} = \mathrm{E}\left[R_{k_i^*}\right]$ and $\mu_{k_i^*} = \mathrm{E}\left[R_{k_i^*+1}\right]$ be the means of the ranks before and after the change-point k_{i^*}. Similarly, we can let $\varsigma^2_{k_i^*-1} = \mathrm{V}ar\left(R_{k_i^*}\right)$ and $\varsigma^2_{k_i^*} = \mathrm{V}ar\left(R_{k_i^*+1}\right)$ be the variances of the ranks before and after the change-point k_{i^*}. We also define $b_{n,k_i^*-1} = \frac{12\varsigma^2_{k_i^*-1}}{n^2} = O(1)$ and $b_{n,k_i^*} = \frac{12\varsigma^2_{k_i^*}}{n^2} = O(1)$. Now, let

$$\tilde{R}_j = \frac{R_j - \mathrm{E}\left[R_j\right]}{\mathrm{V}ar\left(R_j\right)}.$$

It follows that

$$\frac{12\vartheta_{i^*}n}{n^2}\left[\frac{1}{n\vartheta_{i^*}}\sum_{j=k_{i^*\text{-}1}+1}^{k_{i^*}}R_j\right]^2 = b_{n,k_i^*}\left[\sqrt{n\vartheta_{i^*}}\sum_{j=k_{i^*\text{-}1}+1}^{k_{i^*}}\tilde{R}_j+\sqrt{n\vartheta_{i^*}}\frac{\mu_{k_i^*}}{\varsigma_{k_i^*}}\right]^2$$
$$= b_{n,k_i^*}n\vartheta_{i^*}\left(\frac{\mu_{k_i^*}}{\varsigma_{k_i^*}}\right)^2+O_p(n^{1/2}).$$

The last line follows from the central limit theorem of (Weber, 1980) and the previous paragraph. Let $a_n=(n+1)/n$. We can produce similar analyses as above to give

$$a_n\left(\mathcal{C}(\mathbf{k})-\mathcal{C}(\mathbf{w}_1')\right) = b_{n,k_i^*-1}n\vartheta_{i^*}\left(\frac{\mu_{k_i^*-1}}{\varsigma_{k_i^*-1}}\right)^2+b_{n,k_i^*}n\vartheta_{i^*+1}\left(\frac{\mu_{k_i^*}}{\varsigma_{k_i^*}}\right)^2-\frac{Cn^r}{4}b_{n,k_i^*-1}\left(\frac{\mu_{k_i^*-1}}{\varsigma_{k_i^*-1}}\right)^2$$
$$-\frac{Cn^r}{4}b_{n,k_i^*}\left(\frac{\mu_{k_i^*}}{\varsigma_{k_i^*}}\right)^2-b_{n,k_i^*-1}(\vartheta_{i^*}-Cn^{r-1}/2)\left(\frac{\mu_{k_i^*-1}}{\varsigma_{k_i^*-1}}\right)^2$$
$$-b_{n,k_i^*}n(\vartheta_{i^*+1}-Cn^{r-1}/2)\left(\frac{\mu_{k_i^*}}{\varsigma_{k_i^*}}\right)^2+O_p(n^{1/2})\xrightarrow{\text{p}}\infty,$$

where the conclusion follows from the fact that $r>1/2$ and the $O_p(n)$ term is positive. From which it follows that

$$\mathcal{T}(\mathbf{k})-\widehat{\mathcal{T}}(\widehat{\mathbf{k}})=\mathcal{C}(\mathbf{k})-\widehat{\mathcal{C}}(\widehat{\mathbf{k}})\geq\mathcal{C}(\mathbf{k})-\widehat{\mathcal{C}}(\mathbf{w}_2')=\mathcal{C}(\mathbf{k})-\mathcal{C}(\mathbf{w}_1')+O_p(1)\xrightarrow{\text{p}}+\infty. \qquad \square$$

The result follows directly from Lemma 5, Lemma 6 and Lemma 7. $\square$

A.2 Simulation on the rank distributions

We show here that more types of changes in the covariance matrix are exhibited by changes in the depth rankings. Consider two samples each from a 6-dimensional multivariate normal distribution. We fix

$$\Sigma_1=\begin{bmatrix}1 & 0.4 & 0.4 & 0 & 0 & 0\\ 0.4 & 1 & 0.4 & 0 & 0 & 0\\ 0.4 & 0.4 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0.4\\ 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0.4 & 0 & 1\end{bmatrix}$$

as the covariance matrix of the first sample. Additionally, let $\sigma_{d_1,d_2,m}$ be the $(d_1,\ d_2)^{th}$ entry of the covariance matrix for sample m (where $d_1,\ d_2 \in \{1,\ldots,6\}$ and $m \in \{1,2\}$). We test four specifications of Σ_2, the covariance matrix of the second sample, and check for a difference in the distribution of ranks:

1. **Submatrix on the diagonal change:** $\sigma_{d_1,d_2,2} = 2\sigma_{d_1,d_2,1}$ for $d_1, d_2 > 3$ and $\sigma_{d_1,d_2,2} = \sigma_{d_1,d_2,1}$ otherwise.
2. **Submatrix off the diagonal change:** $\sigma_{6,4,2} = \sigma_{4,6,2} = 2\sigma_{6,4,1}$ and $\sigma_{d_1,d_2,2} = \sigma_{d_1,d_2,1}$ otherwise.
3. **Mixed change scenario:** $\sigma_{6,4,2} = \sigma_{4,6,2} = -\sigma_{6,4,1}$, $\sigma_{4,4,2} = 0.2\sigma_{4,4,1}$, $\sigma_{d_1,d_2,2} = 2\sigma_{d_1,d_2,1}$ for $d_1, d_2 \leq 3,\ d_1 \neq d_2$ and $\sigma_{d_1,d_2,2} = \sigma_{d_1,d_2,1}$ otherwise.
4. **Offsetting Expansion and Contraction:** $\sigma_{4,4,2} = 0.5\sigma_{4,4,1}$, $\sigma_{6,6,2} = 2\sigma_{6,6,1}$ and $\sigma_{d_1,d_2,2} = \sigma_{d_1,d_2,1}$ otherwise.

We drew samples of size $n = 5000$ from each population and computed the combined sample depth ranks. We then repeated this 100 times for each scenario. Figure A.1 shows histograms of each samples' depth ranks, one graph for each scenario. We see that expansions and contractions of submatrices correspond to changes in the rank distribution. Scenario three represents a mixture of these expansions and contractions (of different submatrices) and a change in the rank distributions is still exhibited. Scenario four shows that if we have two simultaneous contractions and expansions that 'perfectly' offset each other, there won't be a change in the rank distributions. We note that if the offset is not perfect, (such as $\sigma_{4,4,2} = 0.49\sigma_{4,4,1}$ instead) a change in the rank distribution will appear. This is fairly intuitive; since depth functions focus on the *magnitude of outlyingness* and not necessarily the *direction of outlyingness*. We can summarize the results as follows:

- Expansions/contractions in the submatrices produce a change in the rank distributions.
- The smaller the submatrix, the smaller the change in rank distribution.
- Certain combinations of expansions/contractions also admit changes in the rank distribution, provided the expansion(s) does not offset the contraction(s).
- Sign changes cannot be detected.

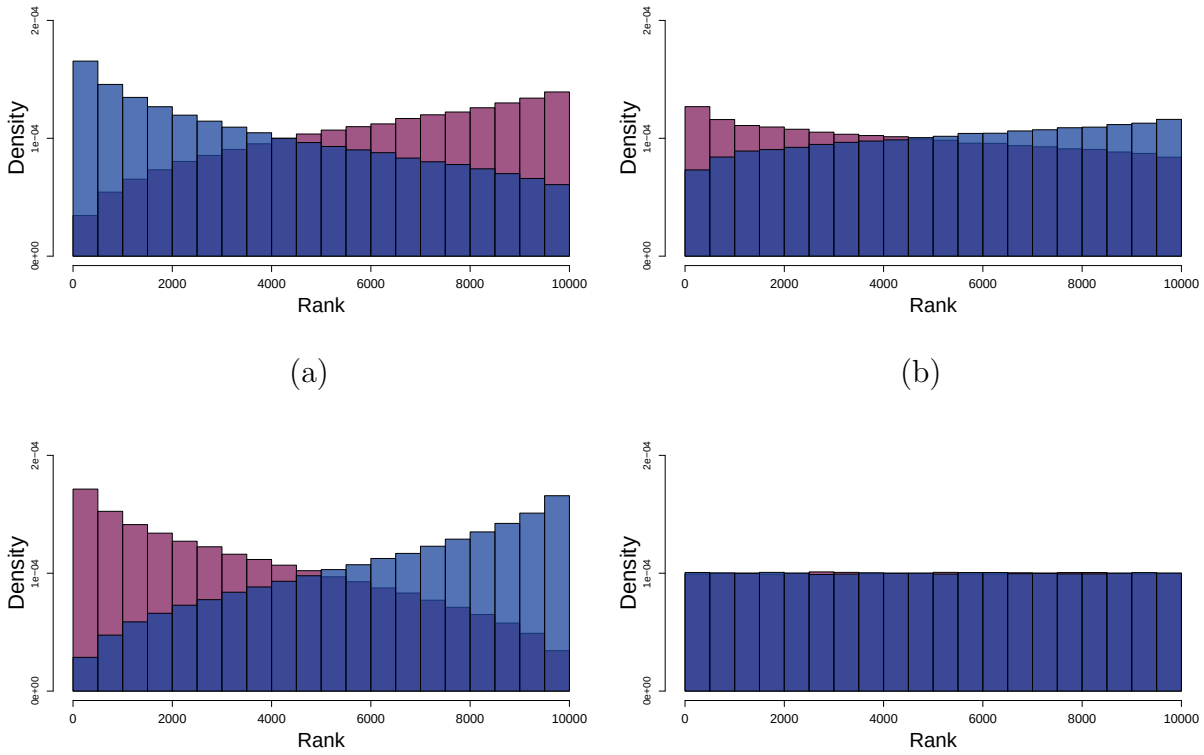


Figure A.1: Normalised histograms of the depth ranks of sample 1 (red) and sample 2 (blue) under a (a) submatrix on the diagonal change, (b) submatrix on the off diagonal change, (c) mixed change and (d) offsetting expansion and contraction.

In conclusion, we aim to detect changes that can be expressed as contractions or expansions of submatrices. Additionally, we remark that many combinations of contractions and expansions can be detected, with the caveat that offsetting combinations of such changes make the change more difficult to detect, or in a special case, impossible.

A.3 L^2-root depth

The following theorem lists some properties of the L^2-root depth.

Theorem 19. *Let $X \sim F$, where F is a measure over $\mathscr{L}^2([0,1], \mathscr{B}, \mu)$, F_n be the empirical*

measure corresponding to a random sample of size n from F, $a, b \in \mathfrak{F}$ and $c, c' \in \mathbb{R}^+$, then LTR *depth satisfies the following properties*

1. *Sample ranks based on* LTR *depth are invariant when the sample is transformed by a linear function h, such that $h(x) = a\,x + b$.*
2. *If $X \overset{d}{=} -X$, then $\sup_x \mathrm{LTR}(x; F) = \mathrm{LTR}(\mathbf{0}; F)$.*
3. *If $X \overset{d}{=} -X$, then $\mathrm{LTR}(c\,x; F)$, is decreasing in c.*
4. $\lim_{c\to\infty} \mathrm{LTR}(c\,x; F) = 0$.
5. *Suppose that $\mathrm{E}\left[\|X\|^2\right] < \infty$, then $\sup_x |\,\mathrm{LTR}(x; F_n) - \mathrm{LTR}(x; F)| = o(1)$ a.s. .*

The proof of this theorem is given in Section A.5. We remark that $\mathrm{LTR}(x; F)$ is not invariant under linear transformations as given by h in property 1. of Theorem 19. It is easy to see that if

$$\mathrm{LTR}(x; F) = \frac{1}{1 + c'} \qquad \text{then} \qquad \mathrm{LTR}(ax; aF) = \frac{1}{1 + \|a\|\, c'},$$

where $aX \sim aF$, with $X \sim F$. This fact implies that hypothesis tests based on LTR depth values themselves won't be invariant under linear transformations. However, Theorem 19 shows that hypothesis tests based on ranks of these depth values invariant under linear transformations. Further, a median based on this depth would be equivariant under linear transformations as given by h in property 1. .

In this setting, we see that ranking the observations based on LTR-depth is equivalent to ranking the observations based on their norms. Under the assumption of zero mean, we have that

$$\mathrm{E}_{F_*}\left[\|X_{ji} - X\|^2\right] = \mathrm{E}_{F_*}\left[\|X_{ji}\|^2\right] + \mathrm{E}_{F_*}\left[\|X\|^2\right] = \|X_{ji}\|^2 + \sum_{j=1}^{J} \vartheta_j \mathscr{K}_j + o(1),$$

from which it is easily seen that the ranks are equivalent to those based on $\mathrm{E}_{F_*}\left[\|X_{ji}\|^2\right]$. We emphasize that this relationship relies on the assumption of zero mean, and the data must be centred. In this context, ranks generated from this depth function do not need to be estimated; we can compute ranks based on $\mathrm{D}(\cdot; F_*)$ directly.

A.4 Additional information surrounding the simulation study in Chapter 3

This section contains some additional simulation results. The finite dimensional models are as follows: Specifically, we ran six scenarios, which, letting λ_{jk} be the eigenvalues of the covariance operator of group j, are

1. Reversed short linear decay: $\lambda_{1k} = k,\ \lambda_{2k} = 3 - k + 1,\ k < 4,\ \lambda_{jk} = 0,\ k \geq 4.$
2. Reversed long linear decay: $\lambda_{1k} = k,\ \lambda_{2k} = 11 - k + 1,\ k < 12,\ \lambda_{jk} = 0,\ k \geq 12.$
3. Reversed long exponential decay: $\lambda_{1k} = 2^k,\ \lambda_{2k} = 2^{11-k+1},\ k < 12,\ \lambda_{jk} = 0,\ k \geq 12.$
4. Scaled short linear decay: $\lambda_{1k} = k,\ \lambda_{2k} = 1.5\lambda_{1k},\ k < 4,\ \lambda_{jk} = 0,\ k \geq 4.$
5. Scaled long linear decay: $\lambda_{1k} = k,\ \lambda_{2k} = 1.5\lambda_{1k}\ k < 12,\ \lambda_{jk} = 0,\ k \geq 12.$
6. Scaled long exponential decay: $\lambda_{1k} = 2^k,\ \lambda_{2k} = 1.5\lambda_{1k},\ k < 12,\ \lambda_{jk} = 0,\ k \geq 12.$

We also have extra tables for imbalanced group sizes:

A.4.1 On the number of directions for the random projection depth

We also studied the effect of the number of directions in the random projection on the test. Figure A.2 shows that the test is stable under the number of directions.

A.4.2 If the curves have missing values at random time points

Figure A.3 shows that the results are essentially the same if the curves have missing points.

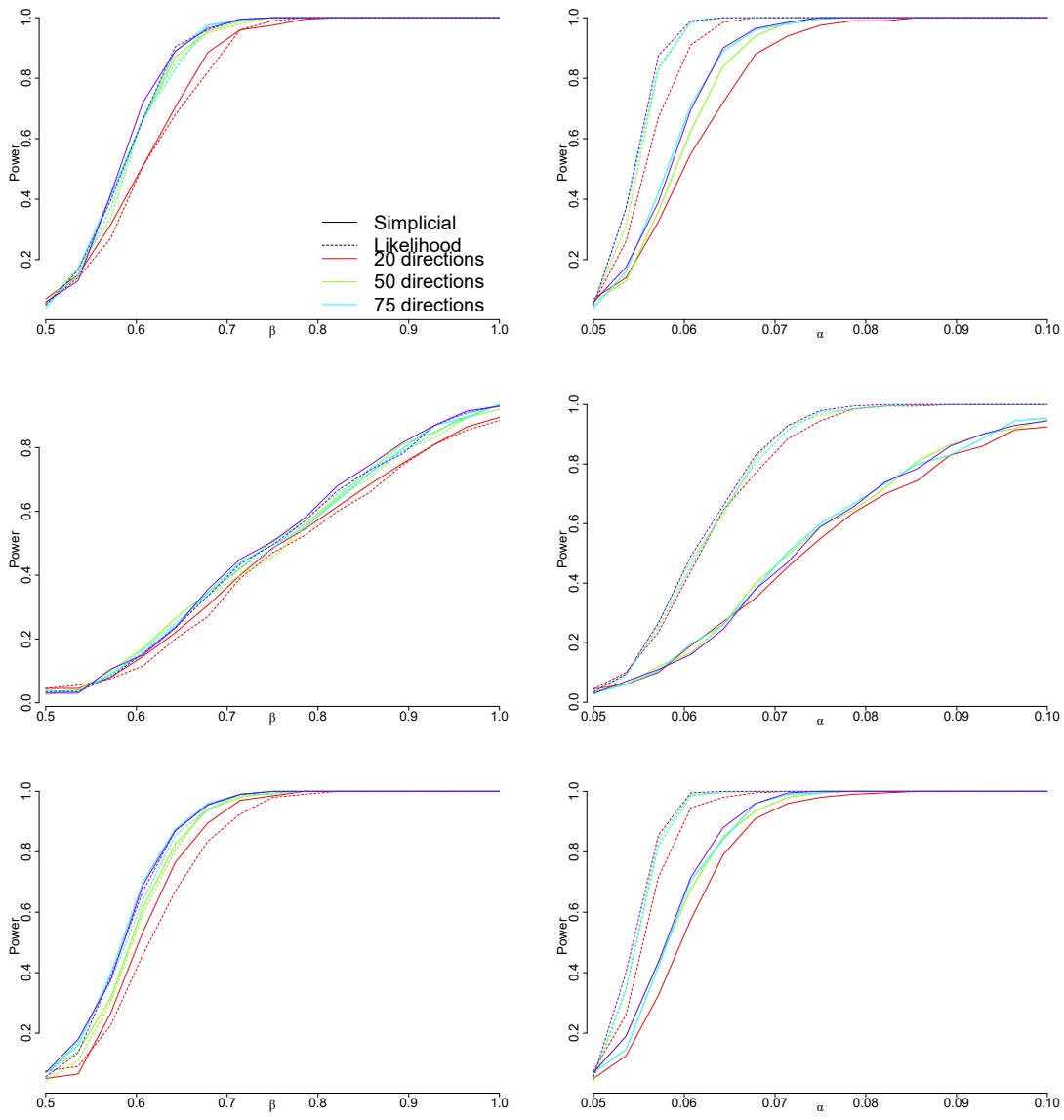


Figure A.2: Power of the FKWC test paired with the random projection depth to detect scale and shape differences for $J = 2$ with $n_1 = n_2 = 50$, using different amounts of sampled directions. In the top row the data were Gaussian, in the middle row the data were student t data with 3 degrees of freedom, and in the final row the data were skewed Gaussian data.

		Gaussian			Student t			Skewed Gaussian		
	n_1/n :	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
FKWC	MFHD	1.00	1.00	1.00	0.85	0.92	0.94	1.00	1.00	1.00
	RP	1.00	1.00	1.00	0.91	0.94	0.97	1.00	1.00	1.00
	MBD	1.00	1.00	1.00	0.88	0.94	0.97	1.00	1.00	1.00
	LTR	1.00	1.00	1.00	0.96	0.97	0.98	1.00	1.00	1.00
	RP†	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Competing	Boen	0.95	0.99	0.99	0.11	0.08	0.04	0.95	0.98	0.99
	L2br	0.81	0.94	0.97	1.00	1.00	1.00	0.78	0.92	0.99
	L2rp	0.78	0.92	0.97	0.04	0.04	0.08	0.73	0.92	0.98
	Tmax	0.34	0.62	0.70	0.06	0.07	0.06	0.35	0.63	0.78
	GPFrp	0.71	0.92	0.97	0.05	0.07	0.13	0.70	0.90	0.96
	Fmax	0.60	0.82	0.90	0.07	0.18	0.32	0.57	0.82	0.92

Table A.1: Empirical power of the different tests for detecting a shape difference with $\alpha_1 = 0.05$ and $\alpha_2 = 0.071$. Here $J = 2$, $n = 500$ and the group sample sizes were unequal. Notice that when the sample sizes differ greatly, the competing tests do not perform as well.

A.5 Proofs from Chapter 3

Proof. Proof of (3.4) Let

$$Z_N = \sup_{x \in \mathfrak{F}} |\,\mathrm{RP}_{M_N}(x; F_N) - \mathrm{RP}_\infty(x; F)|.$$

Observe that

$$\begin{aligned}
\mathrm{E}\,[Z_n] &= \mathrm{E}\left[\left| M_n^{-1} \sum_{m=1}^{M_n} \mathrm{D}\langle x, u_m\rangle; F_{n,u_m} - \int_S \mathrm{D}\langle x, u\rangle; F_u d\nu(u)\right|\right] \\
&\leq \mathrm{E}\left[M_n^{-1} \sum_{m=1}^{M_n} |\mathrm{D}\langle x, u_m\rangle; F_{n,u_m} - \mathrm{D}\langle x, u_m\rangle; F_{u_m}|\right] + O(n^{-1/2}) \\
&\leq \mathrm{E}\left[4M_n^{-1} \sum_{m=1}^{M_n} \mathrm{E}\left[\sup_{z\in\mathbb{R}} |F_{n,u_m}(z) - F_{u_m}(z)| \middle| u_1, \ldots u_{M_n}\right]\right] + O(n^{-1/2}) \\
&= O(n^{-1/2}),
\end{aligned}$$

	$n_1/n = 0.2$										
Scen.	MFHD	RP	MBD	LTR	RP†	Boen	L2br	L2rp	Tmax	GPFrp	Fmax
1	0.90	0.85	0.72	0.68	0.92	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	0.98	0.98	0.98	0.98	0.97	0.97	0.82	0.78	0.68	0.74	0.41
5	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.72	0.51	0.41	0.01
6	0.97	0.98	0.97	0.96	0.97	0.97	0.80	0.74	0.68	0.74	0.31
	$n_1/n = 0.3$										
Scen.	MFHD	RP	MBD	LTR	RP†	Boen	L2br	L2rp	Tmax	GPFrp	Fmax
1	0.93	0.88	0.78	0.70	0.97	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.97	0.92	0.97	0.91
5	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.97	0.73
6	0.99	1.00	0.99	0.99	0.99	0.99	0.97	0.94	0.92	0.96	0.88
	$n_1/n = 0.4$										
Scen.	MFHD	RP	MBD	LTR	RP†	Boen	L2br	L2rp	Tmax	GPFrp	Fmax
1	0.94	0.89	0.80	0.76	0.97	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.98
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
6	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.98	0.96	0.99	0.98

Table A.2: Empirical power of the different tests for $n = 200$ when the group sample sizes were unequal, under the finite dimensional models.

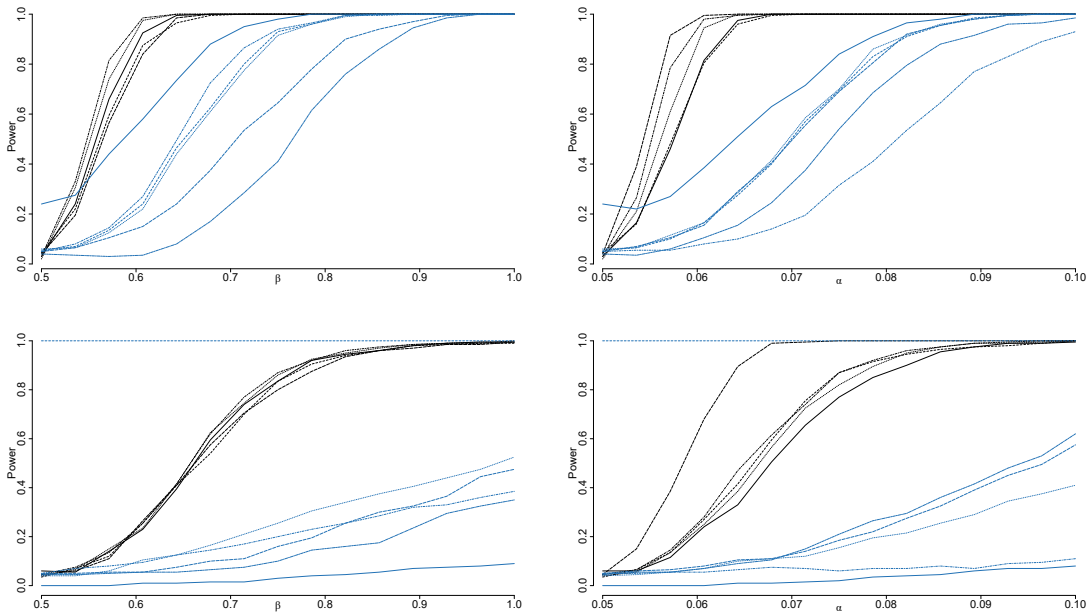


Figure A.3: Power of the two sample versions of the tests when $n_1 = n_2 = 100$ when 20% of the curves were uniformly, randomly missing. The missing values were first interpolated with splines. The top row is the infinite dimensional Gaussian process model and the bottom row is the infinite dimensional student t model. Note that the legend follows that of Figure 3.3.

where the first inequality is from the triangle inequality and Hoeffding's inequality, and the last equality results from the Dvoretzky–Kiefer–Wolfowitz inequality. □

Proof of Theorem 19. Let $aF + b$ be the measure associated with $aX + b$. We have that

$$\mathrm{LTR}(ax + b; aF + b) = \left(1 + \mathrm{E}_F\left[\|ax + b - aX + b\|^2\right]^{1/2}\right)^{-1} = \left(1 + \|a\|\,\mathrm{E}_F\left[\|x - X\|^2\right]^{1/2}\right)^{-1}.$$

The function $(1 + c'x)^{-1}$ is monotonic for any $c' > 0$. Therefore, for any $x, y \in \mathfrak{F}$ such that $\mathrm{LTR}(x; F) < \mathrm{LTR}(y; F)$, it holds that $\mathrm{LTR}(ax + b; aF + b) < \mathrm{LTR}(ay + b; aF + b)$.

