Topics in High-Dimensional Data Analysis

By

Junwen Yao

Dissertation

Submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Office of Graduate Studies

of the

University of California

Davis

Approved:

Jane-Ling Wang, Co-Chair

Miles E. Lopes, Co-Chair

Hans-Georg Müller

Committee in Charge

2023

To my family.

Contents

Abstract

Topics in High-Dimensional Data Analysis

With the advancement of technologies, high-dimensional data have become more and more popular. One aspect of these data is that each subject often contains many features and has a possibly complicated intrinsic structure. It has motivated enormous research to tackle these challenges. This dissertation works on three different but related problems. Two of them are rooted in functional data, an extreme case of high-dimensional data, where each subject is a smooth curve and has infinite many values. For the third problem, we analyze the bootstrap method in the context of high-dimensional principal component analysis (PCA).

In the first proportion of this dissertation, we discuss the scalar-on-function linear regression model in the sparse observation case and propose two methods to estimate the linear relationship. This also leads to a new framework for prediction with sparse functional covariates. Rates of convergence in estimation and prediction are established for both approaches.

Next, we look at the emerging field of applying modern deep learning to functional data. We focus on both regression and classification tasks with functional input and finite (possible non-linear) index relations. The new methodology learns to apply parsimonious dimension reduction to functional inputs and focus only on information relevant to the target rather than irrelevant variation in the input function in an end-to-end fashion, removing the manual selection of principal components in classical solutions.

Lastly, we turn to the bootstrap method and analyze how well it can approximate the joint distribution of the leading eigenvalues of the sample covariance matrix in high dimensions and establish non-asymptotic convergence rates of approximation. It is also demonstrated that applying a transformation to the sample eigenvalues prior to bootstrapping can lead to inference benefits.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisors Professor Miles E. Lopes and Professor Jane-Ling Wang. Their deep knowledge and extensive advice on research have helped me work through many obstacles during my study. It is inevitable for a junior researcher like me to make mistakes in the process of research. It is their open-mindedness and patient training that encourage me never to give up to my failures. It was also my fortunate to work in the open and free research environment that they have created for students. The distinct topics I choose to present in this dissertation is a perfect example of how flexible my research can be. Their thinking and attitude towards research have had positive influences on me on both academic and personal level. It is hard to imagine a Ph.D. life without their generous support and guidance.

I also want to thank my qualifying exam and dissertation committee members for their insightful discussion and suggestions, besides Miles and Jane-Ling, including Professor Alexandar Aue, Professor Krishna Balasubramanian, Professor Debashis Paul, Professor Hans-Georg Müller, Professor Wolfgang Polonik, and Professor Qinglan Xia. In addition, I would like to express my special thanks to Professor Jie Peng and three professors from math department, Professor John K. Hunter, Professor Bruno Nachtergaele, and Professor Roland Freund. Professor Peng was my supervisor for my masters degree and offered many helpful suggestions on my academic choices. I met three math professor during my first two years at UC Davis. Their outstanding lectures and perspicacious discussion in real analysis and numerical analysis advanced my understanding of these topics and helped me build a good foundation, which turns out to be extremely useful in my doctoral study.

I was very fortunate to be surrounded by a group of helpful department staff and talented friends, colleagues, and peers. I want to give thanks to Pete, Olga, Cristeta, Sarah, and Andi for running the department smoothly and always being willing to offer help. I am very grateful for knowing Jilei Yang, Minjie Fan, Xiaokang Wang, and Yang Zhou at an early stage of my Ph.D. study. I owe them for deep discussions on academia and career. Chang Shu, Haotian Li and Karry Wong are my friends from math department,

and I thank them for insightful discussions on research problems and being supportive at all times. I give thanks to my excellent colleagues, Xingmei Lou, Yucheng Liu, Tongyi Tang, Satarupa Bhattacharjee and Alvaro Eduardo Gajardo Catald, for inspiring and motivating me. Also, I would like to thank my peers, Lingyou Pang, Yuanyuan Li, Shuting Liao, Xiangbo Mo, Jue Wang, Xiawei Wang, Xiaoliu Wu, Tesi Xiao, Jingwei Xiong, Zitong Zhang, Yuxuan Zhang, Xiner Zhou, Yejiong Zhu for being my companion and making my Ph.D. journey happy and joyful.

Most of all, I thank my family for their unconditional love and support.

Chapter 1

Introduction

1.1 Functional Data Analysis

Functional data is a type of data observed continuously on a compact interval and takes the form as a random function. The statistical analysis of functional data is called functional data analysis (FDA). In an abstract sense, typical examples of functional data are random trajectories over a real-valued closed interval. Assume that the interval is $[a, b]$ with $a < b$.[1] Functional data defined on $[a, b]$ are often denoted as $X_i(t)$, $t \in [a, b]$, and i is the index of a sample trajectory. The subject $X_i(t)$ is intrinsically infinite-dimensional, which is a fundamental property that distinguishes functional data from other data types. In the literature, it is commonly assumed that $X_i(t)$ are smooth functions (i.e., continuously differentiable up to some order), which brings benefits when encountering the challenges posed by high-dimensionality in statistical analysis. In scientific studies and real world application, functional data appear frequently in datasets of air pollution, fMRI scans, growth curves, and sensors like wearable devices.

Except for some ideal settings where functional data are fully observed, they are usually sampled intermittently at discrete time points, e.g., $t_1, \ldots, t_m \in [a, b]$. Depending on the number m of sampling points, functional data are often categorized into dense, sparse or neither. For example, with the advance of technology, a machine records data regularly and intensely with $t_{j+1} - t_j = t_j - t_{j-1}$ and the sampling number m is sufficiently large

[1] Often the domain interval is called time interval, which refers to the random curves observed during a period of time. But time is not the only domain measurement. It can be spatial as well.

often grows with the sample size n. On the other extreme, the case when m is bounded by a finite small integer is called the sparse sampling scheme. Each subject can have its own sampling plan, in other words, they have different number of observations. Although there is no uniform consensus on which case should be considered dense or sparse, Zhang and Wang (2016) developed a unified theory for functional data and criteria characterizing sampling plans based on the average number of each curve's observations. We refer readers to the theorems and discussions in the paper and the references therein.

In the literature, functional data are also considered smooth stochastic processes. In fact, there are two different perspectives on functional data. The first view is to treat functional data as random elements in a Hilbert space, which allows us to develop core statistical properties, such as mean and covariance. The second view analyzes functional data from the stochastic process perspective. Classical results such as Mercer's theorem and Karhunen-Lòeve theorem are useful in this perspective to study the covariance of functional data and related problems. The difference between these two perspectives are subtle and worth clarification. Often not mentioned, a Hilbert space with specific characterization of measurability is necessary to study functional data from both perspectives simultaneously. We refer readers to Chapter 7 in Hsing and Eubank (2015) for a rigorous treatment and more discussions. In the rest chapters, we will always assume that this requirement holds so that tools from both views are available for us to do theoretical analysis.

Following the random element perspective, we define the mean of functional data as $\mu(t) = \mathbb{E}[X_i(t)]$, and then the covariance is computed as $\Gamma(s,t) = \mathbb{E}[(X_i(s)-\mu(s))(X_i(t)-\mu(t))]$. Unlike the finite-dimensional case where mean and covariance can be estimated as sample averages, most of the time, the estimation for functional mean and covariance requires careful handle of observations, as functional trajectories are discretely observed and even contaminated by measurement errors. For dense functional data, we can pre-smooth individual curves and then treat smoothed versions as fully observed data. For sparse functional data, pre-smoothing is infeasible, but the estimation is still possible by borrowing information from neighboring data and across all subjects, such as local poly-

nomial smoother (Fan and Gijbels, 2018; Yao et al., 2005a; Zhang and Wang, 2016). One powerful tool that connects functional data to the finite-dimensional world is functional principal component analysis (FPCA). It works like principal component analysis (PCA) for finite-dimensional data. Under mild regularity conditions, the covariance $\Gamma(s,t)$ can be expressed in a summation of countable eigen-components thanks to Mercer's theorem. To be more specific, a typical decomposition is

$$\Gamma(s,t) \;=\; \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t), \tag{1.1.1}$$

where λ_k are eigenvalues and $\phi_k(t)$ are corresponding eigenvectors of the operator defined by $\Gamma(s,t)$. Applying Karhunen-Lòeve theorem, a random trajectory $X_i(t)$ can be written as a series $X_i(t) = \mu(t) + \sum_{k=1}^{\infty} A_{ik}\phi_k(t)$, where the random scores A_{ik} are uncorrelated and uniquely characterize the properties of $X_i(t)$. Then the functional dimension reduction can be achieved through the approximate process $X_{i,\mathrm{M}}(t) = \mu(t) + \sum_{k=1}^{\mathrm{M}} A_{ik}\phi_k(t)$ for an appropriate integer M. The selected M scores can also be viewed as a vector summary of the original process and used in downstream tasks.

There is a wide range of studies on applications and methodologies in FDA, such as regression (e.g., Brumback and Rice (1998); Cai and Hall (2006); Cardot et al. (1999, 2003); Cardot and Sarda (2005); Chen et al. (2011); Dou et al. (2012); Hall and Horowitz (2007); Hilgert et al. (2013); Hu et al. (2004); James (2002); Malfait and Ramsay (2003); Müller and Stadtmüller (2005); Ramsay and Dalzell (1991); Wang et al. (2010); Yao et al. (2005b); Zhu et al. (2014) and many others), classification (e.g., Araki et al. (2009); Chang et al. (2014); Delaigle and Hall (2013); Ferraty and Vieu (2003); Hall et al. (2001); James (2002); James and Hastie (2001); Leng and Müller (2006b); Matsui et al. (2011); Müller and Stadtmüller (2005); Rincón and Ruiz-Medina (2012); Wang et al. (2007); Zhu et al. (2012, 2010)) and clustering (e.g., Abraham et al. (2003); Chiou (2012); Chiou and Li (2007); Coffey et al. (2014); Garcia-Escudero and Gordaliza (2005); Giacofci et al. (2013); Heinzl and Tutz (2014); Jacques and Preda (2013, 2014); Kayano et al. (2010); Li and Chiou (2011); Peng and Müller (2008); Serban and Wasserman (2005) and others). Furthermore, the development in software focusing on FDA makes these methods accessible to various real problems. For instance, the `R` project has a page dedicated to various packages of

FDA tools (CRAN Task View: Functional Data Analysis), including `fda`, `fdapace`, and many others.

Contributing to the literature, we devote two chapters to two different yet interesting problems in FDA. In Chapter 2, we study the estimation and predictions problems in the functional linear model for sparse functional covariates. While the dense case has been well studied, the sparse case, due to limited observations, is more challenging and not carefully examined. Two methods are introduced, one through normal equation and the other through reproducing kernel Hilbert space, to estimate slope functions in functional linear models. In addition, our work also proposes a solution to an open problem: the prediction task in the same setup. The methodology is accompanied with numerical simulations and real data analysis. In Chapter 3 we explore some possibilities of applying deep learning to FDA. More recently, thanks to rapid advancements in computing, neural networks and deep learning become popular again, and they have shown success in various tasks that are previously considered hard, including computer vision and natural language processing. However, their applications in FDA remain scarce. In this chapter, we introduce a network architecture called Adaptive Functional Neural Networks (AdaFNN), which operates on functional input directly and is trained in an end-to-end manner. We establish the universal approximation theory for this network and empirically illustrate the advantages of this approach over numerous classification/regression tasks with functional data.

1.2 The Bootstrap Method

Statistics is the science of empirical experience. Population quantities are estimated from the limited amount of data collected from the real world. Besides, a fundamental task in statistics is to understand how accurate the estimation is. If we were able to repeat the estimation on newly collected data many times, the estimate would vary. To know how trustworthy the estimate from one sample is, we would like to quantify its uncertainty. Many statistics and probability concepts are developed to serve this purpose. One simple measure is the variance, which gives a quantitative measurement how far an estimate deviates from its true value on average. A more informative tool is confidence interval,

or the more general notion confidence region, for a population quantity, and it contains finer evidence on the estimate's distribution. Unlike many textbook examples where these uncertainty measures can computed analytically, the real world problems are more cruel, and simple formulas seldom exist. In many cases, people rely on asymptotic distributions to make such quantification, based on a beautiful wish that the sample collected is large enough and the sample distribution is close to its asymptotic limit. With the same purpose but in a different direction, the bootstrap methods are developed to tackle these problems directly without access to exact mathematical derivation. They are computationally intense algorithms based on the concept of resampling from empirical data. Since the introduction of the method (Efron, 1979), it has gained great popularity across the subject of statistics, from classical multivariate analysis to high-dimensional problems. (The original bootstrap paper is among the top papers downloaded from Project Euclid.) More recently, it has been successfully applied to study the error estimates and uncertainty in randomized algorithms (Chen and Lopes, 2020; Lopes et al., 2020, 2018, 2019; Lunde et al., 2021).

We illustrate the core of bootstrap in the following example on constructing confidence intervals for scalar parameters. Let X be a random variable with a distribution function $F(\cdot)$, and $\phi(\cdot)$ is a functional defined in the space of X. Given a sample $\{X_1, X_2, \ldots, X_n\}$, of i.i.d. copies of X, we want to estimate the quantity $t = \mathbb{E}[\phi(X)] = \int \phi(x) \mathrm{d}F(x)$. For example, if $\phi(x) = x$, then t is the mean $\mathbb{E}[X]$. If we denote the empirical distribution function as $\widehat{F} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$, where $\delta_{X_i}(\cdot)$ is a unit point mass at X_i, then a sample estimate of t is $\widehat{t} = \int \phi(x) \mathrm{d}\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)$. A level $(1-\alpha)$ $(\alpha \in (0,1))$ confidence interval for t can be constructed as $[\widehat{t} - q_{1-\alpha/2}, \widehat{t} - q_{\alpha/2}]$, where $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are the $\alpha/2$-th and $(1-\alpha/2)$-th quantiles of the random quantity $\widehat{t} - t$ respectively. Since both quantiles are often unknown in practice, they need to be estimated from the data. The bootstrap can be used to achieve this goal. Let $\{X_1^*, \ldots, X_n^*\}$ be a bootstrap sample by sampling with replacement from $\{X_1, \ldots, X_n\}$. The bootstrap distribution is written as $\widehat{F}^*$. Then a bootstrap estimate of t is $\widehat{t}^* = \int \phi(x) \mathrm{d}\widehat{F}^*(x) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i^*)$. Repeating the procedure many times yields multiple realizations of $\widehat{t}^* - \widehat{t}$, whose $\alpha/2$-th and $(1-\alpha/2)$-th quantiles,

written as $\widehat{q}_{\alpha/2}$ and $\widehat{q}_{1-\alpha/2}$, are available. Under certain regularity conditions, we can use $\widehat{q}_{\alpha/2}$ and $\widehat{q}_{1-\alpha/2}$ as estimates for $q_{\alpha/2}$ and $q_{1-\alpha/2}$. So a bootstrapped confidence interval for t is $[\widehat{t} - \widehat{q}_{1-\alpha/2}, \widehat{t} - \widehat{q}_{\alpha/2}]$. This example is only a glimpse of how to use bootstrap procedures to quantify estimation uncertainty. Its practical application is much wider. More generally, bootstrap methods allow us to approximate the distribution of $\widehat{t} - t$ using the bootstrap distribution of $\widehat{t}^* - \widehat{t}$.

The classical theory of bootstrap dates back to 1980s. Bickel and Freedman (1981); Freedman (1981); Singh (1981) are among the first to establish the theoretical foundation for applying bootstrap to standard problems, such as mean estimation and linear regression models. Since then many tools have been developed to understand the bootstrap method and improve its accuracy in various cases. For example, the Edgeworth expansion (Hall, 2013) uses an approximate series to analyze and improve the performance of bootstrap. The empirical processes theory (Vaart and Wellner, 1996) is also helpful for establishing some foundations of bootstrap methods. However, the classical theory studies bootstrap in low-dimensional cases where the sample size n grows to infinity but the data dimension remains constant. Conclusions are often developed in asymptotics even though the method works in finite sample scenarios. It is of theoretical and practical importance to extend the analysis to higher dimensional scenarios and give finite sample guarantees for bootstrap methods. Indeed, there has been research studying bootstrap performance in various high-dimensional cases. To give a few examples, Chernozhukov et al. (2014, 2017a, 2022); Lopes et al. (2020) analyze the performance of bootstrap for the max statistics in high dimensions; others (Bunea and Xiao, 2015; Jung et al., 2018; Koltchinskii et al., 2020; Koltchinskii and Lounici, 2017; Lounici, 2014; Naumov et al., 2019) study bootstrapping statistics related to spectral decomposition and projections; and Han et al. (2018); Lopes et al. (2023) look at bootstrapping the operator norm error for high-dimensional covariance estimation. More recently, Lopes (2022) obtains the near-$1/\sqrt{n}$ rate for the Berry-Esseen bounds for bootstrap approximation in the context of a sum of n random vectors that are p-dimensional (the data dimension p is allowed to grow exponentially in n) and have sub-Gaussian or sub-exponential entries.

Contributing to the research on high dimensional PCA, we provide some theoretical supports for bootstrapping the leading eigenvalues of covariance matrices in high dimensions in Chapter 4. We start with a review on the current state of related research and give an overview of our contributions. Under some reasonable assumptions, we show that the vanilla and transformed bootstrap can achieve a dimension-free rate. As an evidence to the theory, several simulation real data analysis studies are provided to demonstrate the accuracy of the bootstrap and its transformed variants.

Chapter 2

Prediction in Linear Models for Sparse Functional Covariates

Estimation and prediction in functional linear models with scalar response and fully or densely observed functional covariates have been widely studied in the literature. However, the extension to sparsely observed functional covariates remains an open problem. We provide a solution by imputing the sample functional path and substituting the imputed path into the original functional linear model. We show why such a substitution method works and propose two estimation approaches: one through the normal equation, which targets the population quantities, and the other estimates the slope function using a reproducing kernel Hilbert space approach. We establish asymptotic properties of these estimators and their corresponding prediction methods. Numerical performance of the proposed methods is evaluated through simulations and illustrated with data from the Framingham Heart Study.

The prediction in functional linear models with scalar response and fully or densely observed functional covariates has been widely studied in the literature. However, the extension to sparsely observed functional covariates remains an open problem. In this chapter, we provide a solution by imputing the sample functional path and substituting the imputed path into the original functional linear model. We show that with this substitution the regression parameters in the original model can still be consistently estimated and propose a prediction procedure based on the imputed path. Two estimation

approaches are studied: one is done through the normal equation which targets the population quantities, and the other estimates the slope function using a reproducing kernel Hilbert space penalty. The estimated slope function can be further applied in the prediction task. In addition, we establish asymptotic properties of these estimators and the prediction method. Numerical performance of the proposed methods is studied through simulations and illustrated with data from the Framingham Heart Study.

2.1 Introduction

Given a random scalar response Y and its functional predictor $X(t)$, the functional linear model refers to the relationship

$$Y = \alpha + \int_a^b \zeta_0(t)X(t)\mathrm{d}t + \epsilon, \tag{2.1.1}$$

where ϵ is a random noise, $\zeta_0(t)$ is the slope function defined on the closed interval $[a, b]$, and α is a real number. Without loss of generality, we can take $[a, b] = [0, 1]$ and write the integral as $\int_a^b \zeta_0(t)X(t)\mathrm{d}t = \langle \zeta_0, X\rangle_2$. Equation (2.1.1) is called functional linear model (FLM), it has been studied by many (Cai and Hall, 2006; Cardot et al., 1999; Crambes et al., 2009; Hall and Horowitz, 2007; Shin, 2009), including its extension to function-on-function linear model, first introduced in Ramsay and Dalzell (1991) and further explore in Yao et al. (2005b). It can also be viewed as an extension of the traditional linear regression model where the inner product is operated on two vectors in a finite-dimensional Hilbert space.

To carry out the estimation of the unknown slope function ζ_0, some constraints are needed. One option is to assume that ζ_0 falls into the space spanned the eigen-functions of the covariance $\Gamma(s, t) = \mathbb{E}[(X(s) - \mu(s))(X(t) - \mu(t))]$, where $\mu(t) = \mathbb{E}[X(t)]$. Then the slope function satisfies the normal equation: $\int \Gamma(s, t)\zeta_0(s)\mathrm{d}s = \mathrm{cov}(X(t), Y)$. When the covariance $\Gamma(s, t)$ admits the spectral decomposition $\Gamma(s, t) = \sum_{k=1}^{\infty} \lambda_k\phi_k(s)\phi_k(t)$, the slope function satisfies: $\zeta_0 = \sum_{k=1}^{\infty} z_k\phi_k(t)$, which can be estimated through the estimates of z_k and ϕ_k. Such an approach has been extensively studied when the sample path X_i is either fully observed or densely sampled. A caveat is that this approach requires delicate regularization for the infinite-dimensional functional covariate X_i and synchronization of

ζ_0 with the eigenfunctions of $\Gamma(s,t)$ (Cardot et al., 1999; Hall and Horowitz, 2007).

An alternative is to assume that ζ_0 belongs to a function space S. Under this condition, the estimation can be conducted through a constrained least-squares optimization problem:

$$\min_{\alpha\in\mathbb{R},\zeta\in S} \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \langle X_i,\zeta\rangle_2 - \alpha\big)^2, \quad \text{subject to} \quad \text{penalty}(\zeta) \;\le\; C,$$

where $\{(X_i,Y_i)\}_{i=1}^n$ are i.i.d. copies of (X_1,Y_1), and C is some fixed positive constant that regularizes the estimated slope function. Solving this optimization problem is equivalent to solving its penalization form:

$$\min_{\alpha\in\mathbb{R},\zeta\in S} \left\{\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \langle X_i,\zeta\rangle_2 - \alpha\big)^2 + \rho\cdot\text{penalty}(\zeta)\right\}, \tag{2.1.2}$$

where $\rho > 0$ is the tuning parameter corresponding to the constraint constant C. Often the penalty term determines which space the estimated slope function falls into. So, choosing a penalty is equivalent to imposing constraints on ζ. For example, using $\text{penalty}(\zeta) = \|\zeta^{(2)}\|_2^2$ is equivalent to assuming that ζ is second-order differentiable and $\zeta^{(2)}$ has bounded $\mathcal{L}_2$-norm, and it is the same as the assumption that S is a Sobolev space. A more general constraint can be that the function space S is a Reproducing kernel Hilbert space (RKHS) and the penalty is the corresponding RKHS norm (Berrendero et al., 2019; Cai and Yuan, 2012; Crambes et al., 2009; Li and Hsing, 2007; Shin and Lee, 2016; Sun et al., 2018; Yuan and Cai, 2010). Such an approach typically requires fully observed sample paths X_i, but can also be employed to densely sampled X_i. see, e.g., Cardot et al. (2003) for error-free observations and Crambes et al. (2009); Li and Hsing (2007) for noisy observations. Alternatively, one can pre-smooth each individual path to get a "fully observed" function, then apply existing methodology for fully observed functional data. However, the sampling frequency needs to be very intensive in order for the subsequent inference to go through, so this does not always work. In addition, it's not obvious how to assess whether the sampling frequency meets the assumption needed for the pre-smoothing method.

How to estimate ζ_0 when sample curves are only sparsely observed has rarely been explored. Furthermore, when making predictions, all existing approaches encounter the

challenge that a reliable estimate of the regression function, i.e. the index $\langle \zeta_0, X_i \rangle_2$, is difficult to attain for subjects with only a few measurements on $X_i(t)$. Recognizing this challenge, Gajardo et al. (2021) focused on predicting the distributions of a truncated conditional process of $\langle X, \zeta \rangle_2$ under the assumption that X is a Gaussian process. In this paper, we target directly the prediction of $\langle X, \zeta \rangle_2$ based on a new observation. To do so, we first need to resolve the estimation problem of ζ_0.

The estimation procedures based on normal equation should still work even when individual paths cannot be densely sampled, as they target population parameters, such as covariance and cross-covariance. We provide supporting theory for the normal equation approach in Section 2.4.1. The estimation is more challenging for the RKHS approach because the estimation of $\zeta_0(t)$ involves the evaluation of the $\langle \zeta_0, X_i \rangle_2$ as triggered by the formulation (2.1.2). For instance, when the penalty term is the RKHS norm of a candidate function ζ, we need to find $\alpha \in \mathbb{R}$ and $\zeta \in \mathcal{H}(K)$ that minimize the loss

$$\ell\big(\alpha, \zeta; \{X_i\}_{i=1}^n\big) \;=\; \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \langle \zeta, X_i \rangle_2 - \alpha\big)^2 + \rho \, \big\|\zeta\big\|^2_{\mathcal{H}(K)}. \tag{2.1.3}$$

In order to evaluate the loss ℓ we need to have good approximations of $\langle \zeta, X_i \rangle_2$ which is not feasible for sparsely observed covariates. Thus, sparse functional covariates bring dual challenges to the RKHS approach in functional linear models (2.1.1), both in terms of estimation and prediction. We address these challenges for the RKHS approach in Section 2.4.2 and show that both consistent estimation for the slope function $\zeta_0(t)$ and prediction are still possible, for sparsely observed covariates $X_i(t)$, possibly contaminated with noise. We also aim at prediction for both approaches In Section 2.5 when only sparsely measured functional covariates are available.

Our approach is simple and aims at imputing each of the unobserved process $X_i(t)$ by $\tilde{X}_i(t)$ so that the integral $\langle \zeta_0, X_i \rangle_2$ is replaced by $\langle \zeta_0, \tilde{X}_i \rangle_2$, which can be evaluated effectively. However, in order for this to work, the linear structure between Y_i and $\tilde{X}_i(t)$ must be retained by the same slope function $\zeta_0(t)$ and intercept α_0, so the surrogate model should be

$$Y_i \;=\; \alpha + \int_0^1 \zeta_0(t) \tilde{X}_i(t) \mathrm{d}t + e_i, \tag{2.1.4}$$

where Y, α_0 and $\zeta_0(t)$ are the same as in the original model (2.1.1), but e_i is a new random noise variable that replaces ϵ_i. One approach to impute $X_i(t)$ is to use the Principal Analysis by Conditional Estimation (PACE) approach in Yao et al. (2005a) on the observed path points, i.e., for each $i = 1, \ldots, n$, $\tilde{X}_i(t) = \mathbb{E}[X_i(t) \,|\, \mathbf{W}_i(\mathbf{T}_i)]$, where $\mathbf{T}_i = (T_{i1}, \ldots, T_{iN_i})$ is the vector of sampling schedule for subject i and $\mathbf{W}_i(\mathbf{T}_i) = (W_i(T_{i1}), \ldots, W_i(T_{iN_i}))$ is the corresponding vector of the noisy observations of $X_i(t)$ at different time points. Note that both the sampling schedule and number of measurements N_i may vary with subjects. Indeed such an imputation approach has been explored in the literature for the RKHS approach. Avery et al. (2014) numerically demonstrates that this method works reasonably well in both simulations and real datasets, but no theoretical justification has been offered. We fill this theoretical gap and show that the imputation approach also works in a more general setting than previously described in Yao et al. (2005a). At first glance such an imputation approach should not work for sparsely sampled $X_i(t)$ because $\tilde{X}_i(t)$ cannot approximate $X_i(t)$ consistently, so there will be non-ignorable prediction errors in the imputed process, which normally would trigger a bias in the resulting estimators. Intriguingly, we show in Theorem 1 that this particular imputation does not induce bias into the regression estimates.

In practice, the imputed process $\tilde{X}_i(t)$ cannot be fully observed either and needs to be approximated from the data. There are two sources of approximation errors, one from approximating $\tilde{X}_i(t)$ by a finite dimensional projection and the other from estimating this finite dimensional projection. We show that under some weak assumptions the estimation of $\tilde{X}_i(t)$ does not affect the estimation of $\zeta_0(t)$ much and establish the consistency of the proposed estimator in the $\mathcal{L}_2$-norm for sparsely observed functional covariates. For the prediction problem, we propose to replace the new covariate $X^*(t)$ with an estimate of its imputed process $\tilde{X}^*(t)$. Let $\widehat{X}^*(t)$ denote its estimate, then the prediction for a new subject with covariate $X^*(t)$ is thus $\langle \zeta_0, \widehat{X}^* \rangle_2$. Theoretical properties of the proposed prediction method are studied in Section 2.5.

To summarize, we consider two different approaches in scalar-on-function linear models for sparsely observed functional covariates. A main novelty of our approach is the

discovery that through a Berkson type of measurement errors, the original functional covariate X_i and the imputed covariate process $\tilde{X}_i$ share the same regression parameters. For each approach, we consider both the estimation and prediction problems. A summary of the main contributions of this paper follows.

1. We resolve a long-standing open problem to make predictions for functional linear model based on sparsely observed data.

2. We provide theoretical support to estimate the regression coefficient functions in sparse case under both the normal equation and RKHS framework.

Section 2.2 contains the model setup and estimation procedures. In Section 2.3 we discuss the imputation method. The estimation of the slope function under the two frameworks is presented in Section 2.4. Numerical implementation of the estimators is provided in Section 2.6. Simulated experiments and data analysis are also conducted to demonstrate the proposed method in the same section. Lastly, proofs are in Section **??**.

2.2 Sampling Plan and Surrogate Model

We will use $\mathcal{L}_2(\mathcal{T})$ and $\mathcal{L}_2(\mathcal{T}^2)$ to denote the space of square-integrable functions defined on $\mathcal{T}$ and $\mathcal{T}\times\mathcal{T}$. If $R(s,t)$ is a function in $\mathcal{L}_2(\mathcal{T}^2)$, the linear operator $L_R : \mathcal{L}_2(\mathcal{T}) \mapsto \mathcal{L}_2(\mathcal{T})$ is defined according to $\mathcal{L}_R(f) = \int_{\mathcal{T}} R(s,t)f(s)\mathrm{d}s$, and $\|L_R\|_{\mathrm{op}}$ denotes its operator norm. For the sake of brevity, we assume $\mathcal{T} = [0,1]$ and write $\int_{\mathcal{T}} f_1(t)f_2(t)\mathrm{d}t = \langle f_1, f_2\rangle_2$ and $\int_{\mathcal{T}^2} h_1(s,t)h_2(s,t)\mathrm{d}s\mathrm{d}t = \langle h_1, h_2\rangle_2$ whenever the context is clear. Lastly, we use $\|f\|_2$ to denote the $\mathcal{L}_2$-norm of f.

Assume that $X_1(t)$ is a mean-square continuous process in $\mathcal{L}_2(\mathcal{T})$ with mean $\mu(t) = \mathbb{E}[X_1(t)]$ and covariance $\Gamma(s,t) = \mathrm{cov}(X_1(s), X_1(t))$, and ϵ_1 is a random variable with mean 0 and finite variance σ^2. The sample $\{(X_i, Y_i)\}_{i=1}^n$ are i.i.d. copies of (X_1, Y_1). We consider the setting where the full process $X_i(t)$ cannot be observed continuously as measurements can only be taken at random discrete time points. In addition, the discrete observations may contain noise. Assumptions on the sampling plan with measurement errors are summarized in Assumption 1 below.

Assumption 1 (Sampling Plan)**.**

(A1.1) The integer-valued random variables $\{N_i\}_{i=1}^n$ *are i.i.d., and there exists a constant* $c > 0$ *such that* $\mathbb{P}(\sup_i N_i \leq c) = 1$.

(A1.2) The i*th sample path* $X_i(t)$ *is sparsely sampled at* N_i *time points*

$$\mathbf{T}_i \;=\; (T_{i1}, \ldots, T_{iN_i})^\top.$$

The observation vector is $(X_i(T_{i1}), \ldots, X_i(T_{iN_i}))^\top$*. Each* $X_i(T_{ij})$ *is contaminated by a noise variable* η_{ij} *and the observed noisy data are*

$$W_i(T_{ij}) \;=\; X_i(T_{ij}) + \eta_{ij},$$

where $\{\eta_{ij}\}_{j=1}^{N_i}$ *are measurement errors with mean zero, finite variance* σ_η^2*, and are independent across* i*. We write*

$$\mathbf{W}_i(\mathbf{T}_i) \;=\; (W_i(T_{i1}), \ldots, W_i(T_{iN_i}))^\top. \tag{2.2.1}$$

(A1.3) For the i*th subject, the time points* $\{T_{ij}\}_{j=1}^{N_i}$ *form an i.i.d. sample from a continuous random variable* T *defined on* $\mathcal{T}$*. The density of* T *is bounded above and below from zero. Furthermore, the second order derivative of the density is also bounded.*

Remark. Functional data that satisfy Assumptions (A1.1) and (A1.2) are usually referred to as sparse functional data (Yao et al., 2005a), which occurs frequently in longitudinal studies. They are harder to handle than fully or densely observed functional data due to the irregular sampling design in Assumption (A1.1) (Wang et al., 2016). □

Since the mean function $\mu(t)$ can be estimated at a faster rates than the covariance $\Gamma(s,t)$ (Yao et al., 2005a; Zhang and Wang, 2016) and the slope function $\zeta_0(t)$ in the sparse case, it is customary to assume $\mu(t) \equiv 0$ and consider the model

$$\begin{aligned} Y_i \;&=\; \int_0^1 \zeta_0(t) X_i(t) \mathrm{d}t + \epsilon_i, \\ &=\; \langle \zeta_0, X_i \rangle_2 + \epsilon_i, \end{aligned} \tag{2.2.2}$$

where both Y_i and $X_i(t)$ are centered, and $\mathcal{T} = [0, 1]$. The results in this paper also apply to the general case when $\mu(t) \neq 0$ with additional technical details.

While the challenges for estimation using sparse functional data have largely been resolved for estimating the mean and covariance function (Yao et al., 2005a; Zhang and Wang, 2016), predictions in functional linear model remains unresolved due to the lack of data to evaluate the index $\langle \zeta_0, X_i \rangle_2$ in (2.1.1). We propose to use the imputed process $\tilde{X}_i = \mathbb{E}[X_i | \mathbf{W}_i(\mathbf{T}_i)]$ to replace the unobservable X_i and re-write (2.2.2) as

$$\begin{aligned} Y_i &= \langle \zeta_0, \tilde{X}_i \rangle_2 + \langle \zeta_0, X_i - \tilde{X}_i \rangle_2 + \epsilon_i, \\ &= \langle \zeta_0, \tilde{X}_i \rangle_2 + e_i, \end{aligned} \tag{2.2.3}$$

where $e_i = \langle \zeta_0, X_i - \tilde{X}_i \rangle_2 + \epsilon_i$. Theorem 1 confirms the validity of the surrogate model (2.2.3) based on the imputed processes.

Theorem 1. *Suppose Assumption 1 holds. The new noise e_i is uncorrelated with $\tilde{X}_i$, and so model* (2.2.3) *is a functional linear model with the same slope function $\zeta_0(t)$ as model* (2.2.2)*.*

Remark. An interesting phenomenon emerges from Theorem 1 as what we did is to replace the original covariate $X_i(t)$ in (2.2.2) with the imputed $\tilde{X}(t)$, which results in an error-in-variable model as the error $(X_i - \tilde{X}_i)(t)$ is non-ignorable in size for sparsely observed $X(t)$. Usually, this will induce a bias but we do not have such a bias issue because we have a special type of error-in-variable of the Berkson type, where the error is uncorrelated with the (imputed) covariate (Berkson, 1950). While it seems that a more efficient imputation method would be to replace $\mathbb{E}[X_i(t)|\mathbf{W}_i(\mathbf{T}_i)]$ in (2.2.3) with $\mathbb{E}[X_i(t)|\mathbf{W}_i(\mathbf{T}_i), Y_i]$ as the latter utilizes the additional information on Y_i. However, this would lead to residuals that are correlated with the covariates, hence this strategy will not work. □

With the help of Theorem 1, we can use the surrogate model (2.2.3) to predict the target $\langle \zeta_0, \tilde{X}_i \rangle_2$. However, the imputed process $\tilde{X}_i(t)$ is unobservable in practice, so we

need to find a good approximation of $\tilde{X}_i(t)$. Luckily, this is feasible and will be elaborated in the next section.

2.3 Imputation

It is common to assume that the covariance $\Gamma(s,t)$ admits the expansion

$$\Gamma(s,t) \;=\; \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t), \tag{2.3.1}$$

where $\{\lambda_k\}$ are eigenvalues of $\Gamma(s,t)$ and satisfy $\lambda_1 > \lambda_2 > \lambda_3 > \ldots$, and $\{\phi_k(t)\}$ are the corresponding eigenfunctions. Hence, for each sample trajectory $X_i(t)$, we have the following Karhunen-Loève expansion

$$X_i(t) \;=\; \mu(t) + \sum_{k=1}^{\infty} A_{ik}\phi_k(t), \tag{2.3.2}$$

where $A_{ik} = \langle X_i, \phi_k \rangle_2$, $\mathbb{E}[A_{ik}] = 0$ and $\mathrm{var}(A_{ik}) = \lambda_k$.

Through conditioning and by assuming that $\mu(t) = 0$, the imputed process is

$$\tilde{X}_i(t) \;=\; \sum_{k=1}^{\infty} \tilde{A}_{ik}\phi_k(t), \tag{2.3.3}$$

where $\tilde{A}_{ik}$ is defined as $\tilde{A}_{ik} = \mathbb{E}[A_{ik}|\mathbf{W}_i(\mathbf{T}_i)]$. Below we introduce some regularity conditions on the process.

Assumption 2 (Regularity on Functional Data)**.**

(A2.1) The process satisfies $\left\|\mathbb{E}[X_1^4(t)]\right\|_1 < \infty$.

(A2.2) Let $\alpha_1 > 1$ *be a fixed constant not depending on* n*. For* $k \geq 1$, $\lambda_k \asymp k^{-\alpha_1}$.

(A2.3) Let $\delta_k = \min_{1\leq l\leq k}(\lambda_l - \lambda_{l+1})$*. There exist fixed constants* $c > 0$ *and* $\gamma > \alpha_1$ *such that* $\delta_k \leq k^{-\gamma}/c$.

2.3.1 Truncation of the imputed process

Since the $\tilde{X}_i$ is defined via an infinite series, we choose a positive integer $\mathrm{M}_\circ$, which increases to infinity as $n \to \infty$, and truncate $\tilde{X}_i(t)$ to

$$\tilde{X}_{i,\mathrm{M}_\circ}(t) \;=\; \sum_{k=1}^{\mathrm{M}_\circ} \tilde{A}_{ik}\phi_k(t). \tag{2.3.4}$$

Proposition 1. *Suppose that Assumption 2 holds, then*

$$\left\|\tilde{X}_i(t) - \tilde{X}_{i,\mathrm{M}_\circ}(t)\right\|_2 \;=\; O_p\big(\mathrm{M}_\circ{}^{-(\alpha_1-1)/2}\big).$$

2.3.2 Estimation of the truncated imputed process

The PACE method aims at estimating $\tilde{A}_{ik}$ through its best linear predictor.

Assumption 3 (Conditional score)**.**

(A3) $\mathbb{E}[A_{ik}|\mathbf{W}_i(\mathbf{T}_i)]$ *is a linear function of* $\mathbf{W}_i(\mathbf{T}_i)$*, i.e. there exists a scalar* a_{ik} *and a vector* b_{ik} *such that*

$$\mathbb{E}[A_{ik}|\mathbf{W}_i(\mathbf{T}_i)] = a_{ik} + b_{ik}^\top \mathbf{W}_i(\mathbf{T}_i).$$

Remark. In the supplement, we provide an example to show that Assumption (A3) holds beyond the Gaussian assumption. In particular, Assumption (A3) holds when X_i and the measurement errors η_{ij} are jointly elliptical (see Bali and Boente (2009); Boente et al. (2014) for the definition of elliptical distributions). Intriguingly, the jointly elliptical assumption may not hold for multivariate t distributions in the presence of measurement errors. Therefore, we propose an alternative assumption and further discussion in the supplement (Section A.11) that replaces $\mathbf{W}_i(\mathbf{T}_i)$ by $\mathbf{X}_i(\mathbf{T}_i)$ in Assumption (A3). This alternative assumption, which only assume that $\mathbb{E}[A_{ik}|\mathbf{X}_i(\mathbf{T}_i)]$ is a linear function of $\mathbf{X}_i(\mathbf{T}_i)$, facilitates consistent estimation of the slope function ζ_0 and only requires that X_i is elliptical.

The fact that all A_{ij}, η_{ij}, $X_i(t)$ are mean zero implies that the scalars a_{ik} are all zero, and the vector b_{ik} is the weight in the least square solutions, which leads to

$$\begin{aligned}\tilde{A}_{ik} \;&=\; \mathrm{cov}(A_{ik}, \mathbf{W}_i(\mathbf{T}_i))^\top \big[\mathrm{cov}(\mathbf{W}_i(\mathbf{T}_i))\big]^{-1}\mathbf{W}_i(\mathbf{T}_i),\\ &=\; \lambda_k \boldsymbol{\phi}_{ik}^\top \boldsymbol{\Sigma}^{-1}_{\mathbf{W}_i(\mathbf{T}_i)}\mathbf{W}_i(\mathbf{T}_i), \qquad (2.3.5)\end{aligned}$$

where $\boldsymbol{\phi}_{1k} = \big(\phi_k(T_{i1}), \ldots, \phi_k(T_{iN_i})\big)^\top$ and $\boldsymbol{\Sigma}_{\mathbf{W}_i(\mathbf{T}_i)}$ denotes the covariance matrix of $\mathbf{W}_i(\mathbf{T}_i)$. Denote $\widehat{\lambda}_k$, $\widehat{\phi}_k$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{W}_i(\mathbf{T}_i)}$ as the estimates of λ_k, ϕ_k, $\boldsymbol{\Sigma}_{\mathbf{W}_i(\mathbf{T}_i)}$. The relation (2.3.5) thus lead to the estimate $\widehat{A}_{ik}$ for $\tilde{A}_{1ik}$, with

$$\widehat{A}_{ik} \;=\; \widehat{\lambda}_k \widehat{\boldsymbol{\phi}}_{ik}^\top \widehat{\boldsymbol{\Sigma}}^{-1}_{\mathbf{W}_i(\mathbf{T}_i)}\mathbf{W}_i(\mathbf{T}_i). \qquad (2.3.6)$$

Consequently, the truncated imputed process $\tilde{X}_{i,\mathrm{M}_\circ}(t)$ will be approximated by

$$\widehat{X}_{i,\mathrm{M}_\circ}(t) \;=\; \sum_{k=1}^{\mathrm{M}_\circ} \widehat{A}_{ik}\widehat{\phi}_k(t). \tag{2.3.7}$$

Proposition 2. *Let the equal-weight-per-observation scheme ((2.3) and (2.4) in Zhang and Wang (2016)) be the weighing scheme used to estimate the mean $\mu(t)$ and covariance $\Gamma(s,t)$. Suppose that Assumptions 1, 2, and 3 hold. Then*

$$\big\|\tilde{X}_{i,\mathrm{M}_\circ}(t) - \widehat{X}_{i,\mathrm{M}_\circ}(t)\big\|_2 \;=\; O_p\big(\mathrm{M}_\circ{}^{\gamma+1}(\log(n)/n)^{1/3}\big).$$

Remark. The proof of Proposition 2 is based on the bound on sample eigenfunctions, $\mathbb{E}[\|\widehat{\phi}_j - \phi_j\|_2^2] \lesssim \eta_j^{-2}\mathbb{E}[\|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}}^2]$, which is commonly applied in the literature and can be found in Bosq (2000); Hsing and Eubank (2015); Li and Hsing (2010). Recently, Zhou et al. (2022) pointed out that this bound can be sharpened by replacing the eigen-gap assumption (A2.3) with a stronger assumption and by further imposing assumptions on the first two derivatives of the eigenfunctions. Under their assumptions we are able to improve the estimation error in Proposition 2 to $\|\tilde{X}_{i,\mathrm{M}_\circ}(t) - \widehat{X}_{i,\mathrm{M}_\circ}(t)\|_2 = O_p(\mathrm{M}_\circ{}^1(\log(n)/n)^{1/3})$. Consequently, the new bounds on $\widehat{\phi}_j$ and $\tilde{X}_{i,\mathrm{M}_\circ}$ can also be applied to improving Theorem 2 and Theorem 3 with considerable technical tedium. However, their approach requires fine tuning of the smoothing parameter for eigenfunctions. Details are thus omitted in this paper.

2.4 Estimation of ζ_0

The previous section establishes the validity of making predictions through the imputed process of conditional expectation and how to estimate the conditional processes. To predict the response from a new sparsely sampled functional trajectory, we need to know or be able to estimate the slope function ζ_0. This section is devoted to two approaches to estimate ζ_0, one through the normal equation and the other through the RKHS approach.

2.4.1 Estimation of ζ_0 through the normal equation

Assume that the slope function is in the space spanned by $\{\phi_k(t)\}_{k=1}^{\infty}$. It the follows from the spectral decomposition of $\Gamma(s,t)$ in (2.3.1) that ζ_0 can be represented as

$$\zeta_0 \;=\; \sum_{k=1}^{\infty} z_k \phi_k(t), \tag{2.4.1}$$

where $z_k = \langle \zeta_0, \phi_k \rangle_2$. We make the following assumption on z_k.

Assumption 4 (Decay rate of z_k)**.**

There exists fixed constants $c > 0$ and $\beta_1 > 0$ such that $|z_k| \leq c\, k^{-\beta_1}$.

Then ζ_0 also satisfies the following population equality:

$$\begin{aligned} \mathbb{E}[Y_i X_i(t)] \;&=\; \mathbb{E}[\langle \zeta_0, X_i \rangle_2 X_i(t)], \\ &=\; \langle \zeta_0(s), \Gamma(t,s) \rangle_2, \\ &=\; \sum_{k=1}^{\infty} \lambda_k z_k \phi_k(t). \end{aligned} \tag{2.4.2}$$

Write $g(t) = \mathbb{E}[Y_1 X_1(t)]$, then it admits the series expansion $g(t) = \sum_{k=1}^{\infty} g_k \phi_k(t)$ with $g_k = z_k \lambda_k$. Based on (2.4.1) and (2.4.2), an estimator can be found through

$$\widehat{\zeta}_{\mathrm{M}}(t) \;=\; \sum_{k=1}^{\mathrm{M}} \frac{\widehat{g}_k}{\widehat{\lambda}_k} \widehat{\phi}_k(t), \tag{2.4.3}$$

where $\widehat{g}_k$, $\widehat{\lambda}_k$, and $\widehat{\phi}_k(t)$ are empirical estimates from a random sample. In particular, $\widehat{\lambda}_k$ and $\widehat{\phi}_k(t)$ can be estimated from a covariance estimate, e.g., the one in Yao et al. (2005a,b); Zhang and Wang (2016). Also note that

$$g_k = \mathbb{E}[Y_1 A_{1k}] = \mathbb{E}[Y_1 \langle X_1, \phi_k \rangle_2] = \langle \mathbb{E}[Y_1 X_1(t)], \phi_k(t) \rangle_2, \tag{2.4.4}$$

where $g(t) = \mathbb{E}[Y X_1(t)]$ can be estimated through a local linear polynomial smoother similar to the estimation of $\mathbb{E}[X_1(t)]$ (Zhang and Wang, 2016). Therefore, g_k is estimated by $\widehat{g}_k = \langle \widehat{g}, \widehat{\phi}_k \rangle_2$.

Theorem 2. *Suppose Assumptions 1, 2, and 4 hold. Then*

$$\big\|\widehat{\zeta}_{\mathrm{M}}(t)-\zeta_0(t)\big\|_2^2 \;=\; O_p\big(\mathrm{M}^{2\alpha_1+1}r_1^2+\mathrm{M}^{2(\alpha_1+\gamma)+1}r_2^2+\mathrm{M}^{-2\beta_1+1}\big),$$

where $r_1=(\log(n)/n)^{2/5}$ *and* $r_2=(\log(n)/n)^{1/3}$. *With the choice*

$$\mathrm{M} \;=\; (n/\log(n))^{2/(5(\alpha_1+\beta_1))},$$

we obtain

$$\big\|\widehat{\zeta}_{\mathrm{M}}(t)-\zeta_0(t)\big\|_2 \;=\; O_p\Big(\big(\log(n)/n\big)^{\frac{1}{3}-\frac{1}{5}\frac{2(\alpha_1+\gamma)+1}{\alpha_1+\beta_1}}\Big).$$

2.4.2 Estimation of ζ_0 in RKHS

For the normal equation approach above, we adopted the conventional approach of using a local linear smoother to estimate the unknown components. Although a different smoother, such as RKHS could also be adopted in principle, the theory would not be comparable as stronger assumptions are typically imposed on the mean and covariance estimation of functional data (Cai and Yuan, 2010, 2011). We thus resorted to the common approach in the literature to assume that ζ_0 belongs to an RKHS $\mathcal{H}(K)$ with kernel $K(s,t)$.

If $X_i(t)$ are fully observed, the RKHS estimator of the slope function can be obtained through

$$\widehat{\zeta}_{\rho}(t) \;=\; \mathrm{argmin}_{\zeta\in\mathcal{H}(K)}\bigg\{\frac{1}{n}\sum_{i=1}^{n}\big(Y_i-\langle\zeta,X_i\rangle_2\big)^2+\rho\,\big\|\zeta\big\|_{\mathcal{H}(K)}^2\bigg\}. \tag{2.4.5}$$

The sparse sampling assumption deprives the index $\langle\zeta,X_i\rangle_2$ a good approximation. Our proposal is to replace X_i by an imputed process $\tilde{X}_i$, which can be estimated well by $\widehat{X}_{i,\mathrm{M}_\circ}(t)$ as shown in Section 2.3. The the slope function can be estimated through the following penalized regression:

$$\widehat{\zeta}_{\rho,\mathrm{M}_\circ}(t) \;=\; \mathrm{argmin}_{\zeta\in\mathcal{H}(K)}\bigg\{\frac{1}{n}\sum_{i=1}^{n}\big(Y_i-\langle\zeta,\widehat{X}_{i,\mathrm{M}_\circ}\rangle_2\big)^2+\rho\,\big\|\zeta\big\|_{\mathcal{H}(K)}^2\bigg\}. \tag{2.4.6}$$

The rest of the section establishes the convergence rate of $\widehat{\zeta}_{\rho,\mathrm{M}_\circ}$.

It is well known, e.g. from Theorem 7.6.4 of Hsing and Eubank (2015), that $\mathcal{H}(K)$ has the property $\mathcal{H}(K)=L_{K^{1/2}}(\mathcal{L}_2[0,1])$, where the operator L_K satisfies $L_K(f)=$

$L_{K^{1/2}}\big(L_{K^{1/2}}(f)\big)$. Hence, there exists a unique element $\mathrm{f}_0 \in \mathcal{L}_2[0,1]/\ker(L_{K^{1/2}})$ such that the true slope function $\zeta_0(t)$ can be expressed as $\zeta_0 = L_{K^{1/2}}(\mathrm{f}_0)$. Denote $\tilde{\Gamma}(s,t) = \mathrm{cov}(\tilde{X}_1(s), \tilde{X}_1(t))$ as the covariance of the imupted process and define the bivariate kernel function $R(s,t)$ so that it satisfies $L_R(f) = L_{K^{1/2}}(L_{\tilde{\Gamma}}(L_{K^{1/2}}(f)))$. Then L_R is compact, and $R(s,t)$ admits the following spectral decomposition

$$R(s,t) \;=\; \sum_{k=1}^{\infty} \theta_k \varphi_k(s)\varphi_k(t),$$

where $\{(\theta_k, \varphi_k(t))\}_{k=1}^{\infty}$ are the eigen-pairs of $R(s,t)$ and $\theta_1 > \theta_2 > \cdots \geq 0$. The function f_0 can be expressed as

$$\mathrm{f}_0(t) \;=\; \sum_{k=1}^{\infty} f_k \varphi_k(t). \tag{2.4.7}$$

The following assumption imposes regularity conditions on θ_k and f_k.

Assumption 5.

(A5.1) There exist constants $c > 0$ and $\beta_2 > 1/2$, not depending on n, such that $|f_k| \leq c\, k^{-\beta_2}$.

(A5.2) There exists $\alpha_2 > 1$, not depending on n, such that $\theta_k \asymp k^{-\alpha_2}$.

Theorem 3. *Under assumptions 1, 2, 3, and 5, we have*

$$\big\|\widehat{\zeta}_{\rho,\mathrm{M}_\circ}(t) - \zeta_0(t)\big\|_2 \;=\; O_p\big(\rho^{\nu} + \rho^{-2}\big(\mathrm{M}_\circ{}^{-(\alpha_1-1)/2} + \mathrm{M}_\circ{}^{\gamma+1}(\log(n)/n)^{1/3}\big)\big),$$

where $\nu = \frac{(2\beta_2-1)/\alpha_2}{2+(2\beta_2-1)/\alpha_2} \in (0,1)$. With the choices

$$\mathrm{M}_\circ = (n/\log(n))^{1/(3(\alpha_1+1)/2+3\gamma)}$$

and

$$\rho = ((\alpha_1-1)/2)/(3(2+\nu)((\alpha_1+1)/2+\gamma)),$$

we have

$$\|\widehat{\zeta}_{\rho,\mathrm{M}_\circ}(t) - \zeta_0(t)\big\|_2 \;=\; O_p\Big(\big(\log(n)/n\big)^{\frac{1}{3}\frac{\nu}{2+\nu}\frac{(\alpha_1-1)/2}{\gamma+(\alpha_1+1)/2}}\Big).$$

2.5 Prediction

Denote $\widehat{\zeta}$ as a generic estimator for ζ_0. For the observation vector $\mathbf{W}^*(\mathbf{T}^*)$ of a new trajectory X^* that is independent of the sample, written as $(X^*(T_1^*), \ldots, X^*(T_N^*))$, its truncated estimate using (2.3.7) is $\widehat{X}^*_{\mathrm{M}_\circ} = \sum_{k=1}^{\mathrm{M}_\circ} \widehat{A}_k^* \widehat{\phi}_k(t)$, where the eigen-components $\widehat{\phi}_k(t)$ are estimated from the existing sample, and the scores $\widehat{A}_k^*$ are calculated according to (2.3.6) using the new observed vector $\mathbf{W}^*(\mathbf{T}^*)$. Then $\langle \widehat{\zeta}, \widehat{X}^*_{\mathrm{M}_\circ} \rangle_2$ can be used to predict $\langle \zeta, \tilde{X}^* \rangle_2$, where $\tilde{X}^* = \mathbb{E}[X^* | \mathbf{W}^*(\mathbf{T}^*)]$ is the imputed process of X^* Our analysis will focus on the prediction error $\mathbb{E}[(\langle \widehat{\zeta}, \widehat{X}^*_{\mathrm{M}_\circ} \rangle_2 - \langle \zeta_0, \tilde{X}^* \rangle_2)^2 \,|\, \widehat{\zeta}\,]$ as we have replaced the linear regression model (2.1.1) with the surrogate model (2.1.4) based on Theorem 1. Theorem 4 establishes the oracle rate for prediction when the slope function is known. Then we show in Theorem 5 the prediction rate when a generic ζ_0 estimator is used.

Theorem 4. *Assume that the slope function ζ_0 is known. Let X^* denote a new functional trajectory, and its conditional process is $\tilde{X}^* = \mathbb{E}[X^* | \mathbf{W}^*(\mathbf{T}^*)]$. Then,*

$$\mathbb{E}\big[\big(\langle \zeta_0, \tilde{X}^* \rangle_2 - \langle \zeta_0, \widehat{X}^*_{\mathrm{M}_\circ} \rangle_2\big)^2\big] \;=\; O_p\Big(\big(\log(n)/n\big)^{\frac{2}{3}\frac{\alpha_1 - 1}{\alpha_1 + 2\gamma + 1}}\Big).$$

The next theorem quantify the prediction error rate when ζ_0 is estimated.

Theorem 5. *Suppose that an estimator $\widehat{\zeta}$ has the asymptotic property:*

$$\|\widehat{\zeta} - \zeta_0\|_2 \;=\; O_p(\delta).$$

Then using this estimator in replacement of ζ_0 in prediction, we have

$$\mathbb{E}\big[\big(\langle \zeta_0, \tilde{X}^* \rangle_2 - \langle \widehat{\zeta}, \widehat{X}^*_{\mathrm{M}_\circ} \rangle_2\big)^2\big] \;=\; O_p\big(\,\delta^2 + {\mathrm{M}_\circ}^{-\alpha_1 + 1} + {\mathrm{M}_\circ}^{2(\gamma+1)}\big(\log(n)/n\big)^{2/3}\,\big).$$

We note that if the goal is prediction, the estimate with the optimal rate of convergence for the regression function ζ_0 does not lead to the optimal rate for prediction. This is because optimal prediction involves some level of over-fitting. Fortunately, given a particular method, it is possible to determine which estimation rate will lead to the optimal prediction rate. We demonstrate this for both the normal equation (Theorem 6) and the RKHS (Theorem 7) methods.

Theorem 6 (Prediction using normal equation)**.** *Suppose that Assumptions 1, 2 and 4 hold. If the normal equation is used to estimate ζ_0, then*

$$\mathbb{E}\big[\big(\langle\zeta_0, \tilde{X}^*\rangle_2 - \langle\widehat{\zeta}_{\mathrm{M}}, \widehat{X}^*_{\mathrm{M}_\circ}\rangle_2\big)^2\big] \;=\; O_p\big((\log(n)/n)^{\frac{2}{3}\frac{\min\{\alpha_1,2\beta_1\}-1}{2(\alpha_1+\gamma)+\min\{\alpha_1,2\beta_1\}}}\big).$$

Theorem 7 (Prediction using RKHS)**.** *Suppose that Assumptions 1, 2, 3, and 5 hold. If the RKHS approach* (2.4.6) *is used to estimate ζ_0, then*

$$\mathbb{E}\big[\big(\langle\zeta_0, \tilde{X}^*\rangle_2 - \langle\widehat{\zeta}_{\rho,\mathrm{M}_\circ}, \widehat{X}^*_{\mathrm{M}_\circ}\rangle_2\big)^2\big] \;=\; O_p\big((\log(n)/n)^{\frac{2}{3}\frac{2\nu_\circ}{4+2\nu_\circ}\frac{\alpha_1-1}{\alpha_1+2\gamma+1}}\big),$$

where $\nu_\circ = \frac{(2\beta_2+3\alpha_2-1)/\alpha_2}{2+(2\beta_2+3\alpha_2-1)/\alpha_2}$.

2.6 Numerical Experiments

2.6.1 Simulation Studies

We consider three different distributions for the functional covariate. The random function $X(t)$ is generated from $X = \sum_{k=1}^{50}\zeta_k Z_k\phi_k(t)$, where $\zeta_k = (-1)^{k+1}k^{-1}$, $\rho_1(t) = 1$ and $\rho_{k+1}(t) = \sqrt{2}\cos(k\pi t)$, $k \geq 1$. The scores Z_k are the random components from one the following cases:

(1) $Z_1, \ldots, Z_{50}$ are i.i.d. uniform random variables defined on $[-\sqrt{3}, \sqrt{3}]$.

(2) $Z_1, \ldots, Z_{50}$ are i.i.d. from standard Gaussian random variables $N(0, 1)$.

(3) The vector $(Z_1, \ldots, Z_{50})^\top$ is a multivariate t-distribution with uncorrelated coordinates and degree of freedom is 5.

Scenarios (2) and (3) are motivated by the elliptical family of distributions. The noise of the responses follows Gaussian distribution $N(0, 0.5^2)$, and the measurement error is another Gaussian random variable $N(0, 0.1^2)$. The true slope function is taken as

$$\zeta_0(t) \;=\; \sum_{k=1}^{50} 4(-1)^{k+1}k^{-2}\rho_k(t).$$

We employ both the normal equation and the reproducing kernel Hilbert space approach to estimate the slope function. For the first approach, we use the estimator defined

in (2.4.3). The cross-covariance $g(t)$ and covariance $\Gamma(s,t)$ are estimated from observed sample. The number of eigen-components used in the approximation is selected according to the elbow rule in the scree plot. Two or three components were selected on average in the simulation.

For the second approach, we first show how to obtain a numerical estimate of the target function in a generic RKHