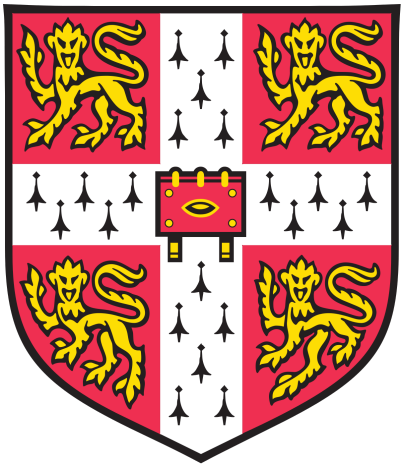
A concentration inequality based statistical methodology for inference on covariance matrices and operators

Adam B Kashlak

Cambridge Centre for Analysis
Clare Hall
University of Cambridge



This dissertation is submitted for the degree of
Doctor of Philosophy

June 2017

Abstract

In the modern era of high and infinite dimensional data, classical statistical methodology is often rendered inefficient and ineffective when confronted with such big data problems as arise in genomics, medical imaging, speech analysis, and many other areas of research. Many problems manifest when the practitioner is required to take into account the covariance structure of the data during his or her analysis, which takes on the form of either a high dimensional low rank matrix or a finite dimensional representation of an infinite dimensional operator acting on some underlying function space. Thus, novel methodology is required to estimate, analyze, and make inferences concerning such covariances.

In this manuscript, we propose using tools from the concentration of measure literature–a theory that arose in the latter half of the 20th century from connections between geometry, probability, and functional analysis–to construct rigorous descriptive and inferential statistical methodology for covariance matrices and operators. A variety of concentration inequalities are considered, which allow for the construction of nonasymptotic dimension-free confidence sets for the unknown matrices and operators. Given such confidence sets a wide range of estimation and inferential procedures can be and are subsequently developed.

For high dimensional data, we propose a method to search a concentration inequality based confidence set using a binary search algorithm for the estimation of large sparse covariance matrices. Both sub-Gaussian and sub-exponential concentration inequalities are considered and applied to both simulated data and to a set of gene expression data from a study of small round blue-cell tumours. For infinite dimensional data, which is also referred to as functional data, we use a celebrated result, Talagrand's concentration inequality, in the Banach space setting to construct confidence sets for covariance operators. From these confidence sets, three different inferential techniques emerge: the first is a k-sample test for equality of covariance operator; the second is a functional data classifier, which makes its decisions based on the covariance structure of the data; the third is a functional data clustering algorithm, which incorporates the concentration inequality based confidence sets into the framework of an expectation-maximization algorithm. These techniques are applied to simulated data and to speech samples from a set of spoken phoneme data.

Lastly, we take a closer look at a key tool used in the construction of concentration based confidence sets: Rademacher symmetrization. The symmetrization inequality, which arises in the probability in Banach spaces literature, is shown to be connected with optimal transport theory and specifically the Wasserstein distance. This insight is used to improve the symmetrization inequality resulting in tighter concentration bounds to be used in the construction of nonasymptotic confidence sets. A variety of other applications are considered including tests for data symmetry and tightening inequalities in Banach spaces. An R package for inference on covariance operators is briefly discussed in an appendix chapter.

Dedicated to Walter B Kashlak, a proponent of education,
and to the prospective progeny on the horizon

Preface

> Naturally, only one who had always been more or less studious, eccentric, and solitary could have pursued this course.
>
> *The Case of Charles Dexter Ward*
> H.P. Lovecraft (1927)

My pathological distrust in asymptotic statistical methodology began long before arriving at the University of Cambridge bus terminal in the autumn of 2013. When, for example, a biologist desires to perform a goodness-of-fit test considering 10,000 possible gene expressions, even an optimistic sample size of 1000 subjects will yield all classical methods erroneous due to such tests' reliance on distributional convergence theorems. In the current era of high and infinite dimensional data, we require a certain amount of bravado and cleverness to, respectively, eschew the old methodologies and fabricate novel techniques. It is, hence, not surprising that I discovered my love of concentration inequalities–the so-called *nonasymptotic theory of independence* (Boucheron et al., 2013)–upon commencing my studies at the University of Cambridge in 2013. While there are many paradigms for high and infinite dimensional statistical inference such as the much touted lasso regularization, the concentration inequality based methodology developed in this manuscript can stand amongst all of them and in some cases supersede them.

This area of my own doctoral research has been preceded by an extensive collection of brilliant works, and thus it is an honour to add to such a compendium of knowledge no matter how much or how little future significance my own results will garner. But unlike much of the past research that lives inside the elegant ivory tower of abstract theory, the results within this manuscript are keenly focused on applications to real data problems even if only as a proof-of-concept at this early stage of the methodology. While it is doubtful that I could even begin to compete with the over forty years of collective works of statisticians, probabilists, and functional analysts who have come before me, I do hope and believe that this manuscript will cleverly collide such lofty abstraction with tangible application to problems of interest both inside and outside of the realm of pure mathematics.

The pages that follow consist of the totality of the research spanning the previous four years of my life concerning the topic of covariance estimation in the high or infinite dimensional setting through the application of concentration inequalities. The problem of high dimensional data arises in Chapter 2. Intuitively, that chapter concerns itself with locating the needles-in-the-haystack, which is to determine which pairs of variables in a high dimensional vector are strongly correlated under the sparsity assumption that most such pairs are not. The problem of infinite dimensional data arises in Chapter 3. That chapter is chronologically the initial chapter of my dissertation set into motion by the polar research topics of my doctoral advisers. It is concerned with using concentration inequalities to construct confidence balls–the infinite dimensional analogue of the ubiquitous confidence intervals–for estimators of the covariance in such settings. Chapter 4 is the product of my inescapable obsession with removing the coefficient of 2 from the symmetrization inequality, a result I discovered for myself when pursuing the methodology of Chapter 3. From the standpoint of pure mathematics, such coefficients are almost wholly ignored as it is generally sufficient to know that a finite coefficient C exists no matter its true value. However, for the sake of statistical methodology such pervasive powers of two lead to overly wide bounds that should not and need not be.

Thus, I invite the reader to enter into this world of nonasymptotic statistics with me. Hopefully, the structure of this dissertation is fine enough to successfully promote the case for the use of such methodology for the estimation of and inference on covariance matrices and operators.

Acknowledgments

Any such endeavour as a PhD cannot be completed in total isolation. With that, I would like to first thank my entire family and specifically my parents, Roger and Bonnie, and younger brother, Jacob, as well as my aunt and uncle, Jane Kashlak and Paul Kerlinger, for their support both tangible and emotional throughout the entirety of my life. Similarly and more so, I would like to thank my wife, Megan, for her support, love, and patience during my four year English escapade.

Academically, I would like to firstly acknowledge my doctoral advisers professor John A D Aston and professor Richard Nickl from the Cambridge Statistical Laboratory for starting me down this research path and for correcting my path when I wandered too far astray. I would also like to thank John and Richard for introducing me to the topics of functional data analysis and concentration of measure, respectively, which collided to form the basis for the research contained in this manuscript.

Beyond my advisers, I would like to thank the many colleagues that I have had the honour to work besides during my years in Cambridge. Thank you to my wonderful

office mates–Harold, Franca, and Davide–as well as to the remainder of the 2013 cohort for maintaining an enthusiastic environment for research. Thank you to my fellow PhD students and now doctors Eoin Devane, Helge Dietert, and Henry Jackson for joining with me to write the winning proposal for the 2015 Royal Statistical Society's Statistical Analytics Challenge. Our article was recently accepted for publication and, though completely unrelated to the material of this manuscript, will be gratuitously cited as follows (Kashlak et al., 2017). Thank you to the Cambridge Centre for Analysis in general and specifically to the original directors James R Norris and Arieh Iserles for providing me with the opportunity to study there. Thank you for the opportunities to learn from teaching others during undergraduate supervisions and the volunteered hours spent at the Cambridge Statistics Clinic. And thank you to my college, Clare Hall, for cultivating an excellent academic environment.

Beyond the University of Cambridge, I would like to thank those from other universities with whom I have had to opportunity to collaborate. Thank you to the statistics department at the Politecnico di Milano for inviting and embracing me and specifically to Alessandra Cabassi for collaborating with me on our joint R package (Cabassi and Kashlak, 2016). Thank you to Significance magazine and the Royal Statistical Society for inviting me to Manchester to present my article regarding the ever increasing frequency with which US politicians speak of America (Kashlak, 2016). Thank you to the University of Alberta and specifically to professor Linglong Kong for inviting me to visit in the autumn of 2016 and for inviting me to return as a faculty member in July 2017.

I would like to acknowledge the amazingly critical work of all of my past colleagues across the Atlantic who continue to solve the most challenging problems on Earth and thank them for all that they taught me. Lastly, I would like to specifically thank Dr. Steven Knox for transforming me from mathematician to statistician and, hence, for effectively catalyzing all of the research contained in this manuscript and all that is to come in the future.

Adam B Kashlak
Cambridge, UK
June 2017

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared below and specified in the text.

This dissertation is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared below and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared below and specified in the text

With respect to specific chapters, the contents of Chapter 1 are mainly expository included to introduce the necessary background material to properly ground and motivate the research contained in the other chapters. Chapter 2 contains original research conducted under the guidance of Professor Linglong Kong from the University of Alberta in the city of Edmonton. Furthermore, Appendix 2.B solely contains past lemmas and theorems used in this chapter. The remaining sections and appendices of this chapter all contain original work unless otherwise noted in the text with explicitly cited sources. Chapter 3 contains original research conducted under the guidance of my two doctoral thesis advisers, Professor John A D Aston and Professor Richard Nickl from the Statistical Laboratory at the University of Cambridge. Furthermore, Appendix 3.C is wholly an exposition of background material to ground the research of this chapter. The remaining sections and appendices of this chapter all contain original work unless otherwise noted in the text with explicitly cited sources. The contents of this chapter have also been slightly guided by the comments and critiques of three anonymous referees during the process of publishing an academic journal article version if this chapter's research. Chapter 4 contains original and self directed research. Furthermore, Appendix 4.A contains a collection of past lemmas and theorems used in the proofs of this chapter. The remaining material in this chapter is all original research unless otherwise noted and properly cited. The contents of this chapter have also been slightly guided by the comments and critiques

of two anonymous referees during the process of publishing an academic journal article version if this chapter's research. The R package referenced and detailed in Appendix A was written jointly with Alessandra Cabassi who, at the time, was a masters student at the Politecnico di Milano in Milan, Italy. Appendix B contains original research that has not been fully explored at the time of the publishing of this thesis.

Contents

Chapter 1

Introduction

As we proceed further into the 21st century, descriptive and inferential statistical problems concerning high and infinite dimensional data are quickly becoming of paramount importance to almost all academic and non-academic sectors. Sophisticated sound recording devices can provide high quality speech data. Medical imaging technology can produce three dimensional data laden images of the human brain. When studying the human genome, it may be of interest to test for correlation among tens of thousands of genes. For such problems, we require novel statistical methods that go beyond the classical methodology.

Classical statistics began to solidify as a discipline in the early 20th century through the brilliant works of such visionaries as Sir Ronald Fisher, Karl and Egon Pearson, Jerzy Neyman, and Sir Harold Jeffreys. The pioneering work of these founding fathers of the field was often in conflict with one another (Berger, 2003). However, most of their respective work is similar in the sense that much of it relies on limit theorems with respect to the size of the data. This feature of classical statistical theory arises in such forms as the law of large numbers, the central limit theorem, and other distributional convergence results for frequentist inference as well as in Berstein–von Mises theorems in the Bayesian context. In the modern setting involving high dimensional parameter spaces–the so called $p \gg n$ setting–such as genomics where a researcher may be faced with data consisting of 10,000 genes tested from a mere sample of 100 patients, such limit based statistical tests such as Pearson's chi-squared goodness of fit test are rendered obsolete. Thus, new methodology is required to supplant classical statistical tests in the modern setting.

Progressing beyond the limit theorems, so much statistical methodology is reliant on the covariance structure of the data. As this structure is generally unknown to the researcher, it must be estimated from the data itself. In the finite dimensional setting, the covariance matrix is ubiquitous. It appears in ordinary least squares regression and other regression tests, linear and quadratic discriminant analysis, principal components analysis, and many other areas of inferential and descriptive

statistics. Furthermore, the precision or inverse covariance matrix also plays a role in such methods. A problem of critical importance is that standard empirical estimates of the covariance matrix in the high dimensional data setting result in estimates that are necessarily singular matrices making it at best a very poor substitute for the unknown true covariance matrix and at worse completely unusable by being non-invertible.

Such problems are only magnified in the functional or infinite dimensional data setting where the covariance matrices are replaced by their infinite dimensional counterparts, the covariance operators. These operators are necessarily trace-class (Horváth and Kokoszka, 2012) and thus numerically unstable to invert even given an excessively large sample size. In some sense, this specific feature of covariance operators has spawned a completely separate subdiscipline of statistics concerned with methodology for functional data (Ramsay and Silverman, 2005). Examples of such functional data include the already mentioned speech samples and medical images. There is also the Berkeley growth curve data set displayed in Figure 1.1, which will be discussed in Chapter 3 as well as a set of weather data charting the daily high temperatures and amount of precipitation for each day of an entire year for multiple locations across Canada.

It is precisely these problems with covariance matrices and operators in the, respectively, high and infinite dimensional settings that motivates the research contained within this manuscript. In the following chapters, we specifically explore how tools from the field of concentration of measure can be utilized to develop rigorous statistical methodology to directly tackle these covariance structures (Ledoux, 2001; Boucheron et al., 2013; Giné and Nickl, 2016). A more significant introduction and overview of the concentration of measure tools is contained in Section 1.2. In brief, these concentration inequalities are concerned with bounding the tail area of a random variable as it moves away from some central reference point such as the mean or median of its distribution. We will use these inequalities to construct confidence sets for covariance matrices and operators. There are two key aspects to this theory that make it extremely palatable for use in statistics: the resulting inequalities are dimension free and non-asymptotic.

Standard asymptotic statistical methodology is impractical and in some cases impossible to implement in these high and infinite dimensional settings. The concentration inequalities that arise from this concentration of measure phenomenon hold for all sample sizes, and in some cases, are proven to give sharp bounds on the tails of the distribution. Thus, even if the bounds provided are not necessarily sharp, it is reasonable to begin a statistical analysis with these slightly too wide concentration inequality based confidence sets that can be adjusted post-hoc to the given data rather than beginning from some asymptotic confidence set and applying a finite

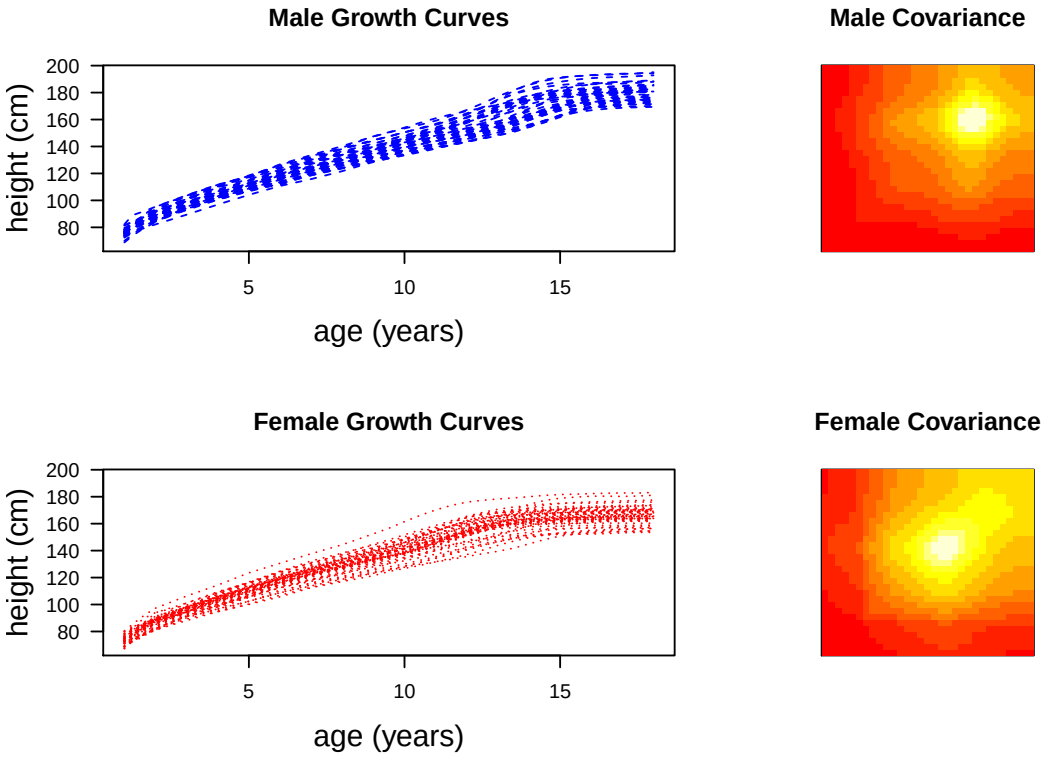


Figure 1.1: As an example of functional data, the Berkeley growth curve data set tracks the heights of 39 male and 54 female children over the first 18 years of their lives. The curves are plotted on the left; the empirical covariance operators are displayed on the right.

sample correction. As a proof-of-concept, we demonstrate in this manuscript making use of diverse sets of simulated and real data that such non-asymptotic methodology can be quite effect in practice.

So much applied mathematics including statistics, optimization, and numerical analysis is beset by the curse of dimensionality. Whether you are trying to numerically integrate a function or find its global minimum, high dimensionality will doom most algorithms to failure. The concentration of measure phenomenon often results in inequalities and bounds that are completely independent of the dimensionality of the space where the data exists. While the property of being dimension free does not guarantee statistical power, this property makes such approaches particularly palatable to the functional or infinite dimensional data settings. Many approaches to statistical problems in that setting rely on a preliminary dimension reduction step where the infinite dimensional data is projected onto a finite dimensional basis. Our approach provides the practitioner with the option of skipping such a dimension reduction step and analyzing the data in its original function space.

Approaching statistical inference via analysis of the covariance structure of the data, whether in the high dimensional matrix setting or infinite dimensional operator setting, has resulted in incredibly useful methodology. There is a plethora of different and diverse data sets that have been considered with such approaches. The work of Panaretos et al. (2010) analyzes the three dimensional wrapping and twisting of DNA microcircles. In the article of Fremdt et al. (2013), they consider the egg-laying trajectories of Mediterranean fruit flies over the lifespans of the insects. Spoken language data is analyzed in order to compare the pronunciation of numbers spoken in five different Latin-based romance languages (Pigoli et al., 2014, 2015; Aston et al., to appear). In similar style, we analyze phoneme data consisting of five different spoken sounds in Section 3.4 using our concentration methodology. A data set of gene expressions for small round blue-cell tumours is looked at in Rothman et al. (2009), Cai and Liu (2011), and in Section 2.4.3 of this manuscript. Bickel and Levina (2008a) consider high dimensional climate data consisting of mean temperatures from January 1850 to June 2006 taken at 2592 different recording stations over planet Earth. Meanwhile, Bickel and Levina (2008b) apply their methodology to call centre data. In the area of evolutionary biology, Cabassi et al. (2017) looks at curves denoting how actively different test sets of mice run on their wheels over a timespan of observation. A variety articles have also examined neuroimaging data collected from fMRI scans or positron emission tomography using the covariance structure of the data (Jiang et al., 2009, 2016; Zhu et al., 2014; Yu et al., 2016; Lila et al., 2017).

In this manuscript, Chapter 1 continues with a collection of definitions and notation in Section 1.1. In that section, we will discuss the setting for this manuscript and the mathematical spaces where the covariance matrices and operators will live.

A variety of norms, metrics, and distributions will also be introduced for later use. Section 1.2 provides a brief introduction to the concentration of measure phenomenon. It contains a collection of preliminary results to set the stage as well as references for further reading. Section 1.3 discusses some past research that was used as a starting point for the work presented in the following chapters.

Chapter 2 considers the application of concentration inequalities for covariance matrices in the setting of high dimensional data. Specifically, the goal of that chapter is to develop a rigorous methodology for the estimation of large sparse covariance matrices. The overall concentration inequality methodology is outlined in Section 2.2. An exposition of specific concentration inequalities for a variety of distributions is considered in the subsections of Section 2.3, which include examples of both sub-Gaussian and sub-exponential concentration. Section 2.4 compares our methodology to other methodologies on simulated multivariate Gaussian and multivariate Laplace data and on the data set of gene expressions for small round blue-cell tumours. The appendices of Chapter 2 contain proofs of the chapter's results, the derivation of Lipschitz constants for the functions used by our proposed methodology, and an expository account of further background on the concentration inequalities utilized by the methodology.

In Chapter 3, concentration inequalities applied to covariance operators in the functional data setting are used to develop a collection of inferential tests. Section 3.2 carefully constructs confidence sets for covariance operators making use of Talagrand's concentration inequality (Talagrand, 1996a). Section 3.3 introduces three different statistical applications for these confidence sets: a k-sample test for the equality of covariance operators; a functional data classifier based on the covariance operator, which can also be used to classify covariance operators directly; and an expectation-maximization style clustering algorithm for functional data. Numerical experiments for all of these methods on both simulated and phoneme data sets are detailed in the subsections of Section 3.4. In the appendices of Chapter 3, the construction of confidence sets using Talagrand's concentration inequality for general Banach space valued random variables is detailed in Appendix 3.A. The weak variance, a key component to constructing these confidence sets for covariance operators, is computed in Appendix 3.B for a collection of p-Schatten norms and for both multivariate Gaussian and the heavier tailed multivariate t-distributed random variables. Appendix 3.C provides some background material on tensor products of Hilbert spaces and Banach spaces. It also specifically connects these abstract notions with the finite dimensional case of data in $\mathbb{R}^n$. Appendix 3.D briefly investigates the consequences of applying our statistical tests to data with noise added and data from heavy tailed distributions. As confidence sets based on concentration inequalities are often larger than desired, Appendix 3.E proposes how the sizes of confidence

sets constructed via Talagrand's concentration inequality can be improved with a cross-validation procedure.

Chapter 4 revisits the symmetrization inequality used in Chapter 3 to propose an improved version of this fundamental result. Section 4.3.2 states and proves the improved Rademacher symmetrization inequality. As this inequality contains a term dependent on the Wasserstein distance W_2, Section 4.4 provides a bootstrap estimator for the empirical estimate of this distance between two probability measures. It also contains numerical simulations to demonstrate that this so-called improved inequality is in fact an improvement on the original symmetrization result. Section 4.5 details a diverse set of applications for this improved inequality from tests of data symmetry and the construction of high dimensional confidence sets via a generalized bootstrap procedure to better bounds for empirical processes and and sharper Nemirovski style inequalities in Banach spaces. An expositional account of some standard results from optimal transport theory are contained in the appendix of Chapter 4.

Appendix A contains a brief summary of the R code written to implement the concentration-based methodology for inference on covariance operators from Chapter 3. This R package, `fdcov`, is available on CRAN (Cabassi and Kashlak, 2016). The included functions are explained in the first part of the appendix. Appendix A ends with some short examples to test the library on real data.

Appendix B discusses three areas where concentration inequality based statistical methodology for covariance operators may prove fruitful given further investigation. The areas considered are longitudinal data, which falls into the category of functional data with often sparse and irregular observations, and data living in a reproducing kernel Hilbert space as the choice of kernel can affect the constructed confidence sets.

1.1 Definitions and notation

1.1.1 Covariance matrices

The covariance matrix is a fundamental statistical object, which is utilized by a high percentage of descriptive and inferential techniques including regression, principal components analysis, and both linear and quadratic discriminant analysis. Let $X \in \mathbb{R}^d$ be a random variable such that $\mathrm{Var}(X_i) < \infty$ for $i = 1, \ldots, d$. The covariance matrix of X, denoted $\Sigma \in \mathbb{R}^{d \times d}$, is the matrix with entries $\Sigma_{i,j} = \mathrm{E}((X_i - \mathrm{E}X_i)(X_j - \mathrm{E}X_j))$. Such matrices have the nice properties of being both symmetric and positive semi-definite.

There is an expansive literature on covariance matrix estimation from both the frequentist and Bayesian perspectives as well as estimation taking into consideration a variety of assumptions and settings for such estimation. We will discuss such

approaches and detail our own methodology for covariance matrix estimation in Chapter 2 specifically in the case of large sparse covariance matrices, but the simplest estimator and a good starting point for more complex estimation is the empirical estimate.

Definition 1.1.1 (Empirical Covariance Matrix)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be independent and identically distributed realizations of some random variable $X \in \mathbb{R}^d$ with unknown covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Then, the sample or empirical estimate for Σ is*

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$$

where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ is the sample mean of the data.

The empirical covariance matrix may be flawed when used in certain estimation settings. However, it does maintain some nice properties such as it being the maximum likelihood estimate for the true Σ in the Gaussian setting. In this form, it is a slightly biased estimator as

$$\begin{aligned}
\mathrm{E}\hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left((X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big((X_i - \mathrm{E}X_i)(X_i - \mathrm{E}X_i)^{\mathrm{T}} + (X_i - \mathrm{E}X_i)(\mathrm{E}X_i - \bar{X})^{\mathrm{T}} + \\
&\quad + (\mathrm{E}X_i - \bar{X})(X_i - \mathrm{E}X_i)^{\mathrm{T}} + (\mathrm{E}X_i - \bar{X})(\mathrm{E}X_i - \bar{X})^{\mathrm{T}}\Big) \\
&= \Sigma - \mathrm{E}\left((\mathrm{E}\bar{X} - \bar{X})(\mathrm{E}\bar{X} - \bar{X})^{\mathrm{T}}\right) \\
&= (1 - n^{-1})\Sigma,
\end{aligned}$$

but regardless, is generally used in practice as we will do in most sections of this manuscript.

When working with covariance matrices and related estimators, we will occasionally desire to find the square roots of such matrices. For a general square matrix $M \in \mathbb{R}^{d\times d}$, a square root can be any matrix $L \in \mathbb{R}^{d\times d}$ such that $M = LL^{\mathrm{T}}$. This is obviously not unique as for any unitary $R \in \mathbb{R}^{d\times d}$, we have that $(RL)RL^{\mathrm{T}} = LL^{\mathrm{T}} = M$, and thus RL is also a square root of M. However, for a positive-definite symmetric matrix, it is possible to find a unique positive-definite square root as is illustrated in the below definition. In our setting, the covariance matrices and estimators of such will always be positive semi-definite symmetric matrices. Hence, the definition can be made more precise.

Definition 1.1.2 (Matrix Square Root)**.** *Let $A \in \mathbb{R}^{d\times d}$ be a symmetric positive semi-definite matrix with eigen-decomposition $M = UDU^{\mathrm{T}}$ where $U = (v_1\ v_2\ \ldots\ v_d)$*

is the orthonormal matrix of eigenvectors and D is the diagonal matrix of eigenvalues, $(\lambda_1, \ldots, \lambda_d)$. Then, $M^{1/2} = UD^{1/2}U^{\mathrm{T}}$ where $D^{1/2}$ is the diagonal matrix with entries $(\lambda_1^{1/2}, \ldots, \lambda_d^{1/2})$.

1.1.2 Covariance operators

When considering functional data, the previous notion of a covariance matrix is generalized to that of the covariance operator. Generally, we will consider functional data to be in the Hilbert space $L^2(I)$ for $I \subset \mathbb{R}$. Our estimated covariance operators of interest are then operator valued random variables. Let

$$Op(L^2) = \left\{ T : L^2(I) \to L^2(I) \,\middle|\, \exists M \geq 0 \text{ s.t. } \|T\phi\|_{L^2} \leq M\|\phi\|_{L^2} \ \forall \phi \in L^2(I) \right\}$$

denote the space of all bounded linear operators mapping L^2 into L^2. This is where our covariance operators will live.

In order to construct a covariance operator from a sample of functional data, the notion of tensor product is required. In the finite dimensional setting, it is sufficient to use the transpose and the so-called outer product. Let $f, g \in L^2(I)$ and ϕ in the dual space $L^2(I)^*$ with inner product $\langle f, \phi \rangle = \phi(f)$. The tensor product, $f \otimes g$, is the rank one operator defined by $(f \otimes g)\phi = \langle g, \phi \rangle f = \phi(g)f$.

Definition 1.1.3 (Covariance Operator)**.** *Let $I \subseteq \mathbb{R}$, and let f be a random function (variable) in $L^2(I)$ with $\mathrm{E}\|f\|_{L^2}^2 < \infty$ and with zero mean. The associated covariance operator $\Sigma_f \in Op(L^2)$ is defined as $\Sigma_f = \mathrm{E} f^{\otimes 2} = \mathrm{E}\left(\langle f, \cdot \rangle f\right)$.*

Covariance operators can be treated as a generalization of covariance matrices. Indeed, if $I = \{i_1, \ldots, i_m\}$ has finite cardinality, then $f = (f_1, \ldots, f_m)$ is a random vector in $\mathbb{R}^m$ and for some fixed vector $v \in \mathbb{R}^m$, $\mathrm{E}\left(\langle f, v \rangle f\right) = \mathrm{E}\left(ff^{\mathrm{T}}\right)v$ where $\Sigma_f = \mathrm{E}\left(ff^{\mathrm{T}}\right)$ is the usual covariance matrix. Covariance operators are integral operators with the kernel function $c_f(s,t) = \mathrm{cov}(f(s), f(t)) \in L^2(I \times I)$. Such operators are characterized by the result that for $f \in L^2(I)$, Σ_f is a covariance operator if and only if it is trace-class, self-adjoint, and compact on $L^2(I)$ where the symmetry follows immediately from the definition and the finite trace norm comes from Parseval's equality (Bosq, 2012, Theorem 1.7)(Horváth and Kokoszka, 2012, Section 2.3).

In the applied statistics setting, such operators will necessarily have a finite representation on a computer. It is possible to treat functional data and the associated operators as vectors and matrices. However, standard computational approaches to covariance matrices will fail in the functional data setting due to the fact that the operators are trace class. Thus, the finite dimensional representation will be nearly singular resulting in a litany of problems for most statistical settings

ranging from numerical instability in the chosen methodology to complete failure due to being unable to invert the matrix. This setting is distinctly different from the high dimensional paradigm where it is often assumed that the very large but finite dimensional covariance matrix is full rank with eigenvalues bounded away from zero, but lacks sufficient data for standard estimation technique to succeed.

In Chapter 3, we will also require tensor powers of covariance operators denoted as $\Sigma^{\otimes 2} : Op(L^2) \to Op(L^2)$, which will necessitate the assumption that $\mathrm{E}\|f\|_{L^2}^4 < \infty$. For a basis $\{e_i\}_{i=1}^{\infty} \in L^2(I)$ with corresponding basis $\{e_i \otimes e_j\}_{i,j=1}^{\infty}$ for $Op(L^2(I))$, the previous definition is extended to $\Sigma^{\otimes 2} = \langle \Sigma, \cdot \rangle \Sigma$ where for $\Sigma_1, \Sigma_2 \in Op(L^2)$ with $\Sigma_1 = \sum_{i,j=1}^{\infty} \lambda_{i,j} e_{i,j}$ and $\Sigma_2 = \sum_{i,j=1}^{\infty} \gamma_{i,j} e_{i,j}$, then $\langle \Sigma_1, \Sigma_2 \rangle = \sum_{i,j}^{\infty} \lambda_{i,j} \gamma_{i,j}$. Specifically for covariance operators, the tensor power takes on a similar integral operator form with kernel $c_{\Sigma}(s,t,u,v) = \mathrm{cov}\left(f(s), f(t)\right) \mathrm{cov}\left(f(u), f(v)\right)$. A further expository look at tensor products and tensor powers of operators and Banach spaces can be found in Appendix 3.C.

In the matrix setting, the standard notion of a matrix transpose will suffice. In the operator setting, we require the definition of the adjoint operator. Given an Hilbert space H with inner product $\langle \cdot, \cdot \rangle$, the adjoint of a bounded linear operator $\Sigma : H \to H$, denoted as Σ^*, is the unique operator such that $\langle \Sigma f, g \rangle = \langle f, \Sigma^* g \rangle$ for $f, g \in H$. The existence of which is given by the Riesz representation theorem (Rudin, 1987, chapter 2). For self-adjoint operators, such as the covariance operators of interest, we have the simplification that $\Sigma = \Sigma^*$.

Similarly to the matrix setting, we are keenly interested in the estimation of covariance operators. As a starting point, we will consider the sample or empirical operator, which has the following form.

Definition 1.1.4 (Empirical Covariance Operator)**.** *Let $f_1, \ldots, f_n \in L^2(I)$ be independent and identically distributed realizations of some random function $f \in L^2(I)$ with unknown covariance operator $\Sigma \in Op(L^2)$. Then, the sample or empirical estimate for Σ is*

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (f_i - \bar{f}) \otimes (f_i - \bar{f}) = \frac{1}{n} \sum_{i=1}^{n} (f_i - \bar{f})^{\otimes 2} = \frac{1}{n} \sum_{i=1}^{n} \left\langle (f_i - \bar{f}), \cdot \right\rangle (f_i - \bar{f})$$

where $\bar{f} = n^{-1} \sum_{i=1}^{n} f_i$ is the sample mean of the data.

As with the matrix setting, this estimator is slightly biased. However, such estimates do have nice convergence properties. From Chapter 2 of Horváth and Kokoszka (2012), we have that

Theorem 1.1.5 (Theorem 2.5 Horváth and Kokoszka (2012))**.** *Let $X, X_1, \ldots, X_n \in L^2(I)$ with $I \subset \mathbb{R}$ be independent and identically distributed. Furthermore, let X*

have a finite fourth moment, which is $\mathrm{E}\|X\|_{L^2}^4 \leq \infty$. *Then,*

$$\mathrm{E}\|\Sigma - \hat{\Sigma}\|_2^2 \leq n^{-1}\mathrm{E}\|X\|_{L^2}^4$$

where $\|\cdot\|_2$ *is the Hilbert-Schmidt norm to be discussed in the following subsection.*

1.1.3 Norms

When defining a space of covariance matrices, there are many matrix norms that can be considered. In this manuscript, the main norms of interest are the p-Schatten norms, which will be denoted $\|\cdot\|_p$. Furthermore, many of the metrics that will be investigated in the following chapters are those that correspond to the p-Schatten norms. When $p \neq 2$, these are not Hilbert norms. The definition of the p-Schatten norm involves taking the square root of a positive semi-definite symmetric matrix, which was defined in Definition 1.1.2.

Definition 1.1.6 (p-Schatten norm for matrices)**.** *For an arbitrary matrix* $\Sigma \in \mathbb{R}^{k\times l}$ *and* $p \in (1, \infty)$*, the p-Schatten norm is*

$$\|\Sigma\|_p^p = \mathrm{tr}\left((\Sigma^{\mathrm{T}}\Sigma)^{p/2}\right) = \|\boldsymbol{\nu}\|_{\ell^p}^p = \sum_{i=1}^{\min\{k,l\}} \nu_i^p$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{\min\{k,l\}})$ *is the vector of singular values of* Σ *and where* $\|\cdot\|_{\ell^p}$ *is the standard* ℓ^p *norm in* $\mathbb{R}^d$*. In the covariance matrix case where* $\Sigma \in \mathbb{R}^{d\times d}$ *is symmetric and positive-definite,* $\|\Sigma\|_p^p = \mathrm{tr}\,(\Sigma^p) = \|\boldsymbol{\lambda}\|_{\ell^p}^p$ *where* $\boldsymbol{\lambda}$ *is the vector of eigenvalues of* Σ*.*

When $p = \infty$*, we have the standard operator norm on Euclidean space*

$$\|\Sigma\|_\infty = \sup_{v\in\mathbb{R}^d, \|v\|_{\ell^2}=1} \|\Sigma v\|_{\ell^2} = \sup_{v\in\mathbb{R}^d, \|v\|_{\ell^2}=1} v^{\mathrm{T}}\Sigma v.$$

In the covariance matrix setting, this coincides with the maximal eigenvalue of Σ*.*

Another family of norms that will be used is the collection of *entrywise* matrix norms denoted $\|\cdot\|_{\ell^p}$. These are the ℓ^p norms of a given matrix treated as a vector in $\mathbb{R}^{kl}$. That is, $\|\Sigma\|_{\ell^p}^p = \sum_{i=1}^k \sum_{j=1}^l \sigma_{i,j}^p$ where $\sigma_{i,j}$ is the ijth entry of Σ. Note that $\|\Sigma\|_2 = \|\Sigma\|_{\ell^2}$, which is referred to as the Frobenius or Hilbert-Schmidt norm. Besides the Hilbert-Schmidt and operator norms, the other main norm of interest in this manuscript is the trace norm or $p = 1$ Schatten norm.

The p-Schatten norms in the matrix setting can be generalized to the operator setting. Thus, we can define the class of p-Schatten operators mapping elements from one Hilbert space to another.

Definition 1.1.7 (p-Schatten norm for operators)**.** *Given two separable Hilbert spaces H_1 and H_2, a bounded linear operator $\Sigma : H_1 \to H_2$, and some $p \in [1, \infty)$, then the p-Schatten norm is $\|\Sigma\|_p^p = \operatorname{tr}\left((\Sigma^*\Sigma)^{p/2}\right)$. For $p = \infty$, the Schatten norm is the operator norm: $\|\Sigma\|_\infty = \sup_{f \in H_1}(\|\Sigma f\|_{H_2}/\|f\|_{H_1})$. In the case that Σ is compact, self-adjoint, and trace-class, then given the associated eigenvalues $\{\lambda_i\}_{i=1}^\infty$, the p-Schatten norm coincides with the standard ℓ^p norm of the eigenvalues:*

$$\|\Sigma\|_p^p = \begin{cases} \|\lambda\|_{\ell^p}^p = \sum_{i=1}^\infty |\lambda_i|^p, & p \in [1, \infty) \\ \max_{i \in \mathbb{N}} |\lambda_i|, & p = \infty \end{cases}$$

In the case that $p \in (0, 1)$, the above defined "p-Schatten norms" become quasinorms as the usual triangle inequality fails to hold. While it is possible that they can produce interesting results in the context of statistical applications, they will not be considered in this manuscript.

1.1.4 Metrics

The choice of metric on the space of covariance matrices or operators is of critical importance to the resulting statistical power achieved by the specific implemented procedures involving such metrics. The following metrics as well as others have been previously investigated in for covariance matrices in Dryden et al. (2009) and for covariance operators in Pigoli et al. (2014).

The main class of metrics considered are those branching from the p-Schatten norms discussed in the previous section. Specifically, for two matrices or operators S_1 and S_2, the p-Schatten metric is $d_p(S_1, S_2) = \|S_1 - S_2\|_p$ for $p \in [1, \infty]$. The resulting Banach space topology of such metrics allows for easy incorporation with a variety of concentration inequalities including the bounded differences inequality (Giné and Nickl, 2016, Section 3.3.4) and Talagrand's inequality (Giné and Nickl, 2016, Section 3.3.3).

Both the square root metric and the more general Procrustes size and shape distance have been shown to offer superior performance in statistical applications when compared with a variety of other metrics (Dryden et al., 2009; Pigoli et al., 2014; Cabassi et al., 2017). However, because these metrics do not correspond to a norm, their incorporation with concentration inequalities will require further and future research. The following definitions are given for real valued matrices, but can be easily applied to bounded linear operators.

Definition 1.1.8 (Square Root Distance)**.** *For two symmetric positive-definite matrices $S_1, S_2 \in \mathbb{R}^{d \times d}$,*

$$d_{sqrt}(S_1, S_2)^2 = \left\| S_1^{1/2} - S_2^{1/2} \right\|_2^2$$

where the matrix square root is defined above.

Definition 1.1.9 (Procrustes Distance)**.** *For two matrices $S_1, S_2 \in \mathbb{R}^{d\times d}$ that each have at least one square root,*

$$d_{Proc}(S_1, S_2)^2 = \inf_{R\in U(\mathbb{R}^{d\times d})} \|L_1 - L_2 R\|_2^2$$

where L_i is any matrix such that $S_i = L_i {L_i}^{\mathrm{T}}$ and where the infimum is taken over the set of $d \times d$ unitary matrices, $U(\mathbb{R}^{d\times d})$.

The Procrustes distance is a generalized form of the Square Root distance. Taking the infimum over all unitary matrices or operators has the effect of transforming L_2 in a unitary fashion to align it as closely as possible with L_1 with respect to the Frobenius / Hilbert-Schmidt distance. This feature is why Procrustes is used heavily in shape statistics and manifold valued data where one may want to first rotate and shift some object onto another before computing the standard distance. Past work has described the geodesics formed via the Procrustes distance as particularly useful in a statistical context. We make use of such paths when searching for sparse covariance matrix estimators in Section 2.2.2.

For two collections of vector, or equivalently two matrices in $\mathbb{R}^{d\times n}$, $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ with $X_i, Y_i \in \mathbb{R}^d$ for all $i = 1, \ldots, n$, we can define the $\ell^{p,q}$ distance as

$$d_{p,q}(X, \mathbf{Y}) = \left(\sum_{i=1}^{n} \|X_i - Y_i\|_{\ell^q}^p \right)^{1/p}. \tag{1.1.1}$$

Such a metric is used in Chapter 2 when investigating Lipschitz functions on covariance matrices.

1.1.5 Distributions

Throughout this work, we will make use of Rademacher random variables, which are alternatively referred to in the literature as symmetric Bernoulli random variables or random signs. Their definition is as follows.

Definition 1.1.10 (Rademacher Distribution)**.** *A random variable $\varepsilon \in \mathbb{R}$ has a Rademacher distribution if* $\mathrm{P}(\varepsilon = 1) = \mathrm{P}(\varepsilon = -1) = 1/2$.

In Chapter 4, we will require a slightly more generalized definition of the Rademacher distribution allowing for asymmetric probability masses on the points ± 1.

Definition 1.1.11 (Rademacher(p) Distribution)**.** *A random variable $\varepsilon \in \mathbb{R}$ has a Rademacher(p) distribution if* $\mathrm{P}(\varepsilon = 1) = p$ *and conversely* $\mathrm{P}(\varepsilon = -1) = 1 - p$ *for some $p \in [0, 1]$.*

The Rademacher distribution is widely applicable in both probabilistic and statistical contexts. It plays a major role in the probability in Banach spaces literature (Ledoux and Talagrand, 1991). The distribution has been used in generalized bootstrap methods (Arlot et al., 2010), for stochastic optimization procedures (Spall, 1992, 2005), and in the machine learning context in the form of *Rademacher Complexities* (Koltchinskii, 2001, 2006; Bartlett et al., 2002; Bartlett and Mendelson, 2003; Kloft and Blanchard, 2011; Cortes et al., 2013). The technique of Rademacher symmetrization used in conjunction with concentration inequalities is applied in Chapter 3. This technique originates from past work including Lounici and Nickl (2011); Kerkyacharian et al. (2012); Fan (2011).

We will also make use of empirical measures, denoted μ_n, with the standard definition as well as the reflected empirical measure, denoted μ_n^-.

Definition 1.1.12 (Empirical Measure and Reflected Empirical Measure)**.** *For independent and identically distributed random variables $X_1, \ldots, X_n \in \mathcal{X}$, the empirical measure is a random measure defined as*

$$\mu_n(A) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \in A}$$

for some measurable set $A \subseteq \mathcal{X}$. We will denote the empirical measure of the reflected variables $-X_1, \ldots, -X_n$ by μ_n^-.

1.1.6 Lipschitz functions

As will be mentioned in the following section and in many other areas throughout this manuscript, the concentration of measure phenomenon usually follows when random variables are combined in a "smooth" way or a "nice" way. This imprecise language usually can be read as referring to Lipschitz or locally Lipschitz functions. The following definitions can be made more general. However, in our case, we will always consider real valued functions either on a high dimensional Euclidean space or on some Hilbert space. Derivations of explicit Lipschitz constants for functions of interest on the space of covariance matrices can be found in Section 2.C.

Definition 1.1.13 (Lipschitz continuity)**.** *Given a metric space $(\mathcal{X}, d)$, then a function $f : \mathcal{X} \to \mathbb{R}$ is Lipschitz continuous if there exists a constant $K \geq 0$ such that*

$$|f(x) - f(y)| \leq K d(x, y)$$

for any $x, y \in \mathcal{X}$. The infimum taken over all $K \geq 0$ that make the above equation valid is referred to as the Lipschitz constant and denoted $K = \|f\|_{Lip}$.

Definition 1.1.14 (Local Lipschitz continuity)**.** *Given a metric space $(\mathcal{X}, d)$, then a function $f : \mathcal{X} \to \mathbb{R}$ is locally Lipschitz continuous if for any $x_0 \in \mathcal{X}$ there exists a constant $K_{x_0} \geq 0$ and a neighbourhood U of x_0 such that for any $x \in U$,*

$$|f(x_0) - f(x)| \leq K_{x_0} d(x_0, x).$$

1.2 Overview of concentration inequalities

Concentration of measure and the vast collection of concentration inequalities traces its history back to the works of Vitali Milman on Banach spaces and Paul Levy in probability theory. The theory has had and continues to make a substantial impact on a multitude of fields from pure analysis to probability and statistics. It ties together ideas from functional and geometric analysis with stochastic analysis, optimal transport, information theory, and many other diverse disciplines. Detailed overviews of the theory can be found in a collection of textbooks and monographs and the references therein (Ledoux and Talagrand, 1991; Steele, 1997; Ledoux, 2001; Milman and Schechtman, 2009; Boucheron et al., 2013; Habib et al., 2013; Giné and Nickl, 2016).

This so-called concentration of measure phenomenon can best be summed up by the words of Michel Talagrand:

> "A random variable that depends (in a 'smooth' way) on the influence of many independent variables (but not too much on any of them) is essentially constant." (Talagrand, 1996b)

In this statement, the "smooth way" generally means a Lipschitz function combining many independent random variables, and the phrase "essentially constant" generally means that the probability of the random variable deviating from some central point–e.g. its mean or median–is bounded by some sub-exponential or sub-Gaussian decay.

Concentration is closely linked to the geometric notion of isoperimetry and how measures concentrate in high dimensions. For example, a point selected uniformly at random from inside the n-dimensional unit sphere will with high probability be close to the surface. Similarly, a point selected uniformly at random from the surface of the n-dimensional unit sphere will with high probability be near the equator–which is paradoxically any equator! These examples motivate the results known as Lévy's inequalities (Levy, 1951).

Theorem 1.2.1 (Lévy's Inequalities)**.** *For a random variable X in some metric measure space $\mathcal{X}$ with metric $d(\cdot, \cdot)$ and probability measure $\mathrm{P}(\cdot)$, define the concentration*

function to be

$$\alpha(r) = \sup_{A \subset \mathcal{X}: \mathrm{P}(A) \geq 1/2} \mathrm{P}\left(d(X, A) \geq r\right)$$

where $d(X, A) = \inf_{a \in A} d(X, a)$*. Then for any 1-Lipschitz function* $f(\cdot)$ *with median denoted by* $\mathrm{M}f(X)$*, we have that*

$$\begin{aligned} \mathrm{P}\left(f(X) \geq \mathrm{M}f(X) + r\right) &\leq \alpha(r), \quad \text{and} \\ \mathrm{P}\left(f(X) \leq \mathrm{M}f(X) - r\right) &\leq \alpha(r). \end{aligned}$$

Thus, controlling the concentration function $\alpha(\cdot)$ allows for controlling nice functions of random variables on a given space.

1.2.1 Examples

As a first example, we consider the sub-Gaussian concentration for bounded real-valued random variables as it is stated in Hoeffding's inequality (Hoeffding, 1963).

Theorem 1.2.2 (Hoeffding's Inequality)**.** *Let* $X_1, \ldots, X_n \in \mathbb{R}$ *be independent random variables such that* $X_i \in [a_i, b_i]$*, and define* $S_n = \sum_{i=1}^n X_i$*. Then,*

$$\mathrm{P}\left(S_n \geq \mathrm{E}S_n + r\right) \leq \exp\left(\frac{-2r^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

This result is achieved by noting that for a bounded real-valued random variable $X_i \in [a_i, b_i]$, the worst possible variance is $\mathrm{Var}\left(X_i\right) \leq (b_i - a_i)^2/4$. As the bound on the right hand side is of the form $\exp(-Cr^2)$ for some fixed constant $C > 0$, we have sub-Gaussian style concentration inequality. This result is extended to the martingale setting via the Azuma-Hoeffding inequality (Azuma, 1967). It is also generalized by the bounded differences inequality (McDiarmid, 1989). We will briefly consider the bounded differences inequality applied to the case of Banach space valued random variables bounded in norm in Section 2.3.2

As noted, Hoeffding's inequality assumes the worst–i.e. largest–possible variance. In the case that an estimate of the variance is possible, we have alternative concentration inequalities known as Bennett's and Bernstein's inequality (Bennett, 1962; Bernstein, 1924).

Theorem 1.2.3 (Bennett's Inequality)**.** *Let* $X_1, \ldots, X_n \in \mathbb{R}$ *be random variables with zero mean and such that* $|X_i| \leq c$ *for* $i = 1, \ldots, n$ *for some fixed* $c > 0$*. Define* $S_n = \sum_{i=1}^n X_i$ *and* $v_n = \sum_{i=1}^n \mathrm{Var}\left(X_i\right)$*. Then, for any* $r > 0$*,*

$$\mathrm{P}\left(S_n \geq r\right) \leq \exp\left(-\frac{v_n}{c^2} h(cr/v_n)\right)$$

where $h(u) = (1 + u)\log(1 + u) - u$*.*

Theorem 1.2.4 (Bernstein's Inequality)**.** *Given the same set up as Bennett's inequality,*

$$\mathrm{P}\left(S_n \geq r\right) \leq \exp\left(-\frac{r^2}{2(v_n + cr/3)}\right)$$

In these concentration inequalities, we have the feature that for small values of $r > 0$, the concentration takes on a sub-Gaussian form whereas for larger values of $r > 0$, the concentration becomes sub-exponential. This type of concentration is demonstrated in the case of self bounding functions (Boucheron et al., 2000). It also occurs in the celebrated result known as Talagrand's concentration inequality (Talagrand, 1996a) and the various extensions and refinements of this result (Ledoux, 1997; Massart, 2000; Panchenko, 2001; Bousquet, 2003; Klein and Rio, 2005). We will rely heavily on Talagrand's inequality to develop the statistical methodology for covariance operators in Chapter 3.

1.3 Connections to past research

One of the main goals of this manuscript, and most research in general, is to build on, extend, and improve past work on such problems. The research contained within this manuscript chronologically began with Chapter 3 building off of the past works of Panaretos et al. (2010); Fremdt et al. (2013); Pigoli et al. (2014), which are all concerned with performing a two sample test for the equality of the covariance operators of each sample. More formally, given two sets of independent and identically distributed functional observations $X_1, \ldots, X_n \in L^2[0,1]$ and $Y_1, \ldots, Y_m \in L^2[0,1]$ with unknown covariance operators Σ_X and Σ_Y, respectively, we wish to test

$$H_0 : \Sigma_X = \Sigma_Y \qquad\qquad H_1 : \Sigma_X \neq \Sigma_Y. \tag{1.3.1}$$

In Panaretos et al. (2010), the data is additionally considered to be observed instances of a Gaussian process. They take advantage of functional principal components and the Karhunen-Loève expansion to represent the data (Adler, 1990; Hall and Hosseini-Nasab, 2006; Horváth and Kokoszka, 2012). Thus, given the set $\{\phi_j^X\}_{j=1}^{\infty}$ of the orthonormal eigenfunctions of the covariance operator Σ_X with the corresponding set of ordered eigenvalues $\{\lambda_j^X\}_{j=1}^{\infty}$ as well as a collection of univariate $Z_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ for $i = 1, \ldots, n$ and $j \geq 1$, then

$$X_i(t) = \sum_{j=1}^{\infty} (\lambda_j^X)^{1/2} Z_{i,j} \phi_j^X(t), \quad \text{for } t \in [0,1].$$

In practice, the eigenfunctions of the empirical estimate of Σ_X are used to construct a finite dimensional representation of the X_i. From this expansion, they construct a

test statistic for testing Hypothesis 1.3.1 making use of Parseval's Theorem (Rudin, 1987, Theorem 4.18) and the properties of the Hilbert-Schmidt norm.

In short, let $\hat{\Sigma}_{XY}$ be the pooled empirical covariance operator from combining both sets of data with eigenfunctions $\{\hat{\phi}_j^{XY}\}_{j=1}^{n+m}$. Furthermore, let $\hat{\lambda}_j^X$ and $\hat{\lambda}_j^Y$ be the jth coefficient in the above expansion of the X_i and Y_i, respectively, with respect to the basis $\{\hat{\phi}_j^{XY}\}_{j=1}^{n+m}$ Then, Theorem 1 of Panaretos et al. (2010) states that the test statistic is

$$T(k) = \frac{nm}{2(n+m)} \sum_{i,j=1}^{k} \left\langle (\hat{\Sigma}_X - \hat{\Sigma}_Y)\hat{\phi}_j^{XY}, \hat{\phi}_j^{XY} \right\rangle^2 \times$$
$$\times \left(\left(\frac{n}{n+m}\hat{\lambda}_i^X + \frac{m}{n+m}\hat{\lambda}_i^Y \right) \left(\frac{n}{n+m}\hat{\lambda}_j^X + \frac{m}{n+m}\hat{\lambda}_j^Y \right) \right)^{-1} \to \chi^2\left(k(k+1)/2\right)$$

for some choice of $k \leq \min\{\text{rank}(\hat{\Sigma}_X), \text{rank}(\hat{\Sigma}_Y)\}$ where the convergence is in distribution as $n+m \to \infty$ with the asymptotics such that $n/(n+m) \to c \in (0,1)$.

In the work of Fremdt et al. (2013), they also project both samples of data onto the eigenfunctions of the pooled empirical covariance operator. However, they do not initially assume that the data arises from a Gaussian process. Their Theorem 1 proposes another asymptotic test statistic converging in distribution to $\chi^2\left(k(k+1)/2\right)$. This test statistic is constructed by projecting the components of the difference $\hat{\Sigma}_X - \hat{\Sigma}_Y$ onto the basis $\phi_i^{XY}\phi_j^{XY}$ for $i,j = 1\ldots,k$, transforming the lower triangular $k(k+1)/2$ entries of that matrix into a vector $\hat{\xi}$, and then estimating the asymptotic covariance matrix of $\hat{\xi}$ denoted as $\hat{L}$. Then, the test statistic is $T = (nm/(n+m))\hat{\xi}^{\mathrm{T}}\hat{L}\hat{\xi}$.

Following these two works, Pigoli et al. (2014) approach the same hypothesis test by further removing the asymptotics and the reliance on the Hilbert-Schmidt topology. In their article, they consider a wide variety of metrics to use to compare the two empirical covariance operators $\hat{\Sigma}_X$ and $\hat{\Sigma}_Y$. As the alternative metrics do not necessarily yield elegant test statistics with asymptotic distributions, the article implements these metrics via an approximate permutation test. Their library of metrics includes the Procrustes, Square-Root, and some based on the p-Schatten norms. Such permutation tests were recently extended to k sample tests for $k > 2$ in Cabassi et al. (2017). Chapter 3 of this manuscript aims to improve upon this methodology by using concentration inequalities to circumvent the computationally costly permutation tests.

From the perspective of concentration inequalities, many of the ideas used in Chapter 3 for constructing confidence sets in Banach spaces can be found in Lounici and Nickl (2011) in the empirical process context. In that article, the statistical deconvolution model, $Y = X + \xi$, is considered where $Y_1, \ldots, Y_n$ are observed real

valued random variable, $X_1, \ldots, X_n$ are unknown, $\xi_1, \ldots, \xi_n$ are errors independent of the X_i, and the goal is to recover the probability density of the X_i from the noisy observed Y_i. Lounici and Nickl (2011) demonstrate a lower bound on the minimax sup-norm risk and show that a wavelet estimator achieves this bound. In the process of constructing such a bound, they prove a Berstein-type inequality for Rademacher processes branching from versions of the upper and lower deviation versions of Talagrand's inequality from Bousquet (2003) and Klein and Rio (2005), respectively.

Specifically, from Proposition 5 of Lounici and Nickl (2011), let $X, X_1, \ldots, X_n$ are independent and identically distributed on a measure space $(S, \mathcal{A})$, and let $\mathcal{F}$ be a countable class of real-valued functions on S such that $|f| \leq 1/2$ for all $f \in \mathcal{F}$ and with some weak variance term $\sigma^2 \geq \sup_{f \in \mathcal{F}} \mathrm{E} f^2(X)$. Consequently,

$$\mathrm{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \mathrm{E} f(X)) \right\|_{\mathcal{F}} \geq 6 \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right\|_{\mathcal{F}} + 10 \left(\frac{(r + \log 2)\sigma^2}{n} \right)^{1/2} + 22 \left(\frac{r + \log 2}{n} \right) \right) \leq \mathrm{e}^{-r}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed Rademacher random variables and where $\|\cdot\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}}(\cdot)$.

Chapter 2

Concentration for covariance matrices

2.1 Introduction

Covariance matrices and accurate estimators of such objects are of critical importance in statistics. Various standard techniques including principal components analysis and linear and quadratic discriminant analysis rely on an accurate estimate of the covariance structure of the data. Applications can range from genetics and medical imaging data to climate and other types of data. Furthermore, in the era of high dimensional data, classical asymptotic estimators perform poorly in applications (Stein, 1975; Johnstone, 2001). To see this, Figure 2.1 displays the rapidly increasing expected distance between the empirical covariance estimator for the identity matrix and the true identity matrix given a sample of $n = 100$ observations from a multivariate Gaussian distribution and the multivariate t distribution with three degrees of freedom. Thus, many alternative estimators for the covariance matrix have been proposed working under the assumption of sparsity (Pourahmadi, 2011), which is, in a qualitative sense, the case when most of the off-diagonal entries are zero. Beyond mere theoretical interest, the assumption of sparsity is widely applicable to real data analysis as it is reasonable for the practitioner to believe that many of the variable pairings to be studied will be uncorrelated. Thus, it is desirable to tailor covariance estimation procedures given this assumption of sparsity.

Sparsity in the simplest sense implies some bound on the number of non-zero entries in the columns of a covariance matrix. Thus, given a $\Sigma \in \mathbb{R}^{d\times d}$ with entries $\sigma_{i,j}$ for $i, j = 1, \ldots, d$, we have that there exists some constant $c > 0$ such that $\max_{j=1,\ldots,d} \sum_{i=1}^{d} \mathbf{1}\left[\sigma_{i,j} \neq 0\right] \leq c$. This can be generalized to "approximate sparsity" as in Rothman et al. (2009) by $\max_{j=1,\ldots,d} \sum_{i=1}^{d} |\sigma_{i,j}|^q \leq c$ for some $q \in [0, 1)$. Furthermore, Cai and Liu (2011) define a broader approximately sparse class by

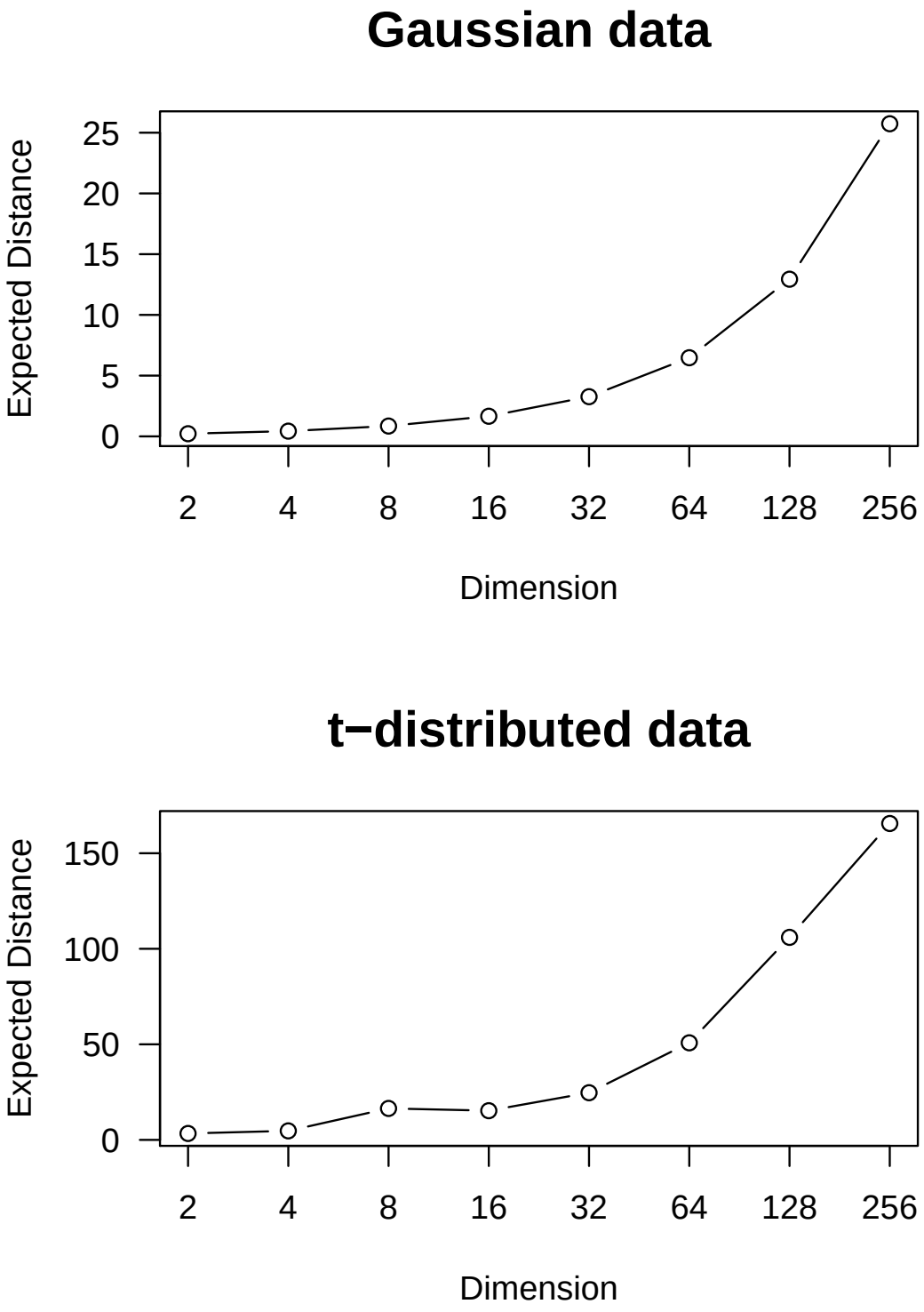


Figure 2.1: A plot of the expected distance from the empirical estimate of the covariance matrix to the true covariance matrix in the Frobenius norm. Here, the true covariance is the d dimensional identity matrix, and the empirical estimator is constructed from a sample of $n = 100$ iid random draws from a multivariate Gaussian distribution (top) and multivariate t-distribution with three degrees of freedom (bottom).

bounding weighted column sums of Σ. In El-Karoui (2008), a similar notion referred to as "β-sparsity" is defined, which stems from associating the covariance with an adjacency matrix for a graph and controlling the number of closed walks of a given length. Such classes of sparse covariance matrices allow for good theoretical performance of estimators.

One class of estimators for large sparse covariance matrices are shrinkage estimators that follow a James-Stein approach by shrinking estimated eigenvalues, eigenvectors, or the matrix itself towards some desired target (Haff, 1980; Dey and Srinivasan, 1985; Daniels and Kass, 1999, 2001; Ledoit and Wolf, 2004; Hoff, 2009; Johnstone and Lu, 2012). Another class of sparse estimators are those that regularize the estimate with lasso-style penalties (Rothman, 2012; Bien and Tibshirani, 2011). Yet another class consists of thresholding estimators, which declare the covariance between two variables to be zero, if the estimated value is smaller than some threshold (Bickel and Levina, 2008a,b; Rothman et al., 2009; Cai and Liu, 2011). Beyond these, there are other methods such as banding and tapering, which apply only when the variables are ordered or a notation of proximity exists–e.g. spatial, time series, or longitudinal data. As we will not assume such an ordering and strive to construct a methodology that is permutation invariant with respect to the variables, these approaches will not be considered. Lastly, there has also been substantial work into the estimation of the precision or inverse covariance matrix. While it is easily possible that our approach can be adapted to this setting, it will not be considered in this manuscript and will, hence, be reserved for future research.

In this chapter, we propose of novel approach to the estimation of sparse covariance matrices making use of concentration inequality based confidence sets. Similar confidence sets will be constructed in Chapter 3 for covariance operators in the functional data setting. This approach takes inspiration from the shrinkage, thresholding, and penalization methods of sparse covariance estimation. In short, consider a sample of real vector valued data $X_1, \ldots, X_n \in \mathbb{R}^d$ with zero mean and unknown covariance matrix Σ_0. Concentration inequalities are used to construct a non-asymptotic confidence set for Σ_0 about the empirical estimate of the unknown covariance matrix, $\hat{\Sigma}^{\text{emp}} = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^{\text{T}}$ where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ is the sample mean. While, it has been noted–for example, see Cai and Liu (2011)–that $\hat{\Sigma}^{\text{emp}}$ may be a poor estimator when the dimension d is large and Σ_0 is sparse, the confidence set is still valid given a desired coverage of $(1 - \alpha)$. To construct a better estimator, we propose to search this confidence set for an estimator $\hat{\Sigma}^{\text{sp}}$ which optimizes some sparsity criterion to be concretely defined later. This estimation method adapts to the uncertainty of $\hat{\Sigma}^{\text{emp}}$ in the high dimensional setting, $d \gg n$, by widening the confidence set and thus allowing our sparse estimator to lie far away from the empirical estimate. Furthermore, given some distributional assumptions, the

concentration inequalities provide us with non-asymptotic dimension-free confidence sets allowing for very desirable convergence results.

Many established methods for sparse estimation make use of a regularization or penalization term incorporated to enforce sparsity (Rothman, 2012; Bien and Tibshirani, 2011). In some sense, our proposed method can be considered to be in this class of estimators. However, we do not enforce sparsity via some lasso-style penalization term, but enforce it through the choice of α. The larger our $(1-\alpha)$-confidence set is, the sparser our estimator is allowed to be. Thus, as with other regularized estimators, the practitioner will have to make a choice of α to achieve the desired level of sparsity or implement the cross-validation selection method for α, which is described in Section 2.2.3. Furthermore, our estimation technique implements a binary search procedure resulting in a highly efficient algorithm especially when compared to the more laborious optimization required by lasso penalization.

In this chapter, the general estimation procedure is outlined in Section 2.2. Two different search methods are proposed as well as a cross-validation technique for tuning the method's parameter. In Section 2.3, three different types of concentration inequalities are considered for specifically log-concave measures, bounded random variables, and sub-exponential distributions. Lastly, Section 2.4 details comprehensive simulations comparing our concentration approach to sparse estimation with standard techniques such as thresholding and penalization. Both extensive simulation experiments and a real data set of gene expressions for small round blue cell tumours are considered.

2.2 Sparse Estimation Procedure

Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be a sample of n independent and identically distributed random vectors with unknown $d \times d$ covariance matrix Σ_0. Define the empirical estimate of Σ_0 to be $\hat{\Sigma}^{\text{emp}} = n^{-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$ where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ is the sample mean. The goal of the following procedure is to construct a sparse estimator, $\hat{\Sigma}^{\text{sp}}$, for Σ_0 by first constructing a confidence set for Σ_0 about the estimator $\hat{\Sigma}^{\text{emp}}$ and then searching this set for the sparsest member. Two different search methods for such a sparse member are outlined in Sections 2.2.1 and 2.2.2.

To construct such a confidence set about $\hat{\Sigma}^{\text{emp}}$, concentration inequalities are employed. Specific inequalities are chosen based on data assumptions and are discussed in subsequent Sections 2.3.1, 2.3.2, and 2.3.3. In general, the inequalities all take a similar form. Let $d(\cdot,\cdot)$ be some metric measuring the distance between two covariance matrices, and let $\psi : \mathbb{R} \to \mathbb{R}$ be monotonically increasing. Then, the

Assumption on X_i	$d(\hat{\Sigma}^{\text{sp}}, \Sigma_0)$	r_α
Log-Concave Measure	$\left\|\hat{\Sigma}^{\text{sp}} - \Sigma_0\right\|_p^{1/2}$	$\sqrt{(-2/c_0 n)\log\alpha}$
Bounded in norm	$\left\|\hat{\Sigma}^{\text{sp}} - \Sigma_0\right\|_p$	$U\sqrt{(-1/2n)\log\alpha}$
Sub-Exponential Measure	$\left\|\hat{\Sigma}^{\text{sp}} - \Sigma_0\right\|_p^{1/2}$	$\max\{-K\log\alpha/\sqrt{n}, \sqrt{-K\log\alpha}\}$

Table 2.1: Specific metrics $d(\cdot,\cdot)$ and deviation thresholds r_α given specific assumptions on the data X_i discussed in subsequent sections.

general form of the concentration inequalities is

$$\mathrm{P}\left(d(\Sigma_0, \hat{\Sigma}^{\text{emp}}) \geq \mathrm{E}d(\Sigma_0, \hat{\Sigma}^{\text{emp}}) + r\right) \leq \mathrm{e}^{-\psi(r)},$$

which is a bound on the tail of the distribution of $d(\Sigma_0, \hat{\Sigma}^{\text{emp}})$ as it deviates above its mean. Thus, to construct a $(1-\alpha)$-confidence set, the variable $r = r_\alpha$ is chosen such that $\exp(-\psi(r_\alpha)) = \alpha$. This r_α will be referred to as the *deviation threshold*. Table 2.1 contains some explicit choices for the metric and deviation threshold given specific assumptions on the X_i, which are used to drive the choice of concentration inequality. These three cases are discussed separately in Section 2.3.

Now, let $\hat{\Sigma}^{\text{sp}}$ be our sparse estimator for Σ_0. We want these two to be close in the sense of the above confidence set and therefore choose a $\hat{\Sigma}^{\text{sp}}$ such that $d(\hat{\Sigma}^{\text{sp}}, \hat{\Sigma}^{\text{emp}}) \leq r_\alpha$. Consequently, we have that

$$\begin{aligned}
&\mathrm{P}\left(d(\hat{\Sigma}^{\text{sp}}, \Sigma_0\ ) \geq \mathrm{E}d(\hat{\Sigma}^{\text{emp}}, \Sigma_0) + 2r_\alpha\right) \\
&\qquad \leq \mathrm{P}\left(d(\hat{\Sigma}^{\text{sp}}, \hat{\Sigma}^{\text{emp}}) + d(\hat{\Sigma}^{\text{emp}}, \Sigma_0) \geq \mathrm{E}d(\hat{\Sigma}^{\text{emp}}, \Sigma_0) + 2r_\alpha\right) \\
&\qquad \leq \mathrm{P}\left(d(\hat{\Sigma}^{\text{emp}}, \Sigma_0) \geq \mathrm{E}d(\hat{\Sigma}^{\text{emp}}, \Sigma_0) + r_\alpha\right) \\
&\qquad \leq \exp(-\psi(r_\alpha)) = \alpha
\end{aligned}$$

To actually identify such a $\hat{\Sigma}^{\text{sp}}$, we require some criterion to optimize over all elements of the confidence set. Two such methods are proposed in the following subsections. The zeroing method in Section 2.2.1 takes inspiration from thresholding techniques for sparse covariance estimation (Bickel and Levina, 2008a; Rothman et al., 2009; Cai and Liu, 2011). It begins with $\hat{\Sigma}^{\text{emp}}$ and attempts to zero as many entries as possible in the empirical estimate while still remaining in the confidence set, which is similar to applying a hard thresholding estimator restricted to the confines of the confidence set. The Procrustes method in Section 2.2.2 is more closely related to the shrinkage estimators (Daniels and Kass, 1999, 2001; Hoff, 2009; Johnstone and Lu, 2012). It chooses $\hat{\Sigma}^{\text{sp}}$ to be a convex combination of $\hat{\Sigma}^{\text{emp}}$ and some sparse

target matrix using the Procrustes size and shape distance, which has been shown to be a useful metric when one is concerned with inference in the space of covariance matrices (Dryden et al., 2009).

2.2.1 Zeroing Method

Beginning with $\hat{\Sigma}^{\text{emp}}$, the goal of this method is to remove as many entries of $\hat{\Sigma}^{\text{emp}}$ as possible while respecting the restriction that $d(\hat{\Sigma}^{\text{sp}}, \hat{\Sigma}^{\text{emp}}) \leq r_\alpha$. Here, the (i,j) entry of $\hat{\Sigma}^{\text{sp}}$ is denoted as $\tilde{\sigma}_{i,j}$ and the (i,j) entry of $\hat{\Sigma}^{\text{sp}}_k$ is denoted as $\tilde{\sigma}^k_{i,j}$

0. Set $\hat{\Sigma}^{\text{sp}}_0 = \hat{\Sigma}^{\text{emp}}$. Choose an α and compute r_α. Later it will be shown that this method is fairly robust to the choice of α. The cross-validation technique described below in Section 2.2.3 can be used to select a desirable α in practice. We also observed in the Gaussian data experiments of Section 2.4.1 that $\alpha \approx 10^{-6}$ gave good performance.

1. While $d(\hat{\Sigma}^{\text{sp}}_k, \hat{\Sigma}^{\text{emp}}) \leq r_\alpha$ and $\hat{\Sigma}^{\text{sp}}_k$ has at least one non-zero off-diagonal entry.

   (a) Choose the smallest non-zero off-diagonal entry in $\hat{\Sigma}^{\text{sp}}_k$ and construct $\hat{\Sigma}^{\text{sp}}_{k+1}$ by setting it equal to zero. That is, determine (i,j) such that $i < j$ and $0 < |\tilde{\sigma}^k_{i,j}| \leq |\tilde{\sigma}^k_{i',j'}|$ for all $i' \neq i$ and $j' \neq j$ such that $|\tilde{\sigma}^k_{i',j'}| > 0$. If the set of such pairs (i,j) has more than one element, then choose one pair uniformly at random and continue.

   (b) Construct $\hat{\Sigma}^{\text{sp}}_{k+1}$ with entries $\tilde{\sigma}^{k+1}_{i,j} = \tilde{\sigma}^{k+1}_{j,i} = 0$ and $\tilde{\sigma}^{k+1}_{i',j'} = \tilde{\sigma}^k_{i',j'}$ for all other $(i',j') \neq (i,j)$.

2. Denote $\hat{\Sigma}^{\text{sp}}$ the final matrix resulting from this recursion. If $\hat{\Sigma}^{\text{sp}}$ is not positive semi-definite, then project it onto the space of positive semi-definite matrices by mapping the negative eigenvalues to zero.

In the case that the metric $d(\cdot,\cdot)$ is a monotonically increasing function of the Hilbert-Schmidt / Frobenius norm $\|\hat{\Sigma}^{\text{sp}}_k - \hat{\Sigma}^{\text{emp}}\|_2$, then the sequence $d(\hat{\Sigma}^{\text{sp}}_k, \hat{\Sigma}^{\text{emp}})$ will be increasing in k. This is true because the Frobenius norm is equivalent to the ℓ^2 norm of the entries in the matrix, and as we run the algorithm, more entries in the difference $\hat{\Sigma}^{\text{sp}}_k - \hat{\Sigma}^{\text{emp}}$ will be non-zero. This property guarantees that the above algorithm will find the sparsest $\hat{\Sigma}^{\text{sp}}$ in the confidence set in the sense of having the most zero entries. However, for an arbitrary metric, this sequence may not necessarily be strictly increasing in k. Another commonly used norm, which will be shown in Section 2.4 to give superior performance on simulated data, is the operator norm $\|\hat{\Sigma}^{\text{sp}}_k - \hat{\Sigma}^{\text{emp}}\|_\infty$, which does not yield a monotonically increasing sequence. Though, this sequence is roughly increasing in the sense that it is lower bounded by definition by the maximum ℓ^2 norm of the columns of $\hat{\Sigma}^{\text{sp}}_k - \hat{\Sigma}^{\text{emp}}$, which is an

increasing sequence. Furthermore, it is upper bounded by the ℓ^1 norm of the columns of $\hat{\Sigma}_k^{\text{sp}} - \hat{\Sigma}^{\text{emp}}$, which follows from the Gershgorin circle theorem (Iserles, 2009), and which is also an increasing sequence. In practice, the operator norm in particular gives superior performance in the numerical simulations of Section 2.4.

From a computational perspective, the above algorithm as stated requires an unacceptable $O(d^2)$ eigenvalue decompositions as thus does not scale well as the dimension of the matrix increases. To account for this, a binary search routine can be incorporated resulting in a reduction to $O(\log_2 d)$ eigenvalue decompositions. In short, set $z_k = \lfloor (d^2 - d)/2^k \rfloor$ to be the number of non-zero off-diagonal entries to set to zero in step (1a). If the resulting $\hat{\Sigma}_{k+1}^{\text{sp}}$ from step (1b) is such that $d(\hat{\Sigma}_{k+1}^{\text{sp}}, \hat{\Sigma}^{\text{emp}}) \leq r_\alpha$, then continue as normal and attempt to remove z_{k+1} more entries. Otherwise, if $d(\hat{\Sigma}_{k+1}^{\text{sp}}, \hat{\Sigma}^{\text{emp}}) \geq r_\alpha$, set $\hat{\Sigma}_{k+1}^{\text{sp}} \leftarrow \hat{\Sigma}_k^{\text{sp}}$, then continue again as normal with z_{k+1} as before.

2.2.2 Procrustes Method

Past research into estimation and hypothesis testing for covariance matrices and operators has highlighted the superior performance of the Procrustes size and shape distance when compared with other metrics (Dryden et al., 2009; Pigoli et al., 2014; Cabassi et al., 2017). The intuition behind this metric and why it is popular in the context of shape analysis is that it allows for unitary transformations to best align the two objects under scrutiny.

In the context of sparse estimation, the Procrustes distance is used to construct $\hat{\Sigma}^{\text{sp}}$ as a convex combination of $\hat{\Sigma}^{\text{emp}}$ and some sparse target matrix Σ^{tar} that presumably lies outside of the confidence set. Hence, this approach attempts to move or shrink from the empirical estimator to the sparse target along a path determined by the Procrustes metric. Specifically, set $L^{\text{emp}} = (\hat{\Sigma}^{\text{emp}})^{1/2}$ and $L^{\text{tar}} = (\Sigma^{\text{tar}})^{1/2}$, and construct the estimator as a function of some $\gamma \in [0, 1]$ to be

$$\hat{\Sigma}^{\text{sp}}(\gamma) = \left(L^{\text{emp}} + \gamma(L^{\text{tar}}R - L^{\text{emp}})\right)\left(L^{\text{emp}} + \gamma(L^{\text{tar}}R - L^{\text{emp}})\right)^{\text{T}}$$

where $R = UV^{\text{T}}$ and U and V are, respectively, the left and right matrices of singular vectors for the matrix $(L^{\text{tar}})^{\text{T}}L^{\text{emp}}$ (Pigoli et al., 2014, Section 3). The argument $\gamma \in [0, 1]$ is chosen to be as large as possible while it still holds that $d(\hat{\Sigma}^{\text{sp}}(\gamma), \hat{\Sigma}^{\text{emp}}) \leq r_\alpha$.

This method finds the estimator closest to Σ^{tar} with respect to the Procrustes distance that is still in some confidence ball about $\hat{\Sigma}^{\text{emp}}$. In practice, a choice of Σ^{tar} must be made based on some assumption regarding the nature of the true Σ_0. In the case of sparse estimation, either I_d, the $d \times d$ identity matrix, or the diagonal of $\hat{\Sigma}^{\text{emp}}$ are reasonable choices for Σ^{tar}. In this way, the Procrustes method has a

semi-Bayesian feel as we are compromising between the empirical estimate and some prior chosen sparse target.

It is easily seen that this distance is an increasing function of γ for any p-Schatten norm. Thus, our goal is to determine the maximal value of γ such that $d_{\text{Proc}}(\hat{\Sigma}^{\text{sp}}(\gamma), \hat{\Sigma}^{\text{emp}}) \leq r_\alpha$. To compute this estimator in practice, a binary search procedure similar to that for the above zeroing method can be implemented. Begin with the initial values $\gamma = \delta = 0.5$. If $d(\hat{\Sigma}^{\text{sp}}(\gamma), \hat{\Sigma}^{\text{emp}}) \leq r_\alpha$, set $\gamma \leftarrow \gamma + \delta/2$ and $\delta \leftarrow \delta/2$. Otherwise set $\gamma \leftarrow \gamma - \delta/2$ and $\delta \leftarrow \delta/2$. This will quickly converge on the optimal choice of γ.

2.2.3 Cross-Validation

In practice, an optimal value of $\alpha \in (0, 1)$ must be chosen to enforce the proper amount of sparsity. Beyond that, many of the concentration inequalities arrive with finite but unknown coefficients that may only have loose upper bounds known. Hence, we propose a cross-validation technique for tuning α, which takes its inspiration from the similar technique proposed in the thresholding literature (Bickel and Levina, 2008a; Rothman et al., 2009; Cai and Liu, 2011).

Given $n = 2m$ observations, we split the data randomly in half to get $X_1^1, \ldots, X_m^1$ and $X_1^2, \ldots, X_m^2$. Then, the two empirical estimators are constructed $\hat{\Sigma}_1^{\text{emp}}$ and $\hat{\Sigma}_2^{\text{emp}}$. The desired sparsifying procedure is applied to $\hat{\Sigma}_2^{\text{emp}}$ for a variety of $\alpha \in \mathcal{A}$ resulting in the collection of estimators $\{\hat{\Sigma}_\alpha^{\text{sp}}\}_{\alpha \in \mathcal{A}}$. The value of α chosen as $\alpha = \arg\min_{\alpha \in \mathcal{A}} d(\hat{\Sigma}_1^{\text{emp}}, \hat{\Sigma}_\alpha^{\text{sp}})$ for some metric $d(\cdot, \cdot)$. This process is repeated k times resulting in the set $\{\alpha_1, \ldots, \alpha_k\}$. Then, the cross-validated choice is the average of the α_i in the log domain, which is $\alpha = \exp(k^{-1} \sum_{i=1}^k \log \alpha_i)$. The reason for the log, as will be seen in the following sections, is that our deviation threshold r_α is often a function of $\log \alpha$ stemming from the application of the concentration inequalities.

2.3 Estimation of sparse covariance

The following three subsections detail different assumptions on the data under scrutiny and the specific concentration results that apply in these cases. We consider sub-Gaussian concentration for both log-concave measures and bounded random variables. We also consider sub-exponential concentration. However, this collection is by no means exhaustive. Given the wide variety of concentration inequalities being researched, our approach can be applied much more widely than to merely these three settings.

2.3.1 Log-Concave Measures

In this section, the general methods from Section 2.2 are specialized for an independent and identically distributed sample $X_1, \ldots, X_n \in \mathbb{R}^d$ whose common measure μ is *strongly log-concave*. This property implies dimension-free sub-Gaussian concentration and includes such common distributions as the multivariate Gaussian, Chi, and Dirichlet distributions.

Definition 2.3.1 (Strongly log-concave measure)**.** *A measure μ on $\mathbb{R}^d$ is strongly log-concave if there exists a $c > 0$ such that $d\mu = \mathrm{e}^{-U(x)}dx$ and Hess$(U) - cI_d \succeq 0$ (i.e. the matrix is non-negative-definite) where Hess(U) is the $d \times d$ matrix of second derivatives.*

The corollary below follows from Corollary 2.B.4 and the other results contained within Appendix 2.B.1. For a detailed exposition of how Gaussian concentration is established for log-concave measures, see Chapter 5 of Ledoux (2001).

Corollary 2.3.2. *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ have measures $\mu_1, \ldots, \mu_n$, which are all strongly log-concave with coefficients $c_1, \ldots, c_n$, respectively. Let $\nu = \mu_1 \otimes \ldots \otimes \mu_n$ be the product measure on $\mathbb{R}^{d \times n}$. Then, for any 1-Lipschitz $\phi : (\mathbb{R}^d)^n \to \mathbb{R}$ and for any $r > 0$,*

$$\mathrm{P}_\nu \left(\phi(X_1, \ldots, X_n) \geq \mathrm{E}\phi(X_1, \ldots, X_n) + r \right) \leq \mathrm{e}^{-\min_i c_i r^2/2}.$$

Example 2.3.3 (Multivariate Gaussian measure)**.** *For an arbitrary Gaussian measure on $\mathbb{R}^d$ with zero mean and covariance Σ, we have that $U(x) = x^{\mathrm{T}}\Sigma^{-1}x/2$. Thus, Hess$(U) = \Sigma^{-1}$, and letting $\{\lambda_i\}_{i=1}^d$ be the eigenvalues of Σ, any $0 < c \leq \min_{i=1,\ldots,d} \lambda_i^{-1} = (\max_i \lambda_i)^{-1}$ satisfies the above definition. Applying Corollary 2.3.2, we have that for $X_1, \ldots, X_n \in \mathbb{R}^d$ independent and identically distributed multivariate Gaussian random variables,*

$$\mathrm{P} \left(\phi(X_1, \ldots, X_n) \geq \mathrm{E}\phi(X_1, \ldots, X_n) + r \right) \leq \mathrm{e}^{-r^2/2\lambda_0}$$

where $\lambda_0 = \max_{i=1,\ldots,d} \lambda_i$.

Example 2.3.4 (Dirichlet Distribution)**.** *For $X \sim$ Dirichlet$((\alpha_1, \ldots, \alpha_d))$ with $\alpha_i > 1$ for all i, the exponent $U = \sum_{i=1}^d (1 - \alpha_i) \log x_i$ and Hess(U) is a diagonal matrix with entries $(\alpha_1 - 1)/x_1^2, \ldots, (\alpha_d - 1)/x_d^2$. Thus, the maximum c to ensure that Hess$(U) - cI_d \succeq 0$ for all values of $x_i \in (0, 1)$ with $\sum_{i=1}^d x_i = 1$ is $c = \min_{i=1,\ldots,d} \alpha_i - 1$.*

This example illustrates one of the necessary conditions for a probability density to be log-concave. That is, the density must be unimodal as is the Dirichlet distribution in the given case where the parameters $\alpha_i > 1$ for all $i = 1 \ldots, d$.

To make use of Corollary 2.3.2, we must choose a suitable Lipschitz function $\phi(\cdot)$. Let $X_1, \ldots, X_n, X \in \mathbb{R}^d$ be independent and identically distributed random

variables with covariance Σ_0 and with a common strongly log-concave measure μ with coefficient $c > 0$. Let $\lambda_1 \geq \ldots \geq \lambda_n$ be the eigenvalues of Σ and $\Lambda = (\lambda_1, \ldots, \lambda_n)$. For some $p \in [1, \infty]$, let $\|\cdot\|_p$ be the p-Schatten norm defined in Section 1.1.3, which in this case is $\|\Sigma\|_p = \|\Lambda\|_{\ell^p}$. Note that $\|XX^{\mathrm{T}}\|_p = \|X\|_{\ell^2}^2$ for any $p \in [1, \infty]$. Define the function ϕ to be

$$\phi(X_1, \ldots, X_n) = \left\| \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathrm{E}X)(X_i - \mathrm{E}X)^{\mathrm{T}} \right\|_p^{1/2}.$$

For all $p \in \{1, 2, \infty\}$, we have that ϕ is Lipschitz with coefficient $\|\phi\|_{\mathrm{Lip}} = n^{-1/2}$ with respect to the Frobenius / Hilbert-Schmidt metric. That is, let the vectors $X_1, \ldots, X_n, Y_1, \ldots, Y_n \in \mathbb{R}^d$, and denote $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$, then, making use of the $\ell^{p,q}$ metric from Equation 1.1.1,

$$|\phi(\mathbf{X}) - \phi(\mathbf{Y})| \leq n^{-1/2} d_{2,2}(\mathbf{X}, \mathbf{Y}) = \left(\frac{1}{n} \sum_{i=1}^{n} \|X_i - Y_i\|_{\ell^2}^2 \right)^{1/2}.$$

This is established in Proposition 2.C.5 for $p \in \{2, \infty\}$. Proposition 2.C.2 establishes that $\phi(\cdot)$ is also Lipschitz with coefficient $\|\phi\|_{\mathrm{Lip}} = n^{-1/2}$ for $p = 1$. From here, the procedure outlined in Section 2.2 can be implemented with the given ϕ and $r_\alpha = \sqrt{(-2/nc_0) \log \alpha}$.

In many cases, including the two examples above, the constructed confidence set is completely dimension-free. Thus, even mild assumptions on the relationship between the sample size n and the dimension d, such as $\log d = o(n^{1/3})$ from the adaptive soft thresholding estimator of Cai and Liu (2011), are not needed to prove consistency in our setting. Furthermore, the concentration inequalities immediately give us a fast rate of convergence as long as $-\log \alpha = o(n)$ with a proof provided in Appendix 2.A.

Proposition 2.3.5. *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be independent and identically distributed with common measure μ. Let μ be strictly log-concave with some fixed constant c_0 from Definition 2.3.1. Then, for $\alpha \in (0, 1)$, $p \in [1, \infty]$, and $r_\alpha = \sqrt{(-2/nc_0) \log \alpha}$,*

$$\sup_{\hat{\Sigma}^{\mathrm{sp}}: \left\| \hat{\Sigma}^{\mathrm{sp}} - \hat{\Sigma}^{\mathrm{emp}} \right\|_p \leq r_\alpha} \mathrm{P}\left(\left\| \hat{\Sigma}^{\mathrm{sp}} - \Sigma_0 \right\|_p \geq O\left(n^{-1/2} (1 + n^{-1/4} \sqrt{-\log \alpha})^2 \right) \right) \leq \alpha.$$

A second issue in the setting of sparse covariance recovery is that of support recovery or "sparsistency" (Lam and Fan, 2009; Rothman et al., 2009). To recover the support of a covariance matrix–that is, to determine which entries $\sigma_{i,j} \neq 0$–we will require a class of sparse matrices similar to those from Bickel and Levina (2008a,b);

Rothman et al. (2009); Cai and Liu (2011). Specifically, let

$$\mathcal{U}(k,\delta) = \Bigg\{ \Sigma \in \mathbb{R}^{d\times d} : \max_{i=1,\dots,d} \sum_{j=1}^{d} \mathbf{1}\left[\sigma_{i,j} \neq 0\right] \leq k, \\ \text{and if } \sigma_{i,j} \neq 0, \text{ then } |\sigma_{i,j}| \geq \delta > 0 \Big\}.$$

In past work, a notation of "approximate sparsity" is considered where the first condition in $\mathcal{U}(k,\delta)$ is replaced with $\max_{i=1,\dots,d} \sum_{i=1}^{d} |\sigma_{i,j}|^q < k$ for $q \in [0,1)$. However, once we bound the non-zero entries away from zero by some fixed δ, such "approximate sparsity" implies standard sparsity, which is when $q = 0$. It is worth emphasizing that the above Proposition 2.3.5 does not require such a sparsity class, because our estimator is forced to remain close enough to $\hat{\Sigma}^{\text{emp}}$ to follow $\hat{\Sigma}^{\text{emp}}$'s convergence to Σ_0. As with the previous proposition, a proof of the following result is provided in Appendix 2.A.

Proposition 2.3.6. *Let $X_1, \dots, X_n \in \mathbb{R}^d$ be independent and identically distributed with common measure μ. Let μ be strictly log-concave with some fixed constant c_0 from Definition 2.3.1. Furthermore, let $\Sigma_0 \in \mathcal{U}(k,\delta)$. Then, for $\hat{\Sigma}^{\text{sp}}$ denoting the concentration estimator using the zeroing method from Section 2.2.1 with the operator norm,*

$$\lim_{n\to\infty} \mathrm{P}\left(\mathrm{supp}(\hat{\Sigma}^{\text{sp}}) \neq \mathrm{supp}(\Sigma_0)\right) = 0$$

where $\mathrm{supp}(\Sigma) = \{(i,j) : \sigma_{i,j} \neq 0\}$.

2.3.2 Bounded Random Variables

In this section, we consider random variables that are bounded in some norm. Consider a Banach space $(B, \|\cdot\|)$ and a collection of independent and identically distributed random variables $X_1, \dots, X_n \in B$ such that, for some finite and fixed U, $\|X_i\| \leq U$ for all $i = 1, \dots, n$. Given only this assumption, the bounded differences inequality, detailed in Appendix 2.B.2 and in Section 3.3.4 of Giné and Nickl (2016), can be applied in this specific setting. It provides sub-Gaussian concentration for such random variables.

Corollary 2.3.7. *Let $X_1, \dots, X_n \in \mathbb{R}^d$ be iid with $\|X_i\|_{\ell^2} \leq U$ for $i = 1, \dots, n$. Then, for any $p \in [1,\infty]$, $\|X_i X_i^{\mathrm{T}}\|_p \leq U^2$, and*

$$\mathrm{P}\left(\left\|\hat{\Sigma}^{\text{emp}} - \Sigma_0\right\|_p \geq \mathrm{E}\left\|\hat{\Sigma}^{\text{emp}} - \Sigma_0\right\|_p + r\right) \leq \mathrm{e}^{-2nr^2/U^2}.$$

Hence, for any collection of real valued random vectors bounded in Euclidean norm, the bounded differences inequality can be applied to the empirical estimate

for any of the p-Schatten norms. The deviation threshold is $r_\alpha = U\sqrt{(1/2n)\log\alpha}$. However, unlike in the previous setting, the bounds may not necessarily be dimension free.

Example 2.3.8 (Distributions on the Hypercube). *For $X_1, \ldots, X_n \in \mathbb{R}^d$ with each component $|X_{i,j}| \leq 1$ for $i = 1, \ldots, n$ and $j = 1, \ldots, d$ such as for multivariate uniform or Rademacher random variables, then $U = d^{1/2}$. Consequently, the deviation threshold $r_\alpha = O(\sqrt{d/n})$ is not dimension free. Hence, this example is not included in the numerical experiments of Section 2.4 as it fails to give adequate performance when $d \gg n$.*

While this example fails to yield a useful concentration methodology, not all distributions restricted to the hypercube are intractable. Specifically, Example 2.3.4 above considers the Dirichlet distribution, which is restricted to a subset of the hypercube and for which useful concentration properties hold. It is reasonable to believe that other more clever methods can be applied to the multivariate Rademacher random variables or uniform random variables on the unit hypercube.

2.3.3 Sub-Exponential Distributions

Compared with the previously discussed measures that have sub-Gaussian concentration, there exists a larger class of measures that have the weaker sub-exponential concentration. Such measures can be specified as those that satisfy the Poincaré or spectral gap inequality (Bobkov and Ledoux, 1997; Ledoux, 2001; Gozlan, 2010).

Corollary 2.3.9 (Ledoux (2001), Corollary 5.15). *Let X, a random variable on $\mathbb{R}^d$ with measure μ, satisfy the Poincaré inequality*

$$\mathrm{Var}\left(f(X)\right) \leq C \int |\nabla f|^2 d\mu$$

for some $C > 0$ and for all locally Lipschitz functions f. Then, for $X_1, \ldots, X_n \in \mathbb{R}^d$ independent and identically distributed copies of X and for some Lipschitz function $\phi : \mathbb{R}^{d\times n} \to \mathbb{R}$,

$$\mathrm{P}\left(\phi(X_1, \ldots, X_n) \geq \mathrm{E}\phi(X_1, \ldots, X_n) + r\right) \leq \exp\left(-\frac{1}{K}\min\left\{\frac{r}{b}, \frac{r^2}{a^2}\right\}\right)$$

where $K > 0$ is a constant depending only on C and

$$a^2 \geq \sum_{i=1}^{n} |\nabla_i \phi|^2, \qquad b \geq \max_{i=1,\ldots,n} |\nabla_i \phi|.$$

As in Section 2.3.1, ϕ is chosen to be

$$\phi(X_1, \ldots, X_n) = \left\| \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathrm{E}X)(X_i - \mathrm{E}X)^{\mathrm{T}} \right\|_p^{1/2},$$

which is Lipschitz with constant $n^{-1/2}$. This results in values of $a^2 = 1$ and $b = n^{-1/2}$ for the above coefficients. Hence, the deviation threshold in this setting is $r_\alpha = \max\{-K \log \alpha / \sqrt{n}, \sqrt{-K \log \alpha}\}$. While an optimal (or reasonable) value for K may not be known, it makes little difference given the proposed cross-validation procedure as the term $-K \log \alpha$ will be tuned to determine the optimal size of the constructed confidence set.

As the deviation threshold in this setting is bounded below by a constant $\sqrt{-K \log \alpha}$, we do not achieve the nice convergence results as in the log-concave setting. However, the dimension-free concentration still allows for good performance in simulation settings as will be seen in the follow section.

2.4 Numerical Simulations

As mentioned before, our proposed concentration confidence set based method has a similar feel to regularized / penalized estimators as the larger the constructed confidence set is, the sparser the returned estimator will be. Thus, we compare our approach with the two following lasso style estimators. The first method is from the R package `spcov` (Bien and Tibshirani, 2012) and uses a majorize-minimize algorithm to determine

$$\hat{\Sigma}^{\mathrm{MMA}} = \operatorname*{arg\,min}_{\Sigma \succeq 0} \left\{ \operatorname{tr}\left(\hat{\Sigma}^{\mathrm{emp}} \Sigma^{-1} \right) - \log \det(\Sigma^{-1}) + \lambda \|\Sigma\|_{\ell^1} \right\}$$

for some penalization $\lambda > 0$. The second method is from the R package `PDSCE` (Rothman, 2013) and optimizes the similar

$$\hat{\Sigma}^{\mathrm{PDS}} = \operatorname*{arg\,min}_{\Sigma \succeq 0} \left\{ \|\Sigma - \hat{\Sigma}^{\mathrm{emp}}\|_2 - \tau \log \det(\Sigma) + \lambda \|\Sigma\|_{\ell^1} \right\}$$

with $\tau, \lambda > 0$. Here, the log det term is used to enforce positive-definiteness of the final solution, and $\|\cdot\|_{\ell^1}$ is the lasso style penalty. The main concern with implementing such methods is the speed of finding an optimal solution. The majorize-minimize algorithm in its current instantiation in `spcov` requires a significant amount of time to run. The algorithm used to compute the $\hat{\Sigma}^{\mathrm{PDS}}$ estimate is much faster.

We will also compare our method against universal thresholding of the empirical covariance matrix (Bickel and Levina, 2008a; Rothman et al., 2009). Such estimators

are constructed by applying a thresholding function, which satisfies some nice properties, to the empirical covariance matrix. The two types of thresholding considered in our numerical experiments will be *Hard*, which zeros any entries smaller than some $\lambda > 0$, and *Soft*, which shrinks the entries by some $\lambda > 0$. In the below definition of the soft thresholding estimator, the notation $(\cdot)_+$ is such that for $x \in \mathbb{R}$, $(x)_+ = \max\{x, 0\}$.

$$\hat{\Sigma}_\lambda^{\text{Hard}} = \{\hat{\sigma}_{i,j}\mathbf{1}\left[\hat{\sigma}_{i,j} > \lambda\right]\}_{i,j} \qquad \hat{\Sigma}_\lambda^{\text{Soft}} = \{\text{sign}(\hat{\sigma}_{i,j})(|\hat{\sigma}_{i,j}| - \lambda)_+\}_{i,j}$$

where $\hat{\sigma}_{i,j}$ is the (i,j)th entry of the empirical covariance estimate and $\lambda > 0$ is some thresholding parameter which is chosen in practice via cross-validation as we explain for the sake of our own method in Section 2.2.3. Briefly, the data is split in half, two empirical estimators are formed, one is thresholded while the other is not modified, and λ is selected to minimize the distance between the one empirical estimate and the other thresholded estimate.

There are four sparsity patterns that will be considered in the following simulation studies. Table 2.3 takes the unknown covariance Σ_0 to be the $d \times d$ identity matrix, which is as sparse as possible. Table 2.4 chooses Σ_0 to be a tri-diagonal matrix with diagonal entries of 1 and off-diagonal entries of 0.25, which is a moving average. Table 2.5 considers the autoregressive matrix with entries $\sigma_{i,j} = \rho^{|i-j|}$ where $\rho = 0.25$. In this case, there are no zero entries in the true covariance matrix, but it can still be considered approximately sparse as entries further from the diagonal quickly become negligible. Table 2.6 gives Σ_0 a random sparse pattern, which is $\sigma_{i,i} = 1$ for $i = 1, \ldots, d$ and $\sigma_{i,j} = B_{i,j}U_{i,j}$ where $B_{i,j} \sim \text{Bernoulli}\,(0.05)$ and $U_{i,j} \sim \text{Uniform}\,[0.3, 0.8]$ for $i \neq j$.

A variety of estimators are considered in the following simulations. They are as follows: `Emp` is merely the empirical estimate; `Diag` is the diagonal of the empirical estimate with all other entries set to zero, which can be considered to be the most extreme thresholding procedure. `Tri` is a tri-diagonal matrix formed from the empirical estimate with all entries set to zero except for the diagonal and immediate off-diagonal entries; `CZ 2` is our concentration estimator using the zeroing method with the Hilbert-Schmidt / Frobenius distance; `CZ` ∞ is our concentration estimator using the zeroing method with the operator norm distance; `CP 2` is our concentration estimator using the Procrustes method with target matrix the diagonal of the empirical estimate and with Hilbert-Schmidt / Frobenius distance; `CP` ∞ is our concentration estimator using the Procrustes method with target matrix the diagonal of the empirical estimate and with operator norm distance; `CP Id` is our concentration estimator using the Procrustes method with target matrix the $d \times d$ identity matrix and with operator norm distance; `MMA`, `PDS`, `Hard`, and `Soft` are the four alternative

sparse estimators described immediately above.

To test the efficacy of these 12 estimators, their respective average distances to the true covariance will be tabulated. The distances considered are `Op`, the operator norm or maximal eigenvalue, `HS`, the Hilbert-Schmidt, Frobenius, or ℓ^2 distance, and `Supp` being the percentage of correct support recovery, which is

$$\mathrm{Supp}(\hat{\Sigma}, \Sigma_0) = \frac{1}{d^2}|\{(i,j) : \sigma_{i,j} = \hat{\sigma}_{i,j} = 0 \text{ or } \sigma_{i,j} \neq 0 \text{ and } \hat{\sigma}_{i,j} \neq 0\}|.$$

This measure of success is excluded from the autoregressive setting whose support is the entire matrix. This setting is still considered approximately sparse due to the rapid decay in and the, hence, negligible effect of the individual covariances as they move away from the diagonal.

The respective average compute times, `Time`, of constructing each estimator is also considered in the tables. The slowest method tested by far was the `MMA` method, which uses a majorize-minimize algorithm and was seen to be computationally infeasible when $d = 200$. Our concentration methods generally took about one to two minutes to run in the $d = 200$ setting. The vast majority of that time is the cross-validation phase, which effectively multiplies the $O(\log d)$ runtime by a large constant depending on the number of iterations performed. Choosing to use a preselected choice of α perhaps chosen based on past cross-validation computations will drastically speed up this algorithm. It is also very possible that with some additional thought a more efficient implementation the cross-validation procedure is possible.

2.4.1 Multivariate Gaussian Data

Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be independent and identically distributed zero mean random vectors with a strictly log-concave measure and covariance matrix Σ. By Corollary 2.3.2, there exists a constant $c_0 > 0$ such that $\mathrm{P}\left(\phi(\mathbf{X}) \geq \mathrm{E}\phi(\mathbf{X}) + r\right) \leq \mathrm{e}^{-nr^2/2c_0}$ where $\phi(\mathbf{X}) = \|\hat{\Sigma}^{\mathrm{emp}} - \Sigma\|_p^{1/2}$ where $\hat{\Sigma}^{\mathrm{emp}} = n^{-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$ is the empirical estimate of the covariance matrix. This results in the size $1-\alpha$ confidence set for Σ

$$\mathcal{C}_{1-\alpha} = \left\{\Sigma : \|\hat{\Sigma}^{\mathrm{emp}} - \Sigma\|_p^{1/2} \leq \mathrm{E}\|\hat{\Sigma}^{\mathrm{emp}} - \Sigma\|_p^{1/2} + \sqrt{(-2c_0/n)\log\alpha}\right\}$$

for $\alpha \in (0,1)$. In the notation of Section 2.2, $r_\alpha = \sqrt{(-2c_0/n)\log\alpha}$

In the multivariate Gaussian case, c_0 is the maximal eigenvalue of the covariance matrix Σ. We avoid the issue of estimating c_0 in practice. This is because choosing c_0 to be the maximal eigenvalue of the empirical estimate $\hat{\Sigma}^{\mathrm{emp}}$ and then applying cross-validation to choose an optimal value for the regularization parameter α negates

the need for an accurate estimate of c_0.

The top halves of Tables 2.3, 2.4, 2.5, and 2.6 chart the results of testing 11 estimators in four settings on multivariate Gaussian data. In short, the `CZ` ∞ method often performed the best in the higher dimensional settings but was surpassed by the `PDS` method in lower dimensions. For the diagonal and tri-diagonal settings, the concentration approach using the zeroing method with operator norm distance gave the best performance competing with hard/soft thresholding for support recovery and competing with the `PDS` method for dominance in operator norm distance. Once again, the `CZ` ∞ method performs best for the AR(1) matrices in the higher, $d = 100, 200$, dimensional settings. For the random sparse matrices, the `CZ` ∞ method does not uniformly dominate the other methods. However, it still is quite good at support recovery and in the Hilbert-Schmidt distance.

2.4.2 Multivariate Laplace Data

There are many possible ways to extend the univariate Laplace distribution, also referred to as the double exponential distribution, onto $\mathbb{R}^d$. For the following simulation study, we choose the extension detailed in Eltoft et al. (2006). Considering the univariate case, let $Z \sim \mathcal{N}(0, \sigma^2)$ and let $V \sim \text{Exponential}(1)$. Then, $X = \sqrt{V}Z \sim \text{Laplace}\left(\sigma/\sqrt{2}\right)$, which has probability density $f(x) = \sqrt{2}\sigma^{-1}\exp(-\sqrt{2}|x|/\sigma)$ and variance $\text{Var}(X) = \sigma^2$. For the multivariate setting, now let $Z \in \mathbb{R}^d$ be multivariate Gaussian with zero mean and covariance Σ_0 and, once again, let $V \sim \text{Exponential}(1)$. Then, we declare $X = \sqrt{V}Z$ to have a multivariate Laplace distribution with zero mean and covariance Σ_0. Applying Corollary 2.3.9 results in the following concentration based confidence set,

$$\mathcal{C}_{1-\alpha} = \Big\{\Sigma : \|\hat{\Sigma}^{\text{emp}} - \Sigma\|_p^{1/2} \leq \mathrm{E}\|\hat{\Sigma}^{\text{emp}} - \Sigma\|_p^{1/2} + \\ + \max\left\{-K\log\alpha/\sqrt{n}, \sqrt{-K\log\alpha}\right\}\Big\},$$

where the term $-K\log\alpha$ is selected via the cross-validation technique detailed in Section 2.2.3. As a starting point for cross-validation, the initial value of K is chosen to be the maximal eigenvalue of the empirical estimator.

The bottom halves of Tables 2.3, 2.4, 2.5, and 2.6 consider the same simulation experiments as in the previous section, but for multivariate Laplace rather than Gaussian data. In summary, the concentration approach using Procrustes targeting the identity matrix with operator norm distance, `CP Id`, generally gives the best results with respect to operator norm distance for the diagonal, tri-diagonal, and AR(1) matrix settings. Meanwhile, the zeroing method with operator norm distance, `CZ` ∞, dominates all four settings in Hilbert-Schmidt distance and in support recovery.

Albeit, the CZ ∞ method succeeds by often choosing the diagonal of the empirical estimator as its choice, but this is because the small sample size and heavier tails does not provide enough information to choose a better estimator.

2.4.3 Small Round Blue-Cell Tumour Data

Following the same analysis performed in Rothman et al. (2009) and subsequently in Cai and Liu (2011), we will consider the data set resulting from the small round blue-cell tumor (SRBCT) microarray experiment (Khan et al., 2001). The data set consists of a training set of 64 vectors containing 2308 gene expressions. The data contains four types of tumors denoted EWS, BL-NHL, NB, and RMS. As performed in the two previous papers, the genes are ranked by their respective amount of discriminative information according to their F-statistic

$$F = \frac{\frac{1}{k-1}\sum_{m=1}^{k} n_m(\bar{x}_m - \bar{x})^2}{\frac{1}{n-k}\sum_{m=1}^{k}(n_m - 1)\hat{\sigma}_m^2}$$

where $\bar{x}$ is the sample mean, $k = 4$ is the number of classes, $n = 64$ is the sample size, n_m is the sample size of class m, and likewise, $\bar{x}_m$ and $\hat{\sigma}_m^2$ are, respectively, the sample mean and variance of class m. The top 40 and bottom 160 scoring genes were selected to provide a mix of the most and least informative genes.

Table 2.2 reports the support recovery of six methods applied to this data set: the empirical estimate; the zeroing estimate using the operator norm; the Procrustes estimate; the PDS regularized estimate; and the hard and soft thresholded estimates. The concentration techniques were implemented using the equations from the log-concave setting. The sub-exponential methodology was also applied to this data set, but resulted in an extremely sparse estimator similar to the results seen in the hard thresholding case. Both the empirical and Procrustes estimates do not, in general, contain zero entries. Hence, for the sake of this real data test, any covariance entries less than 0.01 were rounded to zero. The support recovery is partitioned into two sections, which are the informative 40×40 block and the remaining uninformative entries. Of the six methods, hard thresholding, as mentioned in Cai and Liu (2011), over thresholds and sets most of the entries to zero. On the converse, the PDS method keeps about 47.3% of the entries in the informative section and 16.5% of the uninformative section. Both the zeroing method and soft thresholding fall in between these extremes by maintaining a respectable amount of entries in the informative section but removing most of the entries outside of that section. The resulting covariance estimators from the six methods are displayed in Figure 2.2.

non-zero (%)	Empirical	Zeroing	Procrustes
Informative	98.1%	25.0%	96.8%
Uninformative	81.8%	2.3%	80.1%
	PDS lasso	Hard Thresh	Soft Thresh
Informative	47.3%	4.5%	29.4%
Uninformative	15.6%	0.2%	3.1%

Table 2.2: The percentages of non-zero off-diagonal entries in the six respective covariance estimates partitioned into two parts: the informative part is the 40×40 block of the highest scoring genes; the uninformative part is the remaining matrix entries.

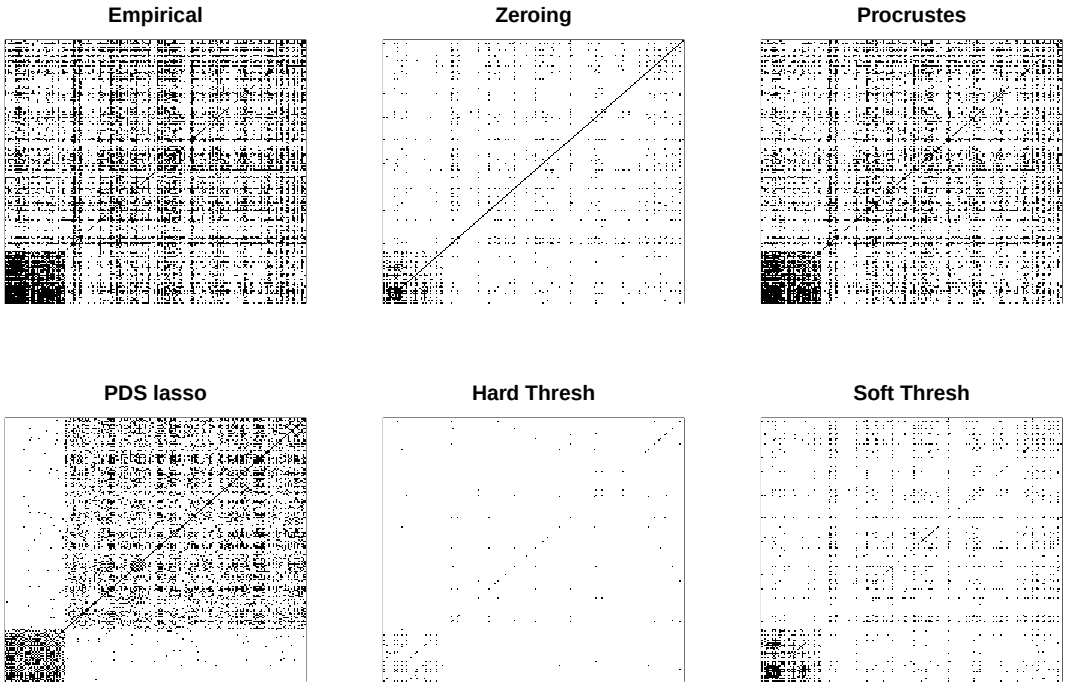


Figure 2.2: Six different methods for covariance estimation applied to the SRBCT data set. White entries are zeros in the covariance estimate while black entries indicate non-zero covariance values.

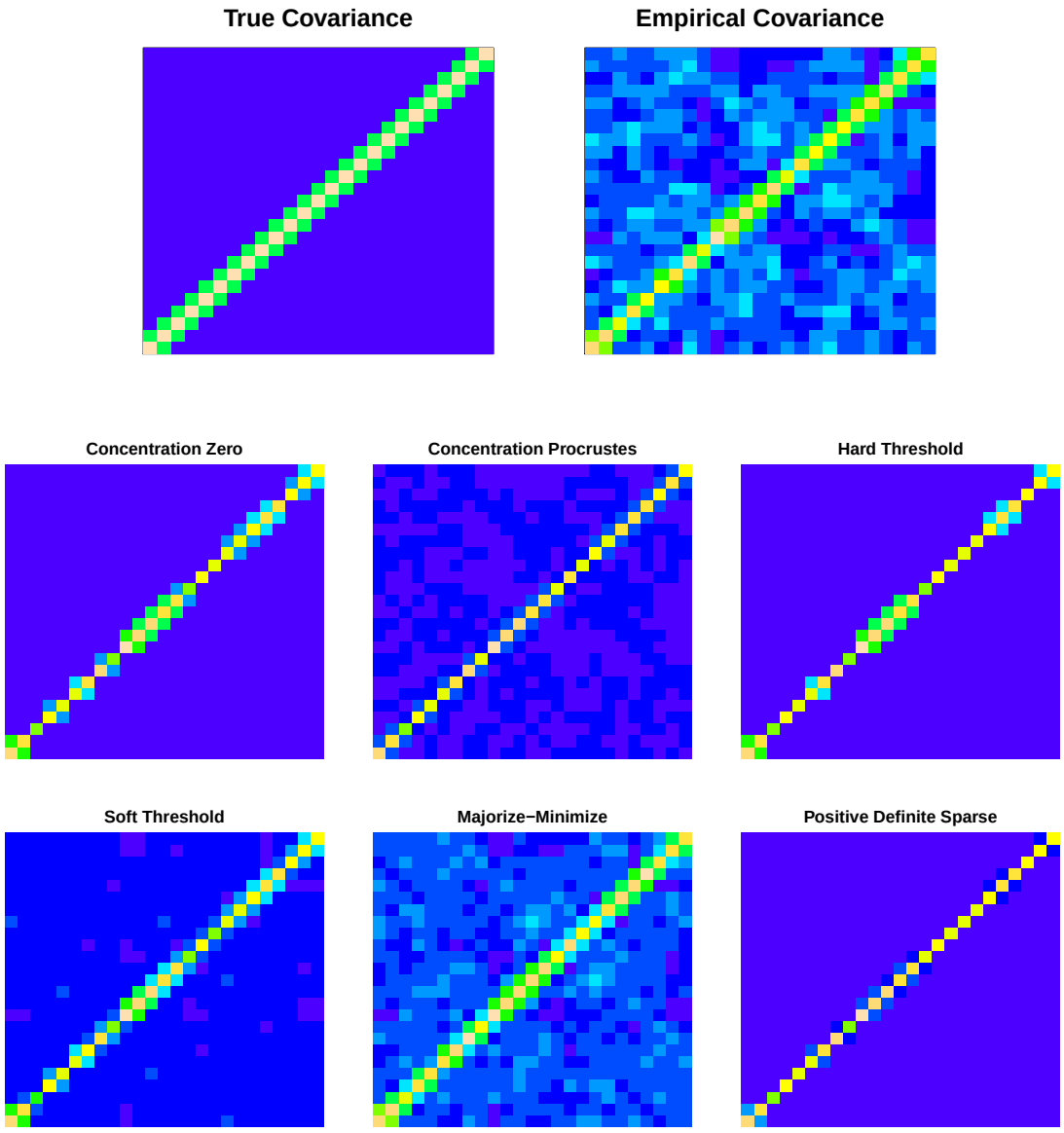


Figure 2.3: The images of seven different covariance estimators with the true covariance in the top left corner. The true covariance is the tri-diagonal matrix with ones on the diagonal and 0.4 on the off-diagonals. These came from a sample of $n = 75$ multivariate Gaussian random vectors with dimension $d = 25$.

Diagonal Matrix

Multivariate Gaussian Data

	$d = 25$				$d = 50$			
	Op	HS	Supp	Time	Op	HS	Supp	Time
Emp	2.37 (0.32)	4.73 (0.31)	4.0%	-	3.95 (0.37)	9.35 (0.38)	2.0%	-
Diag	0.61 (0.17)	1.30 (0.20)	100.%	-	0.70 (0.17)	1.85 (0.21)	100.%	-
Tri	0.80 (0.18)	1.83 (0.21)	92.3%	-	0.91 (0.18)	2.61 (0.22)	96.1%	-
CZ 2	0.93 (0.22)	1.62 (0.23)	99.5%	0.353	2.19 (0.25)	5.32 (0.53)	6.4%	1.27
CZ ∞	0.90 (0.22)	1.56 (0.20)	**99.7%**	0.500	1.04 (0.22)	**2.09** (0.20)	**99.9%**	2.28
CP 2	**0.69** (0.16)	1.58 (0.19)	4.0%	0.288	1.03 (0.07)	4.35 (0.25)	2.0%	1.71
CP ∞	0.84 (0.17)	1.82 (0.21)	4.0%	0.273	1.02 (0.06)	4.06 (0.21)	2.0%	1.34
MMA	1.79 (0.56)	4.04 (0.63)	9.5%	0.988	1.30 (0.09)	5.54 (0.09)	9.7%	2.70
PDS	**0.69** (0.17)	**1.51** (0.19)	89.7%	0.001	**0.86** (0.17)	2.40 (0.20)	89.5%	0.01
Hard	1.01 (0.10)	2.23 (0.45)	99.4%	0.153	1.02 (0.09)	4.09 (0.44)	99.4%	0.42
Soft	0.85 (0.08)	2.46 (0.15)	97.1%	0.245	0.96 (0.05)	3.94 (0.16)	98.9%	0.77
	$d = 100$				$d = 200$			
Emp	6.57 (0.47)	18.60 (0.56)	1.0%	-	11.54 (0.52)	37.22 (0.67)	0.5%	-
Diag	0.78 (0.16)	2.61 (0.19)	100.%	-	0.89 (0.16)	3.72 (0.20)	100.%	-
Tri	1.01 (0.17)	3.71 (0.21)	98.0%	-	1.11 (0.17)	5.24 (0.21)	99.0%	-
CZ 2	4.67 (0.26)	13.29 (0.51)	1.0%	6.13	9.08 (0.30)	28.98 (0.69)	0.5%	32.1
CZ ∞	**0.78** (0.16)	**2.61** (0.19)	**100.%**	9.22	**0.96** (0.33)	**3.83** (0.67)	98.0%	52.2
CP 2	2.05 (0.13)	9.21 (0.39)	1.0%	11.65	5.06 (0.12)	19.99 (0.60)	0.5%	91.2
CP ∞	1.03 (0.08)	6.94 (0.30)	1.0%	8.32	1.04 (0.13)	10.58 (0.59)	0.5%	62.8
MMA	1.41 (0.07)	8.99 (0.07)	4.3%	16.35	-	-	-	-
PDS	1.13 (0.15)	4.03 (0.18)	89.4%	0.02	1.56 (0.12)	7.22 (0.18)	89.3%	0.13
Hard	1.03 (0.11)	6.99 (0.35)	99.5%	1.47	1.03 (0.12)	11.26 (0.30)	99.7%	5.90
Soft	1.00 (0.01)	6.22 (0.14)	99.6%	2.91	1.00 (0.01)	9.77 (0.16)	**99.9%**	11.87

Multivariate Laplace Data

	$d = 25$				$d = 50$			
	Op	HS	Supp	Time	Op	HS	Supp	Time
Emp	4.04 (1.44)	6.48 (1.51)	4.0%	-	7.77 (2.56)	13.04 (2.79)	2.0%	-
Diag	0.94 (0.42)	1.94 (0.58)	100.%	-	1.12 (0.50)	2.83 (0.73)	100.%	-
Tri	1.24 (0.51)	2.64 (0.66)	92.3%	-	1.47 (0.59)	3.80 (0.85)	96.1%	-
CZ ∞	1.43 (0.65)	2.32 (0.65)	**99.7%**	0.533	1.83 (0.73)	**3.23** (0.81)	**99.9%**	2.47
CP ∞	1.24 (0.53)	2.54 (0.79)	4.0%	0.232	1.38 (0.49)	4.71 (0.64)	2.0%	0.97
CP Id	**0.56** (0.06)	**1.61** (0.08)	4.0%	0.233	**1.00** (0.01)	4.01 (0.27)	2.0%	0.97
MMA	3.92 (1.45)	6.34 (1.51)	4.4%	0.205	1.89 (0.43)	6.17 (0.41)	7.1%	3.03
PDS	1.43 (0.69)	2.65 (0.75)	77.2%	0.002	2.40 (1.14)	4.66 (1.24)	77.1%	0.01
Hard	1.53 (0.95)	4.02 (0.82)	97.8%	0.168	2.13 (1.63)	6.41 (1.20)	98.6%	0.44
Soft	1.05 (0.26)	3.69 (0.42)	97.7%	0.261	1.14 (0.51)	5.72 (0.50)	98.7%	0.81
	$d = 100$				$d = 200$			
Emp	13.59 (3.52)	25.02 (4.76)	1.0%	-	28.56 (9.46)	51.23 (11.1)	0.5%	-
Diag	1.22 (0.44)	3.89 (0.73)	100.%	-	1.45 (5.69)	5.74 (1.39)	100.%	-
Tri	1.52 (0.53)	5.24 (0.92)	98.0%	-	1.84 (7.18)	7.68 (1.72)	99.0%	-
CZ ∞	1.22 (0.44)	**3.89** (0.73)	**100.%**	9.65	1.45 (5.69)	**5.74** (1.39)	**100.%**	53.0
CP ∞	1.30 (0.36)	7.44 (0.44)	1.0%	5.50	1.43 (4.66)	11.27 (0.77)	0.5%	37.9
CP Id	**1.00** (0.00)	6.99 (0.43)	1.0%	5.53	**1.00** (0.00)	10.67 (0.28)	0.5%	37.9
MMA	2.00 (0.35)	9.28 (0.35)	3.5%	22.97	-	-	-	-
PDS	3.62 (1.32)	8.03 (1.77)	76.9%	0.05	7.60 (3.90)	15.83 (4.48)	76.6%	3.7
Hard	2.49 (1.71)	9.36 (1.30)	99.3%	1.57	5.26 (5.46)	15.59 (5.52)	99.5%	6.4
Soft	1.10 (0.34)	8.70 (0.52)	99.3%	3.05	1.74 (2.24)	12.88 (1.35)	99.5%	12.5

Table 2.3: Listed are the distances from a variety of estimators to the true $\Sigma_0 = I_d$ with sample size $n = 30$, dimensions $d = 25, 50, 100, 200$, and considering both multivariate Gaussian and Laplace data.

Tri-Diagonal Matrix

Multivariate Gaussian Data

	$d = 25$				$d = 50$			
	Op	HS	Supp	Time	Op	HS	Supp	Time
Emp	2.43 (0.36)	4.74 (0.34)	11.7%	-	4.05 (0.41)	9.36 (0.44)	5.9%	-
Diag	0.83 (0.13)	2.17 (0.12)	92.3%	-	0.89 (0.13)	3.09 (0.12)	96.1%	-
Tri	0.86 (0.21)	1.86 (0.24)	100.%	-	0.96 (0.22)	2.64 (0.25)	100.%	-
CZ 2	0.99 (0.20)	2.31 (0.16)	**92.4%**	0.346	2.12 (0.29)	5.31 (0.51)	12.7%	1.26
CZ ∞	0.97 (0.20)	2.28 (0.14)	**92.4%**	0.492	1.11 (0.23)	**3.21** (0.15)	**96.1%**	2.25
CP 2	**0.84** (0.11)	2.15 (0.13)	11.7%	0.279	1.31 (0.04)	4.65 (0.19)	5.9%	1.67
CP ∞	1.14 (0.16)	2.85 (0.17)	11.7%	0.276	1.30 (0.05)	4.79 (0.17)	5.9%	1.33
MMA	1.92 (0.61)	4.14 (0.66)	15.7%	0.953	1.34 (0.03)	5.56 (0.09)	13.2%	2.62
PDS	**0.84** (0.13)	**2.13** (0.14)	86.0%	0.001	**0.98** (0.15)	3.23 (0.15)	87.6%	0.01
Hard	1.24 (0.13)	2.88 (0.35)	91.7%	0.151	1.39 (0.10)	4.86 (0.39)	95.5%	0.42
Soft	1.10 (0.07)	2.90 (0.15)	90.8%	0.241	1.22 (0.05)	4.56 (0.15)	95.3%	0.76
	$d = 100$				$d = 200$			
Emp	6.86 (0.55)	18.71 (0.57)	3.0%	-	11.91 (0.59)	37.25 (0.68)	1.5%	-
Diag	0.98 (0.14)	4.38 (0.12)	98.0%	-	1.03 (0.15)	6.21 (0.11)	99.0%	-
Tri	1.10 (0.20)	3.76 (0.25)	100.%	-	1.22 (0.23)	5.31 (0.22)	100.%	-
CZ 2	4.70 (0.28)	13.06 (0.54)	3.0%	6.19	9.14 (0.34)	28.59 (0.76)	1.5%	31.4
CZ ∞	**1.04** (0.25)	**4.44** (0.26)	96.3%	9.25	**1.27** (0.54)	**6.58** (0.87)	88.8%	51.0
CP 2	1.92 (0.14)	9.20 (0.34)	3.0%	11.57	4.91 (0.14)	19.62 (0.67)	1.5%	88.5
CP ∞	1.39 (0.02)	7.78 (0.31)	3.0%	8.35	1.43 (0.03)	11.74 (0.52)	1.5%	61.2
MMA	1.43 (0.02)	9.15 (0.06)	6.2%	17.21	-	-	-	-
PDS	1.21 (0.16)	5.07 (0.18)	88.4%	0.02	1.61 (0.15)	8.40 (0.17)	88.8%	0.2
Hard	1.45 (0.07)	7.89 (0.33)	97.6%	1.50	1.49 (0.11)	12.27 (0.27)	98.7%	5.8
Soft	1.31 (0.03)	7.06 (0.13)	**97.7%**	2.94	1.38 (0.02)	10.88 (0.13)	**98.9%**	11.5

Multivariate Laplace Data

	$d = 25$				$d = 50$			
	Op	HS	Supp	Time	Op	HS	Supp	Time
Emp	4.14 (1.44)	6.52 (1.41)	11.7%	-	7.63 (2.46)	12.84 (2.75)	5.9%	-
Diag	1.11 (0.32)	2.62 (0.38)	92.3%	-	1.28 (0.44)	3.76 (0.60)	96.1%	-
Tri	1.32 (0.56)	2.67 (0.63)	100.%	-	1.54 (0.67)	3.83 (0.92)	100.%	-
CZ ∞	1.53 (0.58)	2.88 (0.47)	**92.3%**	0.532	1.87 (0.74)	**4.05** (0.69)	**96.0%**	2.72
CP ∞	1.49 (0.46)	3.37 (0.60)	11.7%	0.233	1.59 (0.42)	5.33 (0.59)	5.9%	1.07
CP Id	**0.92** (0.04)	**2.68** (0.08)	11.7%	0.234	**1.29** (0.04)	4.72 (0.19)	5.9%	1.07
MMA	4.09 (1.47)	6.47 (1.45)	11.8%	0.111	1.81 (0.38)	6.20 (0.37)	10.6%	3.36
PDS	1.53 (0.66)	3.03 (0.62)	74.9%	0.001	2.36 (1.10)	5.03 (1.11)	76.0%	0.01
Hard	1.85 (0.89)	4.40 (0.76)	90.3%	0.169	2.38 (1.49)	6.91 (1.11)	94.8%	0.50
Soft	1.38 (0.29)	4.02 (0.41)	90.6%	0.263	1.52 (0.41)	6.24 (0.45)	95.0%	0.87
	$d = 100$				$d = 200$			
Emp	14.50 (4.44)	25.21 (5.03)	3.0%	-	30.34 (10.61)	52.5 (10.75)	1.5%	-
Diag	1.40 (0.49)	5.32 (0.61)	98.0%	-	1.68 (0.63)	7.7 (1.00)	99.0%	-
Tri	1.74 (0.70)	5.43 (1.00)	100.%	-	2.19 (0.93)	8.0 (1.69)	100.%	-
CZ ∞	1.40 (0.49)	**5.32** (0.61)	**98.0%**	9.56	1.68 (0.63)	**7.7** (1.00)	**99.0%**	52.0
CP ∞	1.59 (0.34)	8.20 (0.47)	3.0%	5.46	1.69 (0.40)	12.4 (0.66)	1.5%	37.2
CP Id	**1.39** (0.03)	7.79 (0.37)	3.0%	5.42	**1.43** (0.01)	11.8 (0.26)	1.5%	36.9
MMA	1.99 (0.39)	9.49 (0.31)	5.4%	22.60	-	-	-	-
PDS	4.04 (1.82)	8.79 (1.81)	76.1%	0.04	8.33 (4.88)	17.0 (4.73)	75.8%	3.6
Hard	3.19 (2.43)	10.42 (1.78)	97.2%	1.58	6.35 (6.97)	17.0 (6.38)	98.5%	6.1
Soft	1.66 (0.66)	9.41 (0.48)	97.3%	3.02	2.42 (3.16)	13.9 (2.10)	98.5%	12.0

Table 2.4: Listed are the distances from a variety of estimators to the tri-diagonal Σ_0 with diagonal entries of 1 and off-diagonal entries of 0.25 with sample size $n = 30$, dimensions $d = 25, 50, 100, 200$, and considering both multivariate Gaussian and Laplace data.

AR(1) Matrix

Multivariate Gaussian Data

	$d = 25$			$d = 50$		
	Op	HS	Time	Op	HS	Time
Emp	2.44 (0.38)	4.73 (0.35)	-	4.08 (0.43)	9.37 (0.44)	-
Diag	0.87 (0.11)	2.22 (0.13)	-	0.92 (0.12)	3.16 (0.12)	-
Tri	0.88 (0.22)	1.91 (0.24)	-	0.99 (0.21)	2.73 (0.23)	-
CZ 2	1.03 (0.21)	2.36 (0.17)	0.348	2.08 (0.29)	5.27 (0.52)	1.25
CZ ∞	1.01 (0.21)	2.33 (0.16)	0.497	1.11 (0.22)	**3.26** (0.14)	2.25
CP 2	**0.88** (0.09)	**2.18** (0.13)	0.279	1.37 (0.05)	4.70 (0.20)	1.67
CP ∞	1.11 (0.16)	2.89 (0.18)	0.280	1.37 (0.05)	4.86 (0.17)	1.35
MMA	1.93 (0.59)	4.15 (0.66)	0.886	1.42 (0.03)	5.58 (0.09)	2.63
PDS	**0.88** (0.12)	**2.18** (0.14)	0.001	**1.00** (0.13)	3.28 (0.15)	0.01
Hard	1.29 (0.14)	2.90 (0.35)	0.151	1.46 (0.08)	4.87 (0.37)	0.41
Soft	1.20 (0.06)	2.93 (0.15)	0.240	1.32 (0.05)	4.59 (0.15)	0.76
	$d = 100$			$d = 200$		
Emp	6.87 (0.46)	18.66 (0.52)	-	11.89 (0.62)	37.3 (0.77)	-
Diag	0.96 (0.11)	4.49 (0.13)	-	1.03 (0.12)	6.4 (0.11)	-
Tri	1.08 (0.21)	3.88 (0.24)	-	1.20 (0.19)	5.5 (0.23)	-
CZ 2	4.67 (0.27)	13.00 (0.52)	6.17	9.13 (0.36)	28.6 (0.73)	32.5
CZ ∞	**1.02** (0.21)	**4.55** (0.28)	9.34	**1.36** (0.60)	**6.8** (0.98)	53.9
CP 2	1.88 (0.15)	9.14 (0.34)	11.72	4.90 (0.14)	19.6 (0.62)	96.2
CP ∞	1.49 (0.04)	7.74 (0.30)	8.57	1.55 (0.02)	11.8 (0.51)	66.1
MMA	1.55 (0.01)	9.18 (0.06)	17.44	-	-	-
PDS	1.21 (0.14)	5.16 (0.17)	0.02	1.62 (0.14)	8.5 (0.18)	0.2
Hard	1.57 (0.05)	8.01 (0.32)	1.48	1.62 (0.05)	12.4 (0.25)	6.1
Soft	1.42 (0.03)	7.14 (0.12)	2.90	1.50 (0.03)	11.0 (0.13)	12.0

Multivariate Laplace Data

	$d = 25$			$d = 50$		
	Op	HS	Time	Op	HS	Time
Emp	4.16 (1.48)	6.57 (1.52)	-	7.70 (2.34)	12.87 (2.56)	-
Diag	1.18 (0.36)	2.70 (0.44)	-	1.30 (0.39)	3.80 (0.48)	-
Tri	1.35 (0.52)	2.74 (0.67)	-	1.56 (0.59)	3.88 (0.78)	-
CZ ∞	1.59 (0.61)	2.96 (0.54)	0.570	1.88 (0.69)	**4.08** (0.57)	2.49
CP ∞	1.52 (0.48)	3.46 (0.68)	0.250	1.61 (0.36)	5.36 (0.51)	0.99
CP Id	**0.95** (0.09)	**2.72** (0.09)	0.250	**1.35** (0.06)	4.75 (0.21)	0.98
MMA	4.07 (1.52)	6.47 (1.56)	0.168	1.84 (0.34)	6.22 (0.32)	3.09
PDS	1.56 (0.66)	3.09 (0.67)	0.002	2.37 (1.03)	5.07 (0.99)	0.01
Hard	1.92 (0.95)	4.46 (0.87)	0.182	2.38 (1.33)	6.88 (1.00)	0.45
Soft	1.50 (0.29)	4.06 (0.43)	0.282	1.61 (0.41)	6.26 (0.49)	0.81
	$d = 100$			$d = 200$		
Emp	14.52 (4.76)	25.50 (5.60)	-	28.02 (7.60)	51.8 (9.24)	-
Diag	1.40 (0.43)	5.46 (0.82)	-	1.60 (0.47)	7.7 (0.89)	-
Tri	1.75 (0.64)	5.60 (1.24)	-	2.03 (0.68)	7.9 (1.46)	-
CZ ∞	**1.40** (0.43)	**5.46** (0.82)	9.57	1.60 (0.47)	**7.7** (0.89)	55.2
CP ∞	1.66 (0.33)	8.30 (0.55)	5.49	1.70 (0.24)	12.5 (0.69)	39.9
CP Id	1.50 (0.04)	7.80 (0.36)	5.50	**1.56** (0.02)	11.9 (0.29)	40.2
MMA	2.00 (0.37)	9.54 (0.36)	22.80	-	-	-
PDS	4.04 (1.98)	8.90 (2.16)	0.06	7.22 (2.96)	16.3 (3.47)	3.9
Hard	3.31 (2.62)	10.56 (2.39)	1.59	4.87 (3.53)	15.8 (3.30)	6.5
Soft	1.78 (0.78)	9.37 (0.66)	3.03	1.90 (0.75)	13.6 (0.61)	12.6

Table 2.5: Listed are the distances from a variety of estimators to the true Σ_0 whose entries are $\sigma_{i,j} = (0.25)^{|i-j|}$ with sample size $n = 30$, dimensions $d = 25, 50, 100, 200$, and considering both multivariate Gaussian and Laplace data. Unlike the other tables, support recovery is not considered in this case.

Random Sparse Matrix

Multivariate Gaussian Data

	$d = 25$				$d = 50$			
	Op	HS	Supp	Time	Op	HS	Supp	Time
Emp	5.12 (0.87)	9.62 (0.81)	15.2%	-	12.05 (1.41)	27.04 (1.37)	13.7%	-
Diag	2.46 (0.19)	5.73 (0.23)	88.8%	-	4.36 (0.17)	11.61 (0.28)	88.3%	-
Tri	2.55 (0.26)	6.16 (0.30)	82.4%	-	4.31 (0.23)	12.50 (0.41)	85.4%	-
CZ 2	2.56 (0.31)	5.86 (0.32)	89.1%	0.397	5.98 (0.84)	15.85 (1.29)	22.9%	1.34
CZ ∞	2.55 (0.31)	5.83 (0.30)	89.0%	0.566	4.41 (0.29)	11.85 (0.36)	**88.3%**	2.40
CP 2	2.29 (0.24)	5.15 (0.27)	15.2%	0.314	4.81 (0.32)	13.97 (0.48)	13.7%	1.77
CP ∞	3.26 (0.33)	7.50 (0.38)	15.2%	0.319	5.36 (0.20)	15.43 (0.47)	13.7%	1.46
MMA	5.12 (0.87)	9.62 (0.81)	15.2%	0.013	5.21 (0.30)	15.56 (0.22)	22.6%	3.27
PDS	**2.21** (0.26)	**5.08** (0.32)	85.3%	0.002	**3.98** (0.28)	**11.32** (0.38)	82.8%	0.01
Hard	2.91 (0.57)	6.35 (0.65)	**89.4%**	0.177	6.76 (1.29)	17.67 (1.57)	86.7%	0.44
Soft	3.01 (0.23)	6.53 (0.36)	88.4%	0.275	5.22 (0.25)	13.78 (0.48)	86.6%	0.80
	$d = 100$				$d = 200$			
Emp	24.7 (2.15)	64.6 (2.03)	10.0%	-	60.7 (3.73)	185.6 (4.03)	10.1%	-
Diag	5.9 (0.14)	19.9 (0.33)	91.0%	-	11.7 (0.11)	41.1 (0.55)	90.4%	-
Tri	6.0 (0.17)	21.6 (0.43)	89.4%	-	11.7 (0.16)	44.9 (0.73)	89.6%	-
CZ 2	16.0 (1.07)	44.5 (2.35)	10.0%	6.35	45.7 (2.09)	141.4 (3.92)	10.1%	31.3
CZ ∞	6.1 (0.60)	**20.4** (1.36)	**85.7%**	9.64	12.2 (1.51)	**44.0** (5.67)	74.8%	51.6
CP 2	7.2 (0.34)	31.9 (1.27)	10.0%	12.35	23.6 (0.86)	96.6 (3.49)	10.1%	90.7
CP ∞	7.6 (0.27)	29.5 (1.11)	10.0%	9.05	14.5 (0.34)	64.2 (2.54)	10.1%	63.1
MMA	7.8 (0.31)	31.9 (0.16)	12.4%	25.89	-	-	-	-
PDS	**5.6** (0.29)	21.1 (0.45)	83.6%	0.02	**11.2** (0.32)	49.4 (0.90)	**82.4%**	0.2
Hard	14.7 (1.97)	45.3 (2.62)	85.1%	1.53	46.9 (3.65)	158.4 (4.87)	72.9%	5.7
Soft	6.8 (0.27)	24.4 (0.50)	85.1%	3.01	18.0 (2.17)	72.0 (2.54)	72.9%	11.7

Multivariate Laplace Data

	$d = 25$				$d = 50$			
	Op	HS	Supp	Time	Op	HS	Supp	Time
Emp	5.94 (2.34)	9.15 (2.25)	12.6%	-	21.34 (7.09)	35.5 (8.07)	11.2%	-
Diag	2.31 (0.56)	5.08 (0.59)	91.4%	-	4.29 (0.93)	12.1 (1.57)	90.8%	-
Tri	2.55 (0.67)	5.62 (0.77)	84.3%	-	4.88 (1.22)	13.9 (2.03)	87.5%	-
CZ ∞	2.61 (0.93)	5.25 (0.79)	**91.5%**	0.503	5.48 (1.79)	**12.8** (1.80)	**90.8%**	2.36
CP ∞	2.88 (0.77)	6.49 (0.97)	12.6%	0.219	**5.06** (0.99)	15.7 (1.66)	11.2%	0.92
CP Id	2.73 (0.13)	5.76 (0.16)	12.6%	0.218	5.50 (0.11)	17.6 (0.41)	11.2%	0.91
MMA	5.94 (2.34)	9.15 (2.25)	12.6%	0.013	4.79 (0.37)	16.0 (0.43)	16.8%	4.12
PDS	**2.48** (1.01)	**4.81** (0.97)	76.9%	0.001	6.68 (2.83)	14.8 (2.98)	73.8%	0.01
Hard	3.68 (1.88)	6.90 (1.71)	90.2%	0.155	15.48 (7.52)	27.7 (9.11)	82.6%	0.42
Soft	2.83 (0.76)	6.12 (0.72)	89.8%	0.243	7.19 (3.82)	16.9 (3.17)	82.2%	0.77
	$d = 100$				$d = 200$			
Emp	55.27 (17.7)	95.7 (20.3)	10.8%	-	149.5 (40.2)	264.3 (44.9)	10.0%	-
Diag	7.06 (0.8)	23.8 (2.3)	90.2%	-	11.8 (1.3)	46.2 (3.6)	90.5%	-
Tri	7.66 (1.4)	27.2 (3.8)	88.4%	-	12.6 (1.9)	53.2 (5.1)	89.6%	-
CZ ∞	**7.06** (0.8)	**23.8** (2.3)	**90.2%**	9.16	**11.8** (1.3)	**46.2** (3.6)	**90.5%**	55.2
CP ∞	8.40 (0.4)	33.2 (1.8)	10.8%	5.21	14.2 (0.8)	67.7 (3.3)	10.0%	39.3
CP Id	9.58 (0.1)	37.4 (0.9)	10.8%	5.16	15.7 (0.2)	73.8 (0.9)	10.0%	39.5
MMA	8.69 (0.4)	35.5 (0.4)	12.8%	38.27	-	-	-	-
PDS	15.54 (7.8)	35.3 (7.8)	72.1%	0.08	39.3 (16.3)	87.1 (16.5)	71.8%	3.9
Hard	45.71 (18.9)	82.9 (23.0)	73.4%	1.49	136.1 (42.1)	247.0 (49.2)	61.8%	6.5
Soft	21.83 (13.3)	44.4 (13.2)	73.4%	2.93	76.5 (32.1)	138.7 (36.9)	61.8%	12.6

Table 2.6: Listed are the distances from a variety of estimators to the true Σ_0 having a diagonal of 1 and off-diagonal entries $\sigma_{i,j} = B_{i,j}U_{i,j}$ such that $B_{i,j} \sim \text{Bernoulli}\,(0.05)$ and $U_{i,j} \sim \text{Uniform}\,[0.3, 0.8]$ with sample size $n = 30$, dimensions $d = 25, 50, 100, 200$, and considering both multivariate Gaussian and Laplace data.

2.5 Summary and Extensions

In this chapter, we were concerned with the recovery of an unknown sparse covariance matrix through the search of non-asymptotic confidence sets constructed via concentration inequalities for a best estimator. This approach was shown to give desirable convergence results in the case of log-concave measures as well as numerical performance similar and often superior to past approaches both in the log-concave and sub-exponential settings. While our focus was on sparsity, the generic procedure of searching the confidence set with some criterion in mind can be applied in much more generality. Given some assumed form or property of the unknown true covariance matrix, a similar search procedure may be formulated to recover such matrices.

Our method is similar to the penalization estimators such as lasso in the sense that the parameter α can be tweaked to determine the amount of sparsity to allow. However, unlike some of the complicated optimizations surrounding lasso penalization, our estimator is fast to compute even in high dimensions where such complicated optimization steps required by lasso penalties can be computationally intractable. This was the case with the majorize-minimization algorithm, which became intractably slow when $d = 200$. The majority of the compute time of our method is delegated to choosing an optimal α via cross-validation. A more clever implementation of this procedure, or merely reusing past values of α, will drastically decrease the computation time.

As was mentioned in the introduction, the empirical covariance is known in cases of sparsity and high dimensionality to be a poor estimator. We nevertheless chose it as our starting point for the search procedure as its form, being a sum of independent and identically distributed random variables, fits nicely into the concentration paradigm. However, this does not imply that other starting points should not be considered. Constructing concentration inequality based confidence sets about more complicated estimators may not be straightforward. But, if some such set can be constructed, then applying our search technique will most likely result in even better estimates of the true covariance matrix.

Even with starting at the empirical covariance matrix, the zeroing method of Section 2.2.1 was shown in the numerical simulations to have superior performance in support recovery when compared with other methods. It is feasible to believe that a hybrid method is possible where the zeroing method is used to recover the support of Σ_0 followed by some other technique for improving upon the non-zero entries.

2.A Proofs

Proof of Proposition 2.3.5. From the derivation in Section 2.2, we have that

$$\mathrm{P}\left(\left\|\hat{\Sigma}^{\mathrm{sp}} - \Sigma_0\right\|_p^{1/2} \geq \mathrm{E}\left\|\hat{\Sigma}^{\mathrm{emp}} - \Sigma_0\right\|_p^{1/2} + 2r_\alpha\right) \leq \alpha$$

for any $\hat{\Sigma}^{\mathrm{sp}}$ such that $\|\hat{\Sigma}^{\mathrm{sp}} - \hat{\Sigma}^{\mathrm{emp}}\| \leq r_\alpha$. Writing $Z = \left\|\hat{\Sigma}^{\mathrm{emp}} - \Sigma_0\right\|_p$ and squaring and rearranging the terms gives,

$$\begin{aligned}
&\mathrm{P}(Z \geq \mathrm{E}Z + 4r_\alpha(\mathrm{E}Z)^{1/2} + 4r_\alpha^2) \\
&\qquad = \mathrm{P}\left(Z \geq \mathrm{E}Z\left(1 + 4r_\alpha(\mathrm{E}Z)^{-1/2} + 4r_\alpha^2(\mathrm{E}Z)^{-1}\right)\right) \\
&\qquad = \mathrm{P}\left(Z \geq \mathrm{E}Z\left(1 + 2r_\alpha(\mathrm{E}Z)^{-1/2}\right)^2\right) \leq \alpha
\end{aligned}$$

Given the standard convergence result for the empirical covariance matrix that $\mathrm{E}\|\hat{\Sigma}^{\mathrm{emp}} - \Sigma_0\|_p = O(n^{-1/2})$ and our definition of $r_\alpha = O(n^{-1/2}\sqrt{-\log\alpha})$, we now have that

$$\mathrm{P}\left(\left\|\hat{\Sigma}^{\mathrm{sp}} - \Sigma_0\right\|_p \geq O\left(n^{-1/2}(1 + n^{-1/4}\sqrt{-\log\alpha})^2\right)\right) \leq \alpha,$$

which holds for any $\hat{\Sigma}^{\mathrm{sp}}$ such that $\|\hat{\Sigma}^{\mathrm{sp}} - \hat{\Sigma}^{\mathrm{emp}}\| \leq r_\alpha$. □

Proof of Proposition 2.3.6. Let $\hat{\Sigma}^\star = \{\hat{\sigma}_{i,j}\mathbf{1}\left[\sigma_{i,j} \neq 0\right]\}$ be the result of a perfect zeroing of the empirical covariance estimate. That is, $\hat{\Sigma}^\star$ has support identical to the true Σ_0 and non-zero entries that coincide with $\hat{\Sigma}^{\mathrm{emp}}$. Furthermore, let $\tilde{\Sigma}$ be some other overly-sparse covariance estimator resulting from zeroing entries in $\hat{\Sigma}^{\mathrm{emp}}$, but with more zeros than Σ_0.

$$\begin{aligned}
&\mathrm{P}\left(\mathrm{supp}(\hat{\Sigma}^{\mathrm{sp}}) \neq \mathrm{supp}(\Sigma_0)\right) = \\
&\qquad = \mathrm{P}\left(\|\hat{\Sigma}^\star - \hat{\Sigma}^{\mathrm{emp}}\|_\infty^{1/2} \geq r_\alpha \text{ or } \|\tilde{\Sigma} - \hat{\Sigma}^{\mathrm{emp}}\|_\infty^{1/2} \leq r_\alpha\right) \\
&\qquad = \mathrm{P}\left(\|\hat{\Sigma}^\star - \hat{\Sigma}^{\mathrm{emp}}\|_\infty^{1/2} \geq r_\alpha\right) + \mathrm{P}\left(\|\tilde{\Sigma} - \hat{\Sigma}^{\mathrm{emp}}\|_\infty^{1/2} \leq r_\alpha\right), \qquad (2.A.1)
\end{aligned}$$

which, assuming a large enough sample size n, are the two mutually exclusive events that the estimator with correct support $\hat{\Sigma}^\star$ is not in the ball of radius r_α and that a sparser estimator $\tilde{\Sigma}$ is in the ball.

For the first term of Equation 2.A.1, we have

$$\begin{aligned}
\mathrm{P}\left(\left\|\hat{\Sigma}^\star - \hat{\Sigma}^{\mathrm{emp}}\right\|_\infty^{1/2} \geq r_\alpha\right) &\leq \mathrm{P}\left(\|\hat{\Sigma}^\star - \Sigma_0\|_\infty^{1/2} + \|\hat{\Sigma}^{\mathrm{emp}} - \Sigma_0\|_\infty^{1/2} \geq r_\alpha\right) \\
&\leq \mathrm{P}\left(\|\hat{\Sigma}^\star - \Sigma_0\|_\infty^{1/2} \geq r_\alpha/2\right) + \mathrm{P}\left(\|\hat{\Sigma}^{\mathrm{emp}} - \Sigma_0\|_\infty^{1/2} \geq r_\alpha/2\right) = \mathrm{(I)} + \mathrm{(II)}
\end{aligned}$$

For (II), we have that $\mathrm{E}\|\hat{\Sigma}^{\text{emp}} - \Sigma_0\| = O(n^{-1/2})$. Let $Z = \|\hat{\Sigma}^{\text{emp}} - \Sigma_0\|_\infty^{1/2}$ for simplicity of notation. Then, using the concentration result already established for Lipschitz functions of log-concave measures,

$$\begin{aligned}
\text{(II)} &= \mathrm{P}\left(Z \geq r_\alpha/2\right) \\
&= \mathrm{P}\left(Z \geq \mathrm{E}Z + (r_\alpha/2 - \mathrm{E}Z)\right) \\
&\leq \exp\left(-n(r_\alpha/2 - \mathrm{E}Z)^2/2c_0\right) \\
&\left[\exp\left(-n(-r_\alpha \mathrm{E}Z + (\mathrm{E}Z)^2)/2c_0\right)\right]\left[\exp\left(-nr_\alpha^2/2c_0\right)\right]^{1/4} \leq C\alpha^{1/4}
\end{aligned}$$

for some positive $C = o(1)$ as $\mathrm{E}Z = O(n^{-1/4})$ and $r_\alpha = O(n^{-1/2})$ making the expression in the first exponent $-n(-r_\alpha \mathrm{E}Z + (\mathrm{E}Z)^2)/2c_0 = O(n^{1/4} - n^{1/2})$.

For (I), applying the Gershgorin circle theorem (Iserles, 2009) to the operator norm gives

$$\begin{aligned}
\text{(I)} &\leq \mathrm{P}\left(\left(\max_{i=1,\ldots,d} \textstyle\sum_{j=1}^d |\hat{\sigma}_{i,j} - \sigma_{i,j}| \mathbf{1}\left[\sigma_{i,j} \neq 0\right]\right)^{1/2} \geq r_\alpha/2\right) \\
&\leq \mathrm{P}\left(\max_{i,j=1,\ldots,d} |\hat{\sigma}_{i,j} - \sigma_{i,j}|^{1/2} |\mathrm{supp}_{\text{col}}(\Sigma_0)|^{1/2} \geq r_\alpha/2\right)
\end{aligned}$$

where $\mathrm{supp}_{\text{col}}(\Sigma) = \max_{j=1,\ldots,d} |\{(i,j) : \sigma_{i,j} \neq 0\}|$ is the maximal number of non-zero entries in any given column. From Proposition 2.C.5, we have that $\|\hat{\Sigma}^{\text{emp}} - \Sigma_0\|_2^{1/2}$ is Lipschitz with constant $n^{1/2}$. As the squared Frobenius norm is equal to the sum of the squares of the entries of the matrix, we in turn have that the entries $|\hat{\sigma}_{i,j} - \sigma_{i,j}|^{1/2}$ are also Lipschitz with constant $n^{1/2}$. As the maximum of d^2 Lipschitz functions is still Lipschitz, we get similarly to case (II) that $\text{(I)} \leq C\alpha^\varepsilon$ for some $\varepsilon > 0$.

For the second term of Equation 2.A.1, let $\mathrm{supp}(\tilde{\Sigma}) \subset \mathrm{supp}(\Sigma_0)$. Then, there exists a pair of indices $(i_0, j_0) \in \mathrm{supp}(\Sigma_0)$ such that $(i_0, j_0) \notin \mathrm{supp}(\tilde{\Sigma})$.

$$\begin{aligned}
\mathrm{P}\left(\|\tilde{\Sigma} - \hat{\Sigma}^{\text{emp}}\|_\infty^{1/2} \leq r_\alpha\right) &\leq \mathrm{P}\left(\max_{i=1,\ldots,d} \sum_{j=1}^d \hat{\sigma}_{i,j}^2 \mathbf{1}\left[\tilde{\sigma}_{i,j} = 0\right] \leq r_\alpha^4\right) \\
&\leq \mathrm{P}\left(\max_{i=1,\ldots,d} \sum_{j=1}^d \hat{\sigma}_{i,j}^2 \mathbf{1}\left[\sigma_{i,j} = 0\right] + \hat{\sigma}_{i_0,j_0}^2 \leq r_\alpha^4\right) \\
&\leq \mathrm{P}\left(\hat{\sigma}_{i_0,j_0} \leq r_\alpha^2\right)
\end{aligned}$$

We have that if $\sigma_{i,j} \neq 0$ then $|\sigma_{i,j}| > \delta > 0$. Hence, $\hat{\sigma}_{i_0,j_0} = (\hat{\sigma}_{i_0,j_0} - \sigma_{i_0,j_0}) + \sigma_{i_0,j_0} \geq o_p(n^{-1/2}) + \delta$. Meanwhile, $r_\alpha = O(n^{-1/2})$. Thus, $\mathrm{P}\left(\hat{\sigma}_{i_0,j_0} \leq r_\alpha^2\right) \to 0$ as $n \to \infty$. $\square$

2.B Concentration Results

2.B.1 Concentration results for log-concave measures

Gaussian concentration for log-concave measures is established via the following theorems. In short, Theorem 2.B.2 states that log-concave measures satisfy a logarithmic Sobolev inequality, which bounds the entropy of the measure; see Definition 2.B.1. Logarithmic Sobolev inequalities were first introduced in Gross (1975), and this result is due to Bakry and Émery (1984). Following that, Theorem 2.B.3 links the logarithmic Sobolev inequality with Gaussian concentration. Finally, Corollary 2.B.4 extends this Gaussian concentration to product measures whose individual components satisfy logarithmic Sobolev inequalities in a dimension-free way due to the subadditivity of the entropy.

Definition 2.B.1 (Entropy)**.** *For a probability measure μ on a measurable space $(\Omega, \mathcal{F})$ and for any non-negative measurable function f on $(\Omega, \mathcal{F})$, the entropy is*

$$\mathrm{Ent}_\mu(f) = \int f \log f d\mu - \left(\int f d\mu\right) \log\left(\int f d\mu\right).$$

Theorem 2.B.2 (Ledoux (2001), Theorem 5.2)**.** *Let μ be strongly log-concave on $\mathbb{R}^d$ for some $c > 0$. Then, μ satisfies the logarithmic Sobolev inequality. That is, for all locally Lipschitz $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathrm{Ent}_\mu\left(f^2\right) \leq \frac{2}{c} \int |\nabla f|^2 d\mu.$$

Theorem 2.B.3 (Ledoux (2001), Theorem 5.3)**.** *If μ is a probability measure on $\mathbb{R}^d$ such that $\mathrm{Ent}_\mu(f^2) \leq \frac{2}{c}\int |\nabla f|^2 d\mu$, then μ has Gaussian concentration. That is, Let $X \in \mathbb{R}^d$ be a random variable with law μ. Then, for all 1-Lipschitz functions $\phi : \mathbb{R}^d \to \mathbb{R}$ and for all $r > 0$,*

$$\mathrm{P}\left(\phi(X) \geq \mathrm{E}\phi(X) + r\right) \leq \mathrm{e}^{-cr^2/2}.$$

Corollary 2.B.4 (Ledoux (2001), Corollary 5.7)**.** *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ have measures $\mu_1, \ldots, \mu_n$, which are all strongly log-concave with coefficients $c_1, \ldots, c_n$. Let $\nu = \mu_1 \otimes \ldots \otimes \mu_n$ be the product measure on $\mathbb{R}^{d \times n}$. Then,*

$$\mathrm{Ent}_\nu\left(f^2\right) \leq \frac{2}{\min_i c_i} \int |\nabla f|^2 d\nu.$$

2.B.2 Concentration results for bounded random variables

The following results can be found in more depth in Giné and Nickl (2016) Section 3.3.4 and specifically in Example 3.3.13 (a). Corollary 2.3.7 is effectively a more general version of Hoeffding's Inequality. To establish the corollary, we begin with the definition of functions of bounded differences.

Definition 2.B.5 (Functions of Bounded Differences)**.** *A function $f : \mathbb{R}^{d\times n} \to \mathbb{R}$ is of bounded differences if*

$$\sup_{x_i, x_i', x_j \in \mathbb{R}^d, j\neq i} |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

Then, Gaussian concentration can be established for functions of bounded differences by the following theorem.

Theorem 2.B.6. *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ and $Z = f(X_1, \ldots, X_n)$ where f has bounded differences with $c = \sum_{i=1}^n c_i$. Then, for all $r > 0$,*

$$\mathrm{P}\left(Z \geq \mathrm{E}Z + r\right) \leq \mathrm{e}^{-2r^2/c^2}.$$

2.B.3 Concentration results for sub-exponential measures

The following exposition can be found in more details in Section 5.3 of Ledoux (2001). In order to achieve the sub-exponential concentration inequality utilized in Section 2.3.3, the role of the logarithmic Sobolev inequality from the log-concave measure setting is replaced with a *modified logarithmic Sobolev inequality*. That is, a probability measure μ on the Borel sets of some metric space (X, d) is said to satisfy such an inequality if there exists a real valued $\beta(\rho) \geq 0$ such that for any f with $\|\nabla f\|_\infty \leq \rho$ and $\int \mathrm{e}^f d\mu \leq \infty$ that

$$\mathrm{Ent}_\mu(\mathrm{e}^f) \leq \beta(\rho) \int |\nabla f|^2 \mathrm{e}^f d\mu$$

where $\mathrm{Ent}_\mu(\cdot)$ is the entropy of the measure μ from Definition 2.B.1.

Given a measure μ that satisfies a modified logarithmic Sobolev inequality, it can be shown that this measure also has sub-exponential concentration as in the conclusion of Corollary 2.3.9. The final link comes from Theorem 5.14 from Ledoux (2001), which states that if a measure μ on the Borel sets of some metric space satisfies the Poincaré inequality for some fixed $C > 0$,

$$\mathrm{Var}\left(f\right) \leq C \int |\nabla f|^2 d\mu$$

then it also satisfies the modified logarithmic Sobolev inequality for any function f such that $\|\nabla f\|_\infty \leq \rho \leq 2/\sqrt{C}$ with

$$\beta(\rho) = \frac{C}{2}\left(\frac{2+\rho\sqrt{C}}{2-\rho\sqrt{C}}\right)^2 \mathrm{e}^{\sqrt{5C}\rho}.$$

2.C Derivations of Lipschitz constants

The following lemmas and propositions establish that specific functions used in the construction of confidence sets are, in fact, Lipschitz functions. This allows such functions to be used in the context of the concentration inequalities of the previous section.

Lemma 2.C.1. *Let A and B be two $d\times d$ real valued symmetric non-negative-definite matrices. Then,*

$$\|A+B\|_1 = \|A\|_1 + \|B\|_1$$

where $\|\cdot\|_1$ is the trace class norm.

Proof. By definition, $\|A\|_1 = \operatorname{tr}\left((A^*A)^{1/2}\right)$. If A is symmetric and non-negative-definite, then $(A^*A)^{1/2} = A$. Hence, if A and B are symmetric and positive-definite, then so is $A+B$. Indeed, if $A = A^{\mathrm{T}}$ and $B = B^{\mathrm{T}}$ then $A+B = A^{\mathrm{T}}+B^{\mathrm{T}} = (A+B)^{\mathrm{T}}$. Also, if for all $x \in \mathbb{R}^d$ we have $x^{\mathrm{T}}Ax \geq 0$ and $x^{\mathrm{T}}Bx \geq 0$, then $x^{\mathrm{T}}(A+B)x = x^{\mathrm{T}}Ax + x^{\mathrm{T}}Bx \geq 0$. Therefore,

$$\|A+B\|_1 = \operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B) = \|A\|_1 + \|B\|_1.$$

□

Proposition 2.C.2 (Lipschitz for $p=1$)**.** *Assume that $X_1,\ldots,X_n \in \mathbb{R}^d$ and that $\mathrm{E}X_i = 0$ for $i = 1,\ldots,n$. The function $\phi : \mathbb{R}^{d\times n} \to \mathbb{R}$ defined as*

$$\phi(X_1,\ldots,X_n) = \left\|\frac{1}{n}\sum_{i=1}^n XX^{\mathrm{T}}\right\|_1^{1/2}$$

is Lipschitz with constant $n^{-1/2}$ with respect to the Euclidean metric $d_{(2,2)}(\mathbf{X},\mathbf{Y}) = \left(\sum_{i=1}^n \|X_i - Y_i\|_{\ell^2}^2\right)^{1/2}$.

Proof. Let $X_1,\ldots,X_n,Y_1,\ldots,Y_n \in \mathbb{R}^d$ with $\mathrm{E}X_i = \mathrm{E}Y_i = 0$ for all i and denote

$\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$. Making use of Lemma 2.C.1, we have

$$\begin{aligned}
n(\phi(\mathbf{X}) - \phi(\mathbf{Y}))^2 &= \left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_1 + \left\|\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}\right\|_1 - 2\left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_1^{1/2}\left\|\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}\right\|_1^{1/2} \\
&= \sum_{i=1}^n \left(\|X_i\|_{\ell^2}^2 + \|Y_i\|_{\ell^2}^2\right) - 2\left[\left(\sum_{i=1}^n \|X_i\|_{\ell^2}^2\right)\left(\sum_{i=1}^n \|Y_i\|_{\ell^2}^2\right)\right]^{1/2} \\
&= \sum_{i=1}^n \left(\|X_i\|_{\ell^2}^2 + \|Y_i\|_{\ell^2}^2\right) - 2\left[\sum_{i,j=1}^n \|X_i\|_{\ell^2}^2 \|Y_j\|_{\ell^2}^2\right]^{1/2} \\
&= \sum_{i=1}^n \left(\|X_i\|_{\ell^2}^2 + \|Y_i\|_{\ell^2}^2\right) - \\
&\quad - 2\left[\sum_{i<j}\left(\|X_i\|_{\ell^2}^2\|Y_j\|_{\ell^2}^2 + \|X_j\|_{\ell^2}^2\|Y_i\|_{\ell^2}^2\right) + \sum_{i=1}^n \|X_i\|_{\ell^2}^2\|Y_i\|_{\ell^2}^2\right]^{1/2} \\
&\leq \sum_{i=1}^n \left(\|X_i\|_{\ell^2}^2 + \|Y_i\|_{\ell^2}^2\right) - \\
&\quad - 2\left[2\sum_{i<j}\left(\|X_i\|_{\ell^2}\|Y_j\|_{\ell^2}\|X_j\|_{\ell^2}\|Y_i\|_{\ell^2}\right) + \sum_{i=1}^n \|X_i\|_{\ell^2}^2\|Y_i\|_{\ell^2}^2\right]^{1/2} \\
&\leq \sum_{i=1}^n \left(\|X_i\|_{\ell^2}^2 + \|Y_i\|_{\ell^2}^2\right) - 2\sum_{i=1}^n \|X_i\|_{\ell^2}\|Y_i\|_{\ell^2} \\
&\leq \sum_{i=1}^n \left(\|X_i\|_{\ell^2} - \|Y_i\|_{\ell^2}\right)^2 \\
&\leq \sum_{i=1}^n \|X_i - Y_i\|_{\ell^2}^2
\end{aligned}$$

The inequality above arises due to the fact that $\left(\|X_i\|_{\ell^2}\|Y_j\|_{\ell^2} - \|X_j\|_{\ell^2}\|Y_i\|_{\ell^2}\right)^2 \geq 0$, and thus $\|X_i\|_{\ell^2}^2\|Y_j\|_{\ell^2}^2 + \|X_j\|_{\ell^2}^2\|Y_i\|_{\ell^2}^2 \geq 2\|X_i\|_{\ell^2}\|Y_j\|_{\ell^2}\|X_j\|_{\ell^2}\|Y_i\|_{\ell^2}$ □

The next two lemmas are used to prove the Lipschitz constant for the p-Schatten norms with $p = 2$ and $p = \infty$, respectively. The first lemma is reminiscent of the Cauchy-Schwarz inequality in the setting of the 2-Schatten norm.

Lemma 2.C.3. *Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n \in \mathbb{R}^d$. Then, for the Frobenius norm,*

$$\left\|\sum_{i=1}^n X_i Y_i^{\mathrm{T}}\right\|_2 \leq \left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_2^{1/2}\left\|\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}\right\|_2^{1/2}.$$

Proof. For any matrix $M \in \mathbb{R}^{d\times d}$, we have that $\|M\|_2^2 = \mathrm{tr}\left(MM^{\mathrm{T}}\right)$. Hence, starting from the left hand side of the desired inequality and applying the Cauchy-Schwarz

inequality gives us

$$\begin{aligned}
\left\|\sum_{i=1}^n X_i Y_i^{\mathrm{T}}\right\|_2 &= \operatorname{tr}\left(\sum_{i,j=1}^n X_i Y_i^{\mathrm{T}} Y_j X_j^{\mathrm{T}}\right)^{1/2} \\
&= \left(\sum_{i,j=1}^n \langle X_i, X_j\rangle \langle Y_i, Y_j\rangle\right)^{1/2} \\
&\leq \left(\left(\sum_{i,j=1}^n \langle X_i, X_j\rangle^2\right)^{1/2} \left(\sum_{i,j=1}^n \langle Y_i, Y_j\rangle^2\right)^{1/2}\right)^{1/2} \\
&\leq \left(\operatorname{tr}\left(\sum_{i,j=1}^n X_i X_i^{\mathrm{T}} X_j X_j^{\mathrm{T}}\right)^{1/2} \operatorname{tr}\left(\sum_{i,j=1}^n Y_i Y_i^{\mathrm{T}} Y_j Y_j^{\mathrm{T}}\right)^{1/2}\right)^{1/2} \\
&\leq \left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_2^{1/2} \left\|\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}\right\|_2^{1/2}
\end{aligned}$$

□

Lemma 2.C.4. *Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n \in \mathbb{R}^d$. Then, for the operator norm,*

$$\left\|\sum_{i=1}^n X_i Y_i^{\mathrm{T}}\right\|_\infty \leq \left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_\infty^{1/2} \left\|\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}\right\|_\infty^{1/2}.$$

Proof. Using the definition of the operator norm and the Cauchy-Schwarz inequality, we have that

$$\begin{aligned}
\left\|\sum_{i=1}^n X_i Y_i^{\mathrm{T}}\right\|_\infty &= \sup_{v\in\mathbb{R}^d,\ \|v\|_{\ell^2}=1} \sum_{i=1}^n \langle X_i, v\rangle \langle Y_i, v\rangle \\
&\leq \left(\sup_{v\in\mathbb{R}^d,\ \|v\|_{\ell^2}=1} \sum_{i=1}^n \langle X_i, v\rangle^2 \sup_{u\in\mathbb{R}^d,\ \|u\|_{\ell^2}=1} \sum_{i=1}^n \langle Y_i, u\rangle^2\right)^{1/2} \\
&= \left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_\infty^{1/2} \left\|\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}\right\|_\infty^{1/2}.
\end{aligned}$$

□

Proposition 2.C.5 (Lipschitz for $p = 2$ or $p = \infty$)**.** *Assume that $X_1, \ldots, X_n \in \mathbb{R}^d$ and that $\mathrm{E}X_i = 0$ for $i = 1, \ldots, n$. Let $p \in [2, \infty]$. The function $\phi : \mathbb{R}^{d\times n} \to \mathbb{R}$ defined as*

$$\phi(X_1, \ldots, X_n) = \left\|\frac{1}{n}\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_p^{1/2}$$

is Lipschitz with constant $n^{-1/2}$ with respect to the Euclidean metric $d_{(2,2)}(\mathbf{X}, \mathbf{Y}) = \left(\sum_{i=1}^n \|X_i - Y_i\|_{\ell^2}^2\right)^{1/2}$.

Proof. To establish that ϕ is Lipschitz with the desired constant, we proceed by bounding the Gâteaux derivative. Let $p \in \{2, \infty\}$. It is conjectured in the paragraph below that this proof can be expanded to all values of $p \in [1, \infty]$ if a little more thought is put into the supporting lemmas. For $h \in \mathbb{R}$ and any $X_1 \ldots, X_n, Y_1, \ldots, Y_n \in \mathbb{R}^d$ such that $\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\|_p \neq 0$ and $\|\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}\|_p \neq 0$,

$$\begin{aligned}
\sqrt{n} d\phi(X_1, \ldots, X_n; Y_1, \ldots, Y_n) &= \\
&= \lim_{h \to 0} \left(\frac{\left\|\sum_{i=1}^n (X_i + hY_i)(X_i + hY_i)^{\mathrm{T}}\right\|_p - \left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_p}{2\left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_p^{1/2} \left(\sum_{i=1}^n \|hY_i\|_{\ell^2}^2\right)^{1/2}} \right) \\
&\leq \lim_{h \to 0} \left(\frac{\left\|\sum_{i=1}^n \left(hY_i X_i^{\mathrm{T}} + hX_i Y_i^{\mathrm{T}} + h^2 Y_i Y_i^{\mathrm{T}}\right)\right\|_p}{2\left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_p^{1/2} \left(\sum_{i=1}^n \|hY_i\|_{\ell^2}^2\right)^{1/2}} \right) \\
&\leq \frac{\left\|\sum_{i=1}^n \left(Y_i X_i^{\mathrm{T}} + X_i Y_i^{\mathrm{T}}\right)\right\|_p}{2\left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_p^{1/2} \left(\sum_{i=1}^n \|Y_i\|_{\ell^2}^2\right)^{1/2}} \\
&\leq \frac{\left\|\sum_{i=1}^n X_i Y_i^{\mathrm{T}}\right\|_p}{\left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_p^{1/2} \left\|\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}\right\|_p^{1/2}}
\end{aligned}$$

where we used the facts that, for $M \in \mathbb{R}^{d \times d}$, $\|M\|_p = \|M^{\mathrm{T}}\|_p$, that $\sum_{i=1}^n \|Y_i\|_{\ell^2}^2 = \sum_{i=1}^n \|Y_i Y_i^{\mathrm{T}}\|_p \geq \|\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}\|_p$, and that

$$\left\|\sum_{i=1}^n \left(Y_i X_i^{\mathrm{T}} + X_i Y_i^{\mathrm{T}}\right)\right\|_p \leq 2\left\|\sum_{i=1}^n X_i Y_i^{\mathrm{T}}\right\|_p.$$

Applying Lemma 2.C.3 in the $p = 2$ case and Lemma 2.C.4 in the $p = \infty$ case shows that $\sqrt{n} d\phi(\cdot) \leq 1$ for all X_i with $\left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_2 \neq 0$. With application of the Mean Value Theorem, we have the desired Lipschitz constant.

In the case that $\left\|\sum_{i=1}^n X_i X_i^{\mathrm{T}}\right\|_p = 0$, we also achieve the same Lipschitz constant. Since, $X_i X_i^{\mathrm{T}}$ is positive-definite, the norm can only be zero if all $X_i = (0, \ldots, 0)^{\mathrm{T}}$. Hence, for any $Y_1, \ldots, Y_n \in \mathbb{R}^d$,

$$\begin{aligned}
\sqrt{n} |\phi(X_1, \ldots, X_n) - \phi(Y_1, \ldots, Y_n)| &= \\
&= \left\|\sum_{i=1}^n Y_i Y_i^{\mathrm{T}}\right\|_p^{1/2} \leq \left(\sum_{i=1}^n \|Y_i\|_{\ell^2}^2\right)^{1/2} = \left(\sum_{i=1}^n \|X_i - Y_i\|_{\ell^2}^2\right)^{1/2}.
\end{aligned}$$

□

It is conjectured that the function $\phi(\cdot)$ is 1-Lipschitz for all $p \in [1, \infty]$, which follows immediately if Lemmas 2.C.3 and 2.C.4 can be expanded to similar results for all $p \in [1, \infty]$.

Conjecture 2.C.6. *Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n \in \mathbb{R}^d$. Then, for any $p \in [1, \infty]$,*

$$\left\| \sum_{i=1}^{n} X_i Y_i^{\mathrm{T}} \right\|_p \leq \left\| \sum_{i=1}^{n} X_i X_i^{\mathrm{T}} \right\|_p^{1/2} \left\| \sum_{i=1}^{n} Y_i Y_i^{\mathrm{T}} \right\|_p^{1/2}.$$

Chapter 3

Concentration for covariance operators

3.1 Introduction

Functional data spans many realms of applications from medical imaging (Jiang et al., 2016) to speech and linguistics (Pigoli et al., 2014, 2015) to the intricate movements of DNA molecules (Panaretos et al., 2010). General inference techniques for functional data are one area of analysis that has received much attention in recent years from the construction of confidence sets to other topics such as k-sample tests, classification, and clustering of functional data. Most testing methodology treats the data as continuous L^2 valued functions and subsequently reduces the problem to a finite dimensional one through expansion in some orthogonal basis such as the often utilized Karhunen-Loève expansion (Horváth and Kokoszka, 2012). However, inference making use of non-Hilbert norms has not been addressed adequately. In this chapter, we propose a novel methodology for performing fully functional inference through the application of concentration inequalities. This is furthermore a single methodology applicable to a wide variety of inference problems; for general concentration of measure results, see Ledoux and Talagrand (1991); Steele (1997); Ledoux (2001); Milman and Schechtman (2009); Boucheron et al. (2013); Habib et al. (2013); Giné and Nickl (2016). Special emphasis is given to inference on covariance operators, which offers a fruitful way to analyze functional data.

Imagine multiple samples of speech data collected from multiple speakers. Each speaker will have his or her own sample covariance operator taking into account the unique variations of his or her speech and language. An exploratory researcher may want to find natural clusters amidst the speakers perhaps corresponding to gender, language, or regional accent. Meanwhile, a linguist studying the similarities between languages may want to test for the equality of such covariances. A computer scientist

may need to implement an algorithm that when given speech data quickly identifies what language is being spoken and furthermore parses the sound clip and identifies each individual phoneme in order to process the speech into text. Our proposed method has the versatility to yield statistical tests that answer all of these questions as well as others.

Past methods for analyzing covariance operators (Panaretos et al., 2010; Fremdt et al., 2013) rely on the Hilbert-Schmidt setting for their inference. However, the recent work of Pigoli et al. (2014) argues that the use of the Hilbert-Schmidt metric ignores the geometry of the covariance operators and that more statistical power can be gained by using alternative metrics. The main drawback of their research is their reliance on permutation based tests, which are computationally intensive and, in some instances, incapable of achieving decent accuracy in a sensible amount of time. Even more, in the age of big data, p-values less than 1/1000 may be desired which are computationally impossible with this method; see Figure 3.1. Hence, we approach such inference for covariance operators by using a non-asymptotic dimension free concentration of measure approach, which can incorporate metrics based on arbitrary norms. This allows us to work in the full generality of Banach spaces where we can choose those norms which provide the most statistical power to our inference. This has previously been used in nonparametric statistics and machine learning, sometimes under the name of *Rademacher complexities* (Koltchinskii, 2001, 2006; Bartlett et al., 2002; Bartlett and Mendelson, 2003; Giné and Nickl, 2010b; Arlot et al., 2010; Lounici and Nickl, 2011; Kerkyacharian et al., 2012; Fan, 2011). These concentration inequalities provide a natural way to construct non-asymptotic confidence regions and, subsequently, statistical tests. Our single approach can classify as well as k-nearest neighbours, cluster as well as k-means, and can test for equality of covariance among multiple samples as well as the permutation test from Pigoli et al. (2014) and Cabassi et al. (2017). These methods are currently available in the R package `fdcov` (Cabassi and Kashlak, 2016).

In this chapter, Section 3.2 details how Talagrand's concentration inequality in the Banach space setting can be used to construct a confidence set similar to as was done for covariance matrices in the previous chapter. Section 3.3 introduces three different inferential techniques that stem from these concentration based confidence sets. They include a k sample test for equality of covariance, a covariance operator classifier, and an expectation-maximization style clustering algorithm for covariance operators. Section 3.4 takes the three mentioned methodologies and applies them to both simulated and phoneme data. In the appendices of this chapter, Appendix 3.A approaches the construction of the confidence set from Section 3.2 in more generality. Appendix 3.B details the weak variance calculations in a number of different settings. Appendix 3.C provides an exposition of how to think about tensor products in

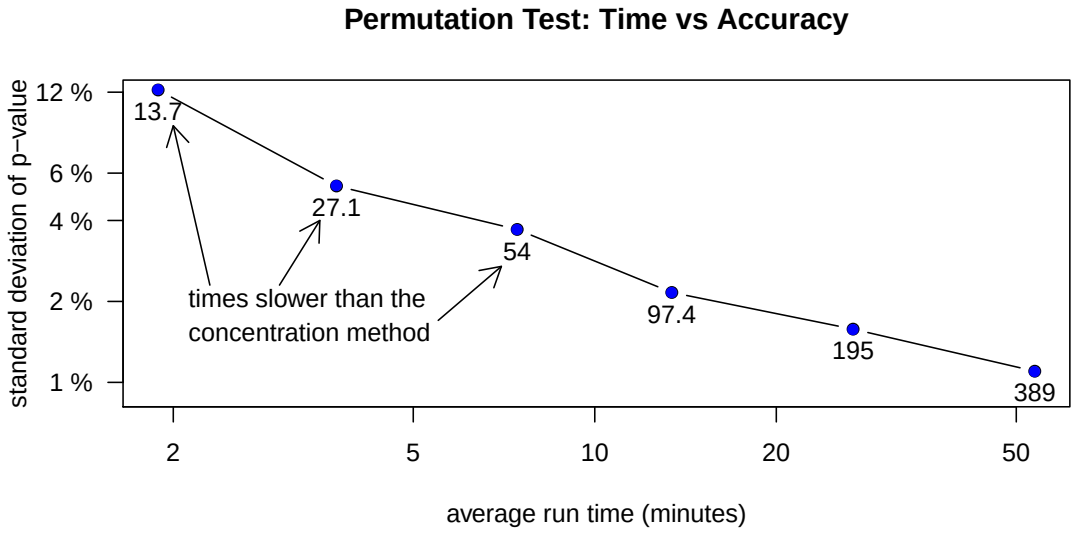


Figure 3.1: Plotted are the run times against the accuracy of the permutation test for testing for equality of covariance given five samples of 30 curves each. The procedure requires over 50 minutes of computation time to get a standard deviation of around 1% for the estimated p-value. Adjacent to each point is the number of times slower the permutation test is when compared to the concentration test. The average run times were clocked on an Intel(R) Core(TM) i3-3217U CPU @ 1.80GHz.

the Hilbert space, Banach space, and finite dimensional settings. Appendix 3.D briefly investigates the consequences of applying our statistical tests to data with noise added and data from heavy tailed distributions. As confidence sets based on concentration inequalities are often larger than desired, Appendix 3.E proposes how the sizes of confidence sets constructed via Talagrand's concentration inequality can be improved with a cross-validation procedure.

3.2 Confidence sets for covariance operators

To construct a confidence set for covariance operators, we let our functional data $f_i \in L^2(I)$ and $f_i^{\otimes 2} = f_i \otimes f_i \in Op(L^2)$, where $Op(L^2)$ is the Hilbert space of bounded linear operators mapping L^2 to L^2, such that $(f_i \otimes f_i)\phi = \langle f_i, \phi \rangle f_i$ for some $\phi \in L^2$. The following construction of our confidence set is based on Talagrand's concentration inequality (Talagrand, 1996a) with explicit constants, which can be thought of as a more general version of Bernstein's inequality (Bernstein, 1924)(Boucheron et al., 2013, Chapter 2). This inequality is typically stated for empirical processes (Giné and Nickl, 2016, Theorem 3.3.9 and 3.3.10), but applies to random variables with values in a separable Banach space $(B, \|\cdot\|_B)$ as well by simple duality arguments (Giné and Nickl, 2016, Example 2.1.6). More details on this construction in the general Banach

space setting can be found in Appendix 3.A. For some desired p-Schatten norm, $\|\cdot\|_p$, with $p \in [1, \infty)$ and with conjugate $q = p/(p-1)$, we require the following terms, which correspond to the distance between the empirical covariance estimate and the true covariance operator and a weak variance term for this random variable:

$$Z = \left\| \frac{1}{n} \sum_{i=1}^{n} (f_i - \bar{f}) \otimes (f_i - \bar{f}) - \mathrm{E}\left(f_i \otimes f_i\right) \right\|_p = \left\| \hat{\Sigma} - \Sigma \right\|_p$$
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\Pi\|_q \leq 1} \mathrm{E} \langle f_i^{\otimes 2} - \mathrm{E} f_i^{\otimes 2}, \Pi \rangle^2.$$

In the above equation, the supremum is to be taken over a countably dense subset of the unit ball of $\Pi \in Op(L^2)$. For some $U \geq \|f_i^{\otimes 2}\|_{L^2}^2$ and $v_n = 2U\mathrm{E}Z + n\sigma^2$, the initial level $(1-\alpha)$ confidence set constructed is

$$C_{n,1-\alpha} = \left\{ \Sigma : Z \leq \mathrm{E}Z + \sqrt{-2v_n \log(2\alpha)/n} - U \log(2\alpha)/(3n) \right\}.$$

To make this confidence set usable on real data, Rademacher random variables, $\varepsilon_1, \ldots, \varepsilon_n$, defined in Section 1.1.5 are incorporated. The Rademacher average defined as $R_n = n^{-1} \sum_{i=1}^{n} \varepsilon_i ((f_i - \bar{f})^{\otimes 2} - \hat{\Sigma})$, will be used as a proxy for the unknown $\mathrm{E}Z$ where the ε_i for $i = 1, \ldots, n$ are independent and identically distributed and furthermore independent of the f_i. This usage is justified by the symmetrization inequality also detailed in Appendix 3.A and further discussed in Chapter 4. Note that R_n is also in $Op(L^2)$, because for any $\phi \in L^2(I)$ and for some $M \in \mathbb{R}$,

$$\|R_n \phi\|_{L^2} = \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i ((f_i - \bar{f})^{\otimes 2} - \hat{\Sigma}) \right) \phi \right\|_{L^2} \leq$$
$$\leq \frac{1}{n} \sum_{i=1}^{n} |\varepsilon_i| \left\| \left((f_i - \bar{f})^{\otimes 2} - \hat{\Sigma} \right) \phi \right\|_{L^2} \leq M \|\phi\|_{L^2}$$

since $((f_i - \bar{f})^{\otimes 2} - \hat{\Sigma})$ is a bounded linear operator and $|\varepsilon_i| = 1$.

Next we look at the bound U. In the case that there exists a fixed $c \in \mathbb{R}$ with $\|f_i\|_{L^2} \leq c$ for all i corresponding to a physical bound on the energy of f_i, $\|f_i^{\otimes 2}\|_p = \|f_i\|_{L^2}^2 \leq c^2 = U$. It will be determined in Appendix 3.B that $U \geq \sigma$ in this case. It may be possible to select U via a cross-validation procedure. In general, setting $U = \sigma$ gives good experimental results when f_i is Gaussian as will be discussed in later sections. This results in $v_n \approx \sigma^2/n$. For any $p \in [1, \infty)$ and $\alpha \in [0, 1/2]$, the proposed $(1-\alpha)$-confidence set for covariance operators is

$$C_{n,1-\alpha} = \left\{ \Sigma : \|\hat{\Sigma} - \Sigma\|_p \leq \|R_n\|_p + \sigma \sqrt{\frac{-2\log(2\alpha)}{n}} - \frac{\sigma \log(2\alpha)}{3n} \right\}. \tag{3.2.1}$$

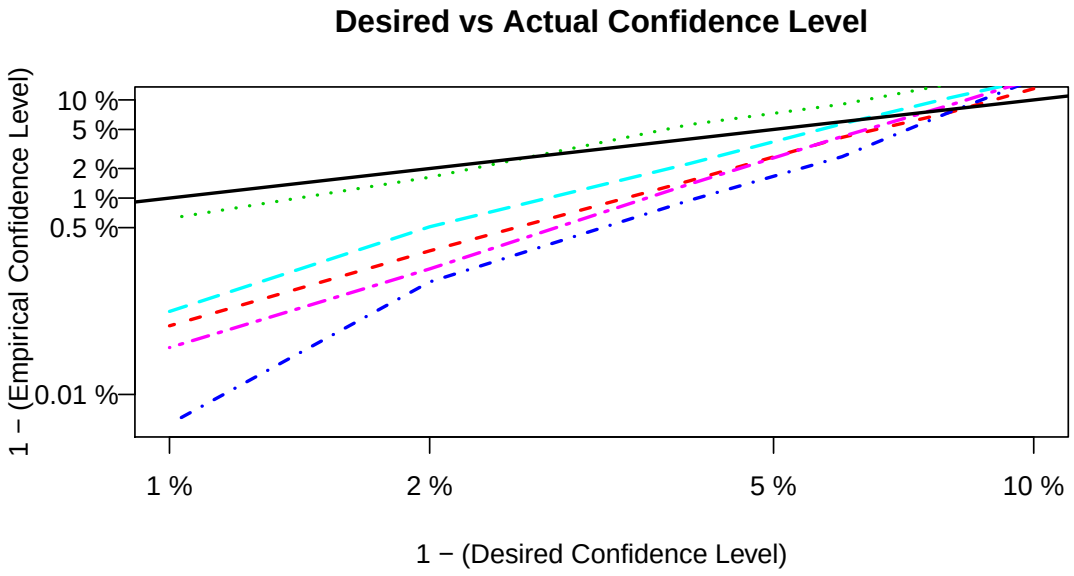


Figure 3.2: The empirical confidence level of the set from Equation 3.2.1 for five different operators given a sample size of 35 curves. The black line is where the desired and empirical levels are equal. The desired level ranges from $\alpha = 1\%$ to $\alpha = 10\%$. 10,000 replications were used to produce these curves. Most lines lie below the black line indicating that the sets are slightly too large.

where σ depends on the distribution on the functional data. As a rule of thumb for the choice of σ^2, as shown in Appendix 3.B, is to note that $\sigma^2 \leq \|\mathrm{E}(f^{\otimes 4}) - \Sigma^{\otimes 2}\|_p$ and to estimate this bound empirically by $\hat{\sigma}^2$. For example, when the f_i are from a Gaussian process $\hat{\sigma} \leq 2^{1/2}\|\Sigma\|_p$ as explained in detail in Appendix 3.B.3. In practice, $\|\Sigma\|_p$ is replaced with the consistent estimator $\|\hat{\Sigma}\|_p$, which follows from the central limit theorem for Banach space valued random variables.

Constructing confidence sets in this way will lead to sets that are too large. That is, our $(1-\alpha)-$confidence set may have a converge greater than the desired $1-\alpha$. While the level increases more quickly than desired, it does not increase too quickly to be useful as will be discussed in the applications of Section 3.3. Figure 3.2 displays the empirical coverage for five different operators. Specifically, for the five operators derived from the phoneme data sets of Section 3.4.1, 35 curves were generated as random realizations of a zero mean Gaussian process with given covariance. Then, the confidence set was constructed, and it was tested whether or not the true covariance operator lied within this set. This was repeated 10,000 to produce the estimates in Figure 3.2.

3.3 Applications

3.3.1 k sample comparison

Testing for the equality of means among multiple sets of data is a common task in data analysis. In the functional setting, there has been recent work on performing such a test on covariance operators in order to test whether or not k sets of curves have similar variation. Panaretos et al. (2010) propose such a method for a two sample test on covariance operators given data from Gaussian processes and apply this method to analyze the bending and twisting of DNA microcircles. Similarly, Fremdt et al. (2013) propose a non-parametric two sample test on covariance operators and apply their method to analyze egg-laying curves for fruit flies. Both of these approaches make use of the Karhunen-Loève expansion and, hence, the underlying Hilbert space geometry. Pigoli et al. (2014) take a comparative look at a variety of metrics to rank their statistical power when used in a two sample permutation test and apply their method to the analysis of samples of spoken words in five romance languages. This permutation test is extended in Cabassi et al. (2017) and applied to exercise curves of mice running on wheels.

Following from the results of Pigoli et al. (2014), our method uses the p-Schatten norms with the concentration inequality based confidence sets of the previous section to compare covariance operators. In the two sample setting, we are able to achieve similar statistical power to that of the permutation test after proper tuning of the coefficients in the inequalities. Furthermore, the analytic nature of the concentration approach leads to a significant reduction in computing time, which offers an even more significant savings for larger values of k as was already displayed in Figure 3.1. Details on permutation tests for comparing covariance operators when $k > 2$ can be found in Cabassi et al. (2017).

From the confidence set constructed in the previous section, we can devise a test for comparing the empirical covariance operators generated from k samples of functional data. Let the k samples be $f_1^{(1)}, \ldots, f_{n_1}^{(1)}, \ldots, f_1^{(k)}, \ldots, f_{n_k}^{(k)}$ where for each sample i and all elements $j = 1, \ldots, n_i$, $f_j^{(i)}$ has common covariance $\Sigma^{(i)}$. Our goal is to design a test for the following two hypotheses:

$$\mathrm{H}_0 : \ \Sigma^{(1)} = \ldots = \Sigma^{(k)} \qquad\qquad \mathrm{H}_1 : \ \exists\, i, j \text{ s.t. } \Sigma^{(i)} \neq \Sigma^{(j)}.$$

To achieve this, a pooled estimate of the weak variance is computed as a weighted average of each sample's individual weak variance in similar style to that of a standard t-test (Casella and Berger, 2002, Chapter 8). Let the total data size be $N = n_1 + \ldots + n_k$ and σ_i^2 be the weak variance for sample i, then the pooled variance is defined as $\sigma_{\text{pool}}^2 = N^{-1} \sum_{i=1}^{k} n_i \sigma_i^2$. Given Gaussian data and the p-Schatten norm, for

example, this reduces to $\sigma^2_{\text{pool}} = 2N^{-1}\sum_{i=1}^k n_i \|\Sigma^{(i)}\|_p^2$. In practice, σ^2_{pool} is estimated from the data for the following confidence regions in order to have those regions only depend on the data.

Taking inspiration from the standard method for analysis of variance (Casella and Berger, 2002, Chapter 11), let $\hat{\Sigma}^{(i)}$ be the empirical estimate of the covariance operator for the ith sample, and let $\hat{\Sigma}$ be the estimate of the covariance operator for the total data set. Making use of the confidence sets for covariance operators from Section 3.2 gives the rejection region

$$\mathcal{R} = \left\{ f : \sum_{i=1}^k \left\| \hat{\Sigma}^{(i)} - \hat{\Sigma} \right\|_p > \sum_{i=1}^k \left\| \sum_{j=1}^{n_i} \varepsilon_{i,j} \left(f_j^{(i)\otimes 2} - \hat{\Sigma} \right) \right\|_p + \right.$$
$$\left. + \sqrt{\sum_{i=1}^k \frac{\sigma^2_{\text{pool}}}{n_i}(-2\log 2\alpha)} + \left(\sum_{i=1}^k \frac{\sigma_{\text{pool}}}{n_i} \right) \frac{\log 2\alpha}{3} \right\},$$

which under the null hypothesis will have size no greater than the desired α.

The size of the test induced by this rejection region is significantly less than the target size α due to the use of multiple concentration inequalities. Hence, tuning the inequalities is required to yield a useful test. Many experiments were run on simulated data sets generated as samples from a Gaussian process with randomly generated covariance operators whose eigenvalues were chosen to decay at a variety of rates. In this setting, the coefficients of $1 - k^{-1/2}$ for the Rademacher term and $(k+2)/(k+3)$ for the deviation term were determined experimentally to improve the size of the confidence region in the Gaussian process data setting:

$$\mathcal{R} = \left\{ f : \sum_{i=1}^k \left\| \hat{\Sigma}^{(i)} - \hat{\Sigma} \right\|_p > \left(1 - \frac{1}{\sqrt{k}} \right) \sum_{i=1}^k \left\| \sum_{j=1}^{n_i} \varepsilon_{i,j} \left(f_j^{(i)\otimes 2} - \hat{\Sigma} \right) \right\|_p + \right.$$
$$\left. + \left(\frac{k+2}{k+3} \right) \left(\sqrt{\sum_{i=1}^k \frac{\sigma^2_{\text{pool}}}{n_i}(-2\log 2\alpha)} + \left(\sum_{i=1}^k \frac{\sigma_{\text{pool}}}{n_i} \right) \frac{\log 2\alpha}{3} \right) \right\}. \quad (3.3.1)$$

The goal of these tweaked coefficients is to achieve to correct empirical size for the rejection region. The values were determined through extensive simulation of Gaussian process data for a variety of operators, sample sizes n, and categories k. Ultimately, they should be used as a heuristic or a starting place for fine tuning this method to a specific problem of interest. It may also be possible to fabricate a data driven choice for tuning via cross-validation, which is briefly considered in Appendix 3.E

3.3.2 Classification of operators

Classification of functional data has been an area of heavy research over the last two decades. James and Hastie (2001) extend linear discriminant analysis to functional data and consider classifying a subject's ethnicity given spinal bone density measurements. Hall et al. (2001) and Glendinning and Herbert (2003) classify with principal components for radar signal discrimination and classifying gray level images, respectively. Ferraty and Vieu (2003) implement kernel estimators for classifying the phoneme data that we will consider later in this chapter. General linear models for functional data are discussed by Müller and Stadtmüller (2005) with respect to the longevity and reproduction of medflys. Delaigle and Hall (2012) analyze the asymptotic properties of the centroid based classifier with application to the protein content of wheat, Australian rainfall data, and the phoneme data set. Wavelet based classification is detailed by Berlinet et al. (2008) looking at the phoneme data and Chang et al. (2014) looking at positron emission tomography images.

One application of our method beyond classification of functional data is the classification of covariance operators. In the setting of speech analysis, consider multiple speakers and multiple samples of speech from each speaker. The speech samples can be combined into a single sample covariance operator for each speaker. Then, our method can be employed, for example, to classify the covariance operators by speaker gender or speaker language. Evidence that this is a fruitful approach can be found in the analysis of Pigoli et al. (2014, 2015) where a variety of metrics are compared for their efficacy when performing inference on covariance operators. These articles detail the discrepancy between sample covariance operators produced by speakers of different romance languages.

Given k possible labels and n samples of labeled data (Y_i, f_i) with label $Y_i \in \{1, \ldots, k\}$ and observation $f_i \in L^2(I)$, our goal is to determine the probability that a newly observed $g \in L^2(I)$ belongs to label $Y = j$ for $j = 1, \ldots, k$. Given such a g, the standard Bayes classifier chooses the label $y = \arg\max_j \mathrm{P}(Y = j \,|\, g)$ where $\mathrm{P}(Y = j \,|\, g) = \mathrm{P}(g \,|\, Y = j)\,\mathrm{P}(Y = j)/\mathrm{P}(g)$.

Beginning with a training set of n samples with n_j samples of label j, the sample mean of each category is computed: $\bar{f}_j = n_j^{-1} \sum_{i:Y_i=j} f_i$. The probability $\mathrm{P}(g \,|\, Y = j)$ above is replaced with $\mathrm{P}\left(\|\bar{f}_j - g\|_{L^2} > \mathrm{E}\|\bar{f}_j - \mathrm{E}\bar{f}_j\|_{L^2} + r\right)$ with the goal of making a decision based on how much more $\bar{f}_j$ differs from g than $\bar{f}_j$ differs from its expectation $\mathrm{E}\bar{f}_j$. Similar techniques to those in Section 3.2 are subsequently used. Define the Rademacher sum, R_j, and the empirical weak variance, $\hat{\sigma}_j^2$, for label

j to be, respectively,

$$R_j = \frac{1}{n_j} \sum_{i:Y_i=j} \varepsilon_i (f_i - \bar{f}_j), \qquad \hat{\sigma}_j^2 = \left\| \frac{1}{n_j} \sum_{i:Y_i=j} f_i^{\otimes 2} - \bar{f}_j^{\otimes 2} \right\|_p$$

where ε_i are independent and identically distributed Rademacher random variables. The tail bound for the above probability is then

$$\mathrm{P}\left(\|\bar{f}_j - g\|_{L^2} - \|R_j\|_{L^2} > r \right) < \exp\left(\frac{-n_j r^2}{4\|R_j\|_{L^2} U + 2\hat{\sigma}_j^2 + 2rU/3} \right), \tag{3.3.2}$$

where U is an upper bound on $\|f_i\|_{L^2}$. However, this can be approximated by the Gaussian tail $\exp\left(-n_j r^2 / 2\sigma_j^2\right)$. In the simulations of Section 3.4.3, this approximation actually achieves a better correct classification rate on both Gaussian and t-distributed data. This specifically works on t-distributed data despite the heavier tails as the estimate in Equation 3.3.3 below is merely concerned with comparing the tail bounds rather than their specific values. Consequently, the tail for every category is underestimated in the t case, but the ratio remains valid for comparison purposes.

Assuming uniform priors on the labels, the estimate for the probability expression in the Bayes classifier is achieved by replacing the r on the right hand side of Equation 3.3.2 with the observed $\|\bar{f}_j - g\|_{L^2} - \|R_j\|_{L^2}$. The result is

$$\begin{aligned} \mathrm{P}\left(Y = j \,|\, g\right) &\approx \frac{\phi_j(g)}{\sum_{l=1}^{k} \phi_l(g)}, \text{ and} \\ \phi_j(g) &= \exp\left[-\frac{n_j}{2} \left(\frac{\|\bar{f}_j - g\|_{L^2} - \|R_j\|_{L^2}}{\hat{\sigma}_j} \right)^2 \right]. \end{aligned} \tag{3.3.3}$$

This can be extended to the case where an unlabeled observation is a collection of curves $g_1, \ldots, g_m$ by replacing $\|\bar{f}_j - g\|_{L^2}$ in the above expression with $\|\bar{\Sigma}_j - \hat{\Sigma}_g\|_p$ where $\bar{\Sigma}_j$ is the sample covariance of the f_i with label j and $\hat{\Sigma}_g$ is the sample covariance of the g_i. The Rademacher and weak variance terms would also be updated accordingly. The result is a classifier that incorporates the covariance structure of the data into the decision and classifies the covariances.

3.3.3 Clustering of operator mixtures

Closely related to the problem of classification is the problem of clustering, which we will approach by combining the concentration inequality based classification framework of the previous section with an expectation-maximization style algorithm. Given a sample of functional data, we want to assign one of a finite collection of labels

to each curve. For example, in speech processing, one may want to cluster sound clips based on the language of the speaker, or, to be discussed in Section 3.4.4, one may want to separate unlabeled phoneme curves into clusters of similar phonemes.

There have been many recently proposed methods for clustering functional data. Many approaches begin by constructing a low dimensional representation of the data in some basis such as modelling the data with a B-spline basis followed by clustering the spline representations with k-means as in Abraham et al. (2003) who apply their method to studying acidification in the process of cheese making. A similar approach makes use of the eigenfunctions of the covariance operator instead of B-splines and is shown to work on online auction data (Peng and Müller, 2008). In contrast, we will attempt to cluster functions or operators directly via a concentration of measure approach similar to the previously described classification procedure.

Consider the same setting to the previous section of multiple observations from multiple categories. However, now the category labels are missing. This is a functional mixture model where each observed functional datum is a stochastic process with one of k possible covariance operators. In the experiments of Section 3.4, the data will be simulated from a Gaussian process. The goal is to correctly separate the data into k sets. To achieve this, an expectation-maximization style algorithm is implemented making use of the concentration inequality based confidence sets.

Let the observed operator data be $S_1, \ldots, S_n \in Op(L^2)$ where each $S_i = \operatorname{cov}(f_1^{(i)}, \ldots, f_{m_i}^{(i)})$ is a rank m_i operator produced from m_i functional observations. Let the latent label variables be $Y_1, \ldots, Y_n \in \{1, \ldots, k\}$. Assuming no prior knowledge on the proportions of data in each category, the algorithm is initialized with the Jeffreys prior for the Dirichlet distribution by randomly generating $\rho_{i,\cdot}^{(0)} = (\rho_{i,1}^{(0)}, \ldots, \rho_{i,n}^{(0)}) \sim \text{Dirichlet}\,(1/2, \ldots, 1/2)$, the initial probability vector that $\rho_{i,j}^{(0)} = \mathrm{P}\,(Y_i = j \mid f_i)$.

Assuming t iterations of the algorithm have completed, we have a label probability vector $\rho_{i,\cdot}^{(t)}$ for each of the n observations. Given this collection of vectors, the expected proportions of each category can be estimated as $\tau_j^{(t+1)} = n^{-1} \sum_{i=1}^n \mathrm{E}\,(\mathbf{1}_{Y_i=j}) = n^{-1} \sum_{i=1}^n \rho_{i,j}^{(t)}$. Similarly, a weighted sum of the data, $\hat{\Sigma}_j^{(t+1)}$, and a weighted Rademacher sum, $R_j^{(t+1)}$, can be used to update the estimated covariance operators for each label j:

$$\hat{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^n \rho_{i,j}^{(t)} S_i}{\sum_{i=1}^n \rho_{i,j}^{(t)}}, \qquad R_j^{(t+1)} = \frac{\sum_{i=1}^n \rho_{i,j}^{(t)} \varepsilon_i \left(S_i - \hat{\Sigma}_j^{(t+1)} \right)}{\sum_{i=1}^n \rho_{i,j}^{(t)}}.$$

Lastly, a pooled weak variance is required, which is used in place of each individual category weak variance. Otherwise, in practice, the single category with largest variance captures all of the data points. By defining the pooled covariance operator

as $\hat{\Sigma}_{\text{pool}}^{(t+1)} = \sum_{j=1}^{k} \tau_j^{(t+1)} \hat{\Sigma}_j^{(t+1)}$, then the pooled weak variance in the Gaussian case, for example, is estimated by $2\|\hat{\Sigma}_{\text{pool}}^{(t+1)}\|_p$.

As a result, the label probability vectors $\rho_{i,\cdot}^{(t)}$ can be updated given the $t+1$st collection of estimated covariance operators, Rademacher sums, and the pooled covariance operator. From the previous section, Equation 3.3.3 can be used to determine $\rho_{i,j}^{(t+1)} = \mathrm{P}\left(Y_i = j \,\middle|\, S_i, \hat{\Sigma}_1^{(t+1)}, \ldots, \hat{\Sigma}_k^{(t+1)}\right)$, the probability that observation i belongs to the jth category. This process can be iterated until a local optimum is reached. This iterative process has been seen to rapidly converge in practice.

3.4 Numerical experiments

3.4.1 Simulated and phoneme data

To test each of the above three applications, experiments were first run on simulated data. These data sets were generated as zero mean observations from Gaussian or t-distributed processes with randomly selected covariance operators. These were selected by choosing a specific decay rate for the eigenvalues in a diagonal operator D, by generating a random orthonormal basis U, and then combining them as $\Sigma = UDU^{\mathrm{T}}$. The random orthonormal basis was generated by first randomly generating a matrix A with independent and identically distributed standard normal entries and then by recovering the eigenvectors of the symmetric matrix AA^{T} to construct U.

Secondly, the phoneme data to be tested (Ferraty and Vieu, 2003; Hastie et al., 1995) is a collection of 400 log-periodograms for each of five different phonemes: /ɑ/ as in the vowel of "dark"; /ɔ/ as in the first vowel of "water"; /d/ as in the plosive of "dark"; /i/ as in the vowel of "she"; /ʃ/ as in the fricative of "she". Each curve contains the first 150 frequencies from a 32 ms sound clip sampled at a rate of 16-kHz. A periodogram measures the density of frequencies in a signal often referred to as the spectral density. A plot of ten of each such curves and the associated covariance operators is displayed in Figure 3.3.

3.4.2 k sample comparison

The above confidence set in Equation 3.3.1 comparing k samples can be used to refute the null hypothesis that all covariance operators are equal. A two sample permutation test was performed in Pigoli et al. (2014). Given two samples of functional data, $f_1^{(1)}, \ldots, f_n^{(1)}$ and $f_1^{(2)}, \ldots, f_m^{(2)}$ with associated covariance operators $\Sigma^{(1)}$ and $\Sigma^{(2)}$,

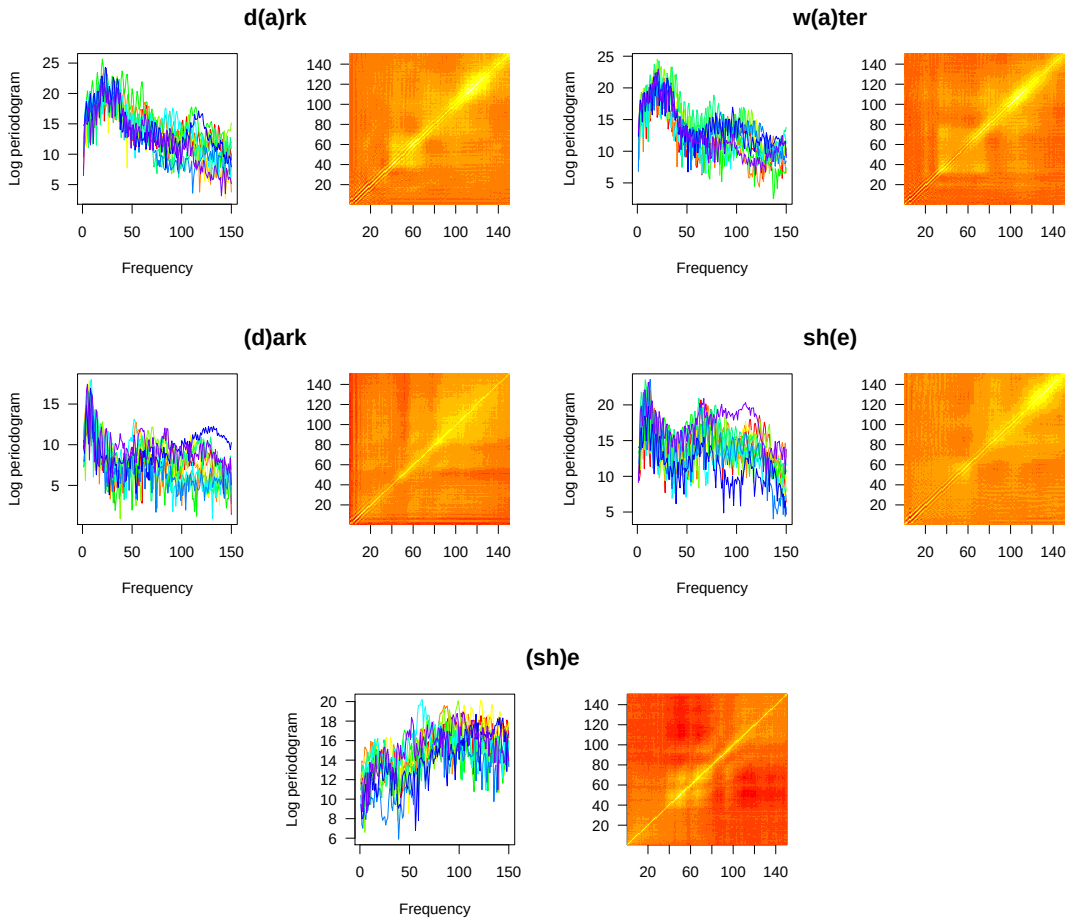


Figure 3.3: Plots of ten log-periodogram curves for each of the five phonemes (left). The sample covariance operators for each of the five phonemes produced from all 400 curves (right). The letters in brackets refer to where the sound is produced by the individual phonemes.

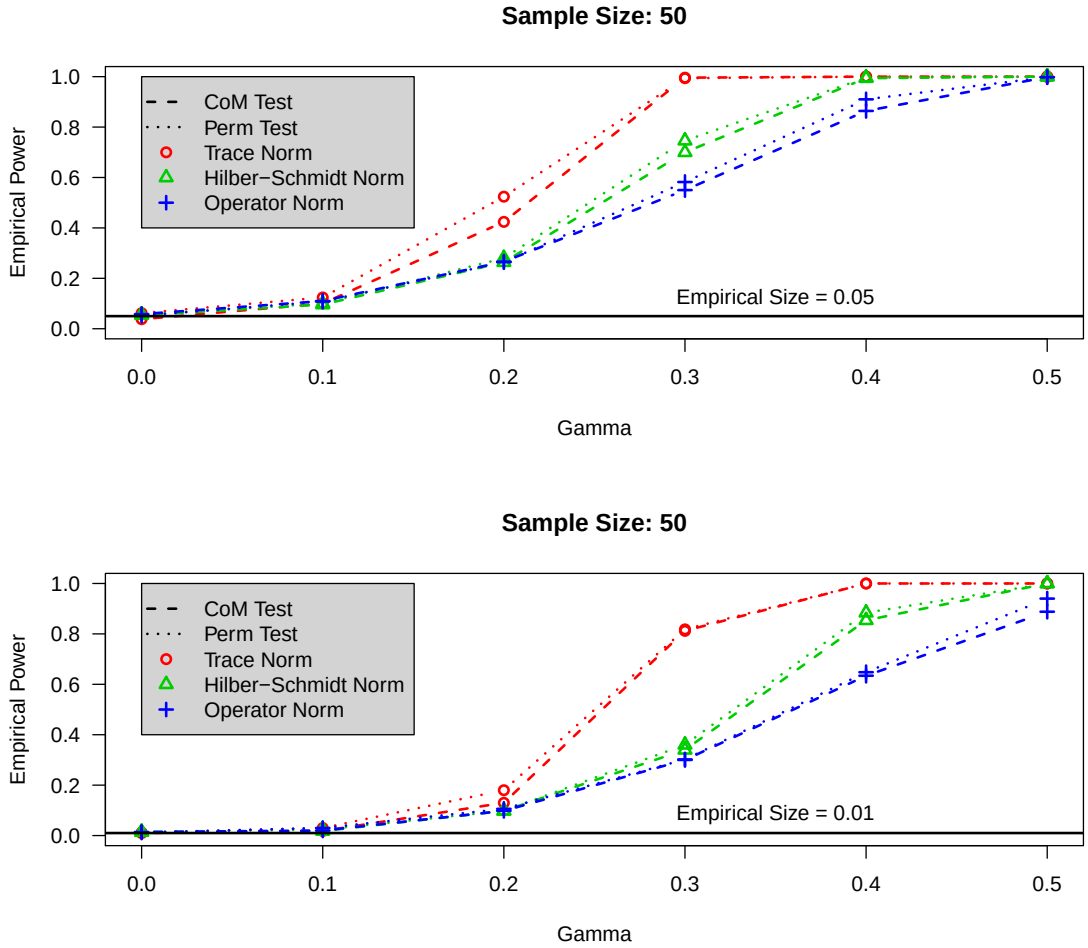


Figure 3.4: A power analysis for testing whether or not operator $\Sigma^{(1)} = \Sigma^{(\gamma)}$ comparing the permutation method (short dashed lines) with the concentration approach (long dashed lines). The size $\alpha = 0.05$ in the top plot, and $\alpha = 0.01$ in the bottom. The eigenvalues of the operators decay at a rate $O(k^{-4})$. The red circle, green triangle, and blue plus lines respectively correspond to the trace class, Hilbert-Schmidt, and operator norms.

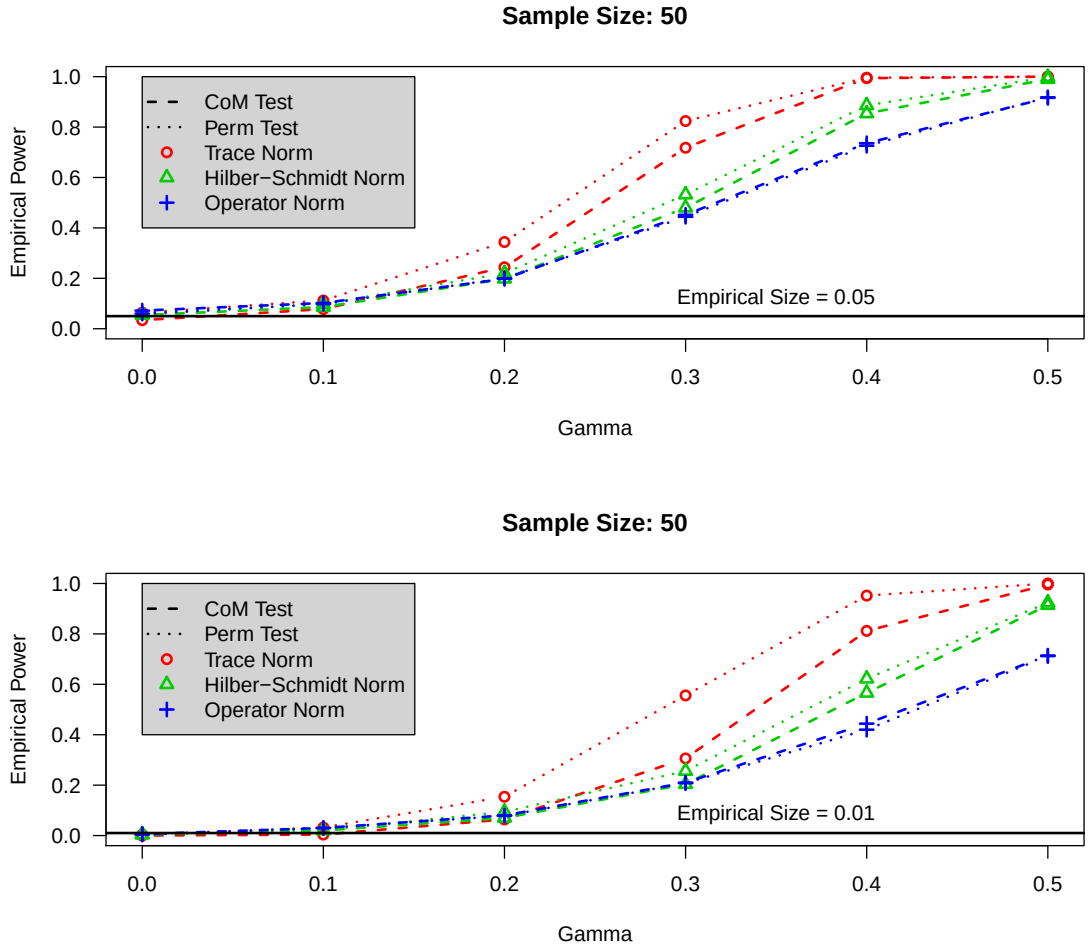


Figure 3.5: A power analysis for testing whether or not operator $\Sigma^{(1)} = \Sigma^{(\gamma)}$ comparing the permutation method (short dashed lines) with the concentration approach (long dashed lines). The size $\alpha = 0.05$ in the top plot, and $\alpha = 0.01$ in the bottom. The eigenvalues of the operators decay at a rate $O(k^{-2})$. The red circle, green triangle, and blue plus lines respectively correspond to the trace class, Hilbert-Schmidt, and operator norms.

respectively, the desired hypotheses to test are

$$\mathrm{H}_0 : \ \Sigma^{(1)} = \Sigma^{(2)} \qquad\qquad \mathrm{H}_1 : \ \Sigma^{(1)} \neq \Sigma^{(2)}.$$

When using a permutation test, the labels are randomly reassigned M times, and each time, the distance between the two new covariance operators is computed. Once aggregated over all possible permutations of the data, this procedure will return the exact significance level of the observations with respect to the data set. In practice, choosing a sufficiently large value for M will suffice.

A power analysis was performed between the permutation method and our proposed concentration approach using Equation 3.3.1. Given two different operators $\Sigma^{(1)}$ and $\Sigma^{(2)}$ and $\gamma \geq 0$, an interpolation between the two operators is constructed as $\Sigma^{(\gamma)} = \Pi\Pi^*$ with

$$\Pi = (\Sigma^{(1)})^{1/2} + \gamma \left(S(\Sigma^{(2)})^{1/2} - (\Sigma^{(1)})^{1/2} \right)$$

where S is an operator minimizing the Procrustes distance from Definition 1.1.9 between $\Sigma^{(1)}$ and $\Sigma^{(2)}$. The Procrustes distance in the operator setting is

$$d_{\mathrm{Proc}}(\Sigma^{(1)}, \Sigma^{(2)})^2 = \inf_{S \in U(L^2(I))} \|R^{(1)} - R^{(2)} S\|_2^2$$

where $\Sigma^{(i)} = (R^{(i)})(R^{(i)})^*$ and $U\left(L^2(I)\right)$ is the space of unitary operators on $L^2(I)$ (Pigoli et al., 2014).

Monte Carlo simulations were run in order to estimate the power of each test. Two operators $\Sigma^{(1)}$ and $\Sigma^{(2)}$ with similar eigenvalue decay were compared with a sample size $n = 50$ and $\gamma \in \{0, .1, .2, .3, .4, .5\}$. For each γ, 5000 samples of size n were generated for $\Sigma^{(1)}$ and $\Sigma^{(\gamma)}$. Equation 3.3.1 and the permutation method (Pigoli et al., 2014) were both implemented to estimate the empirical power.

Figures 3.4 and 3.5 display the results for operators whose eigenvalues decay at a quartic rate, $\lambda_k = O(k^{-4})$, and quadratic rate, $\lambda_k = O(k^{-2})$, respectively. The short dashed lines indicate the power of the permutation test, and the long dashed lines indicate the power of our concentration approach. The colors red, green, and blue and the points circle, triangle, and plus correspond to the three p-Schatten norms for $p = 1, 2, \infty$ being the trace, Hilbert-Schmidt, and operator, respectively. Definitions of and details concerning these norms can be found in Section 1.1.3.

In most cases, the concentration approach is able to achieve the same power to reject the null as does the permutation test. The notable exception is for the trace norm when the eigenvalues decay slowly, which is the lower plot in Figure 3.5. The added benefit to the concentration approach is the speed with which it executes. Across all of the Monte Carlo simulations, our concentration approach ran on average

140.7 times faster than the permutation method based on running the method with 500 permutations. This was computed by tracking the amount of computation time each method spent while producing the plots in Figures 3.4 and 3.5, which corresponds to 6 values of γ, 2 values of α, 3 different norms, and 5000 replications each resulting in 180,000 function calls for both the permutation and concentration methods. Unlike the other norms, the Hilbert-Schmidt norm can be calculated without explicit computation of the eigenvalues and hence results in faster compute times for all statistical methods considered. For each evaluation of the permutation test, 500 permutations of the data were generated, which corresponds to 500 random draws and 500 eigenvalue computations. More accuracy would require even more permutations. In comparison, our concentration approach requires only $3k$ eigenvalue computations and no random draws and hence is only dependent on the number of samples regardless of n, the data size, or α, the test size.

The proposed k-sample test was also used to compare samples of log-periodogram curves from the spoken phonemes /ɑ/ and /ɔ/ . As one can imagine, these vowels can be hard to distinguish both by statistics and by the human ear–see Section 3.4.4 for further evidence of this–as in certain regions of the English speaking world, such as in Canada, the vowels in "dark" and "water" are identical (Bickis, 2016). For $k \in \{2, 3, 4, 5, 6\}$, $k-1$ disjoint sets of 40 /ɑ/ curves and one set of 40 /ɔ/ curves were randomly sampled from the data set. This was replicated 500 times, and each time Equation 3.3.1 was used to decide whether or not the k covariance operators were equivalent at the $\alpha = 0.05$ level. The resulting estimated statistical power for each k is

k	2	3	4	5	6
Power	0.00	0.018	0.228	0.656	0.936

In the null setting, the above experiment was rerun except that every disjoint set of curves came from the /ɑ/ set. The resulting experimentally computed test sizes are

k	2	3	4	5	6
Size	0.00	0.00	0.00	0.004	0.072

3.4.3 Binary and trinary classification

Our concentration of measure (CoM) method is implemented on covariance operators making use of the trace norm $\|\cdot\|_{\text{tr}}$ where for a covariance operator Σ with eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$, $\|\Sigma\|_{\text{tr}} = \sum_{i=1}^{n} |\lambda_i|$. The trace norm was chosen based on the analysis of the preceding section as well as that of Pigoli et al. (2014) where it achieved the best performance when compared with the other p-Schatten norms. The CoM approach to classification of operators is tested in a variety of simulations against other standard

approaches to functional classification. The methods used for comparison are k-nearest neighbours (Ferraty and Vieu, 2006), classification using kernel estimators (Ferraty and Vieu, 2003), general linear model (Müller and Stadtmüller, 2005), and regression trees (Breiman et al., 1984).

The first simulation asks each method to classify observed zero mean Gaussian process data or zero mean t-process data with 4 degrees of freedom. The two covariance operators in question, Σ_1 and Σ_2, are the sample covariances of the male and of the females of the Berkeley growth curve data (Ramsay and Silverman, 2005). In particular, n collections of k curves were generated from each of Σ_1 and Σ_2 as a training set, and m collections of k curves were generated as a test set. The CoM method was trained on the set of n sample covariances and used to classify each of the m test covariances. The remaining classification methods were trained and tested in two separate ways: By treating each sample covariance as a function and classifying as usual, and by training on all $n \times k$ observations and testing each of the m collections by classifying each constituent curve individually and taking a majority vote with ties settled by a uniform random draw.

For group sizes $k = 1, 2, 4, 8, 16$, one hundred simulations were run with $n = 100$ sets of k training curves. To compare the accuracy of each approach $m = 100$ sets of k testing curves were generated for each operator. The accuracy of each method is tabulated in Table 3.1.

The concentration method performed well against the alternatives. Its performance was on par with the kernel method applied to each covariance operator as a function. Our method was only consistently outperformed by the kernel method implementing the majority vote approach, but still displays competitive performance when taking the standard deviations listed in brackets into consideration. However, the two operators in question have very similar weak variances. The next simulation demonstrates how the concentration method adapts naturally when the variances of each label significantly differ.

Continuing from the previous simulation, a third operator is constructed from Σ_1 and Σ_2 by averaging these two and then scaling up the non-principal eigenvalues by a factor of 5. This, in some intuitive sense, creates a third operator between the first two, but with higher variance. The simulation is carried out precisely as before, but incorporating all three operators. In this setting, our concentration approach demonstrates the best performance in the Gaussian setting and still maintains respectable performance in the t-distributed setting when taking the standard deviations into account. The results are listed in Table 3.2.

These five methods tested on simulated data were also tested against phoneme data. Across 50 iterations, each set of 400 curves was partitioned at random into an 100 curve training set and a 300 curve testing set. The five classifiers were trained

Gaussian

k	1	2	4	8	16
CoM	62 (5)	62 (5)	76 (8)	87 (6)	96 (3)
KNN	52 (4)	44 (4)	57 (4)	76 (4)	91 (2)
KNN′	·	47 (3)	59 (4)	74 (3)	89 (2)
Kernel	**65** (4)	64 (5)	75 (3)	87 (3)	96 (2)
Kernel′	·	**70** (4)	**82** (2)	**92** (2)	**99** (1)
GLM	51 (4)	62 (4)	74 (4)	86 (2)	94 (2)
GLM′	·	50 (4)	50 (3)	50 (4)	50 (4)
Tree	57 (4)	54 (4)	59 (3)	66 (4)	75 (3)
Tree′	·	55 (4)	59 (4)	60 (5)	60 (9)

t-distributed

k	1	2	4	8	16
CoM	59 (5)	62 (5)	75 (7)	86 (6)	95 (4)
KNN	42 (4)	45 (4)	58 (4)	76 (3)	92 (2)
KNN′	·	45 (4)	58 (4)	72 (3)	89 (3)
Kernel	**65** (4)	64 (5)	75 (4)	87 (2)	96 (2)
Kernel′	·	**68** (4)	**80** (3)	**92** (2)	**99** (1)
GLM	50 (3)	62 (3)	74 (4)	86 (3)	94 (2)
GLM′	·	50 (3)	50 (3)	51 (4)	50 (3)
Tree	54 (4)	54 (4)	59 (3)	67 (3)	75 (3)
Tree′	·	54 (4)	57 (4)	59 (5)	57 (6)

Table 3.1: A comparison of the performances of five classification methods including our concentration of measure approach (CoM), k-nearest-neighbours (KNN), kernel method (Kernel), generalized linear model (GLM), and regression trees (Tree), on a binary classification problem. The first entry for each method corresponds to classifying the covariance operators as functions. The prime entry corresponds to classifying curves with a majority vote. The estimated percent of correct classification is listed in the table with the sample standard deviation in brackets. The top block comes from Gaussian process data, and the bottom comes from t-process data with 4 degrees of freedom. The highest percentage of each column is marked in bold.

Gaussian

k	1	2	4	8	16
CoM	51 (4)	55 (4)	**75** (5)	**89** (5)	**97** (3)
KNN	50 (3)	52 (3)	61 (3)	75 (3)	90 (2)
KNN′	·	55 (3)	68 (3)	80 (2)	90 (2)
Kernel	**54** (3)	52 (3)	64 (3)	77 (3)	92 (2)
Kernel′	·	**58** (3)	69 (2)	81 (3)	92 (2)
GLM	36 (4)	41 (4)	49 (4)	57 (3)	65 (3)
GLM′	·	35 (4)	36 (4)	36 (5)	35 (5)
Tree	44 (3)	44 (3)	45 (3)	50 (3)	55 (3)
Tree′	·	46 (3)	51 (4)	51 (7)	47 (7)

t-distributed

k	1	2	4	8	16
CoM	46 (5)	50 (6)	63 (5)	75 (8)	85 (7)
KNN	46 (3)	49 (3)	57 (3)	67 (3)	78 (2)
KNN′	·	50 (3)	64 (3)	**77** (2)	**87** (2)
Kernel	**50** (3)	48 (3)	57 (3)	68 (3)	80 (2)
Kernel′	·	**53** (3)	**66** (3)	**77** (2)	85 (2)
GLM	35 (3)	41 (4)	46 (4)	53 (3)	58 (4)
GLM′	·	35 (3)	36 (4)	36 (4)	36 (6)
Tree	42 (3)	44 (3)	45 (3)	46 (3)	49 (3)
Tree′	·	43 (3)	47 (4)	48 (6)	46 (8)

Table 3.2: A comparison of the performances of five classification methods including our concentration of measure approach (CoM), k-nearest-neighbours (KNN), kernel method (Kernel), generalized linear model (GLM), and regression trees (Tree), but, differing Table 3.1, with three potential classes from which to choose. The estimated percent of correct classification is listed in the table with the sample standard deviation in brackets. The top block comes from Gaussian process data, and the bottom block comes from t-process data with 4 degrees of freedom. The highest percentage of each column is marked in bold.

	/ɑ/	/ɔ/	/d/	/i/	/ʃ/
CoM	76.9	76.8	96.6	**98.5**	99.4
KNN	72.4	79.1	**98.5**	97.4	**100.**
Kernel	72.0	**80.5**	98.4	97.2	99.9
GLM	**79.0**	72.3	98.2	95.9	99.2
Tree	70.8	69.4	95.6	87.8	92.6

Table 3.3: Percentage of correct classification of the five phonemes against the five methods: our concentration of measure approach (CoM); k-nearest-neighbours (KNN); kernel method (Kernel); generalized linear model (GLM); and regression trees (Tree). The highest percentage of each column is marked in bold.

and run on each of the 300×5 curves individually. For our concentration of measure approach, the rank one operator associated to each individual curve was compared with the covariance operator formed from the 100×5 training curves. The results are detailed in Table 3.3. Our concentration of measure approach only uniformly outperforms the regression tree classifier, but has comparable performance to the other three methods, and none of the competing methods uniformly outperforms ours.

3.4.4 The expectation-maximization algorithm in practice

The experiments described and depicted below make use of the trace norm only. It was determined through experimentation that the expectation-maximization algorithm we propose in Section 3.3.3 does not perform well under the topology of either the Hilbert-Schmidt or operator norms as they give more emphasis to the principal eigenvalue at the expense of the others. The usual behavior under these norms is for all estimates to converge to the average of all of the data points. This is in contrast to the better performance of the algorithm making use of the trace norm, which is somewhat more uniform in its treatment of the eigenstructure. Hence, we only consider the trace norm in this section.

As a first test case, this algorithm was run given three target covariance operators, which were constructed by taking three randomly generated orthonormal bases U_i for $i = 1, 2, 3$ and a diagonal operator D of eigenvalues decaying at a rate $\lambda_k = O(k^{-4})$ and multiplying $\Sigma_i = U_i D {U_i}^{\mathrm{T}}$. Let the three target covariance operators be denoted as Σ_a, Σ_b, and Σ_c. For each of these operators, 500 rank four data points were generated from a zero mean Gaussian process. From the data, the algorithm initializes three estimates $\hat{\Sigma}_1^{(t)}$, $\hat{\Sigma}_2^{(t)}$, and $\hat{\Sigma}_3^{(t)}$, which attempt to locate the three target operators as the method iterates. After 15 iterations, the original 1500 data points were perfectly separated into three groups. To make the problem harder, a second test case was run identical to the first except that the observed operators are all of rank one. Here the algorithm had a harder time separating the data. The inaccuracy in the rank one

	Concentration					
	Rank 4 Operators			Rank 1 Operators		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Label a	500	0	0	318	0	182
Label b	0	500	0	0	335	165
Label c	0	0	500	295	0	205
	k-means					
Label a	261	0	239	219	0	281
Label b	0	290	210	0	179	321
Label c	0	0	500	211	0	289

Table 3.4: Clustering of simulated operators. The concentration approach performs better than k-means as it takes better account of the covariance structure present in each cluster.

setting is equivalent to the poor performance of classification of rank one operators detailed in Tables 3.1 and 3.2.

The resulting clusters from both tests as well as a comparison with the k-means method are in Table 3.4. The k-means algorithm was run with 50 iterations and 10 random starts. It still performed much worse than the concentration based method in the rank 4 setting. This is because the concentration approach focuses its clustering heavily on the covariance structure of the data whereas k-means does not. The concentration method arguably did better in the rank 1 case as well specifically in the cluster 2 column which more thoroughly captured the label b data.

For the phoneme data, all 400 sample curves from each of the five phoneme sets were clustered individually as curves. The algorithm was run for 20 iterations and told to partition the data into five clusters. The results are in Table 3.5. Clusters A and B captured almost all of the vowels /ɑ/ and /ɔ/ , which, recalling their definition in Section 3.4.1, are quite similar in sound. Clusters C, D, and E contain the majority of /d/ , /i/ , and /ʃ/ curves, respectively. Very similar results were achieved by the tried and true k-means clustering algorithm running with 50 iterations and 10 random starts. The proposed concentration based expectation-maximization algorithm is hence an effective method for the unsupervised clustering of phonemes, which is shown in the tables to perform as well if not better than the trusted k-means algorithm.

3.A Confidence sets for the mean in Banach spaces

The goal of this section is to construct a non-asymptotic confidence region in the general Banach space setting. This is specialized in Section 3.2 to our case of interest, covariance operators, when the X_i below are replaced with $f_i^{\otimes 2}$. The construction of our confidence set begins with Bousquet's upper deviation version of Talagrand's

	Concentration					k-means				
Cluster	A	B	C	D	E	A	B	C	D	E
/ɑ/	281	119	0	0	0	281	119	0	0	0
/ɔ/	125	273	1	1	0	126	272	1	1	0
/d/	0	0	384	15	1	0	2	386	10	2
/i/	1	0	1	393	5	1	3	2	381	13
/ʃ/	0	0	0	3	397	0	0	0	2	398

Table 3.5: Clustering 2000 phoneme curves into 5 clusters. Similar results achieved by both the concentration and k-means methods.

inequality (Bousquet, 2003). This inequality is typically stated for empirical processes (Giné and Nickl, 2016, Theorem 3.3.9 and 3.3.10), but applies to random variables with values in a separable Banach space $(B, \|\cdot\|_B)$ as well by simple duality arguments (Giné and Nickl, 2016, Example 2.1.6).

Let $X_1, \ldots, X_n \in (B, \|\cdot\|_B)$ be zero mean independent and identically distributed Banach space valued random variables with $\|X_i\|_B \leq U$ for all $i = 1, \ldots, n$ where U is some positive constant. Furthermore, let $\langle \cdot, \cdot \rangle : B \times B^* \to \mathbb{R}$ such that for $X \in B$ and $\phi \in B^*$ then $\langle X, \phi \rangle = \phi(X)$. Define

$$Z = \sup_{\|\phi\|_{B^*} \leq 1} \sum_{i=1}^{n} \langle X_i, \phi \rangle = \left\| \sum_{i=1}^{n} X_i \right\|_B, \qquad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\phi\|_{B^*} \leq 1} \mathrm{E} \langle X_i, \phi \rangle^2,$$

where the supremum is taken over a countably dense subset of the unit ball of B^*. Furthermore, define $v = 2U\mathrm{E}Z + n\sigma^2$. Then, $\mathrm{P}\left(Z > \mathrm{E}Z + r\right) \leq \exp\{-r^2/(2v + 2rU/3)\}$. Rewriting Z as $n\|\bar{X} - \mathrm{E}\bar{X}\|_B$ results in

$$\mathrm{P}\left(\|\bar{X} - \mathrm{E}\bar{X}\|_B > \mathrm{E}\|\bar{X} - \mathrm{E}\bar{X}\|_B + r\right) < \exp\left(\frac{-n^2 r^2}{2v + 2nrU/3}\right)$$

where $\|X_i\|_B < U$ and $v = 2nU\mathrm{E}\|\bar{X} - \mathrm{E}\bar{X}\|_B + n\sigma^2$.

The above tail bound incorporates the unknown $\mathrm{E}\left\|\bar{X} - \mathrm{E}\bar{X}\right\|_B$. Consequently, a symmetrization technique is used: This term is replaced by the norm of the Rademacher average $R_n = n^{-1} \sum_{i=1}^{n} \varepsilon_i (X_i - \bar{X})$ where the ε_i are independent and identically distributed Rademacher random variables also independent of the X_i. This substitution is justified by invoking the symmetrization inequality–see (Ledoux and Talagrand, 1991, Chapter 4), (Giné and Nickl, 2016, Theorem 3.1.21), and Chapter 4 of this manuscript and the references therein. The standard symmetrization inequality is

$$\mathrm{E}Z = \mathrm{E}\left\| \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathrm{E}\bar{X}) \right\|_B \leq 2\mathrm{E}\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (X_i - \bar{X}) \right\|_B = 2\mathrm{E}\|R_n\|_B.$$

If the data are symmetric about their mean, which is when $X_i - \mathrm{E}X_i$ and $\mathrm{E}X_i - X_i$ are equidistributed, the coefficient of 2 is unnecessary and can be dropped. This is

	Trace		**Hilbert-Schmidt**		**Operator**	
	$\mathrm{E}Z$	$\mathrm{E}\|R_n\|$	$\mathrm{E}Z$	$\mathrm{E}\|R_n\|$	$\mathrm{E}Z$	$\mathrm{E}\|R_n\|$
/ɑ/	618.3	554.8	112.4	119.5	76.5	81.1
/ɔ/	591.3	525.2	108.7	112.2	70.9	74.2
/d/	506.8	450.7	105.6	115.0	83.3	92.1
/i/	610.4	545.9	107.6	111.2	63.5	72.3
/ʃ/	419.3	363.1	67.4	71.1	40.1	43.0

Table 3.6: A comparison of the left and right hand sides of the symmetrization inequality and, hence, a justification for safely dropping the coefficient of 2 in the construction of confidence sets. These numbers were computed for a sample size of $n = 60$ from the phoneme data set. The computation was repeated 100 times and averaged to approximate the following expectations.

because $X_i - \mathrm{E}X_i$ and $\varepsilon(X_i - \mathrm{E}X_i)$ are also equidistributed. In practice, the data may not be symmetric. However, averaging even a moderately sized data set has a smoothing and symmetrizing effect on the sample mean. Assuming the data is not highly skewed, the coefficient of 2 can be safely dropped in practice to tighten the confidence set. In fact, considering the phoneme data from Section 3.4.1 in this setting results in the values displayed in Table 3.6, which shows that in the trace norm setting, the Rademacher average is much greater than half the size of $\mathrm{E}Z$, and that in the Hilbert-Schmidt and operator norm settings, the Rademacher average is actually marginally less than $\mathrm{E}Z$. A deeper look at symmetrization can be found in Chapter 4.

This symmetrization result allows us to replace the original expectation with the expectation of the Rademacher average. Furthermore, Talagrand's inequality also applies to R_n. This fact is highlighted in Lounici and Nickl (2011) and used to construct global risk bounds for wavelet convolution estimators. Hence, the Rademacher average concentrates strongly about its expectation, which justifies dropping the expectation. In practice, one can use the intermediary $\mathrm{E}_\varepsilon\|R_n\|_B = \mathrm{E}\left(\|R_n\|_B | X_1, \ldots, X_n\right)$, which can be approximated for reasonable sized data sets via Monte Carlo simulations of the ε_i. However, this is not strictly necessary, and for large data sets, a single random draw of ε_i will suffice–see Lounici and Nickl (2011) and (Giné and Nickl, 2016, Section 3.4.2).

The resulting concentration inequality based $(1-\alpha)$-confidence set is

$$\mathcal{C}_{n,1-\alpha} = \Big\{X : \|X - \bar{X}\|_B \leq \|R_n\|_B + \\ + \sqrt{\frac{2}{n}\log(2\alpha)\left(\sigma^2 + 2U\|R_n\|_B\right)} + \frac{U\log(2\alpha)}{3n}\Big\}. \quad (3.A.1)$$

To make use of these results in practice, the weak variance σ^2 must be estimated for the data and furthermore a reasonable choice of U must be made. A main contribution

of this present chapter is to propose such theoretically motivated but practically useful non-asymptotic choices for these constants that work for the functional data applications we are investigating.

3.B Calculation of the weak variance

3.B.1 The weak variance for $p \in [1, \infty)$

To calculate the weak variance σ^2, define $f^{\otimes n} = f \otimes \ldots \otimes f$ to be the n-fold tensor product of f with itself and extend the definition of the bilinear form $\langle \cdot, \cdot \rangle : (L^2)^{\otimes 4} \times \{(L^2)^{\otimes 4}\}^* \to \mathbb{R}$ such that $\langle f^{\otimes 4}, \phi^{\otimes 4} \rangle = \langle f^{\otimes 2}, \phi^{\otimes 2} \rangle^2 = \langle f, \phi \rangle^4 = \phi(f)^4$. For operators $\Pi \in \{(L^2)^{\otimes 2}\}^*$ and $\Xi \in \{(L^2)^{\otimes 4}\}^*$, the weak variance is

$$\begin{aligned}\sigma^2 &= \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\Pi\|_q \leq 1} \mathrm{E}\left(\left\langle f_i^{\otimes 2} - \mathrm{E} f_i^{\otimes 2}, \Pi \right\rangle^2 \right) \\ &\leq \frac{1}{n} \sum_{i=1}^{n} \sup_{\|\Xi\|_q \leq 1} \left\langle \mathrm{E} f_i^{\otimes 4} - \left\{ \mathrm{E} f^{\otimes 2} \right\}^{\otimes 2}, \Xi \right\rangle \leq \left\| \mathrm{E} f^{\otimes 4} - \Sigma^{\otimes 2} \right\|_p\end{aligned}$$

where the inequality stems from the fact that the supremum is being taken over a larger set. However, in the Hilbert space setting, the dual of the tensor product space does coincide with the tensor product of the dual space, and thus the above inequality can be replaced with an equality if the Hilbert-Schmidt norm or 2-Schatten norm is used. More information on tensor products of Banach and Hilbert spaces can be found below in Section 3.C. Given a bound $\|f_i\|_{L^2}^2 \leq c^2 = U$, then $\sigma^2 \leq \left\| \mathrm{E} f^{\otimes 4} - \Sigma^{\otimes 2} \right\|_p \leq \left\| \mathrm{E} f^{\otimes 4} \right\|_p \leq \mathrm{E} \|f\|_{L^2}^4 \leq c^4 = U^2$. Thus, the optimistic choice of $U = \sigma^2$ is used in practice.

3.B.2 The weak variance for $p = \infty$

Let E be a countable dense subset of the unit ball of $L^2(I)$. In the case $p = \infty$, we cannot use duality, but can still write Z and σ^2 as suprema over the countable set

and achieve the same results as above.

$$
\begin{aligned}
Z &= \frac{1}{n}\sup_{e\in E}\sum_{i=1}^{n}\left\langle\left\{f_i^{\otimes 2}-\mathrm{E}f_i^{\otimes 2}\right\}e,e\right\rangle = \sup_{e\in E}\left\langle(\hat{\Sigma}-\Sigma)e,e\right\rangle = \|\hat{\Sigma}-\Sigma\|_\infty,\\
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}\sup_{e_1\in E}\mathrm{E}\left(\left\langle(f_i^{\otimes 2}-\Sigma)e_1,e_1\right\rangle^2\right)\\
&\le \frac{1}{n}\sum_{i=1}^{n}\sup_{e_1,e_2\in E}\mathrm{E}\left(\left\langle f_i^{\otimes 2}-\Sigma,e_1\otimes e_2\right\rangle^2\right)\\
&\le \frac{1}{n}\sum_{i=1}^{n}\sup_{e_1,e_2\in E}\left\langle\left\{\mathrm{E}f_i^{\otimes 4}-\Sigma^{\otimes 2}\right\}(e_1\otimes e_2),e_1\otimes e_2\right\rangle = \|\mathrm{E}f_i^{\otimes 4}-\Sigma^{\otimes 2}\|_\infty.
\end{aligned}
$$

As before, if $\|f_i^{\otimes 2}\|_\infty = \|f_i\|_{L^2}^2 \le c^2 = U$, then $\sigma^2 \le U^2$. Thus, the optimistic choice of $U = \sigma^2$ is again used in practice.

3.B.3 The weak variance for Gaussian data

Similarly to the bounded case, we estimate $\|\mathrm{E}f^{\otimes 4}-\Sigma^{\otimes 2}\|_p$ for Gaussian data. Consider f from a Gaussian process with zero mean and covariance Σ and define $f_s = f(s)$ for notational convenience. Strictly speaking these variables are not norm bounded, but similar to concentration results for Gaussian random variables in $\mathbb{R}^d$ (Giné and Nickl, 2016, Theorem 3.1.9), we found below that our methods still work well. The integral kernel can be written as (Isserlis, 1918)

$$
\begin{aligned}
\mathrm{E}\left(f_sf_tf_uf_v\right) &= \mathrm{E}\left(f_sf_t\right)\mathrm{E}\left(f_uf_v\right)+\mathrm{E}\left(f_sf_u\right)\mathrm{E}\left(f_tf_v\right)+\mathrm{E}\left(f_sf_v\right)\mathrm{E}\left(f_tf_u\right)\\
&= c_f(s,t)c_f(u,v)+c_f(s,u)c_f(t,v)+c_f(s,v)c_f(t,u).
\end{aligned}
$$

Hence, we have that $\mathrm{E}\left(f_sf_tf_uf_v\right)-\Sigma_{s,t}\Sigma_{u,v} = \Sigma_{s,u}\Sigma_{t,v}+\Sigma_{s,v}\Sigma_{t,u}$, and we also have that the operator $\mathrm{E}f^{\otimes 4}-\Sigma^{\otimes 2}$, which can be thought of as an Hilbert-Schmidt operator on the space $Op(L^2)$, can be represented by the integral kernel $c_f(s,u)c_f(t,v)+c_f(s,v)c_f(t,u)$. These two terms are merely relabeled versions of $\Sigma^{\otimes 2}$. Consequently, using the subadditivity of the norm, $\|\mathrm{E}f^{\otimes 4}-\Sigma^{\otimes 2}\|_p \le \|\Sigma^{\otimes 2}\|_p+\|\Sigma^{\otimes 2}\|_p = 2\|\Sigma^{\otimes 2}\|_p$. For example, for the Hilbert-Schmidt norm,

$$
\begin{aligned}
\|\mathrm{E}f^{\otimes 4}-\Sigma^{\otimes 2}\|_{HS}^2 &= \iiiint\left\{c_f(s,u)c_f(t,v)+c_f(s,v)c_f(t,u)\right\}^2 dsdtdudv\\
&= 2\|\Sigma\|_{HS}^4+2\iiiint c_f(s,u)c_f(s,v)c_f(t,v)c_f(t,u)dsdtdudv \le 4\|\Sigma\|_{HS}^4.
\end{aligned}
$$

Lemma 5.1 of Horváth and Kokoszka (2012) gives an explicit form of a covariance operator of Σ in terms of the eigenfunctions of Σ for Gaussian data in the Hilbert-Schmidt setting.

Given $\{\lambda_i\}_{i=1}^{\infty}$, the eigenvalues of Σ, the spectrum of $\Sigma^{\otimes 2}$ is $\{\lambda_i\lambda_j\}_{i,j=1}^{\infty}$. Hence, for any of the p-Schatten norms, $\|\Sigma \otimes \Sigma\|_p = \|\Sigma\|_p^2$. Note that in the above calculations, the weak variance depends on the unknown Σ. In practice, this can be replaced by the empirical estimate $\hat{\Sigma}$, which is a consistent estimator.

3.B.4 The weak variance for t-distributed data

In the case of t-distributed functional data with $\nu > 4$ degrees of freedom–required for the existence of finite fourth moments–and covariance Σ, the required fourth mixed moments can be rewritten as

$$\mathrm{E}\left(f_s f_t f_u f_v\right) = \mathrm{E}\left(Z_s Z_t Z_u Z_v / \sqrt{(V/\nu)^4}\right)$$

for Gaussian process $Z_s = Z(s)$ with covariance $\Pi = \left(\frac{\nu-2}{\nu}\right)\Sigma$ and independent univariate chi-squared random variable $V \sim \chi^2(\nu)$. Now,

$$\begin{aligned}\nu^2 \mathrm{E}V^{-2} &= \nu^2 \int_0^{\infty} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1-2}\mathrm{e}^{-x/2}dx \\ &= \frac{\nu^2}{(\nu-2)(\nu-4)} \int_0^{\infty} \frac{1}{2^{\frac{\nu}{2}-2}\Gamma(\frac{\nu}{2}-2)} x^{\frac{\nu}{2}-3}\mathrm{e}^{-x/2}dx\end{aligned}$$

Hence, using the results for Gaussian data from the previous section,

$$\mathrm{E}f^{\otimes 4} = \left(\frac{\nu-2}{\nu-4}\right)\left(\mathrm{E}\left(f_s f_t\right)\mathrm{E}\left(f_u f_v\right) + \mathrm{E}\left(f_s f_u\right)\mathrm{E}\left(f_t f_v\right) + \mathrm{E}\left(f_s f_v\right)\mathrm{E}\left(f_t f_u\right)\right)$$

The resulting weak variance given $\nu > 4$ is

$$\|\mathrm{E}f^{\otimes 4} - \Sigma^{\otimes 2}\| \leq \left(\frac{2\nu}{\nu-4}\right)\|\Sigma^{\otimes 2}\|,$$

However, this weak variance is quite large for small values of ν, which leads to unnecessarily large confidence sets. Using the weak variance from the Gaussian case leads to good performance in practice as the above t-distribution weak variance is proportional to the Gaussian $2\|\Sigma^{\otimes}\|$.

3.C Tensor products of Banach spaces

For the weak variance calculations of Section 3.B, we must extend the definition of the norm and bilinear form for a given Banach space B to the n-fold tensor product space $B^{\otimes n} = B \otimes \ldots \otimes B$. For a detailed introduction to tensor products of Banach spaces, see Ryan (2013). We begin with the more straight forward Hilbert space setting as functional data is generally considered to be L^2.

Let H be a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$ and let $Op(H)$ be the space of bounded linear operators mapping H into H. Furthermore, let the tensor product $\otimes : H \times H \to Op(H)$ be a bilinear map such that for $f, g, h \in H$, $f \otimes g$ is a rank one operator such that $(f \otimes g)h = \langle h, g \rangle f$. Given an orthonormal basis for H, $\{e_i\}_{i=1}^{\infty}$, consider the space of finite rank operators

$$\Psi = \left\{ \sum_{i,j=1}^{m,n} \lambda_{i,j}(e_i \otimes e_j) \,\middle|\, e_i \in H, \text{ and } \langle e_i, e_j \rangle = \delta_{ij} \right\} \subseteq Op(H)$$

where $\delta_{i,j}$ is the Kronecker delta function. Rewriting $(e_i \otimes e_j)$ as $e_{i,j}$ for notational convenience, the inner product on H can be extended to the finite rank tensor product space Ψ in order to have $\{e_{i,j}\}_{i,j=1}^{\infty}$ be an orthonormal set. Specifically, define

$$\langle e_{i,j}, e_{k,l} \rangle_{\Psi} = \langle e_i, e_k \rangle_H \langle e_j, e_l \rangle_H = \delta_{i,k}\delta_{j,l}.$$

Thus, for $u, v \in \Psi$ and $\lambda_{i,j}, \gamma_{i,j} \in \mathbb{R}$ with $u = \sum_{i,j=1}^{m,n} \lambda_{i,j} e_{i,j}$ and similarly $v = \sum_{i,j=1}^{m,n} \gamma_{i,j} e_{i,j}$, the inner product is

$$\langle u, v \rangle_{\Psi} = \sum_{i,j=1}^{m,n} \lambda_{i,j}\gamma_{i,j}.$$

One can check that the symmetry, linearity, and positive-definiteness of $\langle \cdot, \cdot \rangle_{\Psi}$ follow from those properties for $\langle \cdot, \cdot \rangle_H$ and the bilinearity of $\otimes$. Finally, the associated norm,

$$\|u\|_{\Psi} = \langle u, u \rangle_{\Psi} = \sum_{i,j=1}^{m,n} \lambda_{i,j}^2 = \|u\|_{HS},$$

is thus equivalent to the Hilbert-Schmidt norm. Defining $H \otimes H = \overline{\Psi}$, the closure of Ψ with respect to this norm, we have that the tensor product space that is naturally isometrically isomorphic to the space of Hilbert-Schmidt operators from H to H defined as

$$HS(H) = \left\{ T : H \to H \,\middle|\, \|T\|_{HS} = \sum_{i \in I} \|Te_i\|_H^2 < \infty \right\}.$$

Using the above inner product, we have for $f, g \in H$, $\langle f^{\otimes 2}, g^{\otimes 2} \rangle_{H \otimes H} = \langle f, g \rangle_H^2$, and thus immediately that

$$\left\langle f^{\otimes 2^k}, g^{\otimes 2^k} \right\rangle_{H^{\otimes 2^k}} = \langle f, g \rangle_H^{2^k}$$

where $H^{\otimes 2^k}$ can be thought of as the space of Hilbert-Schmidt operators on $H^{\otimes 2^{k-1}}$, which is $HS\left(H^{\otimes 2^{k-1}}\right)$.

Next, we consider a more general setting. Let B be a separable Banach space with

dual space B^* of linear functionals on B and with bilinear form $\langle \cdot, \cdot \rangle : B \times B^* \to \mathbb{R}$ defined by $\langle f, \phi \rangle = \phi(f)$ for $f \in B$ and $\phi \in B^*$. The tensor product $\otimes : B \times B \to Op(B^*, B)$ is a bilinear map taking $(f, g) \to \langle g, \cdot \rangle f$. There are many different norms that can be extended to the tensor product space from the norms of the constituent Banach spaces. The norm chosen for the tensor product space $B \otimes B$ for the sake of this chapter's proposed methodology is the projective norm defined as

$$\|u\| = \inf \left\{ \sum_{i=1}^{n} \|f_i\| \|g_i\| \,\middle|\, f_i, g_i \in B \text{ for } i = 1, \ldots, n \text{ such that } u = \sum_{i=1}^{n} f_i \otimes g_i \right\},$$

which has the property that for $f, g \in B$, $\|f \otimes g\| = \|f\| \|g\|$ (See Ryan (2013) Proposition 2.1). Taking the completion under this norm gives the projective tensor product space denoted $B \otimes B$. For the dual space,

$$B^* \otimes B^* \subseteq (B \otimes B)^*,$$

and for $\phi \in B^*$, the linear functional $\phi \otimes \phi$ on $B \otimes B$ is the unique linear functional such that

$$\langle \phi \otimes \phi, f \otimes f \rangle = \phi \otimes \phi (f \otimes f) = \phi(f)^2$$

for $f \in B$.

3.C.1 Tensors and covariance matrices in $\mathbb{R}^n$

In this section, we calculate the weak variance from Section 3.B in the Euclidean setting. First, the tensor product, often referred to in this context as the outer product, of two vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ is the $(n \times m)$-matrix

$$u \otimes v = uv^{\mathrm{T}} \in \mathbb{R}^{n \times m}.$$

For $m = n$, the inner product is $\langle u, v \rangle = u^{\mathrm{T}} v \in \mathbb{R}$.

In the matrix setting, let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{k \times l}$ be the matrix representation of linear maps, respectively, from $\mathbb{R}^m$ to $\mathbb{R}^n$ and from $\mathbb{R}^l$ to $\mathbb{R}^k$. Here, the tensor product of linear maps can be represented as the Kronecker product

$$A \otimes B = \begin{pmatrix} a_{1,1}B & a_{1,2}B & \ldots & a_{1,m}B \\ a_{2,1}B & a_{2,2}B & \ldots & a_{2,m}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}B & a_{n,2}B & \ldots & a_{n,m}B \end{pmatrix} \in \mathbb{R}^{ml \times nk}.$$

For $n = k$ and $m = l$, the associated inner product is sum of the entries of the

Hadamard product of A and B,

$$\langle A, B\rangle = \sum_{i,j=1}^{n,m} a_{i,j} b_{i,j}.$$

Hence, the induced norm $\|A\|^2 = \langle A, A\rangle = \sum a_{i,j}^2$ is the Frobenius norm, which is the finite dimensional analogue of the Hilbert-Schmidt norm.

Now, let $X \in \mathbb{R}^n$ be a real valued random vector with zero mean and finite fourth moments. The covariance matrix of $X = (X_1, \ldots, X_n)$ is defined to be

$$\Sigma = \mathrm{cov} X = \mathrm{E} X X^{\mathrm{T}}.$$

The four-fold tensor product is

$$X^{\otimes 4} = X X^{\mathrm{T}} \otimes X X^{\mathrm{T}},$$

which is the $n^2 \times n^2$ matrix with entries of the form $X_i X_j X_k X_l$. For $v \in \mathbb{R}^n$, the weak variance is

$$\begin{aligned}
\sigma^2 &= \sup_{\|vv^{\mathrm{T}}\| \le 1} \mathrm{E} \left\langle X X^{\mathrm{T}} - \Sigma, vv^{\mathrm{T}} \right\rangle^2 \\
&= \sup_{\|vv^{\mathrm{T}}\| \le 1} \mathrm{E} \left(\sum_{i,j=1}^{n} (X_i X_j - \Sigma_{i,j}) v_i v_j \right)^2 \\
&= \sup_{\|vv^{\mathrm{T}}\| \le 1} \mathrm{E} \sum_{i,j,k,l=1}^{n} (X_i X_j - \Sigma_{i,j})(X_k X_l - \Sigma_{k,l}) v_i v_j v_k v_l \\
&= \sup_{\|vv^{\mathrm{T}}\| \le 1} \mathrm{E} \sum_{i,j,k,l=1}^{n} \left(X_i X_j X_k X_l - \Sigma_{i,j}\Sigma_{k,l}\right) v_i v_j v_k v_l \\
&= \sup_{\|vv^{\mathrm{T}} \otimes vv^{\mathrm{T}}\| \le 1} \left\langle \mathrm{E} X^{\otimes 4} - \Sigma^{\otimes 2}, v^{\otimes 4} \right\rangle \\
&= \|\mathrm{E} X^{\otimes 4} - \Sigma^{\otimes 2}\|
\end{aligned}$$

with the penultimate equality due to the fact that for vectors v

$$\|vv^{\mathrm{T}} \otimes vv^{\mathrm{T}}\|^2 = \sum_{i,j,k,l=1}^{n} v_i^2 v_j^2 v_k^2 v_l^2 = \left(\sum_{i,j=1}^{n} v_i^2 v_j^2 \right)^2 = \|vv^{\mathrm{T}}\|^4.$$

Therefore, taking the supremum over v such that $\|vv^{\mathrm{T}}\| \le 1$ is equivalent to taking the supremum for $\|vv^{\mathrm{T}} \otimes vv^{\mathrm{T}}\| \le 1$.

In the case that X is multivariate Gaussian, $\mathrm{E} X_i X_j X_k X_l = \Sigma_{i,j}\Sigma_{k,l} + \Sigma_{i,k}\Sigma_{j,l} +$

$\Sigma_{i,l}\Sigma_{j,k}$, and hence

$$\|\mathrm{E}X^{\otimes 4} - \Sigma^{\otimes 2}\| = \|\{\Sigma_{i,k}\Sigma_{j,l}\} + \{\Sigma_{i,l}\Sigma_{j,k}\}\| \leq 2\|\Sigma^{\otimes 2}\| = 2\|\Sigma\|^2$$

where $\{\Sigma_{i,k}\Sigma_{j,l}\}$ is the tensor product $\Sigma^{\otimes 2}$ with relabeled indices.

3.D Heavy Tails and Noisy Measurements

As often in practice functional data comes from noisy measurements, consider data of the form $Y_i = X_i + \xi_i$ where X_i is a mean zero Gaussian process with covariance operator Σ and ξ_i is Gaussian white noise with covariance $c^2 I$ for some $c^2 > 0$. Figure 3.6 repeats the previous power analysis for the two sample test but in the moderately noisy settings. The results demonstrate that the proposed concentration inequality based methodology is robust when moderate amounts of noise are added to the simulated data.

Secondly, heavier tailed data–specifically t-distributed data with 6 degrees of freedom–can also be handled by this method. Figure 3.7 repeats the earlier two sample power analysis but with the heavier tailed distribution in place of the Gaussian. Once again, the methodology performs well in this setting after being properly tuned to account for the heavier tails of the t-distribution. With respect to such tuning, the coefficient of $(k+2)/(k+3)$ in Equation 3.3.1 was replaced with simply 1 in order to achieve the correct empirical size. In general, given arbitrary data, one can simulate null data and adjust the tuning parameters to match the desired empirical size of the test.

Lastly, the empirical coverage of the concentration based confidence set is still comparable to the desired coverage in the heavy tailed case. Consider t-distributed data with six degrees of freedom; Nine operators were randomly generated and data was simulated from each. Figure 3.8 recreates the simulated confidence sets from Figure 3.2, but with the t-distributed data. To achieve these empirical converges, the Gaussian weak variance, previously calculated to be $\sigma^2 = 2\|\Sigma\|_p^2$, is scaled by a factor of $\nu/(\nu-4)$ where ν is the degrees of freedom. The confidence sets are all slightly larger than desired. However, they are not too large for practical use.

3.E Tuning confidence sets with cross-validation

As noted in the previous sections and immediately visible in Figure 3.8, the concentration based confidence sets for covariance operators are usually too conservative in the sense that $\mathcal{C}_{n,1-\alpha}$, our $(1-\alpha)$ confidence set constructed from a sample of size n, actually has a larger coverage that desired–i.e. $\mathrm{P}\left(\Sigma \in \mathcal{C}_{n,1-\alpha}\right) > 1-\alpha$. This

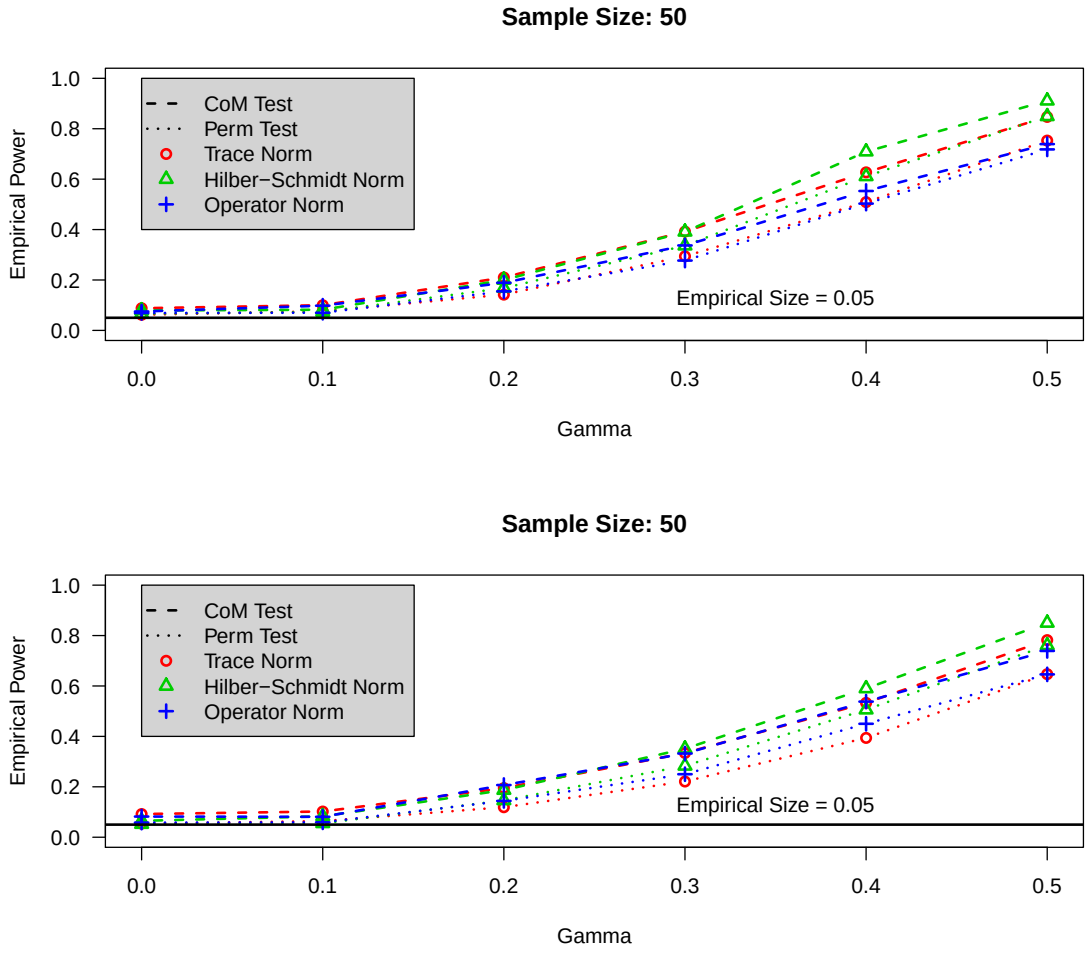


Figure 3.6: Similar to Figures 3.4 and 3.5 except white noise with variance $c^2 = 100$ is added to each functional data observation. In the top plot, the eigenvalues of Σ decay as $O(k^{-4})$ as in Figure 3.4; in the bottom plot, the eigenvalues of Σ decay as $O(k^{-2})$ as in Figure 3.5.

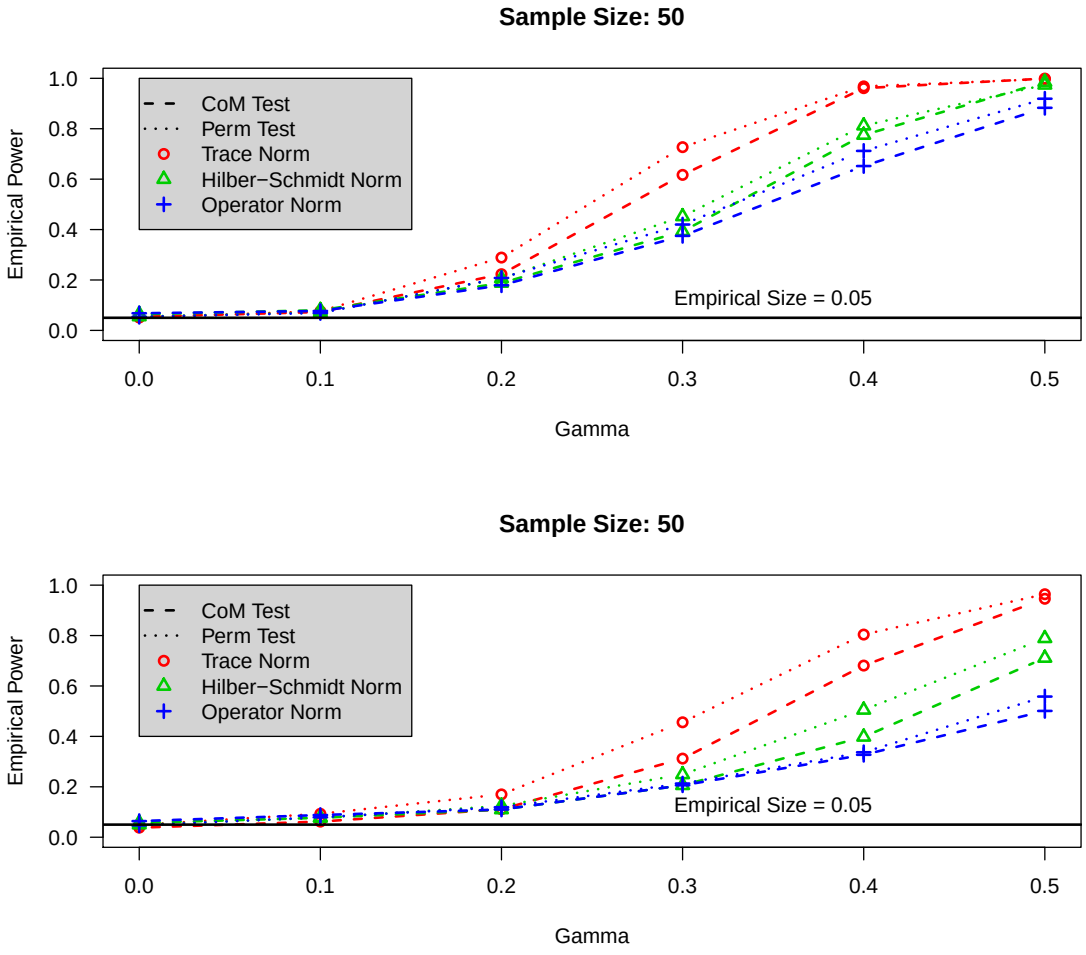


Figure 3.7: A repetition of the experiments from Figures 3.4 and 3.5 but with data simulated from a multivariate t-distribution with 6 degrees of freedom. In the top plot, the eigenvalues of Σ decay as $O(k^{-4})$ as in Figure 3.4; in the bottom plot, the eigenvalues of Σ decay as $O(k^{-2})$ as in Figure 3.5.

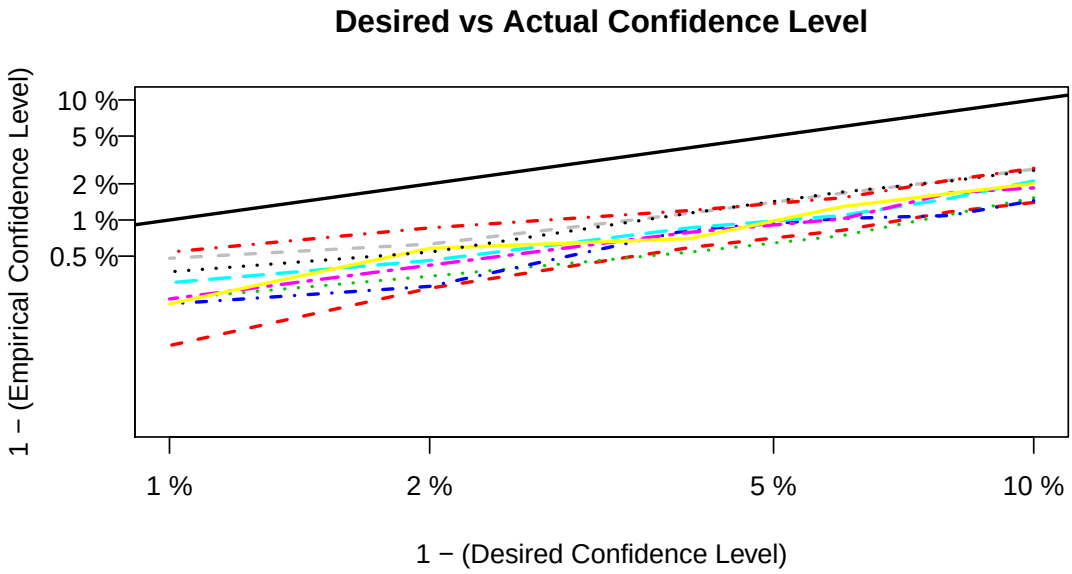


Figure 3.8: The empirical confidence level of the set from Equation 3.2.1 for nine different operators given a sample size of 35 curves generated from a t-distributed process with 6 degrees of freedom. The black line is where the desired and empirical levels are equal. The desired level ranges from $\alpha = 1\%$ to $\alpha = 10\%$. 10,000 replications were used to produce these curves.

fact similarly effects the power of the k sample test for equality of covariance from Section 3.3.1 where coefficients are manually tweaked in order to achieve the desired empirical size of the test. Those tweaks were arrived at through simulations specifically with Gaussian process data. In this appendix, we propose a cross-validation approach to formalize such tweaking by offering a data driven tuning of the radius of the confidence sets.

We begin with the confidence set from Equation 3.2.1 with the addition of a tuning parameter $\kappa > 0$ attached to the deviation term:

$$\mathcal{C}_{n,1-\alpha,\kappa} = \left\{ \Sigma : \|\hat{\Sigma} - \Sigma\|_p \leq \|R_n\|_p + \kappa \left(\sigma \sqrt{\frac{-2\log(2\alpha)}{n}} - \frac{\sigma \log(2\alpha)}{3n} \right) \right\}. \quad (3.E.1)$$

The goal is to scale the deviation term by this tuning parameter κ such that $\mathrm{P}\left(\Sigma \in \mathcal{C}_{n,1-\alpha,\kappa}\right) \approx 1 - \alpha$. In general, κ will depend not only on the distribution of the underlying functional data, but also on the desired coverage of the confidence set of $1 - \alpha$.

To determine a suitable $\kappa > 0$ from the data, we propose the following cross-validation procedure in line with the classic bootstrap technique (Efron, 1979). Let the sample of n functional data observations be $\{f_1, \ldots, f_n\}$. First, the Rademacher

average $\|R_n\|_p$ and the deviation term $D_n = \sigma(-2\log(2\alpha)/n)^{1/2} - \sigma\log(2\alpha)/(3n)$ are computed as usual from the entire data set. Let M be the total number of times to repeat the simulation. For each $m = 1, \ldots, M$, draw two random samples of points from the empirical measure $\mu_n = n^{-1}\sum_{i=1}^{n}\delta_{f_i}$ where δ_{f_i} is the Dirac measure and construct the two covariance estimates $\Sigma_1^{(m)}$ and $\Sigma_2^{(m)}$ from these two samples. Then, the following term is computed for $m = 1, \ldots, M$,

$$Z_m = \|\hat{\Sigma}_1^{(m)} - \hat{\Sigma}_2^{(m)}\|_p.$$

Given the collection of $\{Z_m\}_{m=1}^{M}$, the $1-\alpha$ quantile of the set can be found. This is the number such that

$$Z_{1-\alpha} = \min\left\{z > 0 : |\{m : Z_m < z\}|/M \geq 1-\alpha\right\}.$$

From here, the tuning parameter κ is computed to be $\kappa = (Z_{1-\alpha} - \|R_n\|_p)/D_n$ to scale the deviation term in the original confidence set of Equation 3.E.1.

This cross-validation procedure was tested via simulations of both Gaussian and t-distributed processes. The results of which are displayed in Figure 3.9. The simulations were run in similar style to those that produced Figures 3.2 and 3.8. Over 10,000 replications and five different covariance operators from the five different sets of phoneme data, 60 curves were randomly generated from either a Gaussian process or a t-distributed process. This was performed for desired confidence levels of $\alpha \in [0.01, 0.1]$.

The main effect that the cross-validation procedure has is to tighten the confidence sets, which, when constructed straight from Equation 3.2.1, are generally too large. For the Gaussian simulated data, the confidence sets constructed for two of the five test operators were over tightened, while the other three were close to the desired size. For the t-distributed data, the proposed cross-validation procedure uniformly over-tightens all of the confidence sets to the point that the coverage is too small. This is in contrast to the results of the previous section detailed in Figure 3.8 where it is shown that the pre-tuned confidence sets for t-distributed data are all slightly too large. Thus, with a little more thought, a better tuning method can be developed. Ultimately, these preliminary simulations show that cross-validation is an potentially effective way to implement a data driven tuning procedure to optimize the coverage of the concentration based confidence sets.

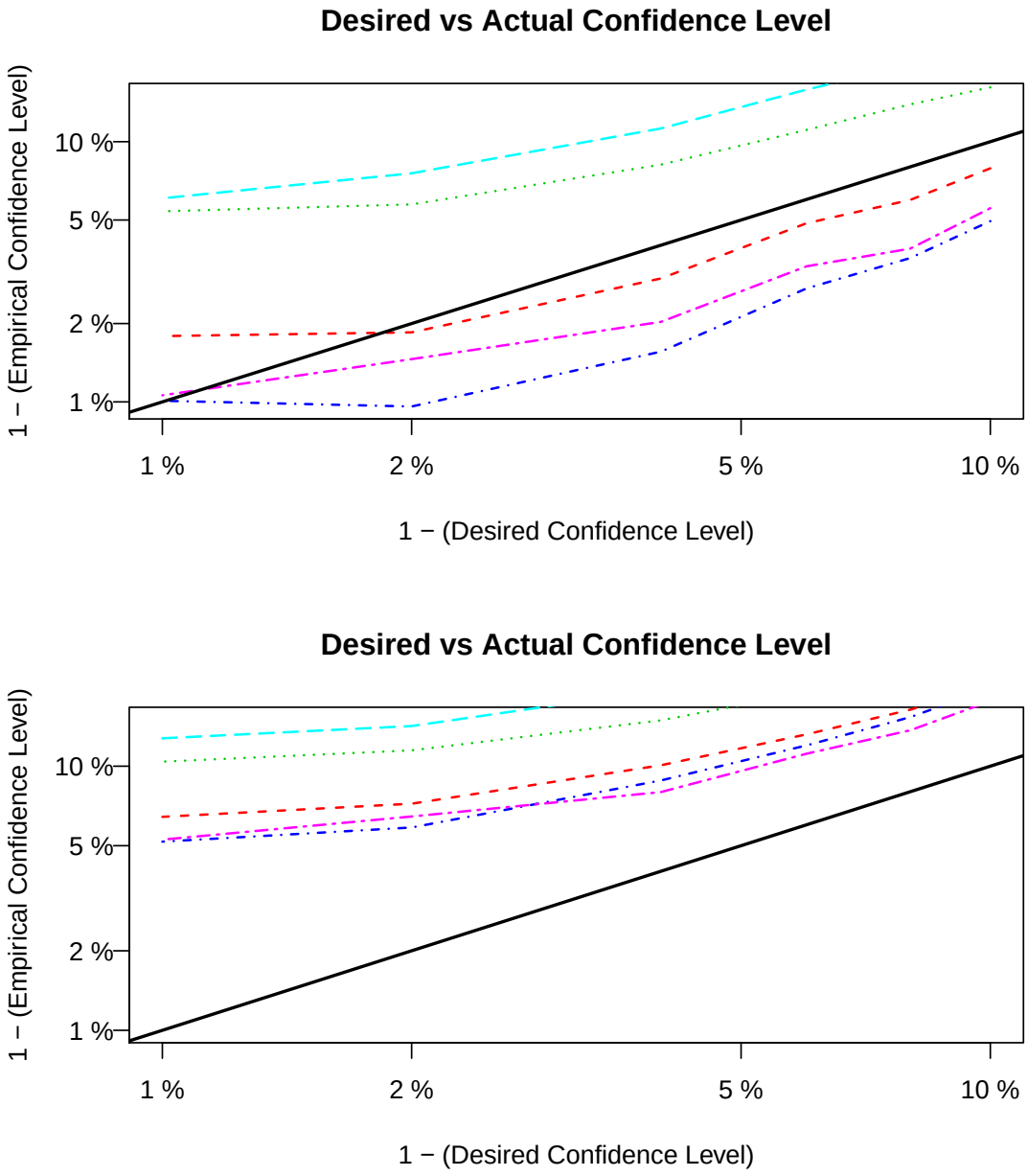


Figure 3.9: The empirical confidence level of the set from Equation 3.2.1 making use of the cross-validation parameter tuning method for five different operators given a sample size of 60 curves. The black line is where the desired and empirical levels are equal. The desired level ranges from $\alpha = 1\%$ to $\alpha = 10\%$. 10,000 replications were used to produce these curves. The top plot represents Gaussian data and the bottom represents t-distributed data with 6 degrees of freedom.

Chapter 4

Improved Rademacher symmetrization

4.1 Introduction

The symmetrization inequality is a ubiquitous result in the probability in Banach spaces literature and in the concentration of measure literature. Dating back at least to Paul Lévy, it is found in the classic text of Ledoux and Talagrand (1991), Section 6.1, and the more recent Boucheron et al. (2013), Section 11.3. Giné and Zinn (1984) use symmetrization in the context of empirical process theory, which is followed by a collection of more recent appearances (Panchenko, 2003; Koltchinskii, 2006; Giné and Nickl, 2010a; Arlot et al., 2010; Lounici and Nickl, 2011; Kerkyacharian et al., 2012; Fan, 2011).

Recalling that ε, a Rademacher random variable or sometimes referred to as a symmetric Bernoulli random variable or a random sign, is such that $\mathrm{P}(\varepsilon = 1) = \mathrm{P}(\varepsilon = -1) = 1/2$, then the symmetrization inequality is as follows.

Proposition 4.1.1. *Let $(B, \|\cdot\|)$ be a Banach space, and let $X_1, \ldots, X_n \in B$ be independent random variables with measure μ. Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent and identically distributed Rademacher random variables, then*

$$\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}X_i)\right\| \leq 2\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(X_i - \mathrm{E}X_i)\right\|.$$

This can be readily proved via Jensen's Inequality and the insight that if Z is a symmetric random variable, that is $Z \overset{\mathrm{d}}{=} -Z$, then $Z \overset{\mathrm{d}}{=} \varepsilon Z$.

Proof. Let $X_1', \ldots, X_n'$ be independent copies of $X_1, \ldots, X_n$ such that X_i and X_i' are

equal in distribution for all $i = 1, \ldots, n$. Then,

$$\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}X_i)\right\| \leq \mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(X_i - X_i')\right\| = \\ = \mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(X_i - X_i')\right\| \leq 2\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(X_i - \mathrm{E}X_i)\right\|.$$

The first inequality comes from Jensen's inequality and the convexity of the norm. The equality results from the fact that $X_i - X_i'$ is a symmetric random variable for all $i = 1, \ldots, n$. The second inequality is just the result of the subadditivity of the norm and the fact that $\mathrm{E}X_i = \mathrm{E}X_i'$. □

Remark 4.1.2. *As the main tool of the previous proof is Jensen's inequality, the result can be generalized with the addition of any convex function $F : \mathbb{R}^+ \to \mathbb{R}^+$ to the following:*

$$\mathrm{E}F\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}X_i)\right\|\right) \leq \mathrm{E}F\left(2\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i - \mathrm{E}X_i\right\|\right).$$

The most notable oversight of this result is that it does not incorporate any measure of the symmetry of the data. Specifically, in the extreme case that the X_i are symmetric about their mean, then the coefficient of 2 can be dropped and the inequality becomes an equality. Taking note of this fact, Arlot et al. (2010) state that

> "it can be shown that this factor of 2 is unavoidable in general for a fixed n when the symmetry assumption is not satisfied, although it is unnecessary when n goes to infinity." (Arlot et al., 2010)

They furthermore

> "conjecture that an inequality holds under an assumption less restrictive than symmetry (e.g., concerning an appropriate measure of skewness of the distribution)." (Arlot et al., 2010)

Hence, in response to this conjecture, we propose an improved symmetrization inequality making use of Wasserstein distance and Hilbert space geometry in order to account for the symmetry, or lack thereof, of the distribution of the X_i under analysis. The main contribution of this chapter is that for some Hilbert space H and $X_1, \ldots, X_n \in H$ independent and identically distributed random variables with common measure μ, there is for a fixed explicit constant $C(\mu)$ depending only on the symmetry of the underlying measure μ of the X_i, which quantifies the symmetry

of μ, such that

$$\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathrm{E}X_i)\right\| \leq \mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(X_i - \mathrm{E}X_i)\right\| + \frac{C(\mu)}{n^{1/2}}.$$

This result is detailed and proved in Section 4.3.2. Furthermore, an empirical bound, $C_n(\mu)$, on the constant C can be calculated as is done in Section 4.4. Such an empirical bound can be further used as a data driven measure of the symmetry of the given sample. In the case that the distribution of the X_i is symmetric, the true $C(\mu) = 0$ and our data driven estimate $C_n(X) = O(n^{-\delta})$ for some $\delta \in (0, 0.5)$ implying a fast rate of convergence to the desired zero for the additive term above: $n^{-1/2}C_n(\mu) = o(n^{-1/2})$. Applications of this result to testing the symmetry of a data set, constructing nonasymptotic high dimensional confidence sets, bounding the variance of an empirical process, and improving coefficients in probabilistic inequalities in the Banach space setting are given in Section 4.5.

4.2 Empirical estimate of the Rademacher sum

Before discussing the main results detailed and proved in Section 4.3, we take a closer look at Rademacher sums to motivate the research in the following sections. These sums arise in the theoretical setting of proving various bounds and inequalities for random variables in Banach spaces. Examples can be found in the many results in the monographs Ledoux and Talagrand (1991) and Boucheron et al. (2013). Alternatively, these sums are used in the applied setting as an analogue for the unknown expectation $\mathrm{E}\|\sum_{i=1}^{n} X_i - \mathrm{E}X_i\|$, which arises when constructing confidence sets using concentration inequalities for such settings as wavelet estimators (Lounici and Nickl, 2011), kernel density estimators (Fan, 2011), and for the covariance operators from Chapter 3 of this manuscript.

In this section, we will consider the practical issue of computing the norm of the Rademacher sum $R_n = \sum_{i=1}^{n} \varepsilon_i(X_i - \bar{X})$ with sample mean $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ to directly estimate the expected value of the norm of the sum $S_n = \sum_{i=1}^{n} X_i - \mathrm{E}X_i$. The Rademacher sum falls into a category of generalized bootstrap techniques. Mainly,

$$\|R_n\| = \left\|\sum_{i=1}^{n}\varepsilon_i(X_i - \bar{X})\right\| = \left\|\sum_{i\in I}(X_i - \bar{X}) - \sum_{j\notin I}(X_j - \bar{X})\right\|$$

for some random subset $I \subseteq \{1, \ldots, n\}$ with cardinality such that $\mathrm{P}(|I| = k) = \binom{n}{k}2^{-n}$. Thus, given some observed $X_1, \ldots, X_n$, the total expectation $\mathrm{E}\|R_n\|$ can be approximated by the conditional expectation $\mathrm{E}_\varepsilon\|R_n\| = \mathrm{E}(\|R_n\||X_1, \ldots, X_n)$.

This conditional expectation can in turn be approximated by randomly drawing M sets of $\{\varepsilon_1^{(m)}, \ldots, \varepsilon_n^{(m)}\}$, computing for each $m = 1, \ldots, M$ the Rademacher sum $\|R_n^{(m)}\| = \|\sum_{i=1}^n \varepsilon_i^{(m)}(X_i - \bar{X})\|$, and averaging over the M sums to get that $\mathrm{E}_\varepsilon\|R_n\| \approx M^{-1}\sum_{m=1}^M \|R_n^{(m)}\|$. However, before continuing, we consider alternative bootstrap techniques to demonstrate the superiority of the Rademacher sum and why the symmetrization inequality matters.

The term $\mathrm{E}\|S_n\|$ cannot be estimated directly, but instead approached via some bootstrap technique. Beyond the Rademacher sum, two other bootstrap estimators for $\mathrm{E}\|S_n\|$ will be considered. Given a sample of size n, $X_1, \ldots, X_n$, the first method is to randomly split the data in half using the first half to estimate $\mathrm{E}X_i$ and the second half to estimate $\mathrm{E}S_n$, which is equivalent to restricting the Rademacher sum bootstrap to index sets $I \subset \{1, \ldots, n\}$ of cardinality $n/2$. Namely, for such sets, we have

$$\hat{S}_n^{\text{half}} = \binom{n}{n/2}^{-1} \sum_{I:|I|=n/2} \left\| \sum_{i \in I} \left(X_i - \frac{2}{n} \sum_{j \in \{1,\ldots,n\} \setminus I} X_j \right) \right\|,$$

which can, of course, be approximated by selecting a reasonable number M of such sets $I_1, \ldots, I_M$.

The second approach is a leave-one-out estimate similar to the jackknife estimator (Efron and Stein, 1981). Once again, given a sample of size n, $X_1, \ldots, X_n$, this method is equivalent to the Rademacher sum bootstrap but restricting the cardinality of the set to $|I| = n - 1$. This results in

$$\hat{S}_n^{\text{LOO}} = \frac{1}{n} \sum_{i=1}^n \left\| X_i - \frac{1}{n-1} \sum_{j \neq i, j=1}^n X_j \right\|.$$

Each of these bootstrap methods are in some sense comparable to each other with respect to accuracy and variance of the estimate for $\mathrm{E}\|S_n\|$. However, the symmetrization inequality allows for us to explicitly bound $\mathrm{E}\|S_n\|$ by the Rademacher sum. Indeed, using the original symmetrization inequality, it is reasonable to bound

$$\mathrm{E}\left\| \sum_{i=1}^n (X_i - \mathrm{E}X_i) \right\| \leq 2\mathrm{E}\|R_n\| \approx 2\mathrm{E}_\varepsilon\|R_n\| \approx \frac{2}{M} \sum_{m=1}^M \|R_n^{(m)}\|.$$

In contrast, the goal of this chapter is to theoretically derive and explicitly compute a small correction term $C_n(\mu)$ to update this bound to the tighter

$$\mathrm{E}\left\| \sum_{i=1}^n (X_i - \mathrm{E}X_i) \right\| \leq \frac{1}{M} \sum_{m=1}^M \|R_n^{(m)}\| + \frac{C_n(\mu)}{\sqrt{2n}}.$$

This is powerful in the construction of non-asymptotic confidence sets for high

dimensional data where one desires to achieve a minimum coverage, say $1 - \alpha$, for such confidence sets as performed in both Arlot et al. (2010) and Chapter 3. Using one of these alternative bootstrap methods does not guarantee such coverage. However, using the Rademacher sum with either the coefficient of 2 or with our correction term proposed in the subsequent section, will, in fact, result in a confidence set with no less than the desired coverage.

4.3 Symmetrization

4.3.1 Overview of Wasserstein spaces

We first require the standard notions of Wasserstein distance and Wasserstein space as stated below. These are defined on Polish spaces, which are complete separable metric spaces. For a thorough introduction to such topics, see Villani (2008).

Definition 4.3.1 (Wasserstein Distance)**.** *Let $(\mathcal{X}, d)$ be a Polish space and $p \in [1, \infty)$. For two probability measures μ and ν on $\mathcal{X}$, the Wasserstein p distance is*

$$W_p(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{X}} d(x, y)^p d\gamma(x, y) \right)^{1/p}$$

where the infimum is taken over all measures γ on $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν.

An equivalent and useful formulation of Wasserstein distance is

$$W_p(\mu, \nu) = \inf_{(X,Y)} (\mathrm{E}\, d(X, Y)^p)^{1/p}$$

where the infimum is taken over all possible joint distributions of X and Y with marginals μ and ν, respectively.

Definition 4.3.2 (Wasserstein Space)**.** *Let $P(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$. The Wasserstein space is*

$$P_p(\mathcal{X}) := \left\{ \mu \in P(\mathcal{X}) \,\middle|\, \int_{\mathcal{X}} d(x_0, x)^p \mu(dx) < \infty \right\}$$

for any arbitrary choice of x_0. This is the space of measures with finite pth moment.

Convergence in Wasserstein space is characterized by weak convergence of measure and convergence in pth moment. From Theorem 6.8 of Villani (2008), convergence in Wasserstein distance is equivalent to weak convergence in $P_p(\mathcal{X})$. Hence, for a sequence of measures μ_n,

$$W_p(\mu_n, \mu) \to 0 \text{ if and only if } \mu_n \xrightarrow{\mathrm{d}} \mu \text{ and } \int_{\mathcal{X}} x^p d\mu_n(x) \to \int_{\mathcal{X}} x^p d\mu(x).$$

4.3.2 Symmetrization result

In the following lemma, we bound the expectation on the left by the sum of a "symmetric" term and an "asymmetric" term.

Lemma 4.3.3. *Let H be an Hilbert space, and let $X_1, \ldots, X_n \in H$ be independent and identically distributed random variables with common law μ. Define μ^- to be the law of $-X$. Furthermore, let $\varepsilon_1, \ldots, \varepsilon_n$ be independent and identically distributed Rademacher random variables also independent of the X_i. Then, for any 1-Lipschitz function ψ,*

$$\mathrm{E}\psi\left(\sum_{i=1}^{n}(X_i - \mathrm{E}X_i)\right) \leq \mathrm{E}\psi\left(\sum_{i=1}^{n}\varepsilon_i(X_i - \mathrm{E}X_i)\right) + \sqrt{\frac{n}{2}}W_2(\mu, \mu^-)$$

where W_2 is the Wasserstein 2 distance.

Proof. For a Polish space $\mathcal{X}$, let $\Pi(\mu, \nu)$ be the space of all product measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν. For $\delta \in (0, 1)$, let $\Pi_\delta(\mu, \nu)$ be the space of all product measures with marginals μ and $\nu_\delta = \delta\mu + (1-\delta)\nu$. For $\gamma \in \Pi(\mu, \nu)$ and $\eta \in \Pi(\mu, \mu)$, the measure $\delta\eta + (1-\delta)\gamma \in \Pi_\delta(\mu, \nu)$. Hence,

$$\begin{aligned}
W_p^p(\mu, \nu_\delta) &= \inf_{\gamma_\delta \in \Pi(\mu, \nu_\delta)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\gamma_\delta(x,y) \\
&\leq \inf_{\eta \in \Pi(\mu,\mu),\, \gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d(\delta\eta + (1-\delta)\gamma)(x,y) \\
&= \inf_{\gamma \in \Pi(\mu,\nu)} (1-\delta) \int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p d\gamma(x,y) \\
&= (1-\delta) W_p^p(\mu, \nu).
\end{aligned}$$

The inequality on the second lines above arises from taking the infimum over a more restrictive set. The law of εX is $\frac{1}{2}(\mu + \mu^-)$. Hence, for our purposes, the above implies that

$$W_2\left(\mu, \frac{\mu + \mu^-}{2}\right) \leq \frac{1}{\sqrt{2}} W_2(\mu, \mu^-).$$

Define μ^{*n} to be the law of $\sum_{i=1}^n (X_i - \mathrm{E}X_i)$ and $\tilde{\mu}^{*n}$ to be the law of $\sum_{i=1}^n \varepsilon_i(X_i - \mathrm{E}X_i)$. Then,

$$\mathrm{E}\psi\left(\sum_{i=1}^{n}(X_i - \mathrm{E}X_i)\right) - \mathrm{E}\psi\left(\sum_{i=1}^{n}\varepsilon_i(X_i - \mathrm{E}X_i)\right) \leq$$

$$
\begin{aligned}
&\leq \sup_{\|\phi\|_{Lip}\leq 1}\left\{\mathrm{E}\phi\left(\sum_{i=1}^{n}(X_i-\mathrm{E}X_i)\right)-\mathrm{E}\phi\left(\sum_{i=1}^{n}\varepsilon_i(X_i-\mathrm{E}X_i)\right)\right\}\\
&\leq W_1\left(\mu^{*n},\tilde{\mu}^{*n}\right)\\
&\leq W_2\left(\mu^{*n},\tilde{\mu}^{*n}\right)\\
&\leq \sqrt{n}W_2\left(\mu,\frac{\mu+\mu^-}{2}\right)\\
&\leq \sqrt{\frac{n}{2}}W_2(\mu,\mu^-)
\end{aligned}
$$

where the second, third, and fourth inequality come respectively from Lemmas 4.A.1, 4.A.2, and 4.A.3 in the appendix. Rearranging the terms gives the desired result. □

This lemma leads immediately to the following theorem. The intuition behind this theorem is that averaging a collection of random variables has an inherent smoothing and symmetrizing effect following from the central limit theorem. Thus, as the sample size n increases, the difference between the expectations of the true average and the Rademacher average become negligible. Of course, we have following from such theorems that, given a finite second moment for the probability measure μ, that $\left|\mathrm{E}\psi\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mathrm{E}X_i)\right)-\mathrm{E}\psi\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(X_i-\mathrm{E}X_i)\right)\right| = O(n^{-1/2})$. However, in the next theorem, we explicitly quantify this error and use it for finite sample empirical estimation in the following sections. This behaviour was shown in the simulations detailed in Table 3.6 of the previous chapter.

Theorem 4.3.4. *Using the setting of Lemma 4.3.3 with either of the following two conditions that*

1. *ψ is additionally positive homogeneous (e.g. a norm), or*
2. *the metric d is positive homogeneous in the sense that for $a\in\mathbb{R}$, $d(ax,ay)=|a|d(x,y)$,*

then

$$
\left|\mathrm{E}\psi\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mathrm{E}X_i)\right)-\mathrm{E}\psi\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(X_i-\mathrm{E}X_i)\right)\right|\leq\frac{1}{\sqrt{2n}}W_2(\mu,\mu^-)
$$

Proof. Running the proof of Lemma 4.3.3 after swapping $\sum_{i=1}^{n}(X_i-\mathrm{E}X_i)$ and $\sum_{i=1}^{n}\varepsilon_i(X_i-\mathrm{E}X_i)$ gives the lower deviation

$$
\mathrm{E}\psi\left(\sum_{i=1}^{n}(X_i-\mathrm{E}X_i)\right)\geq\mathrm{E}\psi\left(\sum_{i=1}^{n}\varepsilon_i(X_i-\mathrm{E}X_i)\right)-\sqrt{\frac{n}{2}}W_2(\mu,\mu^-).
$$

Under condition 1, the result is immediate.

Under condition 2, let μ be the law of $(X_i - \mathrm{E}X_i)$ as before. Then, redefining μ^{*n} to be the law of $\sum_{i=1}^n \frac{1}{n}(X_i - \mathrm{E}X_i)$ and $\tilde{\mu}^{*n}$ to be the law of $\sum_{i=1}^n \frac{1}{n}\varepsilon_i(X_i - \mathrm{E}X_i)$ results in

$$\begin{aligned} W_2(\mu^{*n}, \tilde{\mu}^{*n}) &\leq \sqrt{n} \inf_{(X,Y)} \left(\mathrm{E}\, d(X/n, Y/n)^2\right)^{1/2} \\ &= \frac{1}{\sqrt{2n}} W_2(\mu, \mu^-) \end{aligned}$$

where the infimum is taken over all joint distributions of X and Y with marginals μ and $\frac{\mu+\mu^-}{2}$, respectively. The desired result follows. □

4.4 Empirical estimate of $W_2(\mu, \mu^-)$

In order to explicitly make use of the above results, an empirical estimate of $W_2(\mu, \mu^-)$ is required. We first establish the following bound.

Proposition 4.4.1. *Let $X_1, \ldots, X_n$ be iid with law μ and let $Y_1, \ldots, Y_n$ be iid with law ν. Furthermore, let μ_n and ν_n be the empirical distributions of μ and ν, respectively. Then,*

$$W_p^p(\mu, \nu) \leq \mathrm{E} W_p^p(\mu_n, \nu_n).$$

Proof. The following infima are taken over all possible joint distributions of the random variables in question given fixed marginal distributions. Let X and Y be random variables of law μ and ν, respectively. Also, let S_n be the group of permutations on n elements.

$$\begin{aligned} W_p^p(\mu, \nu) &= \inf_{(X,Y)} \mathrm{E} d(X, Y)^p \\ &= \inf_{(X_1,\ldots,X_n,Y_1,\ldots,Y_n)} \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^n d(X_i, Y_i)^p\right\} \\ &\leq \mathrm{E} \min_{\rho \in S_n} \left\{\frac{1}{n}\sum_{i=1}^n d(X_i, Y_{\rho(i)})^p\right\} \\ &= \mathrm{E} W_p^p(\mu_n, \nu_n) \end{aligned}$$

where the above inequality arises by replacing the infimum over all possible joint distributions of the X_i and Y_i with a specific joint distribution. □

The following subsections establish that it is reasonable to replace $W_2(\mu, \mu^-)$ with a data driven estimate of $\mathrm{E} W_2(\mu_n, \mu_n^-)$ in Lemma 4.3.3 and Theorem 4.3.4. Rates of convergence of $W_2(\mu_n, \mu_n^-)$ are presented, and a bootstrap estimator for $\mathrm{E} W_2(\mu_n, \mu_n^-)$ is proposed and tested numerically.

4.4.1 Rate of convergence of empirical estimate

As $W_p(\cdot, \cdot)$ is a metric, the triangle inequality and the fact that $W_p(\mu, \mu_n) = W_p(\mu_n^-, \mu^-)$ implies that

$$\begin{aligned} W_p(\mu, \mu^-) &\leq W_p(\mu, \mu_n) + W_p(\mu_n, \mu_n^-) + W_p(\mu_n^-, \mu^-) \\ &\leq 2W_p(\mu, \mu_n) + W_p(\mu_n, \mu_n^-), \end{aligned}$$

and therefore,

$$|W_p(\mu, \mu^-) - W_p(\mu_n, \mu_n^-)| \leq 2W_p(\mu, \mu_n).$$

By Lemma 4.A.4, $W_p(\mu, \mu_n) \to 0$ with probability one making the discrepancy negligible for large data sets. However, it is also possible to get a hard upper bound on this term; specifically, the recent work of Fournier and Guillin (2015) proposes explicit moment bounds on $W_p(\mu, \mu_n)$. Their result can be used to demonstrate the speed with which our empirical measure of asymmetry, $W_2(\mu_n, \mu_n^-)$, converges to zero when μ is symmetric.

In the case that μ is symmetric, $W_2(\mu, \mu^-) = 0$, the ideal correction term is equal to zero. This implies that our empirical bound

$$W_2(\mu_n, \mu_n^-) = \left|W_2(\mu, \mu^-) - W_2(\mu_n, \mu_n^-)\right| \leq 2W_2(\mu, \mu_n).$$

Therefore, the moment bound from Theorem 1 of Fournier and Guillin (2015) implies that $W_2(\mu_n, \mu_n^-) = O(n^{-\delta})$ where $\delta \in (0, 0.5]$ depending on the specific moment used and the dimensionality of the measure. Thus, the empirical bound on the correction term in our improved inequality, $W_2(\mu_n, \mu_n^-)/\sqrt{n}$, achieves a faster convergence rate in the symmetric case than the general rate of $n^{-1/2}$.

The tightness of the bounds proposed in Fournier and Guillin (2015) was tested experimentally. While the moment bounds are certainly of theoretical interest, implementing these bounds resulted in an inequality less sharp than the original symmetrization inequality. However, the bootstrap procedure detailed in the following section does produce a practically useful estimate of the expected empirical Wasserstein distance.

4.4.2 Bootstrap estimator

We propose a bootstrap procedure to estimate the expected Wasserstein distance between the empirical measure and its reflection, $\mathrm{E}W_2(\mu_n, \mu_n^-)$. Given observations $x_1, \ldots, x_n$, let $\hat{\mu}_n$ be the empirical measure of the data. Then, for some specified m, two sets $Y_1, \ldots, Y_m$ and $Z_1, \ldots, Z_m$ can be sampled as independent draws from $\hat{\mu}_n$. The goal is to move a mass of $1/m$ from each of the Y_i to each of the negated $-Z_i$

in an optimal fashion. Hence, the $m \times m$ matrix of pairwise distances is constructed with entries $A_{i,j} = d(Y_i, -Z_j)$, which can be accomplished in $O(m^2)$ time. From here, the problem reduces to a *linear assignment problem*, a specific instantiation of a *Minimum-cost flow problem* from linear programming (Ahuja et al., 1993). That is, given a complete bipartite graph with vertices $L \cup R$ such that $|L| = |R| = m$ and with weighted edges, we wish to construct a perfect matching minimizing the total sum of the edge weights. Here, the weights are the pairwise distances $A_{i,j}$. This linear program can be efficiently solved in $O(m^3)$ time via the *Hungarian algorithm* (Kuhn, 1955). For more on linear programs in the probabilistic setting, see Steele (1997).

This estimated distance can be averaged over multiple bootstrapped samples. Though, in general, only a few replications are necessary to achieve a stable estimate as the bootstrap estimator has a very small variance. Indeed, to see this, consider the bounded difference inequality detailed in Section 3.2 of Boucheron et al. (2013) and in Section 3.3.4 of Giné and Nickl (2016), which is a direct corollary of the Efron-Stein-Steele inequality (Efron and Stein, 1981; Steele, 1986; Rhee and Talagrand, 1986).

Definition 4.4.2 (A function of bounded differences)**.** *For $\mathcal{X}$ some measurable space and a real valued function $f : \mathcal{X}^n \to \mathbb{R}$, f is said to have the bounded differences property if for all $i = 1, \ldots, n$,*

$$\sup_{x_1,\ldots,x_n,x_i'} |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i.$$

Proposition 4.4.3 (Corollary 3.2 of Boucheron et al. (2013))**.** *If f has the bounded differences property with constants $c_1, \ldots, c_n$, then* $\operatorname{Var}(f(X_1, \ldots, X_n)) \leq \frac{1}{4}\sum_{i=1}^n c_i^2$.

In our setting, Y_i and Z_i for $i = 1, \ldots, m$ are independent random variables with law $\hat{\mu}_n$. The function $f(Y_1, \ldots, Y_m, Z_1, \ldots, Z_m)$ is the value of the optimal matching from the $\{Y_i\}$ to the $\{-Z_i\}$. This f is, in fact, a function of bounded differences. This is because modifying a single argument will at most change the optimal value by $c = m^{-1}(\max_{i,j=1,\ldots,n}\{d(x_i, -x_j)\} - \min_{i,j=1,\ldots,n}\{d(x_i, -x_j)\}) = C/m$. Thus, from the bounded differences theorem,

$$\operatorname{Var}(f(Y_1, \ldots, Y_m, Z_1, \ldots, Z_m)) \leq \frac{C^2 n}{4m^2}.$$

Therefore, if m is chosen to be of order n, as in the numerical experiments below, then the variance of the bootstrap estimate decays at rate of $O(n^{-1})$.

The proposed bootstrap procedure was experimentally tested on both high dimensional Rademacher and Gaussian data as will be seen in Section 4.4.3. For each replication, the observed data was randomly split in half. That is, given a

random permutation $\rho \in S_n$, the symmetric group on n elements, the Hungarian algorithm was run to calculate the cost of an optimal perfect matching between $\{X_{\rho(1)}, \ldots, X_{\rho(\frac{n}{2})}\}$ and $\{-X_{\rho(\frac{n}{2}+1)}, \ldots, -X_{\rho(n)}\}$.

4.4.3 Numerical experiments

From Proposition 4.4.1, there is an obvious positive bias in our new symmetrization inequality when using the Wasserstein distance between the empirical measures, $W_2(\mu_n, \mu_n^-)$, in lieu of the Wasserstein distance between the unknown underlying measures, $W_2(\mu, \mu^-)$. This is specifically troublesome when μ is symmetric or nearly symmetric. That is, if $W_2(\mu, \mu^-) = 0$, then barring trivial cases, the distance between the empirical measures will be positive with positive probability. However, as stated in Lemma 4.A.4, $W_2(\mu_n, \mu_n^-) \to 0$ with probability one, which will still make this approach superior to the standard symmetrization inequality. In the following subsections, we will compare the magnitude of the expected symmetrized sum and the asymmetric correction term, which are, respectively,

$$R_n = n^{-1/2}\mathrm{E}\left\|\sum_{i=1}^{n} \varepsilon_i(X_i - \mathrm{E}X_i)\right\| \quad \text{and} \quad C_n = W_2(\mu_n, \mu_n^-)/\sqrt{2}.$$

The goal is to demonstrate through numerical simulations that the latter is smaller than the former and thus that newly proposed $R_n + C_n$ is a sharper upper bound than the original $2R_n$ for $n^{-1/2}\mathrm{E}\|\sum_{i=1}^{n}(X_i - \mathrm{E}X_i)\|$.

Rademacher data

For a dimension k and a sample size $n = \{2, 4, 8, \ldots, 256\}$, the data for this first numerical test was generated from a multivariate symmetric Rademacher distribution. That is, for a size n iid sample from this distribution, $X_1, \ldots, X_n$, let $X_{i,j}$ be the jth entry of the ith random variable with $X_{i,1}, \ldots, X_{i,k}$ iid Rademacher(1/2) random variables. Across 10,000 replications, random samples were drawn and used to estimate the expected Rademacher average, R_n, and the expected empirical Wasserstein distance, C_n, under the ℓ_1-norm. The dimensions considered were $k = \{2, 20, 200\}$. The results are displayed on the left column of Figure 4.1. As the sample size n increases with respect to k, we get closer to an asymptotic state and the bound based on the empirical Wasserstein distance becomes more attractive.

Gaussian data

For a dimension k and a sample size $n = \{2, 4, 8, \ldots, 256\}$, the data for this second numerical test was generated from a multivariate Gaussian mixture distribution.

Specifically, $\frac{1}{2}\mathcal{N}(-\mathbf{1}, I_k) + \frac{1}{2}\mathcal{N}(\mathbf{1}, I_k)$, which is a symmetric distribution. Over 10,000 replications, random samples were drawn and used to estimate the expected Rademacher average, R_n, and the expected empirical Wasserstein distance, C_n, under the ℓ_2-norm. The dimensions considered were $k = \{2, 20, 200\}$. The results are displayed on the right column of Figure 4.1. Similarly to the multivariate Rademacher setting, as the sample size n increases, the bound based on the empirical Wasserstein distance becomes sharper than the original symmetrization bound.

4.5 Applications

In the following subsections, a collection of applications of the improved symmetrization inequality are detailed to demonstrate the usefulness of this result. Such applications range from those of theoretical interest to those of practical application to statistical testing. These include a test for data symmetry, the construction of nonasymptotic high dimensional confidence sets, bounding the variance of an empirical process, and Nemirovski's inequality for Banach space valued random variables.

4.5.1 Permutation test for data symmetry

In the previous sections, we proposed the Wasserstein distance $W_2(\mu, \mu^-)$ to quantify the symmetry of a measure μ. Now, given n independent and identically distributed observations $X_1, \ldots, X_n$ with common measure μ, we propose a procedure to test for whether or not μ is symmetric. Unlike other tests for data symmetry which may be restricted to finite dimensional Euclidean space, this testing procedure applies to general Hilbert space valued random variables. Thus, it is applicable to many diverse settings such as, notably, functional data analysis.

The bootstrap approach from Section 4.4 for estimating the empirical Wasserstein distance is applied, and a permutation test is applied to the bootstrapped sample. Note that while the Wasserstein-2 metric is specifically used in our improved symmetrization inequality, for this test, any Wasserstein-p metric can be utilized as is done in the numerical simulations below.

The bootstrap-permutation test proceeds as follows:

0. Choose a number r of bootstrap replications to perform.

1. For each bootstrap replication, permute the data by some uniformly randomly drawn $\rho \in S_n$, the symmetric group on n elements.

2. Use the Hungarian algorithm to compute the optimal assignment cost, ω_0, between the data sets $\{X_{\rho(1)}, \ldots, X_{\rho(n/2)}\}$ and $\{-X_{\rho(n/2+1)}, \ldots, -X_{\rho(n)}\}$.

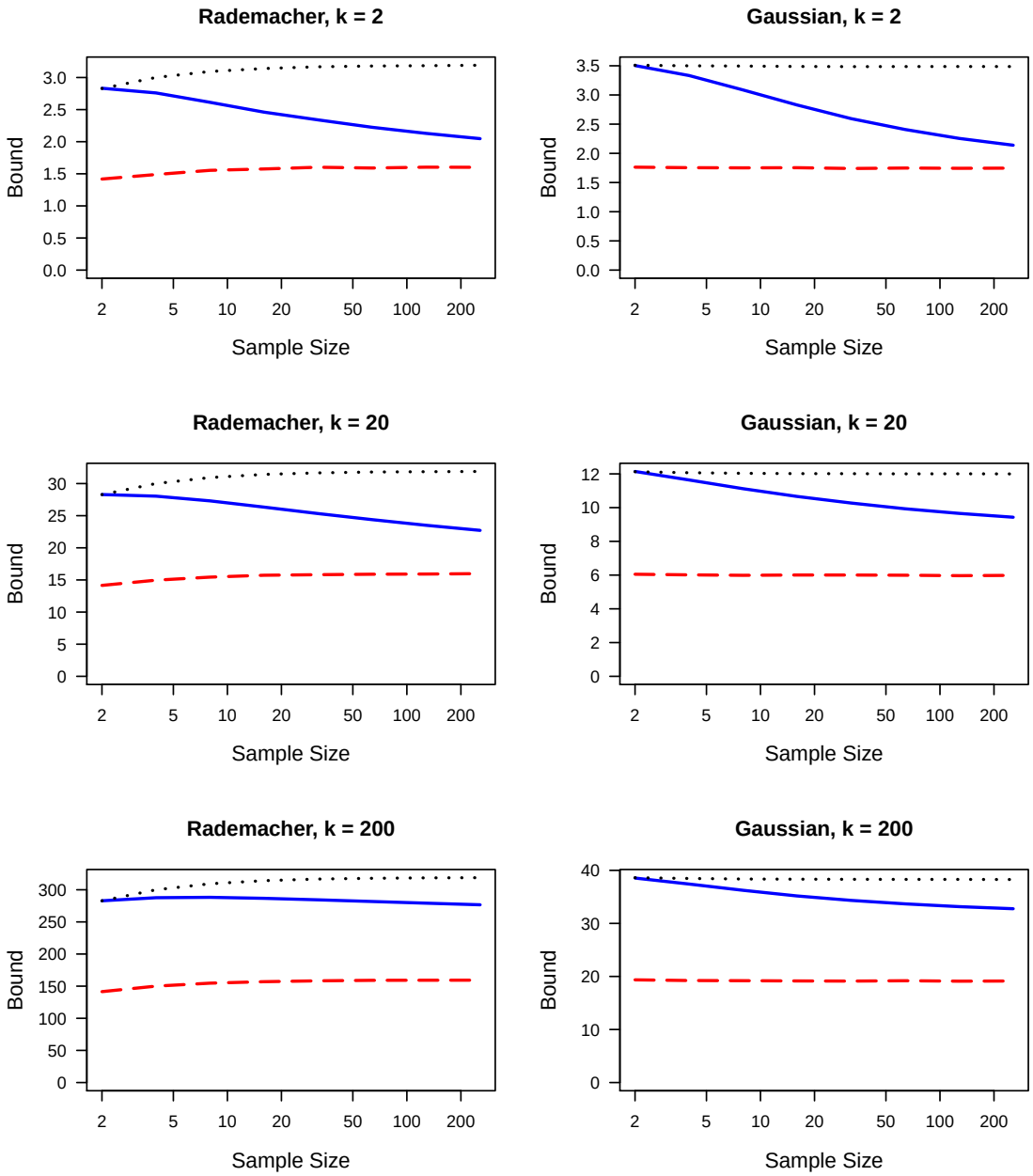


Figure 4.1: For multivariate Rademacher (left) and Gaussian mixture (right) data, the average $n^{-1/2}\mathrm{E}\|\sum_{i=1}^{n}(X_i - \mathrm{E}X_i)\|$ (red dashed lines), twice the Rademacher average $2R_n = 2n^{-1/2}\mathrm{E}\|\sum_{i=1}^{n}\varepsilon_i(X_i - \mathrm{E}X_i)\|$ (black dotted lines), and the bound using the scaled empirical Wasserstein distance, $R_n + W_2(\mu_n, \mu_n^-)/\sqrt{2}$ (blue solid lines) were estimated over 10,000 replications. The dimension of the data is $k = \{2, 20, 200\}$. For the Rademacher setting, the ℓ_1-norm was used. For the Gaussian setting, the ℓ_2-norm was used. As the sample size increases, the Wasserstein term converges to zero thus sharpening the upper bound.

3. Denote this new half-negated data set Y where $Y_i = X_{\rho(i)}$ for $i \leq n/2$ and $Y_i = -X_{\rho(i)}$ for $i > n/2$.

4. Draw m random permutations $\rho_1, \ldots, \rho_m \in S_n$. For each ρ_i, compute ω_i, the optimal assignment cost between $\{Y_{\rho_i(1)}, \ldots, Y_{\rho_i(n/2)}\}$ and $\{Y_{\rho_i(n/2+1)}, \ldots, Y_{\rho_i(n)}\}$.

5. Return the p-value, $p_j = \#\{\omega_i > \omega_0\}/m$.

6. Average the r p-values to get an overall p-value, $p = r^{-1} \sum_{j=1}^{r} p_j$.

Note that for very large data sets, it may be computationally impractical to find a perfect matching between two sets of $n/2$ nodes as performing this test as stated has a computational complexity of order $O(mn^3)$. In that case, randomly draw $n' < n$ elements from the data set in step 1, draw a $\rho \in S_{n'}$, and proceed as before but with the smaller sample size.

This permutation test was applied to simulated multivariate Rademacher(p) data in $\mathbb{R}^5$. For sample sizes $n = 10$ and $n = 100$, let $X_1, \ldots, X_n$ be independent and identically distributed multivariate Rademacher(p) random variables defined in Definition 1.1.11 where each X_i is comprised of a vector of independent univariate Rademacher(p) random variables. For values of $p \in [0.5, 0.8]$, the power of this test was experimentally computed over 1000 simulations. The results are displayed in Figure 4.2. For the ℓ^1 and ℓ^2 metrics and Wasserstein distances W_1 and W_2, the performances of the permutation test were comparable except for the (ℓ^2, W_2) case, which performed poorer in both the large and small sample size settings. For the large sample size, $n = 100$, Mardia's test for multivariate skewness (Mardia, 1970, 1974) was included, which uses the result that

$$\frac{6}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(X_i - \bar{X})^{\mathrm{T}} \hat{\Sigma}^{-1} (X_j - \bar{X}) \right]^3 \xrightarrow{\mathrm{d}} \chi^2 \left(k(k+1)(k+2)/6 \right)$$

where $\hat{\Sigma}$ is the empirical covariance matrix of the data. However, this is shown to be less powerful than the proposed permutation test. Furthermore, as this test is asymptotic in design, it gave erroneous results in the $n = 10$ case and was thus excluded from the figure.

4.5.2 High dimensional confidence sets

A method for constructing nonasymptotic confidence regions for high dimensional data using a generalized bootstrap procedure was proposed in the article of Arlot et al. (2010). Beginning with a sample of independent and identically distributed $Y_1, \ldots, Y_n \in \mathbb{R}^K$ and the assumptions that the Y_i are symmetric about their mean–that is, $Y_i - \mu \stackrel{\mathrm{d}}{=} \mu - Y_i$ for all i–and are bounded in L_p-norm–that is, $\|Y_i - \mu\|_p \leq M$

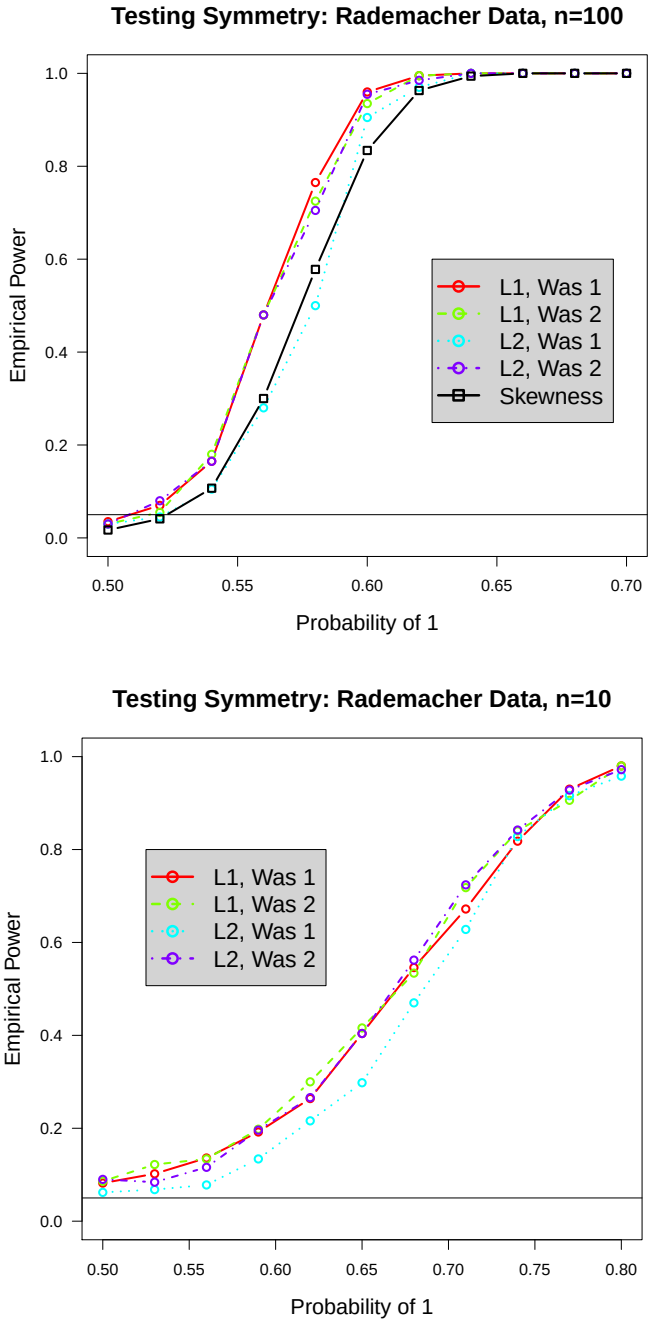


Figure 4.2: For data in $\mathbb{R}^5$, the ℓ^1 and ℓ^2 metrics, and the Wasserstein distances W_1 and W_2, the experimentally computed power of the permutation test is plotted for Rademacher(p) data as p, the probability of 1, increases thus skewing the distribution. The sample size is $n = 100$ on the left plot and is $n = 10$ on the right plot. The $n = 100$ case includes an asymptotic test for skewness. This test fails in the nonasymptotic $n = 10$ case and thus is not included.

almost surely for all i and some $M > 0$–they prove, among many other results, that for some fixed $\alpha \in (0,1)$, the following holds with probability $1-\alpha$:

$$\phi\left(\bar{Y}-\mu\right) \leq \left(\frac{n}{n-1}\right) \mathrm{E}_{\varepsilon}\phi\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(Y_i-\bar{Y})\right) + \frac{2M}{\sqrt{n}}\sqrt{\log(1/\alpha)}$$

where $\phi : \mathbb{R}^K \to \mathbb{R}$ is a function that is subadditive, positive homogeneous, and bounded by L_p-norm. By substituting our Theorem 4.3.4 for their Proposition 2.4 allows us to drop the symmetry condition and achieve a more general $(1-\alpha)$ confidence region.

Proposition 4.5.1. *For a fixed $\alpha \in (0,1)$ and $p \in [1,\infty]$, let $\phi : \mathbb{R}^k \to \mathbb{R}$ be subadditive, positive homogeneous, and bounded in L_p-norm. Then, for some $M > 0$, the following holds with probability at least $1-\alpha$.*

$$\phi\left(\bar{Y}-\mu\right) \leq \mathrm{E}_{\varepsilon}\phi\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(Y_i-\bar{Y})\right) + \\ + (2n)^{-1/2}\left(2\sqrt{2}M\sqrt{\log(1/\alpha)} + W_2(\mu,\mu^{-})\right).$$

4.5.3 Bounds on empirical processes

Symmetrization arises when bounding the variance of an empirical process. In Boucheron et al. (2013), the following result is stated as Theorem 11.8 and is subsequently proved using the original symmetrization inequality resulting in suboptimal coefficients.

Theorem 4.5.2 (Boucheron et al. (2013), Theorem 11.8)**.** *For $i \in \{1,\ldots,n\}$ and $s \in \mathcal{T}$, a countable index set, let $X_i = (X_{i,s})_{s\in\mathcal{T}}$ be a collection of real valued random variables. Furthermore, let $X_1,\ldots,X_n$ be independent. Assume $\mathrm{E}X_{i,s} = 0$ and $|X_{i,s}| \leq 1$ for all $i = 1,\ldots,n$ and for all $s \in \mathcal{T}$. Defining $Z = \sup_{s\in\mathcal{T}}\sum_{i=1}^{n} X_{i,s}$, then*

$$\mathrm{Var}\left(Z\right) \leq 8\mathrm{E}Z + 2\sigma^2$$

where $\sigma^2 = \sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\mathrm{E}X_{i,s}^2$.

The given proof uses the symmetrization inequality twice as well as the contraction inequality–see Ledoux and Talagrand (1991) Theorem 4.4, and Boucheron et al. (2013) Theorem 11.6–to establish the bounds

$$\mathrm{E}\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}X_{i,s}^2 \leq \sigma^2 + 2\mathrm{E}\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\varepsilon_i X_{i,s}^2 \quad \text{and} \quad \mathrm{E}\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\varepsilon_i X_{i,s}^2 \leq 4\mathrm{E}Z.$$

Making use of the improved symmetrization inequality cuts the coefficient of $\mathrm{E}Z$ by a factor of 4 to the tighter

$$\mathrm{Var}\,(Z) \leq 2\mathrm{E}Z + 2\sigma^2 + O(\sqrt{n}).$$

Beyond this textbook example of bounding the variance of an empirical process, symmetrization arguments are used to construct confidence sets for empirical processes in Giné and Nickl (2010a); Lounici and Nickl (2011); Kerkyacharian et al. (2012); Fan (2011). The coefficients in all of their results can be similarly improved using the improved symmetrization inequality.

4.5.4 Type, cotype, and Nemirovski's inequality

In the probability in Banach spaces setting, let $X_i \in (B, \|\cdot\|)$ for $i = 1, \ldots, n$ be a collection of independent zero mean Banach space valued random variables. A collection of results referred to as *Nemirovski inequalities* (Nemirovski, 2000; Dümbgen et al., 2010) are concerned with whether or not there exists a constant K depending only on the norm such that

$$\mathrm{E}\left\|\sum_{i=1}^{n} X_i\right\|^2 \leq K \sum_{i=1}^{n} \|X_i\|^2.$$

For example, in the Hilbert space setting, orthogonality allows for $K = 1$ and the inequality can be replaced by an equality.

One such result requires the notion of type and cotype. A Banach space $(B, \|\cdot\|)$ is said to be of *Rademacher type* p for $1 \leq p < \infty$ (respectively, of *Rademacher cotype* q for $1 \leq q < \infty$) if there exists a constant T_p (respectively, C_q) such that for all finite non-random sequences $(x_i) \in B$ and (ε_i), a sequence of independent Rademacher random variables,

$$\mathrm{E}\left\|\sum_i \varepsilon_i x_i\right\|^p \leq T_p^p \sum_i \|x_i\|^p, \quad \left(\text{respectively, } \sum_i \|x_i\|^q \leq C_q^{-q} \mathrm{E}\left\|\sum_i \varepsilon_i x_i\right\|^q\right).$$

These definitions and the original symmetrization inequality lead to the following proposition.

Proposition 4.5.3 (Ledoux and Talagrand (1991) Proposition 9.11, Dümbgen et al. (2010) Proposition 3.1)**.** *Let $X_i \in B$ for $i = 1, \ldots, n$ and $S_n = n^{-1} \sum_{i=1}^n X_i$. If $(B, \|\cdot\|)$ is of type $p \geq 1$ with constant T_p (respectively, of cotype $q \geq 1$ with constant*

C_q), then

$$\mathrm{E}\|S_n\|^p \leq (2T_p)^p n^{-p} \sum_{i=1}^n \mathrm{E}\|X_i\|^p$$
$$\mathrm{E}\|S_n\|^q \geq (2C_q)^{-q} n^{-q} \sum_{i=1}^n \mathrm{E}\|X_i\|^q.$$

The proposition can be refined by applying our improved symmetrization inequality along with the Rademacher type p condition if the X_i are additionally norm bounded. If the X_i have a common law μ, let $W_2 = W_2(\mu, \mu^-)$ be the Wasserstein distance between μ and its reflection.

Proposition 4.5.4. *Under the setting of Proposition 4.5.3, additionally assume that $\|X_i\| \leq 1$ for $i = 1, \ldots, n$. Then,*

$$\mathrm{E}\|S_n\|^p \leq T_p^p n^{-p} \sum_{i=1}^n \mathrm{E}\|X_i\|^p + \frac{pW_2}{\sqrt{2n}}$$
$$\mathrm{E}\|S_n\|^q \geq C_q^{-q} n^{-q} \sum_{i=1}^n \mathrm{E}\|X_i\|^q - \frac{qW_2}{\sqrt{2n}}.$$

Proof. In the context of Theorem 4.3.4, set $\psi(\cdot) = \|\cdot\|^p$. Given the bound $\|X_i\| \leq 1$, we have that $\|\psi\|_{Lip} = p$. Scale by p, and the first result follows. □

Note that for identically distributed $X_i \in B$, the order of the original bound for a type p Banach space is $O(n^{1-p})$ while the Wasserstein correction term is $O(n^{-1/2})$. This correction will give an obvious benefit for spaces of type $p < 3/2$. However, even for spaces of type 2, the new bound can be tighter specifically in the high dimensional setting when $d \gg n$. Indeed, consider $\ell_\infty(\mathbb{R}^d)$, which is discussed in particular in Section 3.2 of Dümbgen et al. (2010) where it is shown to be of type 2 with constant $T_p = \sqrt{2\log(2d)}$. For independent and identically distributed $X_i \in \ell_\infty(\mathbb{R}^d)$, the bounds to compare are

$$\frac{8\log(2d)}{n}\mathrm{E}\|X_i\|_\infty^2 \quad \text{and} \quad \frac{2\log(2d)}{n}\mathrm{E}\|X_i\|_\infty^2 + \sqrt{\frac{2}{n}}W_2(\mu, \mu^-).$$

Figure 4.3 displays such a comparison for $n = 10$, $d \in \{5, 25, 50\}$, and iid $X_{i,j} + \alpha/(1+\alpha) \sim \mathrm{Beta}\,(\alpha, 1)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Hence, the X_i are Beta random variables that are shifted to have zero mean. $W_2(\mu, \mu^-)$ is approximated by $\mathrm{E}W_2(\mu_5, \mu_5^-)$, which is computed via the bootstrap procedure outlined in Section 4.4. The new bound can be seen to have better performance than the old one specifically in the cases of $d = 25$ and $d = 50$ when α is not too large.

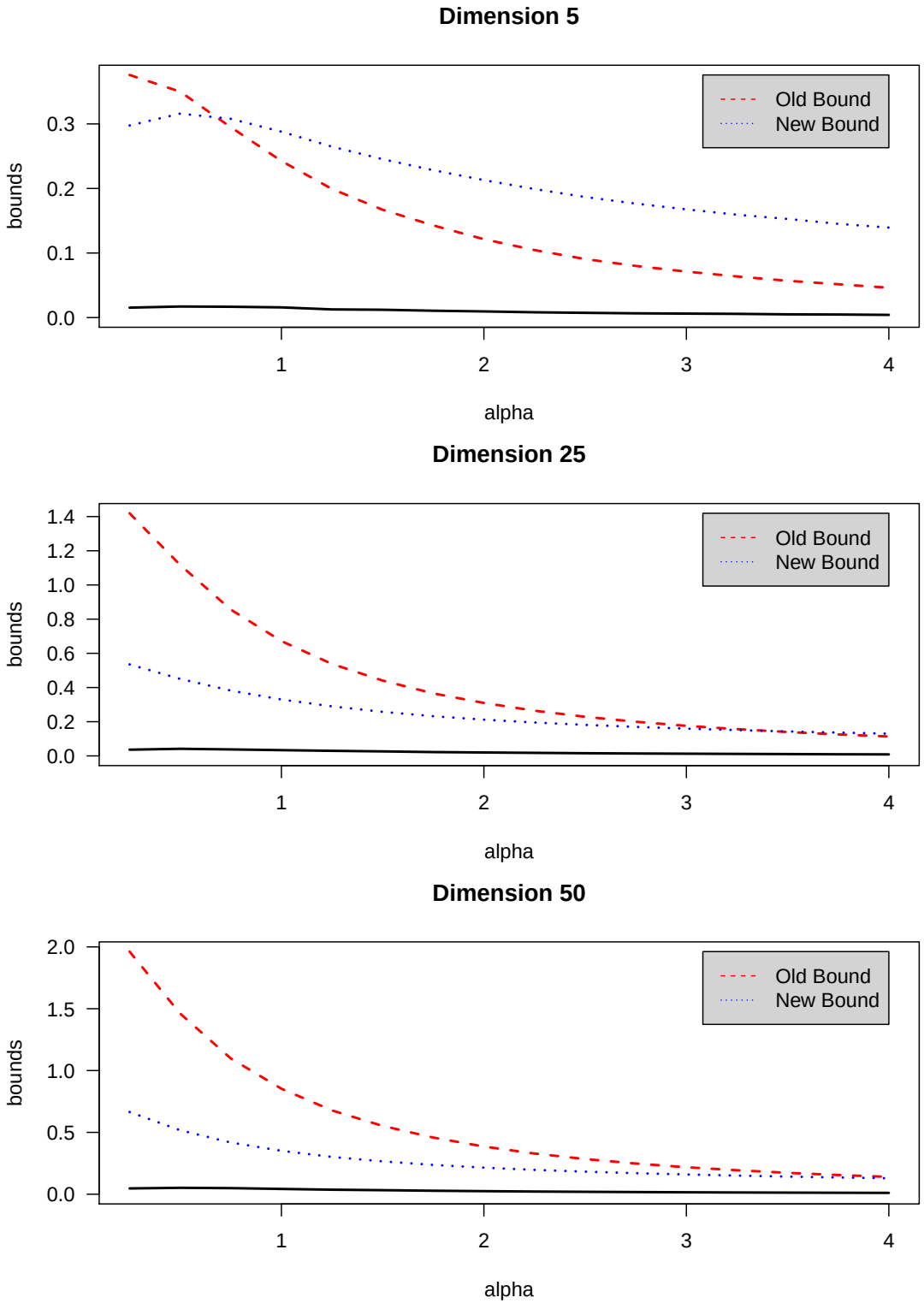


Figure 4.3: A comparison of the old bound from Proposition 4.5.3, the red dashed line, and the new bound from Proposition 4.5.4, the blue dotted line, for a sample $n = 10$, and $X_i \in \ell_\infty(\mathbb{R}^d)$ for dimensions $d \in \{5, 25, 50\}$. Each $X_i = (X_{i,1}, \ldots, X_{i,d})$ where each $X_{i,j} + \alpha/(1+\alpha) \overset{\text{iid}}{\sim} \text{Beta}(\alpha, 1)$. The solid black line indicates the left hand side in the two propositions of $\mathrm{E}\|S_n\|_\infty^2$.

A Nemirovski variant with weak variance

As one further example of improved symmetrization, a variation of Nemirovski's inequality found in Section 13.5 of Boucheron et al. (2013) is proved via a similar symmetrization argument for the ℓ_p norm with $p \geq 1$. Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be independent zero mean random variables. Let $B_q = \{x \in \mathbb{R}^d : \|x\|_q \leq 1\}$, and define the weak variance $\Sigma_p^2 = n^{-2}\mathrm{E}\sup_{t \in B_q} \sum_{i=1}^n \langle t, X_i \rangle^2$. The resulting inequality is

$$\mathrm{E}\|S_n\|_p^2 \leq 578 d \Sigma_p^2.$$

Replacing the old symmetrization inequality with the improved version reduces the coefficient of 578 roughly by a factor of 4 resulting in

$$\mathrm{E}\|S_n\|_p^2 \leq 146 d \Sigma_p^2 + O(n^{-1/2}).$$

4.6 Discussion

The symmetrization inequality is a fundamental result for probability in Banach spaces, concentration inequalities, and many other related areas. However, not accounting for the amount of asymmetry in the given random variables has led to pervasive powers of two throughout derivative results. Our improved symmetrization inequality incorporates such a quantification of asymmetry through use of the Wasserstein distance. Besides being theoretically sound, it is shown in simulations to provide a tightness superior to that of the original result. Going beyond the inequality itself, this Wasserstein distance offers a novel and powerful way to analyze the symmetry of random variables or lack thereof. It can and should be applied to countless other results that were not considered in this current work.

4.A Past results used

Lemma 4.A.1 (Kantorovich-Rubinstein Duality, see Villani (2008))**.** *Under the setting of Definition 4.3.1,*

$$W_1(\mu, \nu) = \sup_{\|\phi\|_{Lip} \leq 1} \left\{ \int_{\mathcal{X}} \phi d\mu - \int_{\mathcal{X}} \phi d\nu \right\}.$$

Lemma 4.A.2. *Under the setting of Definition 4.3.1, for $p < q$,*

$$W_p(\mu, \nu) \leq W_q(\mu, \nu).$$

Proof. Jensen or Hölder's Inequality □

Lemma 4.A.3 (Convolution property of W_2, see Bickel and Freedman (1981))**.** *For Hilbert space valued random variables X_i with law μ_i and Y_i with law ν_i for $i = 1, \ldots, n$, define μ^{*n} to be the law of $\sum_{i=1}^n X_i$ and similarly for ν^{*n}. Then,*

$$W_2^2(\mu^{*n}, \nu^{*n}) \leq \sum_{i=1}^n W_2^2(\mu_i, \nu_i).$$

Lemma 4.A.4 (Convergence of Empirical Measure, see Bickel and Freedman (1981))**.** *Let $X_1, \ldots, X_n$ be independent and identically distributed Banach space valued random variables with common law μ. Let μ_n be the empirical distribution of the X_i. Then,*

$$W_p(\mu_n, \mu) \to 0, \quad \text{as } n \to \infty.$$

Appendix A

R package: `fdcov`

A.1 Code overview

The statistical techniques outlined in Chapter 3 for inference on covariance operators are available in `fdcov`, an R package uploaded to CRAN (R Core Team, 2016). Beyond the methodology of Chapter 3, this package also contains a routine for performing a k sample test for equality of covariance via a permutation test (Cabassi et al., 2017). It is planned to eventually contain the two sample tests from Panaretos et al. (2010) and Fremdt et al. (2013) as well. In this chapter, we briefly outline the functionality of this software package.

The function `ksample.com` applies the k sample test for equality of covariance operator from Section 3.3.1. This function is complimented by `ksample.perm`, which answers the same statistical question but with a permutation test. The arguments are

```
ksample.com(dat, grp, p, alpha, scl1, scl2)
```

where `dat` is an $n \times m$ matrix of data with one entry per row, `grp` is the n long vector of group labels, `p` refers to the choice of p-Schatten norm, `alpha` is the desired empirical size of the test, and `scl1` and `scl2` are the scale factors for tweaking the coefficients as described in Section 3.3.1. Future updates aim to include a cross-validation procedure for selecting a data driven choice for `scl1` and `scl2` such as the procedure described in Appendix 3.E.

The function `classif.com` trains a covariance operator classifier using the method from Section 3.3.2. The arguments are

```
classif.com(datGrp, dat)
```

where `datGrp` is the n long vector of group labels and `dat` is the $n \times m$ matrix of data with one entry per row. The classifier can be used for prediction via the generic

S3 function

```
predict(object, dat, SOFT, ...)
```

where `object` is the classifier, `dat` is the data matrix to be classified with one entry per row, `SOFT` is a Boolean flag to tell the routine to either return hard or soft classification. In this context, hard classification makes the program return a class label while soft classification makes the program return a probability vector with entries $\mathrm{P}(\texttt{dat[i,]} \in \texttt{class[j]})$.

The function `cluster.com` applies the expectation-maximization algorithm from Section 3.3.3 to a set of functional data observations to cluster them based on their covariances. The arguments are

```
cluster.com( dat, labl, grpCnt, iter,
                                SOFT, PRINTLK, LOADING, IGNORESTOP )
```

where `dat` is the $n \times m$ data matrix with one entry per row, `labl` is optional and used to group curves together in order to cluster operators that are not all rank one, `grpCnt` sets the number of groups into which to partition the data, `iter` tells the program the maximum number of iterations to run, `SOFT` is a Boolean flag to tell the program whether or not to return category probabilities, `PRINTLK` is a Boolean flag to tell the program whether or not to print the likelihoods used by the EM algorithm, `LOADING` is a Boolean flag to tell the program whether or not to print a loading bar while running, and `IGNORESTOP` is a Boolean flag to tell the program whether or not to ignore early stopping conditions.

A.2 Examples

A.2.1 k sample test

```
# Load in phoneme data
library(fds)
# Setup data arrays
dat1 = rbind( t(aa$y)[1:20,], t(sh$y)[1:20,] );
dat2 = rbind( t(aa$y)[1:20,], t(ao$y)[1:20,] );
dat3 = rbind( dat1, t(ao$y)[1:20,] );
# Setup group labels
grp1 = gl(2,20);
grp2 = gl(2,20);
grp3 = gl(3,20);
```

```
# Compare two disimilar phonemes (should return TRUE)
ksample.com(dat1,grp1);
# Compare two similar phonemes (should return FALSE)
ksample.com(dat2,grp2);
# Compare three phonemes (should return TRUE)
ksample.com(dat3,grp3);
```

A.2.2 Classifying operators

```
library(fds);
# Setup training data
dat1 = rbind(
  t(aa$y[,1:100]), t(ao$y[,1:100]),
  t(dcl$y[,1:100]), t(iy$y[,1:100]),
  t(sh$y[,1:100])
);
# Setup testing data
dat2 = rbind(
  t(aa$y[,101:400]), t(ao$y[,101:400]),
  t(dcl$y[,101:400]), t(iy$y[,101:400]),
  t(sh$y[,101:400])
);

datgrp = gl(5,100);
clCom = classif.com( datgrp, dat1 );
grp = predict( clCom, dat2, LOADING=TRUE );
acc = c(
  sum( grp[1:300]==1 ), sum( grp[301:600]==2 ),
  sum( grp[601:900]==3 ), sum( grp[901:1200]==4 ),
  sum( grp[1201:1500]==5 )
)/300;
print(rbind(gl(5,1),signif(acc,3)));
```

A.2.3 Clustering operators

```
# Load phoneme data
library(fds);
# Setup data to be clustered
dat  = rbind(
  t(aa$y[,1:20]),t(iy$y[,1:20]),
  t(sh$y[,1:20])
```

```
);
# Cluster data into three groups
clst = cluster.com(dat,grpCnt=3);
matrix(clst,3,20,byrow=TRUE);

# cluster groups of curves
dat  = rbind(
  t(aa$y[,1:40]),t(iy$y[,1:40]),
  t(sh$y[,1:40])
);
lab  = gl(30,4);
# Cluster data into three groups
clst = cluster.com(dat,labl=lab,grpCnt=3);
matrix(clst,3,10,byrow=TRUE);
```

Appendix B

Future Considerations

No single manuscript can contain a complete compendium of all information, extensions, and applications of a single topic. In this appendix, we briefly discuss some of the topics, which have yet to be fully considered under the auspices of the non-asymptotic concentration inequality based statistical methodology promoted by the previous chapters.

The methodology of this manuscript is keenly focused around the construction of dimension-free non-asymptotic confidence sets for a wide range of statistical objects. It is in this paradigm of high or infinite dimensions that so much of modern data lies. In this manuscript, we briefly considered some real data sets as a proof of concept for our proposed methodology. In Chapter 2, we considered high dimensional genomics data. In Chapter 3, we considered infinite dimensional phoneme data. In this appendix chapter, we briefly discuss two areas of data analysis that could be enhanced with the introduction of similar concentration inequality based methodology to that of the rest of this manuscript. The areas to be considered are longitudinal data and data living in a reproducing kernel Hilbert space.

B.1 Longitudinal data

An area of data analysis that is closely related to functional data analysis is that of longitudinal data. In this setting, a collection of subjects are observed at specific instances over a long timespan such as a pharmaceutical trial where patients are examined by doctors at regular intervals. However, such studies are rife with potential problems including long time spans between observations, missing data, and irregular time intervals for subject observations. As a results, much recent research has gone into the study of the analysis of such data from the functional data perspective (Müller, 2005; Yao et al., 2005; Hall et al., 2008; Serban et al., 2013). Furthermore, the estimation of the covariance operator and related inference is often of interest in longitudinal data analysis. Thus, such testing can be considered with

the concentration inequality based methodology promoted in this manuscript.

In the ideal case where the data is observed at regular intervals and is not missing, we can use the methodology developed in Chapter 2 for the estimation of large sparse covariance matrices. The assumption of sparsity is reasonable in this setting as it can usually be assumed that observations further apart in time will be less correlated than those observed in quick temporal succession. Furthermore, the time information can be incorporated into an estimator as meta-information similar to the techniques of tapering and banding in covariance matrix estimation (Wu and Pourahmadi, 2003; Furrer and Bengtsson, 2007; Bickel and Levina, 2008b).

Once the incorporation of sparse irregular observations is added to the framework, further sophistication is required. Thus, longitudinal analysis is often considered from the functional data perspective. The addition of irregular and missing observations leads to more theoretical and practical problems that must be addressed.

A particularly challenging problem of interest is the estimation of the covariance structure of longitudinal data when the responses are binary valued (Avery et al., 2014). While many standard techniques fail in this setting, it is reasonable to attempt a concentration based estimation framework. Specifically, most of the concentration inequalities available require some absolute bound on the data under observation. Hence, binary valued longitudinal data falls nicely into this paradigm.

B.2 Reproducing Kernel Hilbert Spaces

Much recent research has demonstrated that statistical inference in the infinite dimensional setting can be extremely fruitful when working within the confines and framework of a Reproducing Kernel Hilbert Space, or an RKHS for short (Hofmann et al., 2008; Yuan and Cai, 2010; Cai and Yuan, 2010, 2012; Blanchard and Mücke, 2016; Qu et al., 2016). An introduction to RKHS's for statistics with Gaussian processes is detailed in Section 2.6 of Giné and Nickl (2016). In line with such past research, we briefly investigated the consequences of constructing concentration inequality based confidence sets in similar style to those from Chapter 3. Specifically, we considered the implications of estimating the weak variance, a key component to such confidence sets, in the RKHS framework. Those calculations are briefly described below.

Let $\mathcal{H}(K)$ be an Hilbert space of functions on a domain $\mathcal{T}$ where $\mathcal{T} = [0, 1]$ for simplicity of exposition. The reproducing kernel K is a symmetric bivariate function $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ such that for every $t \in \mathcal{T}$ and every $f \in \mathcal{H}(K)$, $K(\cdot, t) \in \mathcal{H}(K)$ and $f(t) = \langle f, K(\cdot, t)\rangle_{\mathcal{H}(K)}$. Equivalently, the RKHS $\mathcal{H}(K)$ can be defined as the Hilbert space in which the evaluation functional $L_t : f \to f(t)$ is continuous (i.e. is a bounded linear functional). Then, $L_t(f) = \langle f, K(\cdot, t)\rangle = f(t)$. The kernel K is symmetric

and positive-definite. Conversely, any symmetric and positive-definite K will be the kernel for a specific RKHS $\mathcal{H}(K)$. This results is known as the Moore-Aronszajn theorem (Aronszajn, 1950).

Assume that K is square integrable, and let $\{\varphi_i\}_{i=1}^{\infty}$ be an orthonormal basis for $L^2(\mathcal{T})$. Then, from Mercer's Theorem (Mercer, 1909),

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(s)\varphi_i(t)$$

where $\lambda_1 \geq \lambda_2 \geq \ldots$ are the eigenvalues of K. For $f, g \in \mathcal{H}(K)$, write $f(t) = \sum_{i=1}^{n} f_i\varphi_i(t)$ and $g(t) = \sum_{i=1}^{n} g_i\varphi_i(t)$. Then, define $\langle f, g\rangle_{\mathcal{H}(K)} := \sum_{i=1}^{\infty} \lambda_i^{-1} f_i g_i$.

The tensor product space $\mathcal{H}(K) \otimes \mathcal{H}(K) = \mathcal{H}(K \otimes K)$ is an RKHS with kernel $K \otimes K : \mathcal{T}^4 \to \mathbb{R}$ such that $K \otimes K((s,t),(u,v)) = K(s,u)K(t,v)$. Furthermore, the collection $\{\varphi_i\varphi_j\}_{i,j=1}^{\infty}$ is an orthonormal basis for $L^2(\mathcal{T} \times \mathcal{T})$. For $f, g \in \mathcal{H}(K)$, then $f \otimes g(s,t) = f(s)g(t)$ and $\|f \otimes g\|_{\mathcal{H}(K\otimes K)} = \|f\|_{\mathcal{H}(K)}\|g\|_{\mathcal{H}(K)}$. For $f_1, \ldots, f_n, g_1, \ldots, g_n \in \mathcal{H}(K)$,

$$\left\| \sum_{i=1}^{n} f_i \otimes g_i \right\|_{\mathcal{H}(K\otimes K)}^2 = \sum_{i,j=1}^{n} \langle f_i, g_i\rangle_{\mathcal{H}(K)} \langle f_j, g_j\rangle_{\mathcal{H}(K)} .$$

For an operator $A \in \mathcal{H}(K \otimes K)$, write $A = \sum_{i,j=1}^{n} a_{i,j}\varphi_i\varphi_j$, then $\|A\|_{\mathcal{H}(K\otimes K)}^2 = \sum_{i,j=1}^{n} \lambda_i^{-1}\lambda_j^{-1}a_{i,j}^2$.

Note that if K is such that $\lambda_i = 1$ for all i, then the norm $\|\cdot\|_{\mathcal{H}(K\otimes K)}$ coincides with the usual Hilbert-Schmidt norm, which is also the Frobenius norm in the finite dimensional setting. Furthermore, let Λ be the diagonal operator with the same eigenvalues as K. Then, $\|A\|_{\mathcal{H}(K\otimes K)}^2 = \|\Lambda^{-1/2}A\Lambda^{-1/2}\|_{\mathrm{HS}}^2$.

Now let $f \in \mathcal{H}(K)$ be a random function such that $f(t) = \sum_{i=1}^{\infty} X_i\varphi_i(t)$ where the X_i are real valued random variables. Then, the mean $\mu = \mathrm{E}f(t) = \sum_{i=1}^{\infty}(\mathrm{E}X_i)\varphi_i(t) \in \mathcal{H}(K)$. Furthermore, letting $c_{i,j} = \mathrm{cov}\,(X_i, X_j)$, the covariance operator is

$$C(s,t) = \mathrm{E}\left((f-\mu)^{\otimes 2}\right) = \sum_{i,j=1}^{\infty} c_{i,j}\varphi_i(s)\varphi_j(t) \in \mathcal{H}(K \otimes K).$$

Next, we compute the weak variance as in the subsections of Appendix 3.B but for functional data in an RKHS. Given $f_1, \ldots, f_n, f \in \mathcal{H}(K)$ independent and identically distributed zero mean random functions, define the empirical covariance operator to be $\hat{C} = n^{-1}\sum_{i=1}^{n} f_i^{\otimes 2}$. The size of a concentration based confidence set is strongly related to the magnitude of the weak variance

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \sup_{\|A\|_{\mathcal{H}(K\otimes K)}\leq 1} \mathrm{E}\left\langle f_i^{\otimes 2} - \mathrm{E}f^{\otimes 2}, A\right\rangle_{\mathcal{H}(K\otimes K)}^2$$

where the supremum is taken over a countable dense subset of the unit ball of $\mathcal{H}(K \otimes K)$. Thus, continuing similarly to the computations of Chapter 3,

$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} \sup_{\|A\|_{\mathcal{H}(K\otimes K)}\leq 1} \mathrm{E}\left\langle f_i^{\otimes 2} - \mathrm{E}f^{\otimes 2}, A\right\rangle^2_{\mathcal{H}(K\otimes K)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \sup_{\|A\|_{\mathcal{H}(K\otimes K)}\leq 1} \left\langle \mathrm{E}f_i^{\otimes 4} - C^{\otimes 2}, A^{\otimes 2}\right\rangle_{\mathcal{H}(K^{\otimes 4})} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \sup_{\|\Pi\|_{\mathcal{H}(K^{\otimes 4})}\leq 1} \left\langle \mathrm{E}f_i^{\otimes 4} - C^{\otimes 2}, \Pi\right\rangle_{\mathcal{H}(K^{\otimes 4})} \\
&= \left\|\mathrm{E}f^{\otimes 4} - C^{\otimes 2}\right\|_{\mathcal{H}(K^{\otimes 4})} \\
&= \left\| \sum_{i,j,k,l=1}^{\infty} \left[\mathrm{E}\left(X_iX_jX_kX_l\right) - \mathrm{E}\left(X_iX_j\right)\mathrm{E}\left(X_kX_l\right)\right] \varphi_i\varphi_j\varphi_k\varphi_l \right\|_{\mathcal{H}(K^{\otimes 4})} \\
&= \left(\sum_{i,j,k,l=1}^{\infty} \left[\mathrm{E}\left(X_iX_jX_kX_l\right) - \mathrm{E}\left(X_iX_j\right)\mathrm{E}\left(X_kX_l\right)\right]^2 \lambda_i^{-1}\lambda_j^{-1}\lambda_k^{-1}\lambda_l^{-1} \right)^{1/2}
\end{aligned}$$

In the Gaussian process setting of Appendix 3.B.3, we have from Isserlis (1918) that $\mathrm{E}\left(X_iX_jX_kX_l\right) = \mathrm{E}\left(X_iX_j\right)\mathrm{E}\left(X_kX_l\right) + \mathrm{E}\left(X_iX_k\right)\mathrm{E}\left(X_jX_l\right) + \mathrm{E}\left(X_iX_l\right)\mathrm{E}\left(X_jX_l\right) = c_{i,j}c_{k,l} + c_{i,k}c_{j,l} + c_{i,l}c_{j,k}$. Hence,

$$\begin{aligned}
\sigma^2 &\leq \left(\sum_{i,j,k,l=1}^{\infty} \left[c_{i,k}c_{j,l} + c_{i,l}c_{j,k}\right]^2 \lambda_i^{-1}\lambda_j^{-1}\lambda_k^{-1}\lambda_l^{-1} \right)^{1/2} \\
&= \left(\sum_{i,j,k,l=1}^{\infty} \left[c_{i,k}^2c_{j,l}^2 + c_{i,l}^2c_{j,k}^2 + 2c_{i,k}c_{j,l}c_{i,l}c_{j,k}\right] \lambda_i^{-1}\lambda_j^{-1}\lambda_k^{-1}\lambda_l^{-1} \right)^{1/2} \\
&= \left(2\left(\sum_{i,j=1}^{\infty} c_{i,j}^2\lambda_i^{-1}\lambda_j^{-1} \right)^2 + 2\sum_{i,j=1}^{\infty} \lambda_i^{-1}\lambda_j^{-1} \left(\sum_{k=1}^{\infty} c_{i,k}c_{j,k}\lambda_k^{-1} \right)^2 \right)^{1/2} \\
&= \left(2\left\|\Lambda^{-1/2}C\Lambda^{-1/2}\right\|_{\mathrm{HS}}^4 + 2\left\|(\Lambda^{-1/2}C\Lambda^{-1/2})^2\right\|_{\mathrm{HS}}^2 \right)^{1/2} \\
&= \left(2\|C\|_{\mathcal{H}(K\otimes K)}^4 + 2\left\|C\Lambda^{-1}C\right\|_{\mathcal{H}(K\otimes K)}^2 \right)^{1/2}
\end{aligned}$$

This choice of the reproducing kernel directly effects our bound on the weak variance of the data. It is noteworthy that in the case that the kernel is the identity operator, that the above bound reduces to our original bound from Appendix 3.B.3.

Bibliography

Christophe Abraham, Pierre-André Cornillon, ERIC Matzner-Løber, and Nicolas Molinari. Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics*, 30(3):581–595, 2003.

Robert J Adler. An introduction to continuity, extrema, and related topics for general Gaussian processes. *Lecture Notes-Monograph Series*, 12:i–155, 1990.

Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Inc, 1993.

Sylvain Arlot, Gilles Blanchard, and Etienne Roquain. Some nonasymptotic results on resampling in high dimension, i: confidence regions. *The Annals of Statistics*, 38(1):51–82, 2010.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

John AD Aston, Davide Pigoli, and Shahin Tavakoli. Tests for separability in nonparametric covariance operators of random surfaces. *The Annals of Statistics*, to appear.

Matthew Avery, Yichao Wu, Hao Helen Zhang, and Jiajia Zhang. RKHS-based functional nonparametric regression for sparse and irregular longitudinal data. *Canadian Journal of Statistics*, 42(2):204–216, 2014.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Dominique Bakry and Michel Émery. Hypercontractivité de semi-groupes de diffusion. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 299(15):775–778, 1984.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3: 463–482, 2003.

Peter L Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

James O Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–32, 2003.

Alain Berlinet, Gérard Biau, and Laurent Rouviere. Functional supervised classification with wavelets. In *Annales de l'ISUP*, volume 52, 2008.

Sergei Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

Peter J Bickel and David A Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, pages 1196–1217, 1981.

Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604, 2008a.

Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008b.

Miķelis G Bickis. personal communication, Royal Statistical Society Conference, Manchester, UK, 2016.

Jacob Bien and Rob Tibshirani. *spcov: Sparse Estimation of a Covariance Matrix*, 2012. URL `https://CRAN.R-project.org/package=spcov`. R package version 1.01.

Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

Gilles Blanchard and Nicole Mücke. Parallelizing spectral algorithms for kernel learning. *arXiv preprint arXiv:1610.07487*, 2016.

Sergey Bobkov and Michel Ledoux. Poincaré's inequalities and Talagrand's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107(3):383–400, 1997.

Denis Bosq. *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media, 2012.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. A sharp concentration inequality with applications. *Random Structures & Algorithms*, 16:277–292, 2000.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Olivier Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic inequalities and applications*, pages 213–247. Springer, 2003.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

Alessandra Cabassi and Adam B Kashlak. *fdcov: Analysis of Covariance Operators*, 2016. R package version 1.0.0.

Alessandra Cabassi, Davide Pigoli, Piercesare Secchi, and Patrick A Carter. Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology. *arXiv preprint arXiv:1701.05870*, 2017.

Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

Tony Cai and Ming Yuan. Nonparametric covariance function estimation for functional and longitudinal data. *University of Pennsylvania and Georgia Institute of Technology*, 2010.

Tony Cai and Ming Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012.

George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Chung Chang, Yakuan Chen, and Todd Ogden. Functional data classification: a wavelet approach. *Computational Statistics*, 29(6):1497–1513, 2014.

Corinna Cortes, Marius Kloft, and Mehryar Mohri. Learning kernels using local Rademacher complexity. In *Advances in neural information processing systems*, pages 2760–2768, 2013.

Michael J Daniels and Robert E Kass. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94(448):1254–1263, 1999.

Michael J Daniels and Robert E Kass. Shrinkage estimators for covariance matrices. *Biometrics*, 57(4):1173–1184, 2001.

Aurore Delaigle and Peter Hall. Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):267–286, 2012.

Dipak K Dey and C Srinivasan. Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics*, pages 1581–1591, 1985.

Ian L Dryden, Alexey Koloydenko, and Diwei Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, pages 1102–1123, 2009.

Lutz Dümbgen, Sara A van de Geer, Mark C Veraar, and Jon A Wellner. Nemirovski's inequalities revisited. *American Mathematical Monthly*, 117(2):138–160, 2010.

Bradley Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

Noureddine El-Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, pages 2717–2756, 2008.

Torbjørn Eltoft, Taesu Kim, and Te-Won Lee. On the multivariate Laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, 2006.

Zhou Fan. Confidence regions for infinite-dimensional statistical parameters. *Part III essay in Mathematics, University of Cambridge*, 2011. `http://web.stanford.edu/ zhoufan/PartIIIEssay.pdf`.

Frédéric Ferraty and Philippe Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1):161–173, 2003.

Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4): 707–738, 2015.

Stefan Fremdt, Josef G Steinebach, Lajos Horváth, and Piotr Kokoszka. Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics*, 40(1):138–152, 2013.

Reinhard Furrer and Thomas Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

Evarist Giné and Richard Nickl. Adaptive estimation of a distribution function and its density in sup-norm loss by wavelet and spline projections. *Bernoulli*, 16(4): 1137–1163, 2010a.

Evarist Giné and Richard Nickl. Confidence bands in density estimation. *The Annals of Statistics*, 38(2):1122–1170, 2010b.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12(4):929–989, 1984.

Richard H Glendinning and RA Herbert. Shape classification using smooth principal components. *Pattern recognition letters*, 24(12):2021–2030, 2003.

Nathael Gozlan. Poincaré inequalities and dimension free concentration of measure. In *Annales de l'institut Henri Poincaré (B)*, volume 46, pages 708–739, 2010.

Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed. *Probabilistic methods for algorithmic discrete mathematics*, volume 16. Springer Science & Business Media, 2013.

LR Haff. Empirical Bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, pages 586–597, 1980.

Peter Hall and Mohammad Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):109–126, 2006.

Peter Hall, DS Poskitt, and Brett Presnell. A functional data-analytic approach to signal discrimination. *Technometrics*, 43(1):1–9, 2001.

Peter Hall, Hans-Georg Müller, and Fang Yao. Modelling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):703–723, 2008.

Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, pages 73–102, 1995.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Peter D Hoff. A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):971–992, 2009.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220, 2008.

Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

Arieh Iserles. *A first course in the numerical analysis of differential equations*. Number 44. Cambridge university press, 2009.

Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2): 134–139, 1918.

Gareth M James and Trevor J Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 533–550, 2001.

Ci-Ren Jiang, John AD Aston, and Jane-Ling Wang. Smoothing dynamic positron emission tomography time courses using functional principal components. *NeuroImage*, 47(1):184–193, 2009.

Ci-Ren Jiang, John AD Aston, and Jane-Ling Wang. A functional approach to deconvolve dynamic neuroimaging data. *Journal of the American Statistical Association*, 111(513):1–13, 2016.

Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, pages 295–327, 2001.

Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 2012.

Adam B Kashlak. The frequency of America in America. *Significance*, 13(5):26–29, 2016.

Adam B Kashlak, Eoin Devane, Helge Dietert, and Henry Jackson. Markov models for ocular fixation locations in the presence and absence of colour. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2017.

Gerard Kerkyacharian, Richard Nickl, and Dominique Picard. Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. *Probability Theory and Related Fields*, 153(1-2):363–404, 2012.

Javed Khan, Jun S Wei, Markus Ringner, Lao H Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R Antonescu, Carsten Peterson, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673–679, 2001.

Thierry Klein and Emmanuel Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

Marius Kloft and Gilles Blanchard. The local Rademacher complexity of lp-norm multiple kernel learning. In *Advances in Neural Information Processing Systems*, pages 2438–2446, 2011.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *Information Theory, IEEE Transactions on*, 47(5):1902–1914, 2001.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):4254, 2009.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Michel Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1997.

Michel Ledoux. *The concentration of measure phenomenon*, volume 89. American Mathematical Soc., 2001.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

Paul Levy. Problèmes concrets d'analyse fonctionnelle. *Gauthier-Villars, Paris*, 1951.

Eardi Lila, John AD Aston, and Laura M Sangalli. Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *The Annals of Applied Statistics*, 10(4):1854–1879, 2017.

Karim Lounici and Richard Nickl. Global uniform risk bounds for wavelet deconvolution estimators. *The Annals of Statistics*, 39(1):201–231, 2011.

Kanti V Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

Kanti V Mardia. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 115–128, 1974.

Pascal Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, pages 863–884, 2000.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209: 415–446, 1909.

Vitali D Milman and Gideon Schechtman. *Asymptotic theory of finite dimensional normed spaces: Isoperimetric inequalities in Riemannian manifolds*, volume 1200. Springer, 2009.

Hans-Georg Müller. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32(2):223–240, 2005.

Hans-Georg Müller and Ulrich Stadtmüller. Generalized functional linear models. *Annals of Statistics*, pages 774–805, 2005.

Arkadi Nemirovski. Topics in non-parametric. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

Victor M Panaretos, David Kraus, and John H Maddocks. Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*, 105(490):670–682, 2010.

Dmitriy Panchenko. A note on Talagrand's concentration inequality. *Electronic Communications in Probability*, 6:55–65, 2001.

Dmitry Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *Annals of Probability*, pages 2068–2081, 2003.

Jie Peng and Hans-Georg Müller. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics*, pages 1056–1077, 2008.

Davide Pigoli, John AD Aston, Ian L Dryden, and Piercesare Secchi. Distances and inference for covariance operators. *Biometrika*, page asu008, 2014.

Davide Pigoli, Pantelis Z Hadjipantelis, John S Coleman, and John AD Aston. The analysis of acoustic phonetic data: exploring differences in the spoken romance languages. *arXiv preprint arXiv:1507.07587*, 2015.

Mohsen Pourahmadi. Covariance estimation: the GLM and regularization perspectives. *Statistical Science*, pages 369–387, 2011.

Simeng Qu, Jane-Ling Wang, and Xiao Wang. Optimal estimation for the functional Cox model. *The Annals of Statistics*, 44(4):1708–1738, 2016.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org`.

James O Ramsay and Bernard W Silverman. *Functional data analysis*. New York: Springer, 2005.

WanSoo T Rhee and Michel Talagrand. Martingale inequalities and the jackknife estimate of variance. *Statistics & probability letters*, 4(1):5–6, 1986.

Adam J Rothman. Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740, 2012.

Adam J. Rothman. *PDSCE: Positive definite sparse covariance estimators*, 2013. URL `https://CRAN.R-project.org/package=PDSCE`. R package version 1.2.

Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485): 177–186, 2009.

Walter Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 1987.

Raymond A Ryan. *Introduction to tensor products of Banach spaces*. Springer Science & Business Media, 2013.

Nicoleta Serban, Ana-Maria Staicu, and Raymond J Carroll. Multilevel cross-dependent binary longitudinal data. *Biometrics*, 69(4):903–913, 2013.

James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3): 332–341, 1992.

James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

J Michael Steele. An Efron-Stein inequality for nonsymmetric statistics. *The Annals of Statistics*, pages 753–758, 1986.

J Michael Steele. *Probability theory and combinatorial optimization*, volume 69. Siam, 1997.

Charles Stein. Estimation of a covariance matrix. *Rietz Lecture*, 1975.

Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996a.

Michel Talagrand. A new look at independence. *The Annals of probability*, pages 1–34, 1996b.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Wei Biao Wu and Mohsen Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, pages 831–844, 2003.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100 (470):577–590, 2005.

Dengdeng Yu, Linglong Kong, and Ivan Mizera. Partial functional linear quantile regression for neuroimaging data analysis. *Neurocomputing*, 195:74–87, 2016.

Ming Yuan and T Tony Cai. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

Hongtu Zhu, Jianqing Fan, and Linglong Kong. Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, 109(507):1084–1098, 2014.