This gives the first property. For the second property, observe that

$$\begin{aligned}
\mathrm{LTR}(x;F) &= \left(1+\mathrm{E}_F\left[\|x-X\|^2\right]^{1/2}\right)^{-1} \\
&= \left(1+2^{-1/2}\mathrm{E}_F\left[\|x-X\|^2+\|x+X\|^2\right]^{1/2}\right)^{-1} \\
&= \left(1+2^{-1/2}\mathrm{E}_F\left[2\|x\|^2+2\|X\|^2\right]^{1/2}\right)^{-1} \\
&= \left(1+2^{-1/2}2^{1/2}\|x\|^2+2^{-1/2}2^{1/2}\mathrm{E}\left[\|X\|^2\right]^{1/2}\right)^{-1} \\
&= \left(1+\|x\|+c'\right)^{-1},
\end{aligned}$$

which is maximized at $x = \mathbf{0}$. For the third and fourth properties, we have in similar fashion

$$\mathrm{LTR}(cx;F) = \left(1+\mathrm{E}_F\left[\|cx-X\|^2\right]^{1/2}\right)^{-1} = \left(1+c\|x\|+c'\right)^{-1},$$

which is decreasing toward 0 as c increases. Lastly, if $X_1,\ldots,X_n$ is a random sample from F, then it holds that

$$\frac{1}{n}\sum_{i=1}^{n}\|x-X_i\|^2 = \|x\|^2+\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2-2\langle x,\frac{1}{n}\sum_{i=1}^{n}X_i\rangle \coloneqq \|x\|^2+\overline{Y}_{x,n}.$$

We have that

$$\begin{aligned}
|\,\mathrm{LTR}(x;F_n)-\mathrm{LTR}(x;F)| &= \left|\left(1+(\|x\|^2+\overline{Y}_{x,n})^{1/2}\right)^{-1}-\left(1+(\|x\|^2+\|\mathscr{K}\|_{TR})^{1/2}\right)^{-1}\right| \\
&= \left|\frac{(\|x\|^2+\|\mathscr{K}\|_{TR})^{1/2}-(\|x\|^2+\overline{Y}_{x,n})^{1/2}}{\left(1+(\|x\|^2+\overline{Y}_{x,n})^{1/2}\right)\left(1+(\|x\|^2+\|\mathscr{K}\|_{TR})^{1/2}\right)}\right| \\
&\leq \left|\frac{|\,\|\mathscr{K}\|_{TR}-\overline{Y}_{x,n}|^{1/2}}{\left(1+(\|x\|^2+\overline{Y}_{x,n})^{1/2}\right)\left(1+(\|x\|^2+\|\mathscr{K}\|_{TR})^{1/2}\right)}\right|,
\end{aligned}$$

where the third line comes from the fact that $\sqrt{x}-\sqrt{y}\leq\sqrt{|x-y|}$. Now, suppose that

$\|x\| < c' n^{1/2} (\log n)^{-1}$.

$$
\begin{aligned}
|\operatorname{LTR}(x; F_n) - \operatorname{LTR}(x; F)| &\leq |\, \|\mathscr{K}\|_{TR} - \overline{Y}_{x,n}|^{1/2} \\
&\leq \left| \|\mathscr{K}\|_{TR} - \frac{1}{n} \sum_{i=1}^{n} \|X_i\|^2 - c' n^{-1/2} (\log n)^{-1} \sum_{i=1}^{n} \int X_i dt \right|^{1/2} \\
&= o(1) \ a.s. \ ,
\end{aligned}
$$

where the last line is from the strong law of large numbers and the law of the iterated logarithm. Note that $\int X_i dt$ has finite variance for all $i \in \{1, \ldots, n\}$ and that the second inequality does not depend on x. Suppose now that $\|x\| \geq c' n^{1/2} (\log n)^{-1}$. Then, it is easy to see that $|\operatorname{LTR}(x; F_n) - \operatorname{LTR}(x; F)| \to 0$ by the vanishing at infinity property. □

Proof of Theorem 3. Observe that $\widehat{R}_{ji}$ are identically distributed under the null hypothesis. This implies that the rank vector has the uniform distribution with probability of each outcome being $1/n!$. This is the same setup as in (Kruskal, 1952) and so it is immediate that $\mathcal{W}_n \xrightarrow{\mathrm{d}} \chi^2_{J-1}$. Similarly, it follows directly from Theorem 2 of (Chenouri et al., 2011) that $\mathcal{M}_{n,r} \xrightarrow{\mathrm{d}} \chi^2_{J-1}$. □

Proof of Equation (3.6). Let

$$
\widetilde{\sigma}_n^2 = \frac{n(n+1)}{12}.
$$

We can rewrite $\widehat{\mathcal{W}}_n$ as follows

$$
\widehat{\mathcal{W}}_n = \frac{1}{\widetilde{\sigma}_n^2} \sum_{j=1}^{J} n_j \overline{\widehat{R}}_j^2 - 3(n+1).
$$

Now, define

$$
\mathcal{W}_n \coloneqq \frac{1}{\widetilde{\sigma}_n^2} \sum_{j=1}^{J} n_j \overline{R}_j^2 - 3(n+1) \qquad \text{with} \qquad \overline{R}_j \coloneqq \frac{1}{n_j} \sum_{i=1}^{n_j} R_{ji}.
$$

Under the alternative hypothesis, Assumption 7 and (3.5), $\mathrm{D}(X_{ji}; F_*)$ are equivalent to the univariate random variables studied by Kruskal (1952). For all $\delta > 0$, it then holds that

$$
\mathrm{P}(\mathcal{W}_n > \delta) \to 1, \quad \text{as } n \to \infty.
$$

We just need $|\widehat{\mathcal{W}}_n - \mathcal{W}_n| = O_p(1)$, which will complete the proof. However, this follows directly from Lemma 4 in Chapter 2. □

Proof of Theorem 4. First, let $u \in S$, where $S = \{u\colon \|u\| = 1,\ u \in \mathfrak{F}\}$ and let $Y_{u,j} = \langle X_{j1}, u\rangle$. Observe that $\mathrm{E}[Y_{u,j}] = 0$ and that

$$\sigma^2_{j,u} \coloneqq \mathrm{E}\left[Y^2_{u,j}\right] = \mathrm{E}\left[\int_{[0,1]}\int_{[0,1]} X_{j1}(t)u(t)\cdot X_{j1}(s)u(s)\; dsdt\right] = \langle \mathscr{K}_j u, u\rangle,$$

where we can take the expectation inside due to Lebesgue's dominated convergence theorem. Namely, $\langle X_{j1}, u\rangle^2 \leq \|X_{j1}\|^2$ which has finite expectation. One should also recall that we take

$$\mathrm{D}\left(\langle x, u\rangle; F_u\right) = F_u(\langle x, u\rangle)(1 - F_u(\langle x, u\rangle))$$

for the univariate depth. For the remainder of the proof, we will suppress the F_* in $\mathrm{RP}(x; F_*)$. In view of (3.6), it is only necessary to verify that

$$\Pr\left(\mathrm{RP}(X_{11}) > \mathrm{RP}(X_{21})\right) \neq \frac{1}{2}. \tag{A.10}$$

Which, under Assumption 10, this is equivalent to showing

$$\mathrm{E}\left[\mathrm{RP}(X_{11}) - \mathrm{RP}(X_{21})\right] \neq 0.$$

We can write

$$\begin{aligned}\mathrm{E}\left[\mathrm{RP}(X_{11}) - \mathrm{RP}(X_{21})\right] &= \mathrm{E}\left[\int_S \mathrm{D}\left(Y_{u,1}; F_{u,*}\right) d\nu(u) - \int_S \mathrm{D}\left(Y_{u,2}; F_{u,*}\right) d\nu(u)\right] \\ &= \mathrm{E}\left[\int_S F_{u,*}(Y_{u,1})(1 - F_{u,*}(Y_{u,1})) - F_{u,*}(Y_{u,2})(1 - F_{u,*}(Y_{u,2}))\; d\nu\right].\end{aligned}$$

Clearly, since $0 < F_{u,*}(Y_{u,1}) < 1$, we have that

$$F_{u,*}(Y_{u,1})(1 - F_{u,*}(Y_{u,1})) - F_{u,*}(Y_{u,2})(1 - F_{u,*}(Y_{u,2})) \leq 1/2.$$

Using Lebesgue's dominated convergence theorem,

$$\mathrm{E}\left[\mathrm{RP}(X_{11}) - \mathrm{RP}(X_{21})\right] = \int_S \mathrm{E}\left[F_{u,*}(Y_{u,1})(1 - F_{u,*}(Y_{u,1})) - F_{u,*}(Y_{u,2})(1 - F_{u,*}(Y_{u,2}))\right] d\nu(u).$$

Using the fact that $F_{*,u}$ is thrice differentiable for all u, we can write

$$\begin{aligned}
\mathrm{E}\left[F_{u,*}(Y_{u,j})\right] &= F_{u,*}(0) + \frac{1}{2} f^{(1)}_{u,*}(0)\sigma^2_{j,u} + \mathcal{R}_{u,j,1} \\
\mathrm{E}\left[F^2_{u,*}(Y_{u,j})\right] &= F^2_{u,*}(0) + (F_{u,*}(0) f^{(1)}_{u,*}(0) + f^2_{u,*}(0))\sigma^2_{j,u} + \mathcal{R}_{u,j,2}
\end{aligned}$$

with

$$\begin{aligned}
\mathcal{R}_{u,j,1} &\coloneqq \mathrm{E}\left[\frac{1}{6}\int_0^{Y_{u,j}} f^{(2)}_{u,*}(t)(Y_{u,j} - t)^3 dt\right] \\
\mathcal{R}_{u,j,2} &\coloneqq \mathrm{E}\left[\frac{1}{3}\int_0^{Y_{u,j}} (3 f_{u,*}(t) f^{(1)}_{u,*}(t) + F_{u,*}(t) f^{(2)}_{u,*}(t))(Y_{u,j} - t)^3 dt\right].
\end{aligned}$$

Note that we expect $\mathcal{R}^i_{u,j}$ to be small from the fact that the mean of $Y_{u,j}$ is 0. It follows that

$$\begin{aligned}
\mathrm{E}\left[\mathrm{D}\left(Y_{u,j}; F_{u,*}\right)\right] &= F_{u,*}(0) + \frac{1}{2} f^{(1)}_{u,*}(0)\sigma^2_{j,u} - F^2_{u,*}(0) - \sigma^2_{j,u}(F_{u,*}(0) f^{(1)}_{u,*}(0) - f^2_{u,*}(0)) + \mathcal{R}_{u,j,3} \\
&= \mathcal{H}(F_{*,u})\sigma^2_{j,u} + F_{u,*}(0) - F^2_{u,*}(0) + \mathcal{R}_{u,j,3},
\end{aligned}$$

where

$$\mathcal{H}(F) \coloneqq \frac{1}{2} f^{(1)}(0) - (F(0) f^{(1)}(0) - f^2(0)) \qquad \text{and} \qquad \mathcal{R}_{u,j,3} = \mathcal{R}_{u,j,1} - \mathcal{R}_{u,j,2}.$$

We can now write

$$\mathrm{E}\left[\mathrm{D}\left(Y_{u,1}; F_{u,*}\right) - \mathrm{D}\left(Y_{u,2}; F_{u,*}\right)\right] = \mathcal{H}(F_{*,u})(\sigma^2_{1,u} - \sigma^2_{2,u}) + \mathcal{R}_{u,1,3} - \mathcal{R}_{u,2,3}.$$

To conclude, under univariate simplicial depth it holds that

$$\begin{aligned}
\mathrm{E}\left[\mathrm{RP}(X_{11}) - \mathrm{RP}(X_{21})\right] &= \int_S \mathcal{H}(F_{*,u})(\sigma^2_{1,u} - \sigma^2_{2,u}) + \mathcal{R}_{u,1,3} - \mathcal{R}_{u,2,3}\; d\nu(u) \\
&= \int_S \mathcal{H}(F_{*,u})\langle \mathscr{K}_1 u - \mathscr{K}_2 u, u\rangle + \mathcal{R}_{u,1,3} - \mathcal{R}_{u,2,3}\; d\nu(u) \\
&= \int_S \mathcal{H}(F_{*,u})\langle \mathscr{K}_1 u - \mathscr{K}_2 u, u\rangle d\nu(u) + \mathcal{R}_1 - \mathcal{R}_2,
\end{aligned}$$

where $\mathcal{R}_j < \infty$ by the fact that the integrand is bounded in u. □

Proof of Theorem 5. In view of (3.6), it is only necessary to show that under the alternative, (3.5) holds. However, by Assumption 10, this is equivalent to

$$\mathrm{E}\left[\|X_{11}\|^2 - \|X_{21}\|^2\right],$$

from which it is well known that $\mathrm{E}\left[\|X_{j1}\|^2\right] = \|\mathscr{K}_j\|_{TR}$ if $\|\mathscr{K}_j\|_{TR} < \infty$ and so

$$\mathrm{E}\left[\|X_{11}\|^2 - \|X_{21}\|^2\right] = \|\mathscr{K}_1\|_{TR} - \|\mathscr{K}_2\|_{TR} \neq 0.$$

We know that by assumption $\mathscr{K}_j$ are trace class. Indeed, X_{ji} are mean square continuous, which means they have a continuous kernel K. This implies immediately that $\|\mathscr{K}_j\|_{TR} < \infty$. □

Proof of Theorem 6. Now, Y_{ji} are univariate observations from a scale family, meaning that $Z_{ji} \coloneqq (1 + \delta_j/\sqrt{n})^{-1} Y_{ji}$ with $Z_{ji} \sim G$. Now, let $\tau \coloneqq \lim_{N\to\infty} \tau_n$. It follows from (Fan et al., 2011) that the test statistic $\mathcal{W}_n \to \chi^2_{J-1}(\tau)$. We have that

$$\tau_n = \frac{12}{n(n+1)} \sum_{j=1}^{J} n_j \left\{ n \sum_{k\neq j} (\vartheta_k + o(1)) \left(\Pr(Y_{k1} \leq Y_{j1}) - 1/2\right) \right\}^2.$$

We have that

$$\Pr(Y_{k1} \leq Y_{j1}) = \Pr\left(Z_{k1} \leq Z_{j1} \left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]\right) = \int_{\mathbb{R}} \Pr\left(Z_{k1} \leq z \left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]\right) g(z) dz. \tag{A.11}$$

Now, note that $z\left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]$ is in a neighborhood of z we can write the Taylor expansion of G about z at the point $z\left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]$ as

$$\Pr\left(Z_{k1} \leq z \left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]\right) = G(z) + z\left[1 - \left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]\right] g(z) + O(n^{-1}).$$

Substituting into (A.11), we have that

$$\begin{aligned}\int_{\mathbb{R}} \Pr\left(Z_{k1} \leq z\left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]\right) g(z)dz &= \int_{\mathbb{R}} \left[G(z) + z\left[1 - \left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]\right] g(z) + O(n^{-1})\right] dG \\ &= 1/2 + \left[1 - \left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]\right] \int_{\mathbb{R}} z g(z)^2 dz + O(n^{-1}) \\ &= 1/2 + \left[1 - \left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]\right] \Delta_G + O(n^{-1}),\end{aligned}$$

where

$$\Delta_G = \int_{\mathbb{R}} z g(z)^2 dz.$$

Now, substituting the above identity into (3.7), gives

$$\tau_n = \frac{12}{n(n+1)} \sum_{j=1}^{J} n_j \left[n \sum_{k \neq j} (\vartheta_k + o(1)) \left(\left[1 - \left[\frac{\sqrt{n}+\delta_j}{\sqrt{n}+\delta_k}\right]\right] \Delta_G + O(n^{-1})\right)\right]^2,$$

which then immediately implies that

$$\lim_{n\to\infty} \tau_n = 12\Delta_G^2 \sum_{j=1}^{J} \vartheta_j \left(\delta_j - \overline{\delta}\right)^2,$$

which completes the proof. □

A.6 Proofs from Chapter 4

Proof of Theorem 7. The first equation follows from (Chenouri et al., 2020b). Therefore, we must only prove that

$$\mathcal{W}_n(\widehat{k}_1, \widehat{k}_2) \xrightarrow{\mathrm{d}} \sup_{t_1, t_2 \in (0,1)} \frac{(B(t_2) - B(t_1))^2}{(t_2 - t_1)(1 - t_2 + t_1)}.$$

Directly from (Chenouri et al., 2020b), see also (Billingsley, 1968), we have that

$$\widehat{Z}_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor tn \rfloor} \frac{\widehat{R}_i - \mu_n}{\sigma_n} \xrightarrow{\mathrm{d}} B(t),$$

for $t \in [0,1]$. Let $q_n = \sigma_n^2/\widetilde{\sigma}_n^2$ and note that $\lim_{n\to\infty} q_n \to 1$. We write $\mathcal{W}_n$ as a function of partial sums which are in a similar form to that $\widehat{Z}_n(t)$. Consider the first term in (4.3). For $0 < t_1 < t_2 < 1$, we have that

$$\begin{aligned}
\mathcal{W}_n(\lfloor t_1 n \rfloor, \lfloor t_2 n \rfloor) &= \frac{1}{\widetilde{\sigma}_n} \sum_{\substack{1 \le j < \lfloor t_1 n \rfloor \\ \lfloor t_2 n \rfloor \le j \le n}}^{n} \frac{\widehat{R}_j - \mu_n}{\sqrt{n - \lfloor t_2 n \rfloor + \lfloor t_1 n \rfloor}} \\
&= \left(\frac{q_n}{1 - \lfloor t_2 n \rfloor / n + \lfloor t_1 n \rfloor / n} \right)^{1/2} \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{\lfloor t_1 n \rfloor - 1} \frac{\widehat{R}_j - \mu_n}{\sigma_n^2} + \sum_{j=k_2}^{n} \frac{\widehat{R}_j - \mu_n}{\sigma_n^2} \right) \\
&= \left(\frac{q_n}{1 - \lfloor t_2 n \rfloor / n + \lfloor t_1 n \rfloor / n} \right)^{1/2} \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{\lfloor t_1 n \rfloor - 1} \frac{\widehat{R}_j - \mu_n}{\sigma_n^2} - \sum_{j=1}^{k_2 - 1} \frac{\widehat{R}_j - \mu_n}{\sigma_n^2} \right) \\
&= - \left(\frac{q_n}{1 - \lfloor t_2 n \rfloor / n + \lfloor t_1 n \rfloor / n} \right)^{1/2} \frac{1}{\sqrt{n}} \left(\sum_{j=k_1}^{\lfloor t_2 n \rfloor - 1} \frac{\widehat{R}_j - \mu_n}{\sigma_n^2} \right).
\end{aligned}$$

For the second term in (4.3), we have that

$$\frac{1}{\widetilde{\sigma}_n} \sum_{j=\lfloor t_1 n \rfloor}^{k_2 - 1} \frac{\widehat{R}_j - \mu_n}{\sqrt{\lfloor t_2 n \rfloor - \lfloor t_1 n \rfloor}} = \left(\frac{q_n}{\lfloor t_2 n \rfloor / n - \lfloor t_1 n \rfloor / n} \right)^{1/2} \frac{1}{\sqrt{n}} \left(\sum_{j=k_1}^{\lfloor t_2 n \rfloor - 1} \frac{\widehat{R}_j - \mu_n}{\sigma_n^2} \right).$$

So, it follows that

$$\mathcal{W}_n(\lfloor t_1 n \rfloor, \lfloor t_2 n \rfloor) = \frac{q_n}{(\lfloor t_2 n \rfloor / n - \lfloor t_1 n \rfloor / n)(1 - \lfloor t_2 n \rfloor / n + \lfloor t_1 n \rfloor / n)} \left(\frac{1}{\sqrt{n}} \sum_{j=k_1}^{\lfloor t_2 n \rfloor - 1} \frac{\widehat{R}_j - \mu_n}{\sigma_n^2} \right)^2.$$

We can then write

$$\mathcal{W}_n(t_1, t_2) \coloneqq g_n(t_1, t_2) \mathcal{W}'_n(t_1, t_2),$$

where

$$g_n(t_1,t_2) = \frac{1}{(t_2-t_1)(1-t_2+t_1)} + o(1) \text{ and } \mathcal{W}'_n(t_1,t_2) = \left(\frac{1}{\sqrt{n}}\left(\sum_{j=\lfloor nt_1\rfloor}^{\lfloor nt_2\rfloor-1} \frac{\widehat{R}_j - \mu_n}{\sigma_n^2}\right)\right)^2.$$

Recall that $\widehat{Z}_n(t) \xrightarrow{\text{d}} B(t)$ and observe that $\mathcal{W}'_n(t_1,t_2) = (\widehat{Z}_n(t_1) - \widehat{Z}_n(t_2))^2$. The quantity $\mathcal{W}'_n(t_1,t_2)$ is a continuous functional of $\widehat{Z}_n(t)$, and so continuous mapping theorem gives that

$$\mathcal{W}'_n(t_1,t_2) = \left(\frac{1}{\sqrt{n}}\left(\sum_{j=\lfloor nt_1\rfloor}^{\lfloor nt_2\rfloor-1} \frac{\widehat{R}_j - \mu_n}{\sigma_n^2}\right)\right)^2 \xrightarrow{\text{d}} (B(t_2) - B(t_1))^2.$$

and

$$\sup_{t_1<t_2} \mathcal{W}(\lfloor nt_1\rfloor, \lfloor nt_2\rfloor) \xrightarrow{\text{d}} \sup_{t_1<t_2} \left(\frac{1}{(t_2-t_1)(1-t_2+t_1)}\right)(B(t_2) - B(t_1))^2.$$

□

Lemma 8. *Suppose that there are n, independent univariate data points, with a change-point at observation k_1. Let $R_1, \ldots, R_{k_1}$ denote the first k_1 combined sample ranks. For $1 \le m < k_1$ and $\delta > 0$, consider $S_m = \sum_{i=1}^m R_i$. Then,*

$$\Pr(|S_m - \mathrm{E}[S_m]| > \delta) \le 4e^{-2\min(m,k_1-m,n-k_1)\left(\frac{\delta}{2\max(k_1-m,n-k_1)m}\right)^2}.$$

Proof. Let $P(X_1 > X_n) = p$. It holds that

$$\begin{aligned}\sum_{j=1}^m R_j &= \frac{m(m+1)}{2} + \sum_{j=1}^m \sum_{i=m+1}^{k_1} \mathbb{1}(X_j \le X_i) + \sum_{j=1}^m \sum_{i=k_1+1}^{n} \mathbb{1}(X_j \le X_i) \\ &:= \frac{m(m+1)}{2} + (k_1-m)mU_{1,n} + m(n-k_1)U_{2,n},\end{aligned}$$

where $U_{1,n}$, $U_{2,n}$ are two sample U-statistics. Note that $\mathrm{E}[U_{1,n}] = 1/2$ and $\mathrm{E}[U_{1,n}] = p$. We make use of (Hoeffding, 1963) which gives, for a two sample U-statistic U related to the sample sizes n_1 and n_2, it holds that

$$\Pr(|U - \mathrm{E}[U]| > \delta) \le 2e^{-2\min(n_1,n_2)t^2}. \tag{A.12}$$

Continuing,

$$\begin{aligned}\Pr(S_m - \mathrm{E}\left[S_m\right] > \delta) &= \Pr\left((k_1 - m)m(U_{1,n} - 1/2) + m(n-k_1)(U_{2,n} - p) > \delta\right) \\ &\leq \Pr\left(U_{1,n} - 1/2 > \frac{\delta}{2(k_1-m)m}\right) + \Pr\left(U_{2,n} - p > \frac{\delta}{2m(n-k_1)}\right) \\ &\leq e^{-2\min(m,k_1-m)\left(\frac{\delta}{2(k_1-m)m}\right)^2} + e^{-2\min(m,n-k_1)\left(\frac{\delta}{2m(n-k_1)}\right)^2} \\ &\leq 2e^{-2\min(m,k_1-m,n-k_1)\left(\frac{\delta}{2\max(k_1-m,n-k_1)m}\right)^2}.\end{aligned}$$

Note that taking

$$\Pr(|S_m - \mathrm{E}\left[S_m\right]| > \delta) \leq \Pr(S_m - \mathrm{E}\left[S_m\right] > \delta) + \Pr(-S_m + \mathrm{E}\left[S_m\right] > \delta),$$

and $-S_m + \mathrm{E}\left[S_m\right]$ has the same U-statistic format as $S_m - \mathrm{E}\left[S_m\right]$. Thus, the previous analysis applies and we get

$$\Pr(|S_m - \mathrm{E}\left[S_m\right]| > \delta) \leq 4e^{-2\min(m,k_1-m,n-k_1)\left(\frac{\delta}{2\max(k_1-m,n-k_1)m}\right)^2}.$$

□

Lemma 9. *Suppose that there are n univariate data points, with a change-point at observation k_1. Let $R_1, \ldots, R_{k_1}$ denote the first k_1 combined sample ranks and let*

$$Z_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor tn \rfloor} \frac{R_i - \mu_n}{\sigma_n}.$$

It holds that

$$\Pr(Z_n(k_1/n) < a_n) \leq e^{-2n^2\min(\theta_1, 1-\theta_1)\left(p - 1/2 - \frac{a_n\sigma_n}{n^2\theta_1(1-\theta_1)}\right)^2},$$

for $a_n < \frac{k_1(n-k_1)(p-1/2)}{\sigma_n}$ and $k_1 = \lfloor n\theta_1 \rfloor$.

Proof. Note that

$$\sqrt{n}\mathrm{E}\left[Z_n(k_1/n)\right] = \frac{k_1(k_1+1)}{2\sigma_n} + \frac{k_1(n-k_1)p}{\sigma_n} - \frac{k_1(n+1)}{2\sigma_n} = k_1(n-k_1)\frac{p-1/2}{\sigma_n}.$$

We then have that

$$\sqrt{n}Z_n(k_1/n) = \frac{k_1(k_1+1)}{2\sigma_n} + \frac{k_1(n-k_1)}{\sigma_n}U' - \frac{k_1(n+1)}{2\sigma_n},$$

where U' is a two sample U-statistic. We then have that

$$\Pr(\mathrm{E}\left[Z_n(k_1/n)\right] - Z_n(k_1/n) > t) \leq e^{-2n^2 \min(\theta,1-\theta)\left(\frac{t\sigma_n}{k_1(n-k_1)}\right)^2},$$

for $t > 0$ (Hoeffding, 1963; Pitcan, 2017). Using this fact,

$$\begin{aligned}
\Pr(Z_n(k_1/n) < a_n) &= \Pr\left(Z_n(k_1/n) - \mathrm{E}\left[Z_n(k_1/n)\right] < a_n - k_1(n-k_1)\frac{p-1/2}{\sigma_n}\right) \\
&= \Pr\left(\mathrm{E}\left[Z_n(k_1/n)\right] - Z_n(k_1/n) > k_1(n-k_1)\frac{p-1/2}{\sigma_n} - a_n\right) \\
&\leq e^{-2n^2 \min(\theta,1-\theta)\left(p-1/2-\frac{a_n\sigma_n}{n^2\theta(1-\theta)}\right)^2}
\end{aligned}$$

where the second line uses the fact that

$$a_n - k_1(n-k_1)\frac{p-1/2}{\sigma_n} < 0.$$

□

Theorem 9. The outline of our proof follows that of the proofs in (Yu and Chen, 2017), but the details differ since we are using ranks. To keep the notation simple, we let $k_1 = k$ and $\widehat{k}_1 = \widehat{k}$ for the remainder of the proof. Let $\sigma_n = \sqrt{(n^2-1)/12}$ and $\mu_n = (n+1)/2$. The aim is to show that $\Pr(|\widehat{k} - k| > tn) < C_1 \exp(-nt^2 C_2)$. To start, note that

$$\Pr(|\widehat{k} - k| > tn) = \Pr(\widehat{k} - k > tn) + \Pr(k - \widehat{k} > tn). \tag{A.13}$$

We start with showing a bound on $\Pr(k - \widehat{k} > tn)$. Note that

$$\begin{aligned}
\Pr(k - \widehat{k} > tn) &\leq \Pr\left(\max_{c_0 n \leq i < k - tn} |Z_n(i/n)| > |Z_n(k/n)|\right) \\
&\leq \Pr\left(\max_{c_0 n \leq i < k - tn} |Z_n(i/n)| > Z_n(k/n)\right) \\
&\leq \Pr\left(\max_{c_0 n \leq i < k - tn} |Z_n(i/n)| - Z_n(k/n) > 0\right) \\
&\leq \Pr(A_1) + \Pr(A_2),
\end{aligned} \tag{A.14}$$

where

$$\Pr(A_1) = \Pr\left(\max_{c_0 n \leq i < k - tn} Z_n(i/n) - Z_n(k/n) > 0\right)$$
$$\Pr(A_2) = \Pr\left(\min_{c_0 n \leq i < k - tn} Z_n(i/n) + Z_n(k/n) < 0\right).$$

First consider event A_1:

$$\begin{aligned}
\Pr(A_1) &= \Pr\left(\max_{c_0 n \leq i < k - tn} Z_n(i/n) - Z_n(k/n) > 0\right) \\
&= \Pr\left(\max_{c_0 n \leq i < k - tn} -\sum_{j=i+1}^{k} \frac{R_j - \mu_n}{\sigma_n} > 0\right) \\
&= \Pr\left(\max_{tn \leq i < k - c_0 n} -Z_n(i/n) > 0\right),
\end{aligned}$$

where the last line follows from exchangeability of $R_1, \ldots, R_k$. Without loss of generality we can assume that

$$\max_{tn \leq i < k - c_0 n} -Z_n(i/n) = -Z_n(i^*/n),$$

occurs for some i^*. Then, in order for $\{\max_{tn \leq i < k - c_0 n} -Z_n(i/n) > 0\}$ to also occur, we must have

$$-Z_n(i^*/n) + \mathrm{E}\left[Z_n(i^*/n)\right] > \mathrm{E}\left[Z_n(i^*/n)\right].$$

We will make use of this fact shortly, but first it helps compute $\mathrm{E}\left[Z_n(i/n)\right]$. At this point, it is helpful to recall from the proof of Lemma 9 that $Z_n(i/n)$ can be written in terms of a generalised, two sample U-statistic. It follows that

$$\begin{aligned}
\sqrt{n}\mathrm{E}\left[Z_n(i/n)\right] &= \frac{i(i+1)}{2\sigma_n} + \frac{i(k-i)}{2\sigma_n} + pi(n-k) - i\frac{n+1}{2\sigma_n} \\
&= \frac{i}{\sigma_n}(n-k)\left(p - \frac{1}{2}\right) \\
&\geq \frac{tn}{\sigma_n}(n-k)\left(p - \frac{1}{2}\right), \qquad \coloneqq b_n.
\end{aligned}$$

for $tn \leq i < k - c_0 n$. Pruss (1998) provides a maximal inequality for sums of exchangeable random variables. Note that we can extend the sequence $R_1, \ldots, R_{k-c_0 n}$ to one of length k. Which, in the notation of Pruss (1998), means that $\gamma = k/(k - c_0 n) = \theta/(\theta - c_0)$ and

we have that

$$c(\gamma) = C_3 \frac{\gamma^2}{(\gamma-1)^2} = C_3 \frac{\theta^2}{c_0^2},$$

where C_3 is an absolute constant. We can now apply Theorem 1 of (Pruss, 1998) and Lemma 8, which gives that

$$\begin{aligned}
\Pr\left(\max_{tn\le i<k-c_0 n} -Z_n(i/n) > 0\right) &\le \Pr\left(\max_{tn\le i<k-c_0 n} -Z_n(i/n) + \mathrm{E}\left[Z_n(i/n)\right] > \frac{b_n}{\sqrt{n}}\right) \\
&\le C_3 \frac{\theta^2}{c_0^2} \Pr\left(\left|Z_n(\theta-c_0) - \mathrm{E}\left[Z_n(\theta-c_0)\right]\right| > \frac{b_n c_0^2}{C_3\theta^2\sqrt{n}}\right) \\
&\le C_3 \frac{\theta^2}{c_0^2} e^{-2\min(n(\theta-c_0),n-k)\left(\frac{\frac{(n-k)tnc_0^2}{C_3\theta^2}\left(p-\frac{1}{2}\right)}{2(n-k)n(\theta-c_0)}\right)^2} \\
&= C_3 \frac{\theta^2}{c_0^2} e^{-2\min(n(\theta-c_0),n-k)\left(\frac{tc_0^2\left(p-\frac{1}{2}\right)}{2C_1\theta^2(\theta-c_0)}\right)^2} \\
&= C_3\theta^2 e^{-C_1 n\min(\theta-c_0,1-\theta)\left(\frac{t\left(p-\frac{1}{2}\right)}{\theta^2(\theta-c_0)}\right)^2} \\
&= C_3\theta^2 e^{-C_1 n t^2\left(p-\frac{1}{2}\right)^2},
\end{aligned}$$

where the constants depend on θ, c_0. We can now show a similar bound for $\Pr(A_2)$. To this end, we have that

$$\begin{aligned}
\Pr(A_2) &= \Pr\left(\max_{c_0 n\le i<k-tn} -Z_n(i/n) - Z_n(k/n) > 0\right) \\
&\le \Pr\left(\max_{c_0 n\le i<k-tn} -Z_n(i/n) > 0\right) + \Pr\left(Z_n(k/n) < 0\right).
\end{aligned}$$

Now, it is easy to see from Lemma 9 that $\Pr\left(Z_n(k/n) < 0\right)$ is quite small, viz.

$$\Pr\left(Z_n(k/n) < 0\right) \le e^{-2n\min(\theta,1-\theta)(p-1/2)^2}.$$

Concerning $\Pr\left(\max_{c_0 n\le i<k-tn} -Z_n(i/n) > 0\right)$, we use Corollary 2 of (Pruss, 1998). Again, in the notation of Pruss (1998), we take $\rho = c_0$ and

$$c(\rho) = c(c_0) = C_4 \frac{1}{c_0(1-c_0)^2},$$

for some absolute constant C_4. It follows that

$$\begin{aligned}
\Pr\left(\max_{c_0 n \le i < k - tn} -Z_n(i/n) > 0\right) &\le \Pr\left(\max_{c_0 n \le i < k - tn} \mathrm{E}\left[Z_n(i/n)\right] - Z_n(i/n) > \frac{c_0 n(n-k)}{\sqrt{n}\sigma_n}\left(p - \frac{1}{2}\right)\right) \\
&\le c(c_0)\Pr\left(\left|Z_n(\theta) - \mathrm{E}\left[Z_n(\theta)\right]\right| > \frac{t c_0 n(n-k)}{\sqrt{n}\sigma_n c(c_0)}\left(p - \frac{1}{2}\right)\right) \\
&\le 2c(c_0) e^{-2n\min(\theta, 1-\theta)\left(\frac{tc_0}{\theta c(c_0)}\right)^2 (p-1/2)^2} \\
&\le C_1 e^{-C_2 n t^2 \left(p - \frac{1}{2}\right)^2},
\end{aligned}$$

where C_1, C_2 depend on c_0 and θ. To conclude this part of the proof, we have that

$$\Pr(k - \widehat{k} > tn) \le C_1 e^{-C_2 n t^2 (p-1/2)^2}. \tag{A.15}$$

It now remains to consider $\Pr(\widehat{k} - k > tn)$. We have that

$$\begin{aligned}
\Pr(\widehat{k} - k > tn) &\le \Pr\left(\max_{k + tn \le i < n} |Z_n(i/n)| > |Z_n(k/n)|\right) \\
&\le \Pr\left(\max_{k + tn \le i < n} |Z_n(i/n)| > Z_n(k/n)\right) \\
&\le \Pr\left(\max_{k + tn \le i < n} |Z_n(i/n)| - Z_n(k/n) > 0\right) \\
&\le \Pr(A_3) + \Pr(A_4),
\end{aligned}$$

where

$$\begin{aligned}
\Pr(A_3) &= \Pr\left(\max_{k + tn \le i < n} Z_n(i/n) - Z_n(k/n) > 0\right) \\
\Pr(A_4) &= \Pr\left(\min_{k + tn \le i < n} Z_n(i/n) + Z_n(k/n) < 0\right).
\end{aligned}$$

Once again, we start with $\Pr(A_3)$. Observe

$$\Pr(A_3) = \Pr\left(\max_{k + tn \le i < n} Z_n(i/n) - Z_n(k/n) > 0\right) = \Pr\left(\max_{k + tn \le i < n} \sum_{j=k+1}^{i} \frac{R_j - \mu_n}{\sigma_n} > 0\right).$$

Now, note that

$$\mathrm{E}\left[-\sum_{j=k+1}^{i}\frac{R_j-\mu_n}{\sigma_n}\right]=(i-k)k\frac{(p-1/2)}{\sigma_n}>tnk\frac{(p-1/2)}{\sigma_n}=b_n'.$$

So, it follows that

$$\begin{aligned}\Pr\left(\max_{k+tn\le i<n}\sum_{j=k+1}^{i}\frac{R_j-\mu_n}{\sigma_n}>0\right)&\le\Pr\left(\max_{k+tn\le i<n}\sum_{j=k+1}^{i}\frac{R_j-\mu_n}{\sigma_n}-\mathrm{E}\left[\sum_{j=k+1}^{i}\frac{R_j-\mu_n}{\sigma_n}\right]>b_n'\right)\\&\le C_1\Pr\left(\left|\sum_{j=k+1}^{k+c_0n}\frac{R_j-\mu_n}{\sigma_n}-\mathrm{E}\left[\sum_{j=k+1}^{i}\frac{R_j-\mu_n}{\sigma_n}\right]\right|>C_2b_n'\right).\end{aligned}$$

We can write

$$\sum_{j=k+1}^{k+c_0n}R_j=\frac{c_0n(c_0n+1)}{2}+c_0n(n-c_0n-k)V_1+c_0nkV_2,$$

where V_1, V_2 are two sample U-statistics. As above, Lemma 8 gives that

$$\begin{aligned}\Pr\left(\left|\sum_{j=k+1}^{k+c_0n}\frac{R_j-\mu_n}{\sigma_n}-\mathrm{E}\left[\sum_{j=k+1}^{i}\frac{R_j-\mu_n}{\sigma_n}\right]\right|>t\right)&\le C_1e^{-C_2n\left(\frac{t\sigma_n}{n(n-c_0n-k)}\right)^2}+2e^{-C_3nc_0\left(\frac{t\sigma_n}{nk}\right)^2}\\&\le C_1e^{-C_2n\left(\frac{t\sigma_n}{n\max(n-k-c_0n,k)}\right)^2}.\end{aligned}$$

It then follows that

$$\Pr\left(A_3\right)\le C_1e^{-C_2nt^2(p-1/2)^2}.$$

Consider now,

$$\begin{aligned}
\Pr(A_4) &= \Pr\left(\max_{k+tn\le i<n} -Z_n(i/n) - Z_n(k/n) > 0\right) \\
&= \Pr\left(\max_{k+tn\le i<n} -\sum_{j=k+1}^{k+i} \frac{R_j-\mu_n}{\sigma_n} > 2Z_n(k/n)\right) \\
&\le \Pr\left(\max_{k+tn\le i<n} -\sum_{j=k+1}^{k+i} \frac{R_j-\mu_n}{\sigma_n} > 2a_n\right) + \Pr\left(Z_n(k/n) < a_n\right) \\
&\le \Pr\left(\max_{k+tn\le i<n} -\sum_{j=k+1}^{k+i} \frac{R_j-\mu_n}{\sigma_n} > 2a_n\right) + e^{-2n\min(\theta,1-\theta)\left(p-1/2-\frac{a_n\sigma_n}{n^2\theta(1-\theta)}\right)^2} \\
&:= \Pr\left(A_5\right) + e^{-C_2n(p-1/2)^2},
\end{aligned}$$

where the second last line is an application of Lemma 9 if $a_n = cn = 3\frac{k(n-k)(p-1/2)}{4\sigma_n}$.

$$e^{-C_2n\left(p-1/2-\frac{a_n\sigma_n}{n^2\theta(1-\theta)}\right)^2} = e^{-C_2n(p-1/2)^2-\frac{a_n\sigma_n}{n^2\theta(1-\theta)}}$$

Looking at $\Pr\left(A_5\right)$, we have that

$$\begin{aligned}
\Pr\left(A_5\right) &= \Pr\left(\max_{k+tn\le i<n} -\sum_{j=k+1}^{k+i} \frac{R_j-\mu_n}{\sigma_n} > cn\right) \\
&\le \Pr\left(\max_{k+tn\le i<n} -\sum_{j=k+1}^{k+i} \frac{R_j-\mu_n}{\sigma_n} + \mathrm{E}\left[\sum_{j=k+1}^{i} \frac{R_j-\mu_n}{\sigma_n}\right] > (n-k)k\frac{(p-1/2)}{2\sigma_n}\right), \\
&\le 2c(c_0)e^{-2nc_0\left(\frac{(n-k)k(p-1/2)}{4c_0c(c_0)n\max(n-k-c_0n,k)}\right)^2} \\
&\le C_1e^{-C_2n(p-1/2)^2}.
\end{aligned}$$

We then have that

$$\Pr(\widehat{k}-k > tn) \le C_1e^{-C_2nt^2(p-1/2)^2} + C_1e^{-C_2n(p-1/2)^2}. \tag{A.16}$$

Therefore, combining (A.15) and (A.16) we have that

$$\Pr(|\widehat{k}-k| > tn) \le C_2e^{-nC_1t^2(p-1/2)^2}.$$

□

Proof of Theorem 10. The proof is complete after proving the following Lemma:

Lemma 10. *Assume the conditions of Theorem 10. We have that*

$$\Pr\left(\left|\widehat{Z}_n(i/n) - \mathrm{E}\left[Z_n(i/n)\right]\right| > t\right) \leq C_1 e^{-C_2 t^2},$$

for some constants C_1, C_2 that depend on b, K', the depth function and the distribution of the data.

Simply replace Lemma 8 with Lemma 10 in the arguments of the proof of Theorem 9 and Lemma 9 and the result follows.

Proof of Lemma 10. We first show that

$$|\mathrm{D}(x;\widehat{F}_n) - \mathrm{D}(x;F)| \leq K \sup_{y\in\mathbb{R}^d,\ a\in A} \left|\widehat{F}_{n,a}(g(y,a)) - F_a(g(y,a))\right|, \tag{A.17}$$

holds for all $x \in \mathfrak{F}$, where A, g are as described in Section 4.3. To begin,

$$\begin{aligned}
|\mathrm{D}(x;\widehat{F}_n) - \mathrm{D}\left(x;F\right)| &= \left|\int_A F_{n,a}(g(x,a)(1-F_{n,a}(g(x,a))) - F_a(g(x,a))(1-F_a(g(x,a)))dP_a\right| \\
&\leq \int_A \left|F_{n,a}(g(x,a))(1-F_{n,a}(g(x,a))) - F_a(g(x,a))(1-F_a(g(x,a)))\right| dP_a \\
&\leq \int_A 3\sup_y \left|F_{n,a}(y) - F_a(y)\right| dP_a \\
&\leq 3\sup_{a\in A}\sup_y \left|F_{n,a}(y) - F_a(y)\right|.
\end{aligned}$$

We can show an analogous inequality if half-space depth is used as the univariate depth function. Continuing with the proof, for brevity, we will now denote C_1 and $\widehat{k}_1$ by k and $\widehat{k}$, respectively. In addition, let

$$h_{ij}(F) = \mathrm{D}\left(X_i;F\right) - \mathrm{D}\left(X_j;F\right).$$

It holds that

$$\begin{aligned} S_m &= \sum_{j=1}^{m} R_j \\ &= \frac{m(m+1)}{2} + \sum_{j=1}^{m} \sum_{i=m+1}^{k} \mathbb{1}\left(h_{ij}(F) \geq 0\right) + \sum_{j=1}^{m} \sum_{i=k+1}^{n} \mathbb{1}\left(h_{ij}(F) \geq 0\right) \\ &:= \frac{m(m+1)}{2} + (k-m)mU_n^1 + m(n-k)U_n^2, \end{aligned}$$

where $U_n^1,\ U_n^2$ are two sample U-statistics. Note that $\mathrm{E}\left[U_n^1\right] = 1/2$ and $\mathrm{E}\left[U_n^2\right] = p$. Similarly,

$$\begin{aligned} \widehat{S}_m &= \sum_{j=1}^{m} \widehat{R}_j \\ &= \frac{m(m+1)}{2} + \sum_{j=1}^{m} \sum_{i=m+1}^{k} \mathbb{1}\left(h_{ij}(\widehat{F}_n) \geq 0\right) + \sum_{j=1}^{m} \sum_{i=k+1}^{n} \mathbb{1}\left(h_{ij}(\widehat{F}_n) \geq 0\right) \\ &:= \frac{m(m+1)}{2} + (k-m)m\widehat{U}_n^1 + m(n-k)\widehat{U}_n^2, \end{aligned}$$

where $\widehat{U}_n^1,\ \widehat{U}_n^2$ are dependent two sample U-statistics. Consider

$$\widehat{U}_n^2 - U_n^2 = \frac{1}{m(n-k)} \sum_{j=1}^{m} \sum_{i=k+1}^{n} \mathbb{1}\left(h_{ij}(\widehat{F}_n) \geq 0\right) - \mathbb{1}\left(h_{ij}(F) \geq 0\right).$$

Let

$$Y_n = \sup_{y \in \mathbb{R}^d,\ a \in A} \left| \widehat{F}_{n,a}(g(y,a)) - F_a(g(y,a)) \right|.$$

It easily follows from (A.17) that

$$\begin{aligned} \mathbb{1}\left(h_{ij}(\widehat{F}_n) > 0\right) - \mathbb{1}\left(h_{ij}(F) > 0\right) &\leq \mathbb{1}\left(h_{ij}(F) > -KY_n\right) - \mathbb{1}\left(h_{ij}(F) > 0\right) \\ &\leq \mathbb{1}\left(-KY_n < h_{ij}(F) < 0\right). \end{aligned}$$

We can now write

$$\begin{aligned}\widehat{U}_n^2 - U_n^2 &\leq \frac{1}{m(n-k)} \sum_{j=1}^{m} \sum_{i=k+1}^{n} \mathbb{1}\left(-KY_n < h_{ij}(F) < 0\right) \\ &\leq \frac{1}{m(n-k)} \sum_{j=1}^{m} \sum_{i=k+1}^{n} \mathbb{1}\left(-Kt < h_{ij}(F) < 0\right) + \mathbb{1}\left(Y_n > t/K\right). \end{aligned} \tag{A.18}$$

Looking only at the left hand terms, notice that this is a U-statistic. Assumption 12 gives that

$$\mathrm{E}\left[\frac{1}{m(n-k)} \sum_{j=1}^{m} \sum_{i=k+1}^{n} \mathbb{1}\left(-Kt < h_{ij}(F) < 0\right)\right] = \Pr(-Kt < h_{ij}(F) < 0) \leq tK'.$$

Using this, and (A.12) we have that

$$\Pr\left(\frac{1}{m(n-k)} \sum_{j=1}^{m} \sum_{i=k+1}^{n} \mathbb{1}\left(-Kt < h_{ij}(F) < 0\right) > t\right) \leq e^{-2\min(m,n-k)(t(1-K'))^2}. \tag{A.19}$$

Looking at the right hand term of (A.18) we have that

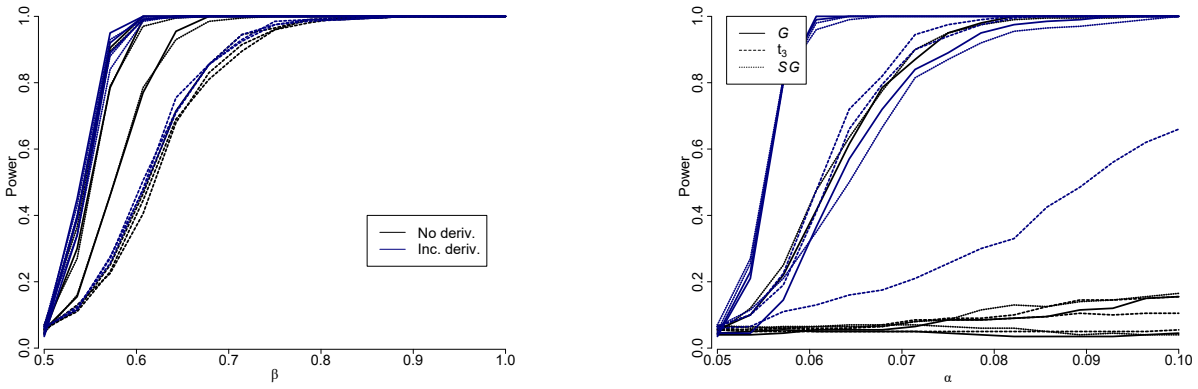
$$\begin{aligned}\Pr\left(\frac{1}{m(n-k)} \sum_{j=1}^{m} \sum_{i=k+1}^{n} \mathbb{1}\left(Y_n > t/K\right) > \delta\right) &\leq \Pr\left(\frac{1}{m(n-k)} \sum_{j=1}^{m} \sum_{i=k+1}^{n} \mathbb{1}\left(Y_n > t/K\right) > 0\right) \\ &= \Pr\left(Y_n > t/K\right) \\ &\leq K_3 e^{-K_4 n\left(\frac{t}{K}\right)^2}, \end{aligned} \tag{A.20}$$

where K_3, K_4 are constants which depend on the depth function and b. The last inequality comes from the finite dimensional assumption on the data. To elaborate, letting $c = (c_1, \ldots, c_b)$, the sets

$$\left\{\left\{\sum_{i=1}^{b} c_i \phi_i(t) \leq z,\ c \in \mathbb{R}^b\right\} t \in [0,1],\ z \in \mathbb{R}\right\}$$

and

$$\left\{\left\{\sum_{i=1}^{b} c_i \langle \phi_i, u\rangle \leq z,\ c \in \mathbb{R}^b\right\} u \in S,\ z \in \mathbb{R}\right\}$$



(a) (b)

Figure A.4: Comparison of the FKWC methods with the derivatives to the FKWC methods without the derivatives under (a) magnitude changes and (b) shape changes.

both have finite VC-dimension, which gives uniform concentration of $\widehat{F}_{n,a}$. If we consider $\widehat{U}_n^1 - U_n^1$, we can use the same technique to provide an analogous bound. Combining (A.19) and (A.19) gives that

$$\begin{aligned}\Pr\left(\left|\widehat{Z}_n(i/n) - \mathrm{E}\left[Z_n(i/n)\right]\right| > t\right) &\leq \Pr\left(\left|\widehat{Z}_n(i/n) - Z_n(i/n)\right| > t/2\right) \\ &\quad + \Pr\left(\left|Z_n(i/n) - \mathrm{E}\left[Z_n(i/n)\right]\right| > t/2\right) \\ &\leq C_1 e^{-C_2 t^2}.\end{aligned}$$

□

As stated above, with the proof of Lemma 10 complete, the proof of Theorem 10 is complete using the logic of the proof of Theorem 9. □

A.7 Additional simulation results from Chapter 4

Below we have some additional simulation results.

A.8 On simplicial depth and privacy

Many of the results from Section 5.4 apply to simplicial depth (Liu, 1988). However, the sensitivity, asymptotics and computation of simplicial depth all scale worse with respect to the dimension than the other considered depth functions. However, we summarize the results in the case of simplicial depth below: The sensitivity of simplicial depth is $(d+1)/n$, see the proof in Section A.9. The proof of Theorem 14 also shows that, for all F,

$$\Pr(\sup_x |\operatorname{SMD}(x; F_n) - \operatorname{SMD}(x; F)| > t) \leq 8e^{(d^2 \log d)\log(n+1) - nt^2/32}.$$

In addition, the proof of Corollary 15 shows that if F is such that the SMD-median θ is unique, then using SMD with Mechanism 5 results in

$$\Pr\left(\left\|\widetilde{T}(\mathbb{X}_n) - \theta\right\| \geq t\right) \leq 8e^{d^2 \log(d)\log(n+1)} e^{-n\alpha_{\mathrm{SMD}}(t)^2/512} + Ke^{-n\epsilon\alpha_{\mathrm{SMD}}(t)/4(d+1)}.$$

The sample complexity can be computed in the same manner as the sample complexities computed in the proof of Corollary 1.

A.9 Proofs from Chapter 5

A.9.1 Proofs related to the properties of the depth functions

Proof of sensitivities of depth functions. The sample halfspace depth value of some point x is the minimum normalised, univariate, centre-outward rank of x's projections amongst the samples' projections, over all univariate directions. Therefore, if a point is exchanged, all the ranks are shifted by at most one, and the global sensitivity of the unnormalised halfspace depth is 1. We get $\mathrm{GS}_n(\mathrm{HD}) = 1/n$. Following the same argument, the IRW depth is the average, normalised, centre-outward rank, and so we can conclude that $\mathrm{GS}_n(\mathrm{IRW}) = 1/n$.

Changing one observation changes F_n by at most 1/n, and (A.21) gives that $|F_n(x)(1 - F_n(-x)) - F_n^*(x)(1 - F_n^*(-x))| \leq 3|F_n(x) - F_n^*(-x)|$ and so we have that $\mathrm{GS}_n(\mathrm{IDD}) = 3/n$. Similarly, for IDD_β, it holds that $\mathrm{GS}_n(\mathrm{IDD}_\beta(x; F_n)) = 3/n$.

For simplicial depth, note that changing one observation can influence a maximum of $\binom{n-1}{d}$ terms of the type $\mathbb{1}\left(x \in \sim (X_{i_1}, \ldots, X_{i_{d+1}})\right)$, and each term has a sensitivity of 1. It follows that $\mathrm{GS}_n(\mathrm{SMD}) = (d+1)/n$. □

Proof of Theorem 14. We use empirical process techniques. First, note that:

$$\Pr(\sup_x |\operatorname{IRW}(x; F_n) - \operatorname{IRW}(x; F)| > t) \leq \Pr(4 \sup_u \|F_{n,u} - F_u\| > t - 1/2n).$$

Consider:

$$\mathcal{F} = \left\{ g(y) = \mathbb{1}\left(y^\top u \leq x^\top u\right) : x \in \mathbb{R}^d, u \in S^{d-1} \right\} \subset \mathcal{H}_d,$$

where

$$\mathcal{H}_d = \left\{ g(y) = \mathbb{1}\left(y^\top u \leq c\right) : c \in \mathbb{R}, u \in S^{d-1} \right\}.$$

It is well-known that $VC(\mathcal{H}_d) = d + 1$. Therefore, we can use results on the concentration of the supremum of an empirical process (Talagrand, 1994). If K is a varying absolute constant, we have that

$$N\left(\epsilon, \mathcal{H}_d, L_1(Q)\right) \leq (K(d+1))(8e/\epsilon)^{d+1} \leq \left(\frac{(K(d+1))^{1/(d+1)} 8e}{\epsilon} \right)^{d+1} \leq \left(\frac{K}{\epsilon} \right)^{d+1}.$$

This gives that

$$\begin{aligned}
\Pr(\sqrt{n} \|F_{n,u} - F_u\| > t) &\leq \frac{K}{t} \left(\frac{Kt^2}{d+1} \right)^{d+1} e^{-2t^2} \\
&\leq t^{2d+1} K^{d+2} \left(\frac{1}{d+1} \right)^{d+1} e^{-2t^2} \\
&\leq e^{(2d+1)\log(t) - (d+1)\log(d+1) + (d+2)\log(K) - 2t^2} \\
&\leq e^{(d+1/2)\log(n) - (d+1)\log(d+1) + (d+2)\log(K) - 2t^2} \\
&\leq e^{(d+1)\log\left(\frac{n}{d+1}\right) + (d+2)\log(K) - 2t^2} \\
&\leq K e^{(d+1)\log\left(K\frac{n}{d+1}\right) - 2t^2}.
\end{aligned}$$

We see that

$$\begin{aligned}
\Pr(\sup_x |\operatorname{IRW}(x; F_n) - \operatorname{IRW}(x; F)| > t) &\leq \Pr(\sup_u \|F_{n,u} - F_u\| > t/4 - 1/8n) \\
&\leq K e^{(d+1)\log\left(K\frac{n}{d+1}\right) - 2\left(\sqrt{n}t/4 - 1/8n\right)^2} \\
&\leq K e^{(d+1)\log\left(K\frac{n}{d+1}\right) - \sqrt{n}t^2/8 - t/16\sqrt{n} + 1/32n} \\
&\leq K e^{(d+1)\log\left(K\frac{n}{d+1}\right) - nt^2/8}.
\end{aligned}$$

For integrated dual depth, we first need the following result. Let $G(x) = F(x)(1 - F(x))$ and $G_n(x) = F_n(x)(1 - F_n(x))$ for some F, F_n such that $\|F\|_\infty \leq 1$ and $\|F_n\|_\infty \leq 1$. It then holds that

$$\begin{aligned}
|G(x) - G_n(x)| &= |F(x)(1 - F(x)) - F_n(x)(1 - F_n(x))| \\
&= |F(x) - F_n(x) - F(x)^2 + F_n(x)^2| \\
&\leq |F(x) - F_n(x) + F_n(x)F(x) - F(x)^2 - F_n(x)F(x) + F_n(x)^2| \\
&= |F(x) - F_n(x) + F(x)(F_n(x) - F(x)) + F_n(x)(F_n(x) - F(x))| \\
&\leq 3|F(x) - F_n(x)|. \qquad \text{(A.21)}
\end{aligned}$$

Using this, we can then write

$$\begin{aligned}
\Pr\left(\sup_x |\operatorname{IDD}(x; F_n) - \operatorname{IDD}(x; F)| > t\right) &\leq \Pr(\sup_u \|F_{n,u} - F_u\| > t/3) \\
&\leq K e^{(d+1)\log\left(K\frac{n}{d+1}\right) - 2nt^2/9}.
\end{aligned}$$

To prove something similar for halfspace depth, we need F_u to be a continuous function in u. It suffices to choose F to be absolutely continuous with a bounded density f. To see this, observe that for some sequence $u_m \to u$, we may have that

$$\left|\int_{H_u} f(x)dx - \int_{H_{u_m}} f(x)dx\right| \leq \int_{H_u \triangle H_{u_m}} f(x)dx \leq \sup_x f(x)\operatorname{vol}(H_u \triangle H_{u_m}) \to 0.$$

We can then say that

$$\Pr(\sup_x |\operatorname{HD}(x; F_n) - \operatorname{HD}(x; F)| > t) = \Pr\left(\sup_x \left|\inf_u F_{n,u} - \inf_u F_u\right| > t\right).$$

Now, we know that $F_{n,u}(x)$ can take a finite number of values as a function of u, and so $\inf_u F_{n,u}(x) = F_{n,u^*_{n,x}}(x)$ for some, non-unique $u^*_{n,x}$. Suppose that $|\inf_u F_{n,u}(x) -$

$\inf_u F_u(x)| = \inf_u F_u(x) - \inf_u F_{n,u}(x)$, it follows that

$$\begin{aligned}
|\inf_u F_{n,u}(x) - \inf_u F_u(x)| &= \inf_u F_u(x) - \inf_u F_{n,u}(x) \\
&= \inf_u F_u(x) - F_{n,u^*_{n,x}}(x) \\
&\leq F_{u^*_{n,x}}(x) - F_{n,u^*_{n,x}}(x) \\
&\leq \sup_u |F_u(x) - F_{n,u}(x)| \\
&\leq \sup_x \sup_u |F_u(x) - F_{n,u}(x)|.
\end{aligned}$$

which holds for all x. Similar to above, if $|\inf_u F_{n,u}(x) - \inf_u F_u(x)| = \inf_u F_{n,u}(x) - \inf_u F_u(x)$ then

$$|\inf_u F_{n,u}(x) - \inf_u F_u(x)| \leq \sup_x \sup_u |F_u(x) - F_{n,u}(x)|.$$

It follows immediately via the same VC-dimension argument that

$$\begin{aligned}
\Pr(\sup_x |\operatorname{HD}(x; F_n) - \operatorname{HD}(x; F)| > t) &\leq \Pr(\sup_{u,x} |F_{n,u}(x) - F_u(x)| > t) \\
&\leq K e^{(d+1)\log\left(K\frac{n}{d+1}\right) - 2t^2}.
\end{aligned}$$

We now prove the bound for the smoothed dual depth. If $t > 1/4$ then it is trivial that

$$\Pr\left(\sup_x |\operatorname{IDD}_\beta(x; F_n) - \operatorname{IDD}_\beta(x; F)| > t\right) = 0,$$

Now, assume that $t < 1/4$. We see that

$$\mathcal{F} = \left\{\sigma(\beta(x - X_i)^\top u) \colon x \in \mathbb{R}^d, u \in S^{d-1}\right\} \subset \left\{\sigma(AX_i + b) \colon A \in \mathbb{R}^d, b \in \mathbb{R}\right\} \coloneqq \mathcal{F}^*,$$

where $\mathcal{F}^*$ is a monotone function applied to a finite dimensional vector space of measurable functions. Therefore $VC(\mathcal{F}^*) = d + 2$ and so $VC(\mathcal{F}) \leq d + 2$. Let

$$Z_n = \sqrt{n} \sup_{\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma(\beta(x - X_i)^\top u) - \operatorname{E}\left[\sigma(\beta(x - X_1)^\top u))\right] \right|.$$

It follows from Talagrand (1994) that

$$\Pr\left(Z_n \geq t\right) \leq \left(\frac{Kt}{\sqrt{2d+4}}\right)^{2d+4} e^{-2t^2} \leq e^{(2d+4)\log\left(\frac{Kt}{\sqrt{2d+4}}\right)-2t^2},$$

where K is a universal constant. Now, we have that

$$\Pr\left(\sup_x |\operatorname{IDD}_\beta(x;F_n) - \operatorname{IDD}_\beta(x;F)| > t\right) \leq \Pr\left(3Z_n \geq \sqrt{n}t\right) \leq e^{(2d+4)\log\left(\frac{K\sqrt{n}}{4\sqrt{2d+4}}\right)-2nt^2/9}.$$

For simplicial depth, a VC-dimension concentration argument and (Dümbgen, 1992) gives that

$$\Pr(\sup_x |\operatorname{SMD}(x;F_n) - \operatorname{SMD}(x;F)| > t) \leq 8(n+1)^{d^2\log d} e^{-nt^2/32} = 8e^{(d^2\log d)\log(n+1)-nt^2/32}.$$

□

A.9.2 LDP theorem, concentration and sample complexity for the private medians

Proof of Theorem 13. Let $E_{n,m} = \{\|\phi_n - \phi\| < m\}$ for short. Then,

$$\begin{aligned}\Pr\left(\left\|\widetilde{T}(\mathbb{X}_n) - \theta\right\| > t\right) &= \Pr\left(E_{n,m}^c\right) + \Pr\left(\left\|\widetilde{T}(\mathbb{X}_n) - \theta\right\| > t \cap E_{n,m}\right) \\ &\leq C_1(\phi,d,n)e^{-C_2(\phi)nm^2} + \Pr\left(\left\|\widetilde{T}(\mathbb{X}_n) - \theta\right\| > t \cap E_{n,m}\right),\end{aligned}$$

We now focus on the right hand term above:

$$\begin{aligned}\frac{1}{\lambda_n}\log\Pr\left(\left\|\widetilde{T}(\mathbb{X}_n) - \theta)\right\| > t \cap E_{n,m}\right) &= \frac{1}{\lambda_n}\log\int_{E_{n,m}} \frac{\int_{\mathcal{B}_t^c(\theta)} \exp\left(\lambda_n\phi_n(x)\right)d\pi}{\int_{\mathbb{R}^d}\exp\left(\lambda_n\phi(x)\right)d\pi} dP \\ &\leq 2m + \frac{1}{\lambda_n}\log\frac{\int_{\mathcal{B}_t^c(\theta)} \exp\left(\lambda_n\phi(x)\right)d\pi}{\int_{\mathbb{R}^d}\exp\left(\lambda_n\phi(x)\right)d\pi}.\end{aligned}$$

First, we lower bound the denominator of the right-hand term:

$$\begin{aligned}\frac{1}{\lambda_n}\log\int_{\mathbb{R}^d}\exp\left(\lambda_n\phi(x)\right)d\pi &\geq \frac{1}{\lambda_n}\log\int_{\mathcal{B}_k(\theta)}\exp\left(\lambda_n\phi(x)\right)d\pi \\ &= \phi(\theta)+\frac{1}{\lambda_n}\log\int_{\mathcal{B}_k(\theta)}\exp\left(\lambda_n(\phi(x)-\phi(\theta))\right)d\pi \\ &\geq \phi(\theta)-\omega_\phi(k)\wedge 1+\lambda_n^{-1}\log\pi(\mathcal{B}_k(\theta)).\end{aligned}$$

Using this,

$$\begin{aligned}\frac{1}{\lambda_n}\log\Pr\left(\mathcal{B}_t^c(\theta)\cap E_{n,m}\right) &\leq \frac{1}{\lambda_n}\log\int_{\mathcal{B}_t^c(\theta)}\exp\left(\lambda_n\phi(x)\right)d\pi-\phi(\theta) \\ &\qquad +\omega_\phi(k)\wedge 1-\lambda_n^{-1}\log\pi(\mathcal{B}_k(\theta))+2m. \\ &\leq \sup_{\mathbf{x}\in\mathcal{B}_t^c(\theta)}\left(\phi(x)-\phi(\theta)\right)+\omega_\phi(k)\wedge 1-\lambda_n^{-1}\log\pi(\mathcal{B}_k(\theta))+2m \\ &:= -f(t)+g(\phi,\pi,\theta,\lambda_n)+2m,\end{aligned}$$

where we set $k=\operatorname{argmin}_{k\in[0,1]}[\omega_\phi(k)\wedge 1-\lambda_n^{-1}\log\pi(\mathcal{B}_k(\theta))]$, where the minimum exists by assumption. We have that

$$g_1(\phi,\pi,\theta,\lambda_n)=\min_{k\in[0,1]}[\omega_\phi(k)\wedge 1-\lambda_n^{-1}\log\pi(\mathcal{B}_k(\theta))].$$

In addition, note that

$$f(t)=\phi(\theta)-\sup_{\mathbf{x}\in\mathcal{B}_t^c(\theta)}\phi(x)>0.$$

It follows then that

$$\Pr\left(\left\|\widetilde{T}(\mathbb{X}_n)-\theta)\right\|>t\cap E_{n,m}\right)\leq\exp(-\lambda_n f(t)+\lambda_n g(\phi,\pi,\theta,\lambda_n)+2\lambda_n m).$$

Letting $m=f(t)/4$, we can write

$$\begin{aligned}\Pr\left(\left\|\widetilde{T}(\mathbb{X}_n)-\theta)\right\|>t\right) &\leq \Pr(\|\phi_n-\phi\|>f(t)/4)+\exp(-\lambda_n f(t)/2+\lambda_n g(\phi,\pi,\theta,\lambda_n)) \\ &\leq C_1(\phi,d,n)e^{-C_2(\phi)nf(t)^2/16}+e^{-\lambda_n f(t)/2+\lambda_n g(\phi,\pi,\theta,\lambda_n)}.\end{aligned}$$

□

Proof of Theorem 15. We first show that our assumptions on F imply the considered depth

functions are Lipschitz continuous, which requires the following auxiliary Lemma:

Lemma 11. *If F has a density, then F_u does.*

Proof. First, we show that F_u has a density if F does. We use the fact that for any real, measurable function g, if

$$\mathrm{E}\left[g(X^\top u)\right] = \int_{\mathbb{R}} g(x) f_u(x) dx,$$

then $f_u(x)$ is the density for $X^\top u$. In order to do this, we start with

$$\mathrm{E}\left[g(X^\top u)\right] = \int_{\mathbb{R}^d} g(\mathbf{x}^\top u) f(\mathbf{x}) d\mathbf{x}.$$

Recall that we can write $\mathbb{R}^d = U \times U^\perp$, where $U = \mathrm{span}(\{u\})$ and $U^\perp$ is the orthogonal compliment of U. Recall from linear algebra that we can write every vector as the sum of two vectors: $x = au + z$, where $a \in \mathbb{R}$, $z \in U^\perp$. Now, $|\mathbf{J}x| = 1$ because u is a unit vector and $z \in U^\perp$. We have that

$$\begin{aligned}
\mathrm{E}\left[g(X^\top u)\right] &= \int_{\mathbb{R}^d} g(\mathbf{x}^\top u) f(\mathbf{x}) d\mathbf{x}. \\
&= \int_{\mathbb{R}} \int_{U^\perp} g((au+z)^\top u) f(au+z) dz da \\
&= \int_{\mathbb{R}} \int_{U^\perp} g(a) f(au+z) dz da \\
&= \int_{\mathbb{R}} g(a) \int_{U^\perp} f(au+z) dz da,
\end{aligned}$$

so then $f_u(x) = \int_{U^\perp} f(xu+z) dz$. □

Lemma 12. *Suppose* $\sup_u f_u < \infty$. *Then all of the considered depth functions are lipschitz functions.*

Proof. Now since, F_u has bounded a density, we can conclude both that the CDF F_u is a lipschitz function (it has a bounded derivative) and that F_u is a continuous function in u (see the proof of Theorem 14). We sometimes will suppress the F in $\mathrm{D}(\cdot, F)$, since it is clear from the context. Note that for halfspace depth consider two points $x, y \in \mathbb{R}^d$. If F_u are lipschitz continuous, then

$$|F_u(x^\top u) - F_u(y^\top u)| \leq \|f_u\|_\infty |(x-y)^\top u| \leq \|f_u\|_\infty \|x-y\|. \tag{A.22}$$

Now, without loss of generality, suppose that $\mathrm{HD}(x, F) > \mathrm{HD}(y, F)$. Suppose further that u^* is such that $F_{u^*}(y^\top u) = \inf_u F_u(y^\top u)$. There exists such a u^* because F is continuous, implying that F_u is continuous in u, thus, F_u is continuous function on a compact set. It follows that

$$|\,\mathrm{HD}(x) - \mathrm{HD}(y)| = \inf_u F_u(x^\top u) - F_{u^*}(y^\top u) \leq F_{u^*}(x^\top u) - F_{u^*}(y^\top u) \leq \|f_u\|_\infty \|x - y\|\,.$$

For IRW depth, it holds that

$$\begin{aligned}
|\,\mathrm{IRW}(x) - \mathrm{IRW}(y)| &\leq \int_{S^{d-1}} |F_u(x^\top u) \vee (1 - F_u(x^\top u)) - F_u(y^\top u) \vee (1 - F_u(y^\top u))| d\nu \\
&\leq \int_{S^{d-1}} 2|F_u(x^\top u) - F_u(y^\top u)| + 2|1 - F_u(x^\top u) - F_u(y^\top u)| d\nu \\
&\leq 4\,\|f_u\|_\infty \|x - y\| \int_{S^{d-1}} 2 d\nu,
\end{aligned}$$

which is a result of (A.22) and the fact that $|1 - F_u(x^\top u) - F_u(y^\top u)| \leq 1$.

For simplicial depth, if F is lipschitz continuous, then we must show that $\Pr(x \in \Delta(X_1, \ldots, X_{d+1})$ is also lipschitz continuous. It is easy to begin with two dimensions. Consider $\Pr(x \in \Delta(X_1, X_2, X_3)) - \Pr(y \in \Delta(X_1, X_2, X_3))$, as per (Liu, 1990), we need to show that $\Pr(\overline{X_1 X_2}$ intersects $\overline{xy}) \leq \|f_u\|_\infty \|x - y\|$. In order for this event to occur, we must have that X_1 is above $\overline{xy}$ and X_2 is below $\overline{xy}$, but both are projected onto the line segment $\overline{xy}$ when projected onto the line running through $\overline{xy}$. The affine invariance of simplicial depth implies we can assume, without loss of generality, that x and y lie on the axis of the first coordinate. Let x_1 and y_1 be the first coordinates of x and y. Suppose that X_{11} is the first coordinate of X_1. It then follows from Lipschitz continuity of F that

$$\Pr(\overline{X_1 X_2} \text{ intersects } \overline{xy}) \leq \Pr(x_1 < X_{11} < y_1) \leq \|f_u\|_\infty |x_1 - y_1| \leq \|f_u\|_\infty \|x - y\|\,.$$

In dimensions greater than two, a similar line of reasoning can be used. We can again assume, without loss of generality, that x and y lie on the axis of the first coordinate. It holds that

$$\Pr(x \in \Delta(X_1, X_2, X_3)) - \Pr(y \in \Delta(X_1, X_2, X_3)) \leq \binom{d+1}{d} \Pr(A_d),$$

where A_d is the event that the $d-1$-dimensional face of the random simplex, formed by d

points randomly drawn from F, intersects the line segment $\overline{xy}$. It is easy to see that

$$\Pr(A_d) \leq \Pr(x_1 < X_{11} < y_1) \leq \|f_u\|_\infty |x_1 - y_1| \leq \|f_u\|_\infty \|x - y\|. \qquad \square$$

We can use these results to write $\omega_{\mathrm{D}}(k) = C' \cdot \sup_u \|f_u\| \cdot k$. Therefore the depth functions satisfy condition 3 of Theorem 13. Note that

$$\Pr\left(\left\|\widetilde{T}(\mathbb{X}_n) - \theta\right\| > t\right) \leq e^{-\lambda_n \alpha_{\mathrm{D}}(t)/2 + \lambda_n \min_{k\in[0,1]}[(C'\cdot \sup_u \|f_u\| \cdot k) \wedge 1 - \lambda_n^{-1} \log \pi(\mathcal{B}_k(\theta))]} + C_1(\mathrm{D}, d, n) e^{-C_2(\mathrm{D}) n \alpha_{\mathrm{D}}(t)^2/16}.$$

By assumption, we have that $\inf_{k\in[0,1]}[(C' \cdot \sup_u \|f_u\| \cdot k) \wedge 1 - \lambda_n^{-1} \log \pi(\mathcal{B}_k(\theta))] < \alpha_{\mathrm{D}}(t)/4$, which gives that

$$\Pr\left(\left\|\widetilde{T}(\mathbb{X}_n) - \theta\right\| \geq t\right) \leq C_1(\mathrm{D}, d, n) e^{-C_2(\mathrm{D}) n \alpha_{\mathrm{D}}(t)^2/16} + e^{-\lambda_n \alpha_{\mathrm{D}}(t)/4.}.$$

The second step is to show that the depth functions satisfy conditions 1 and 2 in Theorem 13. All of the considered depth functions satisfy condition 1 as a result of Theorem 14 with $C_1(\mathrm{D}, d, n) = K e^{(d+1)\log\left(K\frac{n}{d+1}\right)}$, $\lambda_n = \epsilon/\mathrm{GS}_n(\mathrm{D}) = n\epsilon/C$, $C_2 = c$. Condition 2 is assumed. $\square$

Proof of Corollary 1. Let $\alpha(t) = \phi(\theta) - \sup_{\|x-\theta\|\geq t} \phi(x)$ and suppress the D in α_{D} for brevity. We will use K to represent constants independent of the dimension and C to represent constants dependent on the dimension. For fixed $t > 0$ and $0 < \gamma < 1$, we want to find n to ensure

$$\gamma \leq C_1(\phi, d, n) e^{-K_1(\phi) n \alpha(t)^2/16} + e^{-K_2(\phi) n \epsilon \alpha(t)/4}.$$

Looking at the first term, for the depth functions, we have that

$$\begin{aligned} C_1(\phi, d, n) e^{-K_1(\phi) n \alpha(t)^2/16} &= K_3 e^{(d+1)\log\left(\frac{K_3}{d+1}\right)} e^{(d+1)\log(n) - K_1(\phi) n \alpha(t)^2/16} \\ &\leq K_3 e^{(d+1)\log(n) - K_1(\phi) n \alpha(t)^2/16}. \end{aligned}$$

We can now write

$$\gamma < K_3 e^{(d+1)\log(n) - K_1(\phi) n\alpha(t)^2/16} \implies n > \frac{K_1 \log(1/\gamma) + K_3(d+1)\log(n)}{\alpha(t)^2}$$
$$\implies n^{1-r} > \frac{K_1 \log(1/\gamma) + K_3(d+1)}{\alpha(t)^2} \implies n > \left(\frac{K_1 \log(1/\gamma) + K_3(d+1)}{\alpha(t)^2}\right)^{\frac{1}{1-r}},$$

for small $r > 0$. The second term gives that

$$n > \frac{4\log(1/\gamma)}{K_2 \epsilon \alpha(t)}.$$

Now, we can clean up the definitions of the constants, and write

$$n(t, \gamma, d, \epsilon) = \left(\frac{K_1 \log(1/\gamma) + K_2(d+1)}{\alpha(t)^2}\right)^{\frac{1}{1-r}} \vee \frac{K_3 \log(1/\gamma)}{\epsilon \alpha(t)}.$$

□

Proof of Corollary 2. Suppress the D in α_{D} for brevity. We need only need to assess N_0. We need:

$$-\frac{C}{n\epsilon} \log \pi(\mathcal{B}_{C\alpha(t)}(\theta)) < \alpha(t)/8 \implies n > -\frac{8C}{\alpha(t)\epsilon} \log \pi(\mathcal{B}_{C\alpha(t)}(\theta)).$$

Therefore, we would like to lower bound $\log \pi(\mathcal{B}_{C\alpha(t)}(\theta))$. First, suppose $\pi \sim \mathcal{N}_d(\theta, I)$. Recall that $C\alpha(t) < C/2$ in the case of HD. Let $Y \sim \chi^2_d$, then we are interested in lower bounding

$$\Pr\left(\|X - \theta\| \leq \alpha(t)/8\right) = \Pr\left(Y \leq \alpha(t)^2/64\right).$$

Suppose that $d > 2$, so the mode of the distribution is at $d - 2 > 1$. Now, for some small $0 < r < \alpha(t)^2/64$, we write ,

$$\begin{aligned}
\Pr\left(Y \leq \alpha(t)^2/64\right) &= \int_0^{\alpha(t)^2/64} \frac{y^{d/2-1} e^{-y/2}}{2^{d/2}\Gamma(d/2)} dy \\
&\geq \frac{1}{2^{d/2}\Gamma(d/2)} \int_r^{\alpha(t)^2/64} y^{d/2-1} e^{-y/2} dy \\
&> (\alpha(t)^2/64 - r) \frac{r^{d/2-1} e^{-r/2}}{2^{d/2}\Gamma(d/2)}.
\end{aligned}$$

Now, choose $r = \alpha(t)^2/128$, this gives

$$-\log \pi(\mathcal{B}_{C\alpha(t)}(\theta)) \leq C \cdot d \log(2/K\alpha(t)) + C\alpha(t) + \log \Gamma(d/2) \leq C \cdot d \log \left(\frac{1}{\alpha(t)} \vee d \right).$$

Now suppose that $\pi \sim \mathcal{N}_d(\mu, I)$, then we can let $Y' \sim \chi^2_d(\|\mu - \theta\|^2)$. First, recall that the density of $X \sim \chi^2_d(\lambda)$ which we denote by $f(y, d, \lambda)$ has the property:

$$f(y) \geq e^{-\|\mu-\theta\|^2/2} \frac{y^{d/2-1} e^{-y/2}}{2^{d/2}\Gamma(d/2)} = e^{-\|\mu-\theta\|^2/2} f(y, d, 0).$$

Using this, it is easy to see that

$$\Pr\left(Y \leq \alpha(t)^2/64\right) = \Pr\left(Y' \leq \alpha(t)^2/64\right) \geq e^{-\|\mu-\theta\|^2/2} (\alpha(t)^2/64 - r) \frac{r^{d/2-1} e^{-r/2}}{2^{d/2}\Gamma(d/2)}.$$

Once again,

$$\begin{aligned} -\log \pi(\mathcal{B}_{C\alpha(t)}(\theta)) &\leq \|\mu - \theta\|^2 /2 + C \cdot d \log(2/K\alpha(t)) + C\alpha(t) + \log \Gamma(d/2) \\ &\leq \|\mu - \theta\|^2 /2 + C \cdot d \log \left(\frac{1}{\alpha(t)} \vee d \right). \end{aligned}$$

This gives the bound:

$$n > \frac{C'}{\alpha(t)\epsilon} \left[\|\mu - \theta\|^2 /2 + C \cdot d \log \left(\frac{1}{\alpha(t)} \vee d \right) \right].$$

To incorporate the scale parameter, replace $\alpha(t)$ with $\alpha(t)/\varsigma$. To prove the second bound we have

$$-\log \pi(\mathcal{B}_{C\alpha(t)}(\theta)) = -\log \frac{(C\alpha(t))^d}{\Gamma(d/2+1)} = \log \Gamma(d/2+1) - d \log C\alpha(t) \geq C \cdot d \log \left(\frac{1}{\alpha(t)} \vee d \right).$$

□

A.9.3 Proofs related to the smoothed dual depth

Proof of Theorem 11. Consider cu for any unit vector u, then

$$\begin{aligned}\lim_{c\to\infty} \mathrm{IDD}_\beta(cu;F) &= \lim_{c\to\infty} \int_{S^{d-1}} \left(\mathrm{E}\left[\sigma\left(\beta(cu-X)^\top u\right)\right]\right)\left(1-\mathrm{E}\left[\sigma\left(\beta(cu-X)^\top u\right)\right]\right) d\nu(u)\\ &= \int_{S^{d-1}} \lim_{c\to\infty} \left(\mathrm{E}\left[\sigma\left(\beta(cu-X)^\top u\right)\right]\right)\left(1-\mathrm{E}\left[\sigma\left(\beta(cu-X)^\top u\right)\right]\right) d\nu(u)\\ &= 0,\end{aligned}$$

by bounded convergence theorem and the fact that $\lim_{x\to\infty}(1+e^{-x})^{-1}(1+e^{x})^{-1} = 0$. For the next property, suppose that $\theta - X \overset{\mathrm{d}}{=} X - \theta$. We use the fact that $(1-\sigma(x-X)) = \sigma(X-x)$. Then,

$$\begin{aligned}\mathrm{IDD}_\beta(\theta;F) &= \int_{S^{d-1}} \left(\mathrm{E}\left[\sigma\left(\beta(\theta-X)^\top u\right)\right]\right)\left(1-\mathrm{E}\left[\sigma\left(\beta(\theta-X)^\top u\right)\right]\right) d\nu(u)\\ &= \int_{S^{d-1}} \left(\mathrm{E}\left[\sigma\left(\beta(X-\theta)^\top u\right)\right]\right)^2 d\nu(u)\\ &= \int_{S^{d-1}} \left(1-\mathrm{E}\left[\sigma\left(\beta(X-\theta)^\top u\right)\right]\right)^2 d\nu(u),\end{aligned}$$

which implies that

$$\mathrm{IDD}_\beta(\theta;F) = 1/4.$$

For the invariance property, it is easy to see that $(Ax+b-AX-b)^\top u = A^\top(x-X)^\top u = A^\top u^\top(x-X) = \tilde{u}^\top(x-X)$, where $\tilde{u}$ is another unit vector. From this fact, and the fact that we are integrating strictly a function of the expression $u^\top(x-X)$ over S^{d-1}, similarity invariance holds.

Assume without loss of generality that $\theta = 0$; we want to show that if $X - \overset{\mathrm{d}}{=} -X$, then $\mathrm{IDD}_\beta(\alpha x;F) > \mathrm{IDD}_\beta(x;F))$ for $0 < \alpha < 1$. We use the same symmetry tools as in the

proof of maximality at center. It follows that

$$\begin{aligned}
\mathrm{IDD}_\beta(\alpha x; F) &= \int_{S^{d-1}} \mathrm{E}\left[\sigma(\alpha\beta x^\top u)\right](1 - \mathrm{E}\left[\sigma(\alpha\beta x^\top u)\right])d\nu \\
&= \int_{S^{d-1}} \mathrm{E}\left[\sigma(\alpha\beta x^\top u)\right]^2 d\nu \\
&\geq \int_{S^{d-1}} \mathrm{E}\left[\sigma(\beta x^\top u)\right]^2 d\nu \\
&= \int_{S^{d-1}} \mathrm{E}\left[\sigma(\beta x^\top u)\right](1 - \mathrm{E}\left[\sigma(\beta x^\top u)\right])d\nu \\
&= \mathrm{IDD}_\beta(x; F).
\end{aligned}$$

□

Proof of Theorem 12. For brevity let

$$G_{n,u,\beta}(x) = \frac{1}{n}\sum_{i=1}^{n} \sigma\left(\beta(x - X_i)^\top u\right),$$

and

$$Z = \sup_x \left| \frac{1}{M}\sum_{m=1}^{M} G_{n,u_m,\beta}(x)(1 - G_{n,u_m,\beta}(x)) - \int G_{n,u,\beta}(x)(1 - G_{n,u,\beta}(x))d\nu(u) \right|.$$

First, we have that

$$Z \leq \sup_x \left| \frac{1}{M}\sum_{m=1}^{M} G_{n,u_m,\beta}(x) - \int G_{n,u,\beta}(x)d\nu(u) \right| + \sup_x \left| \frac{1}{M}\sum_{m=1}^{M} G_{n,u_m,\beta}(x)^2 - \int G_{n,u,\beta}(x)^2 d\nu(u) \right|.$$

This is the sum of two supremums of two empirical processes. First, for the left-hand process, we have that

$$\mathcal{F} = \left\{ \frac{1}{n}\sum_{i=1}^{n} \sigma\left((\beta(x - X_i)^\top u)\right) : x \in \mathbb{R}^d \right\} \subset \left\{ \frac{1}{n}\sum_{i=1}^{n} \sigma(A^\top u + c_i) \colon A \in \mathbb{R}^d,\ c_1, \ldots, c_n \in \mathbb{R} \right\} \coloneqq \mathcal{F}^*.$$

The set $\mathcal{F}^*$ is a vector space of functions, so we have that $V(\mathcal{F}^*) = d + n + 1$. This implies

that $V(\mathcal{F}) \leq d+n+1$. Assume that $t < 1$. We then have that

$$\Pr\left(\sup_x \left|\frac{1}{M}\sum_{m=1}^{M} G_{n,u_m,\beta}(x) - \int G_{n,u,\beta}(x)d\nu(u)\right| > t/\sqrt{M}\right) \leq e^{2(d+n+1)\log\left(\frac{Kt}{\sqrt{2(d+n+1)}}\right)-2t^2}$$
$$\leq e^{2(d+n+1)\log\left(\frac{K\sqrt{M}}{\sqrt{2(d+n+1)}}\right)-2t^2},$$

which implies that

$$\Pr\left(\sup_x \left|\frac{1}{M}\sum_{m=1}^{M} G_{n,u_m,\beta}(x) - \int G_{n,u,\beta}(x)d\nu(u)\right| > t\right) \leq e^{2(d+n+1)\log\left(\frac{K\sqrt{M}}{\sqrt{2(d+n+1)}}\right)-2Mt^2}.$$

The same logic gives that

$$\Pr\left(\sup_x \left|\frac{1}{M}\sum_{m=1}^{M} G_{n,u,\beta}(x)^2 - \int G_{n,u,\beta}(x)^2 d\nu(u)\right| > t\right) \leq e^{2(d+n+1)\log\left(\frac{K\sqrt{M}}{\sqrt{2(d+n+1)}}\right)-2Mt^2}.$$

Therefore,

$$\Pr\left(Z > t\right) \leq 2e^{2(d+n+1)\log\left(\frac{K\sqrt{M}}{\sqrt{2(d+n+1)}}\right)-Mt^2/2} = 2e^{(d+n+1)\log\left(\frac{KM}{2(d+n+1)}\right)-Mt^2/2}.$$

If $t \geq 1$ then the inequality is trivial. □

A.9.4 Propose-test-release proofs

Proof of Theorem 16. Note that it follows from the assumptions that $k_1 < \text{MAD}(F_u) < k_2$ and $\text{MED}(F_u) < k_3 < \infty$, for all $u \in S^{d-1}$. We start by proving or listing several concentration results about the sample median and the median absolute deviation. We will use $\xi_{q,u}$ to represent the q^{th} (left-continuous) quantile of the projected distribution F_u. We require the following result on concentration of the sample median:

Lemma 13. *Suppose that $X_1, \ldots, X_n$ is a univariate, i.i.d. sample from F whose median is uniquely $\xi_{1/2}$. For all $t > 0$, the sample median $\hat{\xi}_{1/2}$ concentrates according to*

$$\Pr(|\hat{\xi}_{1/2} - \xi_{1/2}| > t) \leq 2e^{-np(t)^2},$$

where $p(t) = (F(\xi_{1/2}+t) - 1/2) \wedge (F(\xi_{1/2}+t) - 1/2)$.

Proof. From sub-additivity we have that

$$\begin{aligned}\Pr(|\hat{\xi}_{1/2} - \xi_{1/2}| > t) &\leq \Pr(\hat{\xi}_{1/2} - \xi_{1/2} > t) + \Pr(\xi_{1/2} - \hat{\xi}_{1/2} > t) \\ &\leq \Pr(Y_1(t) > n/2) + \Pr(Y_2(t) > n/2),\end{aligned}$$

where Y_1 and Y_2 are the number of points in the sample which are greater than $\xi_{1/2} + t$ and less than $\xi_{1/2} - t$ respectively. It follows from a Hoeffding inequality that

$$\Pr(Y_1(t) > n/2) = \Pr(Y_1(t)/n - (1 - F(\xi_{1/2} + t)) > 1/2 - (1 - F(\xi_{1/2} + t))) \leq e^{-n(F(\xi_{1/2}+t)-1/2)^2},$$

where one notes that $F(\xi_{1/2} + t) > 1/2$ and so $-1/2 + F(\xi_{1/2} + t) \geq 0$. Similarly,

$$\Pr(Y_2(t) > n/2) = \Pr(Y_2(t)/n - F(\xi_{1/2} - t) > 1/2 - F(\xi_{1/2} - t)) \leq e^{-n(F(\xi_{1/2}-t)-1/2)^2}.$$

Now, let $p(t) = (1/2 - F(\xi_{1/2} - t)) \wedge (F(\xi_{1/2} + t) - 1/2)$. It is clear that

$$\Pr(|\xi_{1/2} - \xi_{1/2}| > t) \leq 2e^{-np(t)^2}.$$

□

Note that F_u has a unique median from the assumption that $\xi_{1/2,u}$ is continuous and increasing in a neighborhood of $\mathrm{MED}(F_u)$. Lemma 13 then gives that

$$\Pr(|\hat{\xi}_{1/2,u} - \xi_{1/2,u}| > t) \leq 2e^{-n\kappa_u(t/2,1/2,0)^2}. \tag{A.23}$$

Let $\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, \rho_n)$ for some $\rho_n > 0$ and $\tilde{\xi}_{1/2,u} = \mathrm{MED}(\mathbb{Y}_n^\top u)$.

Lemma 14. *Suppose that* $\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, \rho_n)$, $F_u(\xi_{1/2,u} + t) - 1/2 > \rho_n'$ *and* $1/2 - F_u(\xi_{1/2,u} - t/2) > \rho_n'$, *then*

$$\Pr\left(\left|\tilde{\xi}_{1/2,u} - \hat{\xi}_{1/2,u}\right| \geq t\right) \leq 4e^{-n\kappa_u(t/2,1/2,\rho_n')^2}.$$

Proof. Recall that $F_{n,u}$ is the empirical distribution corresponding to $\mathbb{X}_n^\top u$. Let $\hat{\xi}_{r,u} = F_{n,u}^{-1}(r)$. It holds that

$$\left|\tilde{\xi}_{1/2,u} - \hat{\xi}_{1/2,u}\right| \leq |\hat{\xi}_{1/2,u} - \xi_{1/2,u}| + |\xi_{1/2,u} - \hat{\xi}_{1/2+\rho_n/n,u}| \vee |\xi_{1/2,u} - \hat{\xi}_{1/2-\rho_n/n,u}|.$$

Let

$$A_n = \left\{|\xi_{1/2,u} - \hat{\xi}_{1/2+\rho_n/n,u}| \vee |\xi_{1/2,u} - \hat{\xi}_{1/2-\rho_n/n,u}| \geq t/2\right\}.$$

We then have that

$$\Pr\left(\left|\tilde{\xi}_{1/2,u} - \hat{\xi}_{1/2,u}\right| \geq t\right) \leq \Pr\left(|\hat{\xi}_{1/2,u} - \xi_{1/2,u}| \geq t/2\right) + \Pr\left(A_n\right).$$

Equation (A.23) gives that:

$$\Pr(|\hat{\xi}_{1/2,u} - \xi_{1/2,u}| > t/2) \leq 2e^{-n\kappa_u(t/2,1/2,0)^2}.$$

We must now show that the quantiles close to the median also concentrate around the median. Note that

$$\Pr\left(A_n \geq t/2\right) = \Pr\left(|\xi_{1/2,u} - \hat{\xi}_{1/2+\rho_n/n,u}| > t/2 \cup |\xi_{1/2,u} - \hat{\xi}_{1/2-\rho_n/n,u}| \geq t/2\right).$$

Define the events

$$B_{1,n,u}(t) = \{|\xi_{1/2,u} - \hat{\xi}_{1/2+\rho_n/n,u}| > t\} \qquad \text{and} \qquad B_{2,n,u}(t) = \{|\xi_{1/2,u} - \hat{\xi}_{1/2-\rho_n/n,u}| > t\}.$$

We see that

$$\Pr\left(B_{1,n,u}(t)\right) = \Pr\left(Y_1(t) > n/2 - \rho_n,\ Y_2(t) > n/2 + \rho_n\right),$$

where $Y_1(t)$ and $Y_2(t)$ are the number of points in the projected, univariate sample which are greater than $\xi_{1/2,u} + t$ and less than $\xi_{1/2,u} - t$, respectively. In addition,

$$\Pr\left(B_{2,n,u}(t)\right) = \Pr\left(Y_1(t) > n/2 + \rho_n,\ Y_2(t) > n/2 - \rho_n\right),$$

We then have that

$$\Pr(B_{1,n,u}(t) \cup B_{2,n,u}(t)) \leq \Pr(Y_1(t) > n/2 - \rho_n, Y_2(t) > n/2 - \rho_n).$$

It follows from a Hoeffding inequality that

$$\begin{aligned}
\Pr(Y_1(t) > n/2 - \lfloor\rho_n\rfloor - 1) &= \Pr\left(Y_1(t) - \mathrm{E}\left[Y_1(t)\right] > 1/2 - \rho_n/n - (1 - F_u(\xi_{1/2,u} + t))\right) \\
&= \Pr\left(Y_1(t)/n - \mathrm{E}\left[Y_1(t)\right]/n > F_u(\xi_{1/2,u} + t) - \rho_n/n - 1/2\right) \\
&\leq e^{-n\left(F_u(\xi_{1/2,u}+t) - \frac{\rho_n}{n} - 1/2\right)^2},
\end{aligned}$$

provided that $F_u(\xi_{1/2,u}+t)-1/2>\frac{\rho_n}{n}$. Similarly,

$$\begin{aligned}\Pr(Y_2(t)>n/2-\rho_n)&=\Pr((Y_2(t)-F_u(\xi_{1/2,u}-t))/n>1/2-\rho_n/n-F_u(\xi_{1/2,u}-t))\\&\leq e^{-n\left(1/2-F_u(\xi_{1/2,u}-t)-\frac{\rho_n}{n}\right)^2},\end{aligned}$$

provided that $1/2-F_u(\xi_{1/2,u}-t)>\frac{\rho_n}{n}$. It follows that

$$\Pr(B_{1,n,u}(t)\cup B_{2,n,u}(t))\leq e^{-n\left(F_u(\xi_{1/2,u}+t)-\frac{\rho_n}{n}-1/2\right)^2}+e^{-n\left(1/2-F_u(\xi_{1/2,u}-t)-\frac{\rho_n}{n}\right)^2}\leq 2e^{-n\kappa_u(t/2)^2},$$

and we must have that $F_u(\xi_{1/2,u}+t/2)-1/2>\frac{\rho_n}{n}$ and $1/2-F_u(\xi_{1/2,u}-t/2)>\frac{\rho_n}{n}$. Putting everything together gives that

$$\Pr\left(\left|\text{MED}\left(\mathbb{Y}_n^\top u\right)-\text{MED}\left(\mathbb{X}_n^\top u\right)\right|\geq t\right)\leq 2e^{-n\kappa_u(t/2,1/2,\rho_n')^2}+2e^{-n\kappa_u(t/2,1/2,\rho_n')^2}\leq 4e^{-n\kappa_u(t/2,1/2,\rho_n')^2}.$$

□

Using the proofs of the previous two lemmas, we can show that:

$$\Pr\left(\hat{\xi}_{1/2,u}>2k_3\right)\leq\Pr\left(Y_1(2k_3-\xi_{1/2,u})>n/2\right)\leq e^{-n(F_u(2k_3)-1/2)^2},\tag{A.24}$$

for all u. We now prove similar concentration results for the MAD. First, we review a concentration inequality for the median absolute deviation:

Lemma 15 (Serfling and Mazumder (2009))**.** *The median absolute deviation $\hat{\gamma}$ of a univariate, i.i.d. sample from F, where $\text{MAD}(F)$ is unique, concentrates according to*

$$\Pr(|\hat{\gamma}-\text{MAD}(F)|>t)\leq 2e^{-n\Delta(t)^2},$$

where $\Delta(t)=\min\{a_0,b_0,c_0,d_0\}$ and

$$\begin{aligned}a_0&=(F(\xi_{1/2}+t/2)-1/2)^+,\\b_0&=1/2-F(\xi_{1/2}-t/2),\\c_0&=(F(\xi_{1/2}+\gamma+t/2)-F(\xi_{1/2}-\gamma-t/2)-\lfloor(n+1)/2\rfloor/n)^+,\\d_0&=1/2-[F(\xi_{1/2}+\gamma-t/2)-F(\xi_{1/2}-\gamma+t/2)].\end{aligned}\tag{A.25}$$

Recall we let $\gamma_u=\text{MAD}(F_u)$, let $\hat{\gamma}_u=\text{MAD}(\mathbb{X}_n^\top u)$ and let $\tilde{\gamma}_u=\text{MAD}(\mathbb{Y}_n^\top u)$. Using

Lemma 15 it is easy to see that

$$\Pr(|\hat{\gamma}_u - \gamma_u| > t) \leq 2e^{-n\Delta_u(t,0)^2}. \tag{A.26}$$

We can then show that the median average deviation of two similar samples concentrates:

Lemma 16. *Suppose that $\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, \rho_n)$. Then, for all $u \in S^{d-1}$, it holds that*

$$\Pr\left(|\hat{\gamma}_u - \tilde{\gamma}_u| > t\right) \leq 12e^{-n\Delta_u(t/4,\rho_n')^2}.$$

Proof. Our proof follows the outline of the proof of Theorem 1 of (Serfling and Mazumder, 2009). To this end,

$$\begin{aligned}\Pr\left(|\hat{\gamma}_u - \tilde{\gamma}_u| > t\right) &\leq \Pr\left(|\hat{\gamma}_u - \gamma_u| > t/2\right) + \Pr\left(|\gamma_u - \tilde{\gamma}_u| > t/2\right) \\ &\leq 2e^{-n\Delta_u(t/2,0)^2} + \Pr\left(|\gamma_u - \tilde{\gamma}_u| > t/2\right).\end{aligned}$$

We must only resolve $\Pr\left(|\gamma_u - \tilde{\gamma}_u| > t/2\right)$. Let $G_{n,u}$ be the empirical distribution of $\mathbb{Y}_n^\top u$. Clearly it holds that

$$\begin{aligned}\Pr(|\gamma_u - \tilde{\gamma}_u| > t/2) &= \Pr(\tilde{\gamma}_u > t/2 + \gamma_u) + \Pr(\tilde{\gamma}_u < \gamma_u - t/2) \\ &:= \Pr(\mathrm{LH}) + \Pr(\mathrm{RH}).\end{aligned}$$

We can start with the left hand term. First, note that $\left(G_{n,u}(x) - \frac{\rho_n}{n}\right)^+ \leq F_{n,u}(x) \leq G_{n,u}(x) + \frac{\rho_n}{n}$. Using this, and page 5 of (Serfling and Mazumder, 2009), we have that

$$\begin{aligned}\Pr(\mathrm{LH}) &\leq \Pr\left(\left\lfloor \frac{n+1}{2} \right\rfloor /n > G_{n,u}\left(\gamma_u + \tilde{\gamma}_u + t/2\right) - G_{n,u}\left(\tilde{\gamma}_u - \gamma_u - t/2\right)\right) \\ &\leq \Pr\left(\left\lfloor \frac{n+1}{2} \right\rfloor /n > G_{n,u}(\gamma_u + \xi_{1/2,u} + t/4) - G_{n,u}(\xi_{1/2,u} - \gamma_u - t/4)\right) \\ &\qquad + \Pr\left(|\tilde{\xi}_{1/2,u} - \xi_{1/2,u}| > t/4\right) \\ &\leq \Pr\left(\left\lfloor \frac{n+1}{2} \right\rfloor /n > -2\frac{\rho_n}{n} + F_{n,u}(\gamma_u + \xi_{1/2,u} + t/4) - F_{n,u}(\xi_{1/2,u} - \gamma_u - t/4)\right) \\ &\qquad + 4e^{-n\kappa_u(t/8,1/2,\rho_n')^2} \\ &\leq \Pr\left(n\left(\left\lfloor \frac{n+1}{2} \right\rfloor /n + 2\frac{\rho_n}{n} - \mathrm{E}\left[Z_i\right]\right) > \sum_{k=1}^{n} Z_i - \mathrm{E}\left[Z_i\right]\right) + 4e^{-n\kappa_u(t/8,1/2,\rho_n')^2} \\ &\leq 2e^{-n\Delta_{1,u}(t/4,\rho_n')^2} + 4e^{-n\kappa_u(t/8,1/2,\rho_n')^2},\end{aligned}$$

where $Z_i = \mathbb{1}\left(\gamma_u - \xi_{1/2,u} - t/4 < X_i \leq \gamma_u + \xi_{1/2,u} + t/4\right)$. Similarly, we have that

$$\begin{aligned}
\Pr(\mathrm{RH}) &\leq \Pr\left(\left\lfloor \frac{n+1}{2} \right\rfloor /n \leq G_{n,u}(\gamma_u + \xi_{1/2,u} - t/4) - G_{n,u}(\xi_{1/2,u} - \gamma_u + t/4)\right) \\
&\qquad + 4e^{-n\kappa_u(t/8,1/2,\rho_n')^2} \\
&\leq \Pr\left(\left\lfloor \frac{n+1}{2} \right\rfloor /n \leq 2\frac{\rho_n}{n} + F_{n,u}(\gamma_u + \xi_{1/2,u} - t/4) - F_{n,u}(\xi_{1/2,u} - \gamma_u + t/4)\right) \\
&\qquad + 4e^{-n\kappa_u(t/8,1/2,\rho_n')^2} \\
&\leq 2e^{-n\Delta_{2,u}(t/4,\rho_n')^2} + 4e^{-n\kappa_u(t/8,1/2,\rho_n')^2}.
\end{aligned}$$

Bringing everything together, we have that

$$\begin{aligned}
\Pr(|\hat{\gamma}_u - \tilde{\gamma}_u| > t) &\leq 2e^{-n\Delta_u(t/2,0)^2} + 2e^{-n\Delta_{1,u}(t/4,\rho_n')^2} \\
&\qquad + 2e^{-n\Delta_{2,u}(t/4,\rho_n')^2} + 8e^{-n\kappa_u(t/8,1/2,\rho_n')^2} \leq 12e^{-n\Delta_u(t/4,\rho_n')^2}.
\end{aligned}$$

□

Lemma 17. *It holds that*

$$\begin{aligned}
\Pr(\mathrm{MAD}(\mathbb{X}_n^\top u) < k_1/2) &\leq 2e^{-n\Delta_u(\gamma_u - k_1/2,0)^2} \\
\Pr(\mathrm{MAD}(\mathbb{X}_n^\top u) > 2k_2) &\leq 2e^{-n\Delta_u(2k_2 - \gamma_u,0)^2} \\
\Pr(\mathrm{MAD}(\mathbb{Y}_n^\top u) < k_1/2) &\leq 2e^{-n\Delta_u((\gamma_u - k_1/2)/4,\rho_n')^2} \\
\Pr(\mathrm{MAD}(\mathbb{Y}_n^\top u) > 2k_2) &\leq 2e^{-n\Delta_u((2k_2 - \gamma_u)/4,\rho_n')^2}.
\end{aligned}$$

Proof. The result follows from Lemma 15 and the proof of Lemma 16. □

We now proceed with the proof of Theorem 16. We will write $A_\eta\left(O_1\left(x; F_n\right); \mathbb{X}_n\right) = \widehat{A}_{n,\eta}$ for brevity. We want to show that

$$\Pr\left(\widehat{A}_{n,\eta} \leq 1 + \frac{\log(2/\delta_n) - W_1}{\epsilon_n}\right) \tag{A.27}$$

is small. To this end, note that

$$\Pr(|W_1| > \log(2/\delta_n)) = 2e^{-\log(2/\delta_n)} = \delta_n,$$

from the properties of the Laplace distribution. We can then write

$$\begin{aligned}\Pr\left(\widehat{A}_{n,\eta} \leq 1 + \frac{\log(2/\delta_n) - W_1}{\epsilon_n}\right) &\leq \Pr\left(\widehat{A}_{n,\eta} \leq 1 + \frac{\log(2/\delta_n) - W_1}{\epsilon_n}, W_1 > -\log(2/\delta_n)\right) \\ &\quad + \Pr\left(\widehat{A}_{n,\eta} \leq 1 + \frac{\log(2/\delta_n) - W_1}{\epsilon_n}, W_1 < -\log(2/\delta_n)\right) \\ &\leq \Pr\left(\widehat{A}_{n,\eta} \leq 1 + 2\frac{\log(2/\delta_n)}{\epsilon_n}\right) + \delta_n. \end{aligned} \tag{A.28}$$

Now, let $\rho_n = \lfloor 2\frac{\log(2/\delta_n)}{\epsilon_n} \rfloor + 1$ and we want to show that

$$\Pr\left(\widehat{A}_{n,\eta} \leq \rho_n\right) \tag{A.29}$$

is small, which, together with (A.28) implies (A.27) is small. To this end, it holds that

$$\begin{aligned}\Pr\left(\widehat{A}_{n,\eta} \leq \rho_n\right) &= \Pr\left(\bigcup_{j=1}^{\rho_n} \left\{ \sup_{\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, j)} |O_1(x; \mathbb{X}_n) - O_1(x; \mathbb{Y}_n)| \geq \eta \right\}\right) \\ &\leq \Pr\left(\sup_{\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, \rho_n)} |O_1(x; \mathbb{X}_n) - O_1(x; \mathbb{Y}_n)| \geq \eta \right), \end{aligned} \tag{A.30}$$

where the last line follows from the fact that

$$\left\{ \sup_{\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, j+1)} |O_1(x; \mathbb{X}_n) - O_1(x; \mathbb{Y}_n)| \geq \eta \right\} \subset \left\{ \sup_{\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, j)} |O_1(x; \mathbb{X}_n) - O_1(x; \mathbb{Y}_n)| \geq \eta \right\},$$

for $j \in \{1, \ldots, \rho_n\}$. We now must only show that

$$\Pr\left(\sup_{\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, \rho_n)} |O_1(x; \mathbb{X}_n) - O_1(x; \mathbb{Y}_n)| \geq \eta \right), \tag{A.31}$$

is small. Combining this with (A.30) and (A.29) implies that (A.27) holds from the previous argument. Define

$$O^u(x; \mathbb{X}_n) = \frac{|x^\top u - \hat{\xi}_{1/2,u}|}{\hat{\gamma}_u}$$

and note that $O_1(x; \mathbb{X}_n) = \sup_{u \in \mathbb{U}_M} O^u(x; \mathbb{X}_n)$. We first look at $|O^u(x; \mathbb{X}_n) - O^u(x; \mathbb{Y}_n)|$.

Let

$$c_{2,u} = \min\left\{F_u(2k_3) - 1/2, \Delta_u((\gamma_u - k_1/2)/4, \rho_n')\Delta_u((2k_2 - \gamma_u)/4, \rho_n')\right\}.$$

As a result of Lemma 17 and (A.24), it holds with probability $\geq 1 - c_1 e^{-nc_{2,u}^2}$ that

$$\begin{aligned}
|O^u(x; \mathbb{X}_n) - O^u(x; \mathbb{Y}_n)| &= \left|\frac{|x^\top u - \hat{\xi}_{1/2,u}|}{\hat{\gamma}_u} - \frac{|x^\top u - \tilde{\xi}_{1/2,u}|}{\tilde{\gamma}_u}\right| \\
&\leq (2k_1)^{-2}\left|\tilde{\gamma}_u|x^\top u - \hat{\xi}_{1/2,u}| - \hat{\gamma}_u|x^\top u - \tilde{\xi}_{1/2,u}|\right| \\
&\leq (2k_1)^{-2}|x^\top u|\left|\tilde{\gamma}_u - \hat{\gamma}_u\right| + (2k_1)^{-2}\left|\tilde{\gamma}_u\hat{\xi}_{1/2,u} - \hat{\gamma}_u\tilde{\xi}_{1/2,u}\right| \\
&\leq (2k_1)^{-2}(|x^\top u| + 2k_3)\left|\tilde{\gamma}_u - \hat{\gamma}_u\right| + (2k_1)^{-2}2k_2\left|\hat{\xi}_{1/2,u} - \tilde{\xi}_{1/2,u}\right|, \\
&\leq (2k_1)^{-2}\left((\|x\|^2 + 2k_3)\left|\tilde{\gamma}_u - \hat{\gamma}_u\right| + 2k_2\left|\hat{\xi}_{1/2,u} - \tilde{\xi}_{1/2,u}\right|\right) \\
&\coloneqq (2k_1)^{-2}\mathcal{E}_n.
\end{aligned}$$

Now write

$$\Pr\left(\mathcal{E}_n > 4k_1^2\eta\right) \leq \Pr\left(\left|\tilde{\gamma}_u - \hat{\gamma}_u\right| > \frac{2k_1^2\eta}{\|x\|^2 + 2k_3}\right) + \Pr\left(\left|\hat{\xi}_{1/2,u} - \tilde{\xi}_{1/2,u}\right| > \frac{k_1^2\eta}{k_2}\right).$$

Looking at each term individually:

$$\Pr\left(\left|\tilde{\gamma}_u - \hat{\gamma}_u\right| > \frac{2k_1^2\eta}{\|x\|^2 + 2k_3}\right) \leq 12e^{-n\Delta_u\left(\frac{k_1^2\eta}{2\|x\|^2 + 4k_3}, \rho_n'\right)^2},$$

holds from Lemma 16 and

$$\Pr\left(\left|\hat{\xi}_{1/2,u} - \tilde{\xi}_{1/2,u}\right| > \frac{k_1^2\eta}{k_2}\right) \leq 4e^{-n\kappa_u\left(\frac{k_1^2\eta}{2k_2}, 1/2, \rho_n'\right)^2},$$

holds from Lemma 14. It then immediately follows that

$$\Pr\left(\mathcal{E}_n > k_1^2\eta\right) \leq 16e^{-n\min\left\{\kappa_u\left(\frac{k_1^2\eta}{2k_2},1/2,\rho_n'\right),\Delta_u\left(2\frac{k_1^2\eta}{2\|x\|^2+4k_3},\rho_n'\right)\right\}^2} := c_3e^{-nh_{1,u}(\epsilon,\delta,\eta,x)^2}.$$

It follows that

$$\Pr(|O^u(x;\mathbb{X}_n) - O^u(x;\mathbb{Y}_n)| \geq \eta) \leq c_1e^{-nc_{2,u}^2} + c_3e^{-nh_{1,u}(\epsilon,\delta,\eta,x)^2}.$$

Note that

$$\left|\sup_{u\in\mathbb{U}_M} O_2^u(x;\mathbb{X}_n) - \sup_{u\in\mathbb{U}_M} O_2^u(x;\mathbb{Y}_n)\right| \leq 2\sup_{u\in\mathbb{U}_M}\left|O_2^u(x;\mathbb{X}_n) - O_2^u(x;\mathbb{Y}_n)\right|.$$

Let $c_2 = \inf_u c_{2,u}$ and $h_1(\epsilon,\delta,\eta,x) = \inf_u h_{1,u}(\epsilon,\delta,\eta,x)$, then a Bonferroni inequality gives that

$$\Pr(|O(x;\mathbb{X}_n) - O(x;\mathbb{Y}_n)| \geq \eta) \leq m\cdot c_1e^{-nc_2^2} + m\cdot c_3e^{-nh_1(\epsilon,\delta,\eta,x)^2}.$$

We must now prove the bound on $\Pr(|\widetilde{O}_1(x;F_n) - O(x;F)| > t)$. First, it is clear that

$$\begin{aligned}\Pr(|\widetilde{O}_1(x;F_n) - O_1(x;F)| > t) &\leq \Pr(|O_1(x;F_n) - O_1(x;F)| > t/2) + \Pr\left(|W| > \frac{t\epsilon}{2\eta}\right)\\ &\leq \Pr(|O_1(x;F_n) - O_1(x;F)| > t/2) + e^{-\frac{t\epsilon}{2\eta}}.\end{aligned}$$

We just need to prove that the outlyingness concentrates. From Lemma 17, we have with probability $1 - c_1e^{-n(\Delta_u(\gamma_u-k_1/2,0)\vee\Delta_u(2k_2-\gamma_u,0))^2}$, that

$$|O_1^u(x;F_n) - O_1^u(x;F)| \leq (2k_1)^{-2}\left((\|x\|^2+2k_3)\left|\gamma_u - \hat{\gamma}_u\right| + 2k_2\left|\hat{\xi}_{1/2,u} - \xi_{1/2,u}\right|\right).$$

We also have that

$$\Pr\left(\left|\hat{\xi}_{1/2,u} - \xi_{1/2,u}\right| \geq \frac{2k_1^2t}{k_2}\right) \leq 2e^{-n\kappa_u\left(\frac{2k_1^2t}{k_2}.1/2,0\right)^2},$$

and that

$$\Pr\left(|\gamma_u - \hat{\gamma}_u| \geq \frac{2k_1^2t}{\|x\|^2+2k_3}\right) \leq ce^{-n\Delta_u\left(\frac{2k_1^2t}{\|x\|^2+2k_3},0\right)^2}.$$

It follows that

$$\Pr(|O_1(x;F_n) - O_1(x;F)| > t/2) \leq m \cdot c_1 e^{-c_2 n h_2(t,x)^2},$$

where

$$h_2 = \inf_u h_{2,u}(t,x) = \inf_u \min\left\{\kappa_u\left(\frac{2k_1^2 t}{k_2}.1/2, 0\right), \Delta_u\left(\frac{2k_1^2 t}{\|x\|^2 + 2k_3}, 0\right)\right\}.$$

So, combining all of the results, we have that

$$\Pr(|\widetilde{O}_1(x;F_n) - O_1(x;F)| \geq t) \leq m \cdot c_1 e^{-c_2 n} + m \cdot c_3 e^{-c_4 n h_2(t,x)^2} + e^{-\frac{t\epsilon}{2\eta}}.$$

□

Proof of Theorem 17. We follow the same outline as that of the proof of Theorem 16. Let $\tau_u = \text{IQR}(F_u)$, $\hat{\tau}_u = \text{IQR}(\mathbb{X}_n^\top u)$ and $\tilde{\tau}_u = \text{IQR}(\mathbb{Y}_n^\top u)$ for brevity. We require upper bounds on the following probabilities:

$$\Pr\left(|\hat{\tau}_u - \tau_u| > t\right), \ \Pr\left(|\hat{\tau}_u - \tilde{\tau}_u| > t\right), \ \Pr\left(\hat{\tau}_u > 2k_2\right), \text{ and } \Pr\left(\hat{\tau}_u < 2k_1\right).$$

We start by bounding the deviation of $\hat{\xi}_{r,u}$ around $\xi_{r,u}$:

Lemma 18. *Suppose that the quantile $F_u^{-1}(r) = \xi_{r,u}$ is unique. It holds that*

$$\Pr(|\hat{\xi}_{r,u} - \xi_{r,u}| > t) \leq e^{-2n(F_u(\xi_{r,u}+t)-r)^2} + e^{-2n(\lfloor nr \rfloor/n - F_u(\xi_{r,u}-t))^2} \leq 2e^{-2n\kappa_u(t,r,0)^2}.$$

Proof. We omit the dependence on u for brevity. It is easy to see that

$$\Pr(|\hat{\xi}_r - \xi_r| > t) = \Pr(\hat{\xi}_r > t + \xi_r) + \Pr(\hat{\xi}_r < \xi_r - t).$$

Starting with the first term,

$$\Pr(\hat{\xi}_r > t + \xi_r) = \Pr(Z_{t,r} > n - \lfloor nr \rfloor),$$

where Z_t is the number of points greater than $t + \xi_r$. A Hoeffding inequality gives that,

$$\begin{aligned}
\Pr(Z_{t,r} > n - \lfloor nr \rfloor) &= \Pr\left(Z_{t,r} - n + nF(\xi_r + t) > nF(\xi_r + t) - \lfloor nr \rfloor\right) \\
&\leq e^{-2n(F(\xi_r+t) - \lfloor nr \rfloor/n)^2} \\
&\leq e^{-2n(F(\xi_r+t)-r)^2}.
\end{aligned}$$

Applying the same technique to the right hand term gives that

$$\Pr(\hat{\xi}_r < \xi_r - t) \leq e^{-2n(\lfloor nr \rfloor / n - F(\xi_r - t))^2}.$$

□

Using Lemma 18, we can bound the deviation of the inter-quartile range:

Lemma 19. *Suppose that $\xi_{1/4,u}$, $\xi_{3/4,u}$ are the unique $1/4$ and $3/4$ quantiles of F_u. Then it holds that*

$$\Pr(|\hat{\tau}_u - \tau_u| > t) \leq 4e^{-2n[\kappa_u(t/2,3/4,0) \wedge \kappa(t/2,1/4,0)]^2}.$$

Proof. Again, omit the dependence on u. It follows from the standard and the reverse triangle inequality that

$$|\hat{\tau} - \tau| \leq |\hat{\xi}_{3/4} - \hat{\xi}_{1/4} - \xi_{3/4} + \xi_{1/4}| \leq |\hat{\xi}_{3/4} - \xi_{3/4}| + |\hat{\xi}_{1/4} - \xi_{1/4}|.$$

Using this fact and Lemma 18, we see that

$$\begin{aligned}
\Pr(|\hat{\tau} - \tau| > t) &\leq \Pr(|\hat{\xi}_{3/4} - \xi_{3/4}| > t/2) + \Pr(|\hat{\xi}_{1/4} - \xi_{1/4}| > t/2) \\
&\leq 2e^{-2n\kappa(t/2,3/4,0)^2} + 2e^{-2n\kappa(t/2,1/4,0)^2} \\
&\leq 4e^{-2n[\kappa(t/2,3/4,0) \wedge \kappa(t/2,1/4,0)]^2}.
\end{aligned}$$

□

We now aim to bound $\Pr\left(|\hat{\tau}_u - \tilde{\tau}_u| > t\right)$. To do this, we need the following intermediate lemma:

Lemma 20. *Suppose that $\xi_{r,u}$ is a unique r-quantile of F_u and $c \in \mathbb{R}$, $t > 0$ such that $0 < r + c < 1$, $\lfloor nr \rfloor / n - F_u(\xi_{r+c} - t) > 0$ and $F_u(\xi_{r+c} + t) - \lfloor nr \rfloor / n > 0$. Then it holds that*

$$\Pr(|\xi_{r,u} - \hat{\xi}_{r+c,u}| > t) \leq 2e^{-2n\kappa_u(t,r,c)^2}.$$

Proof. Omit the dependence on u for brevity. It is easy to see that

$$\Pr(|\hat{\xi}_{r+c} - \xi_r| > t) = \Pr(\hat{\xi}_{r+c} > t + \xi_r) + \Pr(\hat{\xi}_{r+c} < \xi_r - t).$$

Starting with the first term,

$$\Pr(\hat{\xi}_{r+c} > t + \xi_r) = \Pr(Z_{t,r} > n - \lfloor n(r+c) \rfloor),$$

where $Z_{t,r}$ is the number of points greater than $t + \xi_r$. A Hoeffding inequality gives that,

$$\begin{aligned}\Pr(Z_{t,r} > n - \lfloor n(r+c) \rfloor) &= \Pr\left(Z_{t,r} - n + nF(\xi_r + t) > nF(\xi_r + t) - \lfloor n(r+c) \rfloor\right) \\ &\leq e^{-2n(F(\xi_r+t) - \lfloor nn(r+c) \rfloor/n)^2} \\ &\leq e^{-2n(F(\xi_r+t)-r-c))^2}.\end{aligned}$$

Applying the same technique to the remaining term, we get that

$$\Pr(\hat{\xi}_{r+c} < \xi_r - t) \leq e^{-2n(\lfloor n(r+c) \rfloor/n - F(\xi_r - t))^2}.$$

□

We now bound $\Pr(|\hat{\tau}_u - \tilde{\tau}_u| > t)$:

Lemma 21. *Suppose that $\mathbb{Y}_n \in \mathcal{D}(\mathbb{X}_n, \rho_n)$. It then holds that*

$$\Pr(|\hat{\tau}_u - \tilde{\tau}_u| > t) \leq 16 e^{-2nh_{2,u}(t)^2},$$

where

$$h_{2,u}(t) = \min\left\{\kappa_u\left(\frac{t}{4}, \frac{3}{4}, -\frac{\rho_n}{n}\right), \kappa_u\left(\frac{t}{4}, \frac{3}{4}, \frac{\rho_n}{n}\right), \kappa_u\left(\frac{t}{4}, \frac{1}{4}, -\frac{\rho_n}{n}\right), \kappa_u\left(\frac{t}{4}, \frac{1}{4}, \frac{\rho_n}{n}\right)\right\}.$$

Proof. Once again, omit the dependence on u. Note that

$$|\hat{\xi}_{3/4-\rho_n/n} - \hat{\xi}_{1/4+\rho_n/n}| \leq \tilde{\tau} \leq |\hat{\xi}_{3/4+\rho_n/n} - \hat{\xi}_{1/4-\rho_n/n}|. \tag{A.32}$$

We see that

$$\Pr(|\hat{\tau} - \tilde{\tau}| > t) \leq \Pr(\hat{\tau} - \tilde{\tau} > t) + \Pr(\tilde{\tau} - \hat{\tau} > t).$$

Starting with the left hand term, using (A.32) and both the standard and the reverse

triangle inequality,

$$\begin{aligned}\Pr(\hat{\tau}-\tilde{\tau}>t) &\leq \Pr(\hat{\tau}-|\hat{\xi}_{3/4-\rho_n/n}-\hat{\xi}_{1/4+\rho_n/n}|>t)\\ &\leq \Pr(|\hat{\xi}_{3/4}-\hat{\xi}_{3/4-\rho_n/n}|-|\hat{\xi}_{1/4}-\hat{\xi}_{1/4+\rho_n/n}|>t)\\ &\leq \Pr(|\hat{\xi}_{3/4}-\hat{\xi}_{3/4-\rho_n/n}|>t/2)+\Pr(|\hat{\xi}_{1/4}-\hat{\xi}_{1/4+\rho_n/n}|>t/2).\end{aligned}$$

Starting with the left-hand term, Lemma 20 gives that

$$\begin{aligned}\Pr(|\hat{\xi}_{3/4}-\hat{\xi}_{3/4-\rho_n/n}|>t/2) &\leq 2e^{-2n\kappa\left(t/4,3/4,-\frac{\rho_n}{n}\right)^2}+2e^{-2n\kappa(t/4,3/4,0)^2}\\ &\leq 4e^{-2n\kappa\left(t/4,3/4,-\frac{\rho_n}{n}\right)^2},\end{aligned}$$

provided that $\lfloor 3n/4\rfloor/n - F(\xi_{3/4-\frac{\rho_n}{n}}-t)>0$ and $F(\xi_{3/4-\frac{\rho_n}{n}}+t)-\lfloor 3n/4\rfloor/n>0$. It is easy to follow the same logic for the remaining terms, such that

$$\Pr(|\hat{\tau}-\tilde{\tau}|>t)\leq 16e^{-2nh_2(t)^2},$$

where

$$h_2(t)=\min\left\{\kappa\left(\frac{t}{4},\frac{3}{4},-\frac{\rho_n}{n}\right),\kappa\left(\frac{t}{4},\frac{3}{4},\frac{\rho_n}{n}\right),\kappa\left(\frac{t}{4},\frac{1}{4},-\frac{\rho_n}{n}\right),\kappa\left(\frac{t}{4},\frac{1}{4},\frac{\rho_n}{n}\right)\right\}.$$

□

Lastly, we have the following Lemma:

Lemma 22. *Let*

$$h_{1,u}(t)=\left[\kappa_u(t/2,3/4,0)\wedge\kappa_u(t/2,1/4,0)\right].$$

It follows that

$$\begin{aligned}\Pr\left(\hat{\tau}_u>2k_2\right) &\leq 4e^{-2nh_{1,u}(k_2)^2}\\ \Pr\left(\hat{\tau}_u<k_1/2\right) &\leq 4e^{-2nh_{1,u}(k_1)^2}\\ \Pr\left(\tilde{\tau}_u<2k_1\right) &\leq 8e^{-2nh_{2,u}(k_1)^2}.\end{aligned}$$

Proof. The first two statements follow from Lemma 19:

$$\Pr\left(\hat{\tau}_u>k_2\right)=\Pr\left(\hat{\tau}_u-\tau_u>k_2-\tau_u\right)\leq 4e^{-2nh_{2,u}(k_2-\tau_u)^2}$$

and

$$\Pr\left(\hat{\tau}_u < k_1\right) = \Pr\left(\tau_u - \hat{\tau}_u > \tau_u - k_1\right) \leq 4e^{-2nh_{2,u}(\tau_u - k_1)^2}.$$

For the last statement, we use Lemma 20

$$\Pr\left(\tilde{\tau}_u < k_1\right) = \Pr\left(\tau_u - \tilde{\tau}_u > \tau_u - k_1\right) \leq 8e^{-2nh_{1,u}(2(\tau_u - k_1))^2}.$$

□

We are now in a position to apply the techniques of the proof of Theorem 16. Define

$$O^u(x; \mathbb{X}_n) = \frac{|x^\top u - \hat{\xi}_{1/2,u}|}{\hat{\tau}_u}$$

and note that $O_2(x; \mathbb{X}_n) = \sup_{u \in \mathbb{U}_M} O^u(x; \mathbb{X}_n)$. From (A.24) and Lemma 22, it holds with probability $\geq 1 - c_1 e^{-c_2 n}$ that

$$|O^u(x; \mathbb{X}_n) - O^u(x; \mathbb{Y}_n)| \leq (2k_1)^{-2} \mathcal{E}'_n,$$

where

$$\mathcal{E}'_n = \left((\|x\|^2 + 2k_3) \left| \tilde{\tau}_u - \hat{\tau}_u \right| + 2k_2 \left| \hat{\xi}_{1/2,u} - \tilde{\xi}_{1/2,u} \right| \right)$$

We have that

$$\Pr\left(\mathcal{E}'_n > 4k_1^2\eta\right) \leq \Pr\left(\left| \tilde{\tau}_u - \hat{\tau}_u \right| > \frac{2k_1^2\eta}{\|x\|^2 + 2k_3} \right) + \Pr\left(\left| \hat{\xi}_{1/2,u} - \tilde{\xi}_{1/2,u} \right| > \frac{k_1^2\eta}{k_2} \right).$$

Looking at each term individually, it follows from Lemma 21 that

$$\Pr\left(\left| \tilde{\tau}_u - \hat{\tau}_u \right| > \frac{2k_1^2\eta}{\|x\|^2 + 2k_3} \right) \leq 16e^{-2nh_{2,u}\left(\frac{2k_1^2\eta}{\|x\|^2 + 2k_3}\right)^2},$$

and from Lemma 14 that

$$\Pr\left(\left| \hat{\xi}_{1/2,u} - \tilde{\xi}_{1/2,u} \right| > \frac{4k_1^2\eta}{k_2} \right) \leq 4e^{-n\kappa_u\left(\frac{k_1^2\eta}{2k_2}, 1/2, \rho'_n\right)^2}.$$

It then immediately follows that

$$\Pr\left(\mathcal{E}'_n > k_1^2\eta\right) \leq 20e^{-n\left[\kappa_u\left(\frac{2k_1^2\eta}{k_2},1/2,\rho'_n\right)\wedge h_{2,u}\left(\frac{8k_1^2\eta}{\|x\|^2+2k_3}\right)\right]^2}.$$

It follows that

$$\Pr(|O^u(x;\mathbb{X}_n) - O^u(x;\mathbb{Y}_n)| \geq \eta) \leq c_1e^{-c_2n} + C_1e^{-n\left[\kappa_u\left(\frac{k_1^2\eta}{2k_2},1/2,\rho'_n\right)\wedge h_{2,u}\left(\frac{2k_1^2\eta}{\|x\|^2+2k_3}\right)\right]^2}.$$

Recall that

$$|\sup O^u(x;\mathbb{X}_n) - \sup O^u(x;\mathbb{Y}_n)| \leq 2\sup|O^u(x;\mathbb{X}_n) - O^u(x;\mathbb{Y}_n)|,$$

since O is continuous in u. It then follows from a simple Bonferroni inequality that

$$\Pr(|O_2(x;\mathbb{X}_n) - O_2(x;\mathbb{Y}_n)| \geq \eta) \leq m\cdot c_1e^{-c_2n} + m\cdot C_1e^{-n\inf_u\left[\kappa_u\left(\frac{k_1^2\eta}{2k_2},1/2,\rho'_n\right)\wedge h_{2,u}\left(\frac{2k_1^2\eta}{\|x\|^2+2k_3}\right)\right]^2}.$$

We must now prove the bound on $\Pr(|\widetilde{O}_2(x;F_n) - O_2(x;F)| > t)$. First, it is clear that

$$\begin{aligned}\Pr(|\widetilde{O}_2(x;F_n) - O_2(x;F)| > t) &\leq \Pr(|O_2(x;F_n) - O_2(x;F)| > t/2) + \Pr\left(|W| > \frac{t\epsilon}{2\eta}\right)\\ &\leq \Pr(|O_2(x;F_n) - O_2(x;F)| > t/2) + e^{-\frac{t\epsilon}{2\eta}}.\end{aligned}$$

We just need to prove that the outlyingness concentrates. From Lemma 22, we have with probability $\geq 1 - c_1e^{-c_2n}$, that

$$|O^u(x;F_n) - O^u(x;F)| \leq (2k_1)^{-2}\left((\|x\|^2+2k_3)\left|\tau_u - \hat{\tau}_u\right| + 2k_2\left|\hat{\xi}_{1/2,u} - \xi_{1/2,u}\right|\right).$$

It then follows from Lemma 19 that

$$\Pr\left(\left|\tau_u - \hat{\tau}_u\right| \geq \frac{2k_1^2t}{\|x\|^2+2k_3}\right) \leq 4e^{-2n\inf_u h_{1,u}\left(\frac{2k_1^2t}{\|x\|^2+2k_3}\right)^2}.$$

Lemma 13 gives that

$$\Pr\left(\left|\hat{\xi}_{1/2,u} - \xi_{1/2,u}\right| \geq \frac{k_1^2t}{k_2}\right) \leq 2e^{-n\inf_u\kappa_u\left(\frac{k_1^2t}{k_2},1/2,0\right)}.$$

We can now conclude via a Bonferroni inequality that

$$\Pr(|O_2(x;F_n) - O_2(x;F)| > t/2) \leq m \cdot c_1 e^{-nc_2} + m \cdot C_1 e^{-n\left(\inf_u\left[h_{1,u}\left(\frac{2k_1^2 t}{\|x\|^2+2k_3}\right)\wedge\kappa_u\left(\frac{k_1^2 t}{k_2},1/2,0\right)\right]\right)^2}.$$

So, combining all of the results, we have that

$$\Pr(|\widetilde{O}_2(x;F_n) - O_2(x;F)| > t) \leq m\cdot c_1 e^{-nc_2} + m\cdot C_1 e^{-n\left(\inf_u\left[h_{1,u}\left(\frac{2k_1^2 t}{\|x\|^2+2k_3}\right)\wedge\kappa_u\left(\frac{k_1^2 t}{k_2},1/2,0\right)\right]\right)^2} + e^{-\frac{t\epsilon}{2\eta}}.$$

□

Proof of Remark 7. First,

$$\Pr\left(k^* \leq 1 + 2\frac{\log(2/\delta_n)}{\epsilon_n}\right) = \delta_n,$$

and note that both

$$W_{1,u}(k^*) = \widehat{O}^u(x) - \mathrm{lo}(\widehat{O}^u(x), k^*) \qquad \text{and} \qquad W_{2,u}(k^*) = \mathrm{up}(\widehat{O}^u(x), k^*) - \widehat{O}^u(x)$$

are increasing in k^*. As such, letting $\rho_n = 1 + 2\frac{\log(2/\delta_n)}{\epsilon_n}$, we have that

$$\Pr\left(W_{1,u}(k^*) \vee W_{2,u}(k^*) \geq \eta\right) \leq \delta_n + \Pr\left(W_{1,u}(\rho_n) \geq \eta\right) + \Pr\left(W_{2,u}(\rho_n) \geq \eta\right).$$

Starting with $W_{1,u}(\rho_n)$ and letting $\hat{m}(u,\rho_n)$ be such that $\mathrm{lo}(\mathrm{MED}, u, k^*) = |x^\top u - \hat{m}(u,\rho_n)|$, we see that:

$$\begin{aligned}
W_{1,u}(\rho_n) &= \frac{|x^\top u - \hat{\xi}_{1/2,u}|}{\hat{\tau}_u} - \frac{\mathrm{lo}\left(\mathrm{MED}, u, \rho_n\right)}{\max \mathcal{E}\left(u,\rho_n\right)} \\
&\leq \frac{|(\max \mathcal{E}\left(u,\rho_n\right) - \hat{\tau}_u)x^\top u + \hat{\tau}_u \hat{m}(u,\rho_n) - \max \mathcal{E}\left(u,\rho_n\right)\hat{\xi}_{1/2,u}|}{\hat{\tau}_u \max \mathcal{E}\left(u,\rho_n\right)} \\
&\leq \frac{(\max \mathcal{E}\left(u,\rho_n\right) - \hat{\tau}_u)|x^\top u - \hat{\xi}_{1/2,u}| + \hat{\tau}_u(\hat{m}(u,\rho_n) - \hat{\xi}_{1/2,u})}{\hat{\tau}_u \max \mathcal{E}\left(u,\rho_n\right)} \\
&\leq \frac{(\max \mathcal{E}\left(u,\rho_n\right) - \hat{\tau}_u)|x^\top u - \hat{\xi}_{1/2,u}| + \hat{\tau}_u(\hat{m}(u,\rho_n) - \hat{\xi}_{1/2,u})}{\hat{\tau}_u^2} \\
&\leq \frac{1}{4k_1^2}\left[(\|x\| + |2k_3|)(\max \mathcal{E}\left(u,\rho_n\right) - \hat{\tau}_u) + 2k_2(\hat{m}(u,\rho_n) - \hat{\xi}_{1/2,u})\right],
\end{aligned}$$

where the last inequality holds with probability $\geq 1 - c_1 e^{-c_2 n}$. This follows from Lemma 22 and (A.24). In order to show the remaining quantity is small, we must bound the following probabilities:

$$\Pr(B_1) = \Pr\left(\max \mathcal{E}\left(u, \rho_n\right) - \hat{\tau}_u > \frac{2k_1^2\eta}{\|x\| + |2k_3|}\right)$$
$$\Pr(B_2) = \Pr\left(\hat{m}(u, \rho_n) - \hat{\xi}_{1/2,u} \geq \frac{k_1^2\eta}{k_2}\right).$$

Stating with the first term:

$$\Pr(\max \mathcal{E}\left(u, \rho_n\right) - \hat{\tau}_u > t) \leq 16 e^{-2nh_{2,u}\left(\frac{2k_1^2\eta}{\|x\|+|2k_3|}\right)^2},$$

from Lemma 21. For the second term, we have that

$$\Pr(|\hat{m}(u, \rho_n) - \hat{\xi}_{1/2,u}| > t) \leq 4 e^{-n\hat{p}_u\left(\frac{k_1^2\eta}{2k_2}\right)^2},$$

which follows from Lemma 14.

Now, we see that

$$\Pr(B_1) + \Pr(B_2) \leq K_1 e^{-K_2 n[h_{2,u}\left(\frac{2k_1^2\eta}{\|x\|+|2k_3|}\right) \wedge \hat{p}_u\left(\frac{k_1^2\eta}{2k_2}\right)]^2},$$

for some absolute constants K_1, K_2. Putting it all together,

$$\Pr(W_{1,u}(\rho_n) \geq \eta) = c_1 e^{-c_2 n} + K_1 e^{-K_2 n[h_{2,u}\left(\frac{2k_1^2\eta}{\|x\|+|2k_3|}\right) \wedge \hat{p}_u\left(\frac{k_1^2\eta}{2k_2}\right)]^2},$$

We can make the same arguments for $W_{2,u}(\rho_n)$, resulting in

$$\Pr(W_{2,u}(\rho_n) \geq \eta) = c_1 e^{-c_2 n} + K_1 e^{-K_2 n[h_{2,u}\left(\frac{2k_1^2\eta}{\|x\|+|2k_3|}\right) \wedge \hat{p}_u\left(\frac{k_1^2\eta}{2k_2}\right)]^2},$$

Now, using a Bonferroni inequality we can write

$$\Pr\left(W_1 \vee W_2 \geq \eta\right) \leq \delta_n + m \cdot c_1 e^{-c_2 n} + m \cdot K_1 e^{-K_2 n[h_{2,u}\left(\frac{2k_1^2\eta}{\|x\|+|2k_3|}\right) \wedge \hat{p}_u\left(\frac{k_1^2\eta}{2k_2}\right)]^2},$$

□

Proof of Differential Privacy of Mechanism 6. The proof has the same outline as that of (Brunel and Avella-Medina, 2020), as well as the proof that the exponential mechanism is differentially private, which can be found in (McSherry and Talwar, 2007; Dwork and Roth, 2014). First, assume that it holds $|\phi_{\mathbb{X}_n}(x) - \phi_{\mathbb{Y}_n}(x)| \leq \eta \ \forall x$, then

$$\begin{aligned} f_{\mathbb{X}_n}(x)/f_{\mathbb{Y}_n}(x) &= \frac{\exp(-\phi_{\mathbb{X}_n}(x)\frac{\epsilon/2}{\eta})}{\exp(-\phi_{\mathbb{Y}_n}(x)\frac{\epsilon/2}{\eta})} \frac{\int \exp(-\phi_{\mathbb{Y}_n}(x)\frac{\epsilon/2}{\eta})dx}{\int \exp(-\phi_{\mathbb{X}_n}(x)\frac{\epsilon/2}{\eta})dx} \\ &\leq e^{\epsilon/2} \frac{\int \exp(-\phi_{\mathbb{Y}_n}(x)\frac{\epsilon/2}{\eta})dx}{\int \exp(-\phi_{\mathbb{X}_n}(x)\frac{\epsilon/2}{\eta})dx} \\ &\leq e^{\epsilon/2} e^{\epsilon/2} \frac{\int \exp(-\phi_{\mathbb{X}_n}(x)\frac{\epsilon/2}{\eta})dx}{\int \exp(-\phi_{\mathbb{X}_n}(x)\frac{\epsilon/2}{\eta})dx} \\ &= e^{\epsilon}. \end{aligned}$$

Note that, for $B \in \mathscr{B}(\mathbb{R}^d)$ (the Borel sets with respect to $\mathbb{R}^d$) this implies that

$$\Pr(\widehat{T}(\mathbb{X}_n) \in B) \leq e^{\epsilon} \Pr(\widehat{T}(\mathbb{Y}_n) \in B). \tag{A.33}$$

It follows from Brunel and Avella-Medina (2020) that $A_\eta(\phi_{\mathbb{X}_n}; \mathbb{X}_n)$ has global sensitivity equal to 1, since changing one point can at most change the breakdown by 1. Then

$$\begin{aligned} \Pr\left(\widetilde{T}(\mathbb{X}_n) \in B\right) &= \Pr\left(A_\eta(\phi_{\mathbb{X}_n}; \mathbb{X}_n) + \frac{1}{\epsilon}V \geq 1 + \frac{\log(2/\delta)}{\epsilon}, \widehat{T}(\mathbb{X}_n) \in B\right) \\ &\leq e^{\epsilon} \Pr\left(A_\eta(\phi_{\mathbb{Y}_n}; \mathbb{Y}_n) + \frac{1}{\epsilon}V \geq 1 + \frac{\log(2/\delta)}{\epsilon}\right) \Pr(\widehat{T}(\mathbb{X}_n) \in B) \\ &\leq e^{2\epsilon} \Pr\left(A_\eta(\phi_{\mathbb{Y}_n}; \mathbb{Y}_n) + \frac{1}{\epsilon}V \geq 1 + \frac{\log(2/\delta)}{\epsilon}\right) \Pr(\widehat{T}(\mathbb{Y}_n) \in B) \\ &= e^{2\epsilon} \Pr\left(\widetilde{T}(\mathbb{Y}_n) \in B\right). \end{aligned}$$

The first inequality is from independence and the fact that $A_\eta(\phi_{\mathbb{X}_n}; \mathbb{X}_n) + \frac{1}{\epsilon}V$ is an ϵ-differentially private estimator. The second inequality is from (A.33). Now what if there

exists an x such that $|\phi_{\mathbb{X}_n}(x) - \phi_{\mathbb{Y}_n}(x)| \geq \eta$? This implies that $A_\eta(\phi_{\mathbb{X}_n}; \mathbb{X}_n) = 1$ and

$$\begin{aligned}
\Pr\left(\widetilde{T}(\mathbb{X}_n) \in B\right) &\leq \Pr\left(A_\eta(\phi_{\mathbb{X}_n}; \mathbb{X}_n) + \frac{1}{\epsilon}V \geq 1 + \frac{\log(2/\delta)}{\epsilon}\right) \\
&= \Pr\left(V \geq \log(2/\delta)\right) \\
&= \delta \\
&\leq \delta + e^{2\epsilon} \Pr\left(\widetilde{T}(\mathbb{Y}_n) \in B\right).
\end{aligned}$$

This implies that we get $(2\epsilon, \delta)$ differential privacy if B is restricted to $\mathscr{B}(\mathbb{R}^d)$. For completeness, we need to include sets of the form $B = B' \cup \{\perp\}$, where $B' \in \mathscr{B}(\mathbb{R}^d)$. Consider

$$\begin{aligned}
\Pr\left(\widetilde{T}(\mathbb{X}_n) \in B\right) &= \Pr\left(\widehat{T}(\mathbb{X}_n) \in B', A_\eta(\phi_{\mathbb{X}_n}; \mathbb{X}_n) + \frac{1}{\epsilon}V \leq 1 + \frac{\log(2/\delta)}{\epsilon}\right) \\
&\qquad + \Pr\left(A_\eta(\phi_{\mathbb{X}_n}; \mathbb{X}_n) + \frac{1}{\epsilon}V > 1 + \frac{\log(2/\delta)}{\epsilon}\right) \\
&\leq e^{2\epsilon}\left(\Pr\left(\widetilde{T}(\mathbb{Y}_n) \in B'\right) + \Pr\left(A_\eta(\phi_{\mathbb{Y}_n}; \mathbb{Y}_n) + \frac{V}{\epsilon} > 1 + \frac{\log(2/\delta)}{\epsilon}\right)\right) + \delta \\
&= e^{2\epsilon} \Pr\left(\widetilde{T}(\mathbb{Y}_n) \in B\right) + \delta.
\end{aligned}$$

The first inequality comes from the fact that we get $(2\epsilon, \delta)$ differential privacy if B is restricted to $\mathscr{B}(\mathbb{R}^d)$ and the fact that $A_\eta(\phi_{\mathbb{Y}_n}; \mathbb{Y}_n) + \frac{1}{\epsilon}V$ is ϵ-differentially private.

Now, suppose that V, a_δ and b_δ correspond to the Gaussian version of PTR. Then, following the same steps as for the Laplace version gives, for $B \in \mathscr{B}(\mathbb{R}^d)$,

$$\Pr\left(\widetilde{T}(\mathbb{X}_n) \in B\right) \leq e^{2\epsilon} \Pr\left(\widetilde{T}(\mathbb{Y}_n) \in B\right) + \delta,$$

when $\|\phi_{\mathbb{X}_n} - \phi_{\mathbb{Y}_n}\|_\infty < \eta$. When $\|\phi_{\mathbb{X}_n} - \phi_{\mathbb{Y}_n}\|_\infty \geq \eta$,

$$\Pr\left(\widetilde{T}(\mathbb{X}_n) \in B\right) \leq \Pr\left(Z \geq \sqrt{2\log(1.25/\delta)}\right) \leq \delta.$$

We then have that

$$\Pr\left(\widetilde{T}(\mathbb{X}_n) \in B\right) \leq e^{2\epsilon} \Pr\left(\widetilde{T}(\mathbb{Y}_n) \in B\right) + \delta.$$

Again, we need to include sets of the form $B = B' \cup \{\perp\}$, where $B' \in \mathscr{B}(\mathbb{R}^d)$. Recall that

$a_\delta = \sqrt{2\log(1.25/\delta)}$ and $b_\delta = 2\log(1.25/\delta)$. Consider

$$\begin{aligned}\Pr\left(\widetilde{T}(\mathbb{X}_n) \in B\right) &= \Pr\left(\widetilde{T}(\mathbb{X}_n) \in B'\right) + \Pr\left(A_\eta(\phi_{\mathbb{X}_n}; \mathbb{X}_n) + \frac{a_\delta}{\epsilon} Z > 1 + \frac{b_\delta}{\epsilon}\right) \\ &\leq e^{2\epsilon}\left(\Pr\left(\widetilde{T}(\mathbb{Y}_n) \in B'\right) + \Pr\left(A_\eta(\phi_{\mathbb{Y}_n}; \mathbb{Y}_n) + \frac{a_\delta}{\epsilon} Z > 1 + \frac{b_\delta}{\epsilon}\right)\right) + 2\delta \\ &= e^{2\epsilon}\Pr\left(\widetilde{T}(\mathbb{Y}_n) \in B\right) + 2\delta.\end{aligned}$$

□

Proof of Theorem 18. The probability of no-reply follows directly from Theorem 16 for $\ell = 1$ or Theorem 17 for $\ell = 2$. Using the techniques of Theorem 13, where $\lambda_n = \epsilon/\eta$, one can show together with the proof of Corollary 2, the condition

$$n \geq C\left(\frac{d\log\left(\frac{1}{\alpha_{\mathcal{O}_\ell}(t)} \vee d\right)}{\epsilon \cdot \alpha_{\mathcal{O}_\ell}(t)}\right)^{\frac{1}{1-r}},$$

for some $r > 0$ and universal constant $C > 0$, gives that

$$-\alpha_{\mathcal{O}_\ell}(t) + g(\mathcal{O}_\ell, \mathcal{B}_\psi, \theta, \epsilon/\eta) \leq -\alpha_{\mathcal{O}_\ell}(t)/4.$$

Following the logic of the proof of Theorem 13, we have that

$$\Pr\left(\left\|\widetilde{T}_n - \theta\right\| > t\right) \leq m \cdot c_1 e^{-nc_2} + m \cdot C_1 e^{-nh_4\left(\alpha_{\mathcal{O}_\ell}(t)/4, \psi\right)^2} + e^{-c'\frac{\epsilon}{\eta}\alpha_{\mathcal{O}_\ell}(t)}.$$

Proof of Corollary 4 follows from the techniques of the previous sample complexity results. □

change type	Magnitude (β)								
dist.		$\mathcal{G}$			t_3			$\mathcal{SG}$	
n	100	200	500	100	200	500	100	200	500
MFHD	0.80	0.99	1.00	0.14	0.45	0.96	0.78	1.00	1.00
RP	0.44	0.89	1.00	0.12	0.44	0.95	0.40	0.90	1.00
LTR	0.90	1.00	1.00	0.16	0.51	0.97	0.90	1.00	1.00
MFHD$'$	0.92	1.00	1.00	0.14	0.49	0.98	0.88	1.00	1.00
RP$'$	0.83	0.99	1.00	0.14	0.48	0.96	0.86	1.00	1.00
LTR$'$	0.93	1.00	1.00	0.16	0.52	0.98	0.93	1.00	1.00
change type	Shape (α)								
dist.		$\mathcal{G}$			t_3			$\mathcal{SG}$	
n	100	200	500	100	200	500	100	200	500
MFHD	0.00	0.00	0.02	0.00	0.00	0.03	0.00	0.00	0.01
RP	0.04	0.26	0.84	0.00	0.00	0.01	0.04	0.28	0.83
LTR	0.01	0.01	0.06	0.00	0.00	0.02	0.00	0.00	0.02
MFHD$'$	0.73	0.96	1.00	0.07	0.37	0.92	0.70	0.98	1.00
RP$'$	0.76	0.96	1.00	0.13	0.49	0.95	0.76	0.97	1.00
LTR$'$	0.04	0.25	0.82	0.00	0.01	0.18	0.05	0.23	0.80
change type	No Change								
dist.		$\mathcal{G}$			t_3			$\mathcal{SG}$	
n	100	200	500	100	200	500	100	200	500
MFHD	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.01
RP	0.01	0.00	0.02	0.00	0.00	0.01	0.00	0.01	0.00
LTR	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01
MFHD$'$	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.01
RP$'$	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.00
LTR$'$	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02

Table A.3: Table of the empirical power for each of the epidemic FKWC tests when there was an epidemic-type magnitude change, an epidemic-type shape change and no change.

(d_1, d_2)	(0,0)	(0.4,0)	(0.8,0)	(0,0.4)	(0,0.8)	(0.4,0.4)
MFHD	0.08	0.59	0.98	0.54	0.94	0.97
RP	0.07	0.77	0.99	0.68	0.99	0.99
LTR	0.09	0.70	0.98	0.57	0.97	0.98
MFHD$'$	0.10	0.98	1.00	0.98	1.00	1.00
RP$'$	0.06	0.97	1.00	0.97	1.00	1.00
LTR$'$	0.09	0.76	0.99	0.66	0.99	0.99

Table A.4: Table of empirical powers at the 5% level of significance for the AMOC FKWC test under the functional autoregressive model discussed in the simulation study of Sharipov and Wendler (2019).

Appendix B

Select Topics from Functional Analysis

To facilitate understanding in Chapter 3 and Chapter 4, it is useful to review some of the results from functional analysis. These materials are closely adapted from Hsing and Eubank (2015) and we refer the read there for more details (Hsing and Eubank, 2015).

B.1 Bochner integrals

The mean and covariance elements of a probability measure on a Hilbert space are written in terms of Bochner integrals. We present the construction of the Bochner integral, and list some properties at the end of this subsection. Define a simple function $f\colon E \to \mathbb{X}$ as

$$f(\omega) = \sum_{i=1}^{k} \mathbb{1}\left(\omega \in E_i\right) g_i$$

for some finite k, $E_i \in \mathscr{B}$ and $g_i \in \mathbb{X}$.

Definition 11 (Bochner integrable)**.** *Any simple function with $\mu(E_i) < \infty$ for all i is said to be Bochner integrable and the Bochner integral is defined as*

$$\int_E f d\mu = \sum_{i=1}^{k} \mu(E_i) g_i.$$

A measurable function is Bochner integrable if there exists a sequence of functions f_n which are simple and Bochner integrable, such that

$$\lim_{n\to\infty}\int_E \|f - f_n\|\, d\mu = 0.$$

The integral is defined as

$$\int_E f d\mu = \lim_{n\to\infty}\int_E f_n d\mu.$$

It is obviously difficult to check if a function is integrable via the definition, and so the following theorem can be useful.

Theorem 20. *Suppose that f is a measurable function with*

$$\int_E \|f\|\, d\mu < \infty.$$

Suppose that for each n there exists a finite dimensional subspace $\mathbb{X}_n \subset \mathbb{X}$ such that

$$\lim_{n\to\infty}\int_E \|f - g_n\|\, d\mu = 0$$

for g_n taking values in $\mathbb{X}_n$. Then f is Bochner integrable.

This is even simpler in a Hilbert space:

Theorem 21. *Suppose that $\mathbb{X}$ is a separable Hilbert space and f is a measurable function from E to $\mathbb{X}$ with $\int \|f\|\, d\mu < \infty$. Then f is Bochner integrable.*

So, if we are working in a Hilbert space in order for f to be Bochner integrable, we must have that the norm of f is Lebesgue integrable. There also exists an extension of the dominated convergence theorem to Bochner integrals.

Theorem 22. *Let $\{f_n\}$ be a sequence of Bochner integrable functions that converges to some f. If there is a non-negative Lebesgue integrable function such that $\|f_n\| \leq g$ for all n a.e. μ then f is Bochner integrable and $\int f d\mu = \lim_{n\to\infty}\int f_n d\mu$.*

Another important feature of the Bochner integral is that

$$\left\|\int_E f d\mu\right\| \leq \int_E \|f\|\, d\mu.$$

B.2 Linear operators and functionals

B.2.1 Operators

The covariance operators discussed in Chapters 3 and 4 are bounded linear transformations. Bounded linear transformations $\mathscr{T}\colon \mathbb{X}_1 \to \mathbb{X}_2$, are uniformly continuous. They form a Banach space if $\mathbb{X}_2$ is a Banach space with the operator norm: $\|\mathscr{T}\| = \sup_{\|u\|=1} \|\mathscr{T}u\|$. The space of all bounded operators will be denoted by $\mathfrak{B}(\mathbb{X}_1, \mathbb{X}_2)$.

- In finite Euclidean space, the operator norm is given by $\max_{x'x=1} x'\mathscr{T}'\mathscr{T}x$
- In $\mathscr{L}^2$, we have that
$$\mathscr{T}f = \int_0^1 \mathcal{K}(\cdot, u)f(u)du$$
is a linear operator. You may notice this looks similar to a covariance operator.

In addition, define the following

1. $Dom(\mathscr{T})$ = the subset of $\mathbb{X}_1$ on which $\mathscr{T}$ is defined
2. $IM(\mathscr{T}) = \{\mathscr{T}x\colon x \in Dom(\mathscr{T})\}$
3. $ker(\mathscr{T}) = \{x \in Dom(\mathscr{T})\colon \ \mathscr{T}x = 0\}$

B.2.2 Adjoint operator

In this section we restrict our attention to Hilbert spaces.

Definition 12. *Suppose $\mathscr{T} \in \mathfrak{B}(\mathbb{H}_1, \mathbb{H}_2)$. Then, there is a unique element $\mathscr{T}' \in \mathfrak{B}(\mathbb{H}_2, \mathbb{H}_1)$ such that*
$$\langle \mathscr{T}x_1, x_2\rangle_2 = \langle \mathscr{T}'x_2, x_1\rangle_1,$$
for all $x_1 \in \mathbb{H}_1,\ x_2 \in \mathbb{H}_2$. $\mathscr{T}'$ is called the adjoint operator.

Definition 13. *Suppose $\mathscr{T} \in \mathfrak{B}(\mathbb{H}, \mathbb{H})$. We say that $\mathscr{T}$ is self-adjoint if $\mathscr{T}' = \mathscr{T}$.*

Any symmetric matrix is self-adjoint in $\mathbb{R}^p$. Note that
$$\langle \mathscr{T}x, y\rangle = x'\mathscr{T}'y = x'\mathscr{T}y = \langle x, \mathscr{T}y\rangle.$$

Recall the integral operators:

$$\mathscr{T} f = \int \mathcal{K}(\cdot, t) f(t) dt.$$

It follows that:

$$\langle \mathscr{T} f, g \rangle = \langle \int \mathcal{K}(\cdot, t) f(t) dt, g \rangle = \int \langle \mathcal{K}(\cdot, t) f(t), g \rangle dt = \int \int \mathcal{K}(s, t) f(t) g(s) ds dt = \langle f, \mathscr{T}' g \rangle,$$

with

$$\mathscr{T} f = \int \mathcal{K}(t, \cdot) f(t) dt.$$

So, if $\mathcal{K}$ is symmetric, then it is self-adjoint. The covariance operator discussed in Chapter 3 and Chapter 4 is an integral operator which a symmetric kernel. Therefore, it is self-adjoint.

Theorem 23. *If $\mathscr{T} \in \mathfrak{B}(\mathbb{H})$ and the sequence x_n converges weakly to x (meaning all linear functionals of x_n converge weakly) then $\mathscr{T} x_n$ also converges weakly.*

We have the following list of properties of the adjoint operator between two Hilbert spaces:

1. $(\mathscr{T}')' = \mathscr{T}$.
2. $\|\mathscr{T}'\| = \|\mathscr{T}\|$
3. $\|\mathscr{T}'\mathscr{T}\| = \|\mathscr{T}\|^2$
4. $ker(\mathscr{T}) = (IM(\mathscr{T}'))^{\perp}$
5. $ker(\mathscr{T}'\mathscr{T}) = ker(\mathscr{T})$ *and* $\overline{IM(\mathscr{T}'\mathscr{T})} = \overline{IM(\mathscr{T}')}$
6. $\mathbb{H}_1 = ker(\mathscr{T}) \oplus \overline{IM(\mathscr{T}')}$
7. $rank(\mathscr{T}^*) = rank(\mathscr{T})$ (recall that $rank(\mathscr{T}) = dim(IM(\mathscr{T}))$.)

It is useful to prove these results, to get a feel for the adjoint operator.

Proof. The first property is very simple. For the second, recall that

$$\|\mathscr{T}' x_2\|_1^2 = |\langle \mathscr{T}' x_2, \mathscr{T}' x_2 \rangle| = |\langle x_2, \mathscr{T}\mathscr{T}' x_2 \rangle_2| \leq \|\mathscr{T}\| \, \|x_2\| \, \|\mathscr{T}' x_2\| .$$

This gives that

$$\|\mathscr{T}' x_2\|_1 \leq \|\mathscr{T}\| \, \|x_2\| \,,$$

which implies that

$$\|\mathscr{T}'\| \leq \|\mathscr{T}\|$$

and vice versa by symmetry. Part 2 gives that $\|\mathscr{T}'\mathscr{T}\| \leq \|\mathscr{T}\|^2$. To show the other way,

$$\|\mathscr{T} x_1\|^2 = \langle \mathscr{T}'\mathscr{T} x_1, x_1 \rangle \leq \|\mathscr{T}'\mathscr{T}\| \, \|x_1\|^2 \,.$$

Suppose that $x_1 \in ker(\mathscr{T})$. Then $\langle x_1, \mathscr{T}' x_2 \rangle = \langle \mathscr{T} x_1, x_2 \rangle = 0$. Which means that x_1 is orthogonal to $IM(\mathscr{T}')$. If $x_1 \in (IM(\mathscr{T}'))^{\perp}$, then $\mathscr{T}'\mathscr{T} x_1 \in IM(\mathscr{T}')$ and so $0 = \langle \mathscr{T}'\mathscr{T} x_1, x_1 \rangle = \|\mathscr{T} x_1\|$. □

B.2.3 Non-negative, square-root and projection operators

Definition 14 (Positive definite)**.** *We say an operator is positive definite if it is self-adjoint, and $\langle \mathscr{T} x, x \rangle > 0$ for all $x \in \mathbb{H}$. We say that $\mathscr{T}_1 < \mathscr{T}_2$ if $\mathscr{T}_2 - \mathscr{T}_1$ is positive definite. If we replace the strict inequalities with non-strict inequalities then we say the operator is non-negative definite.*

For example, for any $\mathscr{T}$, $\mathscr{T}'\mathscr{T}$ is non-negative definite because it is self-adjoint and $\langle \mathscr{T}'\mathscr{T} x, x \rangle = \|\mathscr{T} x\|^2$. Covariance operators are non-negative definite. Non-negative definite operators admit a square-root-type decomposition.

Theorem 24 (Existence of a square root operator)**.** *Let $\mathscr{T} \in \mathfrak{B}(\mathbb{H})$. If $\mathscr{T}$ is non-negative, then there is a unique non-negative operator $\mathscr{T}^{1/2}$ such that $\mathscr{T}^{1/2}\mathscr{T}^{1/2} = \mathscr{T}$ and $\mathscr{T}^{1/2}$ commutes with any operator that commutes with $\mathscr{T}$.*

We can re-imagine the projection of x onto some subspace M as an operator. Suppose that $\mathscr{P}_M x$ is the projection of x onto M. If M is a closed subspace, then $\mathscr{P}_M x$ is a self-adjoint linear operator such that $\mathscr{P}_M \mathscr{P}_M = \mathscr{P}_M$. The projection operator is also non-negative. To see this, note that $\langle \mathscr{P}_M x, x \rangle = \langle \mathscr{P}_M x, \mathscr{P}_M x \rangle \geq 0$. We have that $\|\mathscr{P}_M\| = 1$. If the subspace M has dimension 1 and is spanned by x, with $\|x\| = 1$, then $\mathscr{P}_M$ can be written as $x \otimes x$, with

$$(x \otimes x) y = \langle y, x \rangle x.$$

This gives rise to the tensor product.

Definition 15. *The tensor product operator* $(x_1 \otimes x_2)$ *between* $\mathbb{H}_1$ *and* $\mathbb{H}_2$ *is defined as*

$$(x_1 \otimes x_2)y = \langle x_1, y \rangle x_2.$$

Note this reduces to the outer product in the finite dimensional, Euclidean case. In $\mathscr{L}^2$, we have that

$$(f_1 \otimes f_2)g = \int \int f_1(s_1)\ g(s_2) ds_1 ds_2 f_2.$$

Theorem 25. *For* $x_1, x_2 \in \mathbb{H}$*, it holds that*

$$\|(x_1 \otimes x_2)\| = \|x_2\| \, \|x_1\| \, .$$

B.2.4 Operator inverses

Recall that a linear mapping is one-to-one if $ker(\mathscr{T}) = \{0\}$ and onto if $IM(\mathscr{T}) = \mathbb{X}_2$. When both of these are satisfied, then the linear map is a bijection and it is invertible.

Theorem 26. *If a bounded linear mapping between two Banach spaces has an inverse, then the inverse is bounded.*

Theorem 27. *Let* $\mathbb{H}$ *be a Hilbert space. If* $\mathscr{T} \in \mathfrak{B}(\mathbb{H})$ *and* $\mathscr{T}$ *is self adjoint, and* $\|\mathscr{T} f\| \geq C \|f\|$ *then* $\mathscr{T}$ *is invertible.*

It follows from the above theorem that

Theorem 28. *It* $\|\mathscr{T}\| < 1$ *then* $I - \mathscr{T}$ *is invertible and*

$$(I - \mathscr{T})^{-1} = I + \sum_{i=1}^{\infty} \mathscr{T}^i$$

B.3 Compact operators and singular value decomposition

B.3.1 Compact operators

Definition 16. *A linear operator is compact if for any bounded sequence* x_n*,* $\mathscr{T} x_n$ *contains a convergent sub-sequence.*

Theorem 29. *The identity operator is not compact on an infinite dimensional normed space.*

Some facts about compact operators are as follows:

- The closure of the range of any compact operator is separable
- Operators with finite rank are compact
- The composition of two operators is compact if either operator is compact
- The set of compact operators that map to any Banach space is closed

We also have that

Theorem 30. *A bijective operator between two infinite dimensional Banach spaces is not compact.*

We now present a result specific to Hilbert spaces.

Theorem 31. *Suppose that $\mathscr{T} \in \mathfrak{B}(\mathbb{H}_1, \mathbb{H}_2)$. $\mathscr{T}$ is compact if there exists a sequence of finite dimensional operators such that $\|\mathscr{T}_n - \mathscr{T}\| \to 0$ as $n \to \infty$ and $\mathscr{T}$ is compact if $\mathscr{T}'$ is compact.*

B.3.2 Eigenvalues of compact operators

We need the following definition

Definition 17 (CONS)**.** *We can define a complete orthonormal system as an orthonormal sequence whose closed span is equal to $\mathbb{H}$. It can be shown that an orthonormal sequence $\{e_i\}$ is a CONS if $\langle x, e_i \rangle = 0$ for all i implies that $x = 0$.*

Definition 18. *For some $\mathscr{T} \in \mathfrak{B}(\mathbb{H})$, we say that λ is an eigenvalue and e is an eigenvector for $\mathscr{T}$ if it holds that*

$$\mathscr{T} e = \lambda e.$$

We call

$$ker(\mathscr{T} - \lambda I)$$

the eigenspace of λ.

Note that the eigenspace is a closed linear space.

Theorem 32. *Suppose λ_j are distinct and non-zero, then, $e_j \in ker(\mathscr{T} - \lambda_j I)$ are linearly independent and, if $\mathscr{T}$ is self-adjoint, they are mutually orthogonal.*

Theorem 33 (Eigenvalues of compact operators)**.** *For some $\mathscr{T} \in \mathfrak{B}(\mathbb{H})$, we have that*

1. *$ker(\mathscr{T} - \lambda I)$ is finite dimensional for all positive λ*
2. *The number of distinct eigenvalues greater than any positive number is finite*
3. *The set of non-zero eigenvalues is countable*

Theorem 34 (Eigenvalue decomposition of compact operators)**.** *If $\mathscr{T}$ is compact and self-adjoint, then the set of non-zero eigenvalues for $\mathscr{T}$ is finite or tends to 0. Each nonzero eigenvector has finite multiplicity, and eigenvectors corresponding to difference eigenvalues are orthogonal. Then the set of eigenvectors obtained from the Gram-Schmidt decomposition is a CONS for $\overline{IM(\mathscr{T})}$ and*

$$\mathscr{T} = \sum_{j \geq 1} \lambda_j e_j \otimes e_j, \text{ which means that } \mathscr{T}x = \sum_{j \geq 1} \lambda_j \langle x, e_j \rangle e_j.$$

Theorem 35. *If $\mathscr{T}$ is compact and non-negative definite, then*

$$\lambda_k = \max_{x \in \{e_1, \ldots, e_{k-1}\}^\perp} \frac{\langle \mathscr{T}x, x \rangle}{\|x\|^2},$$

where $\{e_1, \ldots, e_{k-1}\}$ is $\mathbb{H}$ when $k = 1$.

Theorem 36. *If $\mathscr{T}$ is compact and self-adjoint then*

$$|\lambda_k| = \max_{x \in \{e_1, \ldots, e_{k-1}\}^\perp} \frac{\|\mathscr{T}x\|}{\|x\|},$$

where $\{e_1, \ldots, e_{k-1}\}$ is $\mathbb{H}$ when $k = 1$.

Theorem 37 (Courant-Fischer minimax)**.** *If $\mathscr{T}$ is compact and non-negative definite, then,*

$$\lambda_k = \max_{v_1, \ldots, v_k} \min_{v \in \text{span}(v_1, \ldots, v_k)} \frac{\langle \mathscr{T}v, v \rangle}{\|v\|^2}$$

$$\lambda_k = \min_{v_1, \ldots, v_k} \max_{v \in \text{span}(v_1, \ldots, v_k)^\perp} \frac{\langle \mathscr{T}v, v \rangle}{\|v\|^2}$$

Theorem 38. *Let $\mathscr{T}, \tilde{\mathscr{T}}$ be non-negative definite, compact operators, then*

$$\sup_{k\geq 0} |\lambda_k - \tilde{\lambda}_k| \leq \left\| \mathscr{T} - \tilde{\mathscr{T}} \right\|.$$

These results can be used on self-adjoint operators, by splitting the operator into positive and negative components.

B.3.3 Singular value decomposition

The following gives singular value decomposition in the context of compact operators:

Theorem 39. *If $\mathscr{T}$ is a compact operator, then,*

$$\mathscr{T} = \sum_j \lambda_j (f_{1j} \otimes f_{2j}) \tag{B.1}$$

where

- *λ_j^2 are the non-ascending eigenvalues of $\mathscr{T}\mathscr{T}'$*
- *$\{f_{1j}\}$ are the eigenvectors of $\mathscr{T}'\mathscr{T}$*
- *$\{f_{2j}\}$ are the eigenvectors of $\mathscr{T}\mathscr{T}'$ satisfying $\mathscr{T}' f_{2j} = |\lambda_j| f_{1j}$.*

We refer to λ_j as the singular values of $\mathscr{T}$. Note that the largest singular value λ_1 satisfies $\|\mathscr{T}\| = \lambda_1$.

Theorem 40. *An operator $\mathscr{T}$ is compact if an only if (B.1) holds.*

B.3.4 Hilbert-Schmidt operators

Definition 19. *Let $\{e_i\}$ be a CONS for $\mathbb{H}_1$ and $\mathscr{T} \in \mathfrak{B}(\mathbb{H}_1, \mathbb{H}_2)$. If $\mathscr{T}$ satisfies*

$$\sum_{i=1}^{\infty} \|\mathscr{T} e_i\|_2^2 < \infty,$$

then $\mathscr{T}$ is called a Hilbert-Schmidt operator. We can denote the collection of Hilbert-Schmidt operators by $\mathfrak{B}_{HS}(\mathbb{H}_1, \mathbb{H}_2)$.

Note that Hilbert-Schmidt operators are compact. In addition, $\mathfrak{B}_{HS}(\mathbb{H}_1, \mathbb{H}_2)$ is a linear space with the following inner product

$$\langle \mathscr{T}_1, \mathscr{T}_2 \rangle_{HS} = \sum_{i=1}^{\infty} \langle \mathscr{T}_1 e_i, \mathscr{T}_2 e_i \rangle_2,$$

where $\{e_i\}$ is a CONS for $\mathbb{H}_1$. See Theorem 4.4.1 in Hsing and Eubank (2015) to see that this result is independent of the chosen CONS. Therefore, we can choose the CONS to be the singular vectors of $\mathscr{T}$ to get that

$$\|\mathscr{T}\|_{HS}^2 = \sum_{i=1}^{\infty} \lambda_i^2,$$

where λ_i are the singular values of $\mathscr{T}$. This is the Frobenius norm for matrices! The following theorem shows that truncating the sve approximates HS operators well:

Theorem 41. *Let $\mathscr{T}$ be a self-adjoint HS operator between $\mathbb{H}$ and itself, with singular system $\{\lambda_j, e_j\}$. Then, for any finite integer k, it holds that*

$$\left\| \mathscr{T} - \sum_{j=1}^{k} x_j \otimes y_j \right\|_{HS} \geq \left\| \mathscr{T} - \sum_{j=1}^{k} \lambda_j e_j \otimes e_j \right\|_{HS},$$

for any set of functions $x_j,\ y_j \in \mathbb{H}$.

B.3.5 Trace class operators

For this section, we focus on self-adjoint, non-negative definite linear operators.

Definition 20. *Let $\mathscr{T} \in \mathfrak{B}(\mathbb{H})$ for a separable $\mathbb{H}$. Then, $\mathscr{T}$ is trace class if for some CONS $\{e_i\}$, the quantity*

$$\|\mathscr{T}\|_{TR} = \sum_{i=1}^{\infty} \langle \mathscr{T} e_i, e_i \rangle < \infty.$$

We call $\|\mathscr{T}\|_{TR}$ the trace norm of $\mathscr{T}$.

Note that trace class operators are HS operators. In addition, we can then write

$$\|\mathscr{T}\|_{TR} = \sum_{i=1}^{\infty} \lambda_i,$$

where λ_i are the singular values of $\mathscr{T}$. Therefore, $\|\mathscr{T}\|_{HS} \leq \sqrt{\lambda_1 \|\mathscr{T}\|_{TR}}$.

B.3.6 Integral operators

Let $(E, \mathscr{B}, \mu)$ be a probability space. Suppose that $\mathcal{K}$ is a measurable function on $E \times E$ and

$$\int \int \mathcal{K}^2 d\mu d\mu < \infty.$$

Then define the integral operator $\mathscr{K}$ by

$$\mathscr{K} f(\cdot) = \int \mathcal{K}(s, \cdot) f(s) d\mu,$$

for $f \in \mathscr{L}^2(E, \mathscr{B}, \mu)$. Note that $\mathcal{K}$ is the kernel of $\mathscr{K}$. Note that $\mathscr{K} f$ is measurable and that $\mathscr{K} \in \mathfrak{B}(\mathscr{L}^2(E, \mathscr{B}, \mu))$. Suppose now that we only consider $\mathcal{K}$ which is continuous. If E is compact then $\mathcal{K}$ is uniformly continuous. This implies that $\mathscr{K} f(\cdot)$ is uniformly continuous. Note that $\mathscr{K}$ is compact. If $\mathcal{K}$ is symmetric, then $\mathscr{K}$ is self-adjoint and admits

$$\mathscr{K} = \sum_{i=1}^{\infty} \lambda_i e_i \otimes e_i.$$

In addition, $\mathscr{K}$ is non-negative definite if and only if $\mathcal{K}$ is. We can finally present Mercer's theorem:

Theorem 42 (Mercer's Theorem)**.** *Let $\mathcal{K}$ be a continuous kernel which is non-negative definite and symmetric. Let $\mathscr{K}$ be the corresponding integral operator. Then if (λ_j, e_j) are the eigenvalue-eigenfunction pairs of $\mathscr{K}$, $\mathcal{K}$ has the representation*

$$\mathcal{K}(s, t) = \sum_{j=1}^{\infty} \lambda_j e_j(s) e_j(t),$$

for all s, t and the sum converges absolutely and uniformly.

We can then use Mercer's Theorem and the results of the previous section to show that $\mathscr{K}$ is trace class with

$$\|\mathscr{K}\|_{TR} = \int \mathcal{K}(s, s) d\mu(s) \qquad \text{and} \qquad \|\mathscr{K}\|_{HS}^2 = \int \int \mathcal{K}(s, t)^2 d\mu(s) d\mu(t).$$

We then have from the previous section that

$$\hat{\mathcal{K}}(s,t) = \sum_{j=1}^{r} \lambda_j e_j(s) e_j(t)$$

approximates $\mathcal{K}(s,t)$ optimally. Lastly, we discuss unitary operators, these are the extension of orthonormal matrices in Euclidean space. A unitary operator (on a Hilbert space) is a bounded linear operator which is surjective and preserves inner products. We can also define a unitary operator as a bounded linear operator such that $\mathscr{U}\mathscr{U}^* = \mathscr{U}^*\mathscr{U} = I$. We have that if $\mathscr{K}_1$ is unitarily equivalent to $\mathscr{K}_2$, then

$$\langle \mathscr{K}_1 u, u \rangle = \langle \mathscr{K}_2 u^*, u^* \rangle,$$

so if the space is such that the uniform measure on S exists, then

$$\int_S \langle \mathscr{K}_1 u, u \rangle du = \int_S \langle \mathscr{K}_2 u^*, u^* \rangle du = \int_S \langle \mathscr{K}_2 u, u \rangle du.$$

B.4 What are the observations in the setting of functional data?

We can either assume that the functions are random elements which lie in some Hilbert space, or we could assume they are continuous time stochastic processes. When do these assumptions coincide?

B.4.1 Probability on a Hilbert space

For this section consider $\mathscr{L}^2$-valued random elements.

Theorem 43. *If X is a mapping from $\Omega \to \mathscr{L}^2$, then*

- *X is measurable if $\langle X, f \rangle$ is measurable for all f in $\mathscr{L}^2$.*
- *If X is measurable then its distribution is uniquely determined by the marginal distributions of $\langle X, f \rangle$ over f in $\mathscr{L}^2$.*

Definition 21. *Suppose that* $\mathrm{E}\left[\|X\|\right] < \infty$*, the mean of* X *is defined as the Bochner integral:*

$$\mathrm{E}\left[X\right] = \int_{\Omega} X dP.$$

We also have that

Theorem 44.

$$\mathrm{E}\left[\|X - \mathrm{E}\left[X\right]\|^2\right] = \mathrm{E}\left[\|X\|^2\right] + \|\mathrm{E}\left[X\right]\|^2.$$

We can then define the covariance operator as follows:

Definition 22. *Suppose that* $\mathrm{E}\left[\|X\|^2\right] < \infty$*, then the covariance operator for* X *is the element in* $\mathfrak{B}_{HS}(\mathscr{L}^2)$*, given by:*

$$\mathscr{K} = \mathrm{E}\left[(X - \mathrm{E}\left[X\right]) \otimes (X - \mathrm{E}\left[X\right])\right] = \int_{\Omega} (X - \mathrm{E}\left[X\right]) \otimes (X - \mathrm{E}\left[X\right]) \, dP.$$

We also have that

$$\mathrm{E}\left[(X - \mathrm{E}\left[X\right]) \otimes (X - \mathrm{E}\left[X\right])\right] = \mathrm{E}\left[X \otimes X\right] - \mathrm{E}\left[X\right] \otimes \mathrm{E}\left[X\right].$$

One may recall that in $\mathscr{L}^2$, we have that

$$(x \otimes y)(\cdot) = \int x(\cdot) ds \; y.$$

Theorem 45. *Suppose that* $\mathrm{E}\left[X\right] = 0$ *and that* $\mathrm{E}\left[\|X\|^2\right] < \infty$*. For* $f, g \in \mathscr{L}^2$*,*

1. $\langle \mathscr{K} f, g \rangle = \mathrm{E}\left[\langle X, f \rangle \langle X, g \rangle\right]$
2. $\mathscr{K}$ *is non-negative definite, trace class operator with* $\|\mathscr{K}\|_{TR} = \mathrm{E}\left[\|X\|^2\right]$
3. $\Pr(X \in \overline{IM(\mathcal{K})}) = 1$

As a result, we have that

$$\mathscr{K} = \sum_{j=1}^{\infty} \lambda_j e_j \otimes e_j.$$

In addition, it holds with probability 1 that

$$X = \sum_{j=1}^{\infty} \langle X, e_j \rangle e_j,$$

where $\langle X, e_j \rangle$ are uncorrelated with mean zero and variance λ_j. In addition, the best approximation of X in terms of mean squared error is

$$\sum_{j=1}^{k} \sum_{j=1}^{k} \langle X, e_j \rangle e_j,$$

see Theorem 7.2.8 in Hsing and Eubank (2015).

B.4.2 Stochastic process viewpoint

Suppose instead we view some observed function as a stochastic process: $X = \{X(t) \colon t \in [0,1]\}$. Recall that we assume that $X(t)$ is defined on a probability space $(\Omega, \mathscr{A}, P)$ and each $X(t)$ are all measurable for fixed t. It is then simple to define $\mathrm{E}[X(t)]$ as the mean and $\mathcal{K}(s,t) = \mathrm{E}[X(t)X(s)]$ as the covariance kernel, where we assume the expectations exist. Note that $\mathcal{K}$ is non-negative definite.

Definition 23 (Mean square continuous process)**.** *We say X is mean square continuous if*

$$\lim_{n \to \infty} \mathrm{E}\left[|X(t_n) - X(t)|^2\right] = 0$$

for any sequence in $[0,1]$ converging to t.

Theorem 46. *X is mean square continuous if and only if its covariance kernel and mean function are continuous.*

Actually, if the mean function is continuous, then $\mathcal{K}$ is continuous at all points if it is continuous on the diagonal. Define $\mathscr{K}$ as the associated integral operator corresponding to $\mathcal{K}$. Then, Mercer's theorem gives that

$$\mathcal{K}(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s) e_j(t).$$

Theorem 47 (Karhunen-Lòeve Theorem)**.** *Let $X(t)$ be a mean square continuous process with mean zero. There exists a random variable $I_X(e_j)$ such that Then*

$$\lim_{n\to\infty} \mathrm{E}\left[\left|\sum_{i=1}^{n} I_X(e_j)e_j(t) - X(t)\right|^2\right].$$

$I_X(e_j)$ are the principle component scores of X.

B.4.3 Combining the viewpoints

One assumption that allows us to consider X as both a random element in a Hilbert space and a mean square continuous stochastic process is joint measurability. We would assume that $X(t,\omega)$ is measurable with respect to the product field $\mathscr{B}([0,1])\times\mathscr{A}$. This assumption is somewhat opaque, so it is useful to provide a simple condition where joint measurability is implied.

Theorem 48. *Suppose that for each t, $X(t,\cdot)$ is measurable and that $X(\cdot,\omega)$ is continuous for each $\omega\in\Omega$. Then, X is jointly measurable, and is defined by its finite dimensional distributions.*

Now one only needs to verify that $X(t,\cdot)$ is continuous. One way to do this is with the Kolmogrov condition:

$$\mathrm{E}\left[|X(t_1) - X(t_2)|^{\alpha}\right] \leq C|t_1 - t_2|^{1+\beta},$$

holds for all $t_1, t_2 \in [0,1]$. If X is a mean square continuous process that is jointly measurable, then

1. The mean function and mean element coincide
2. The covariance operator coincides with the integral operator described previously
3. The scores satisfy $I_X(f) = \langle X, f\rangle$

Restating the Karhunen-Lòeve theorem, we have that if $X(t)$ is a jointly measurable mean square continuous stochastic process with mean zero, then it holds that Then

$$\lim_{n\to\infty} \mathrm{E}\left[\left|\sum_{j=1}^{n} \langle X, e_j\rangle e_j(t) - X(t)\right|^2\right].$$

Now, one may recall that in Chapter 3 and Chapter 4 we assumed that the the observed functions are mean zero, mean square continuous stochastic processes that satisfy the conditions of Theorem 48. Therefore, the covariance operators considered are trace class and the trace norm exists. Further, we may make use of Mercer's theorem to expand the covariance kernel in terms of its eigenfunctions and eigenvalues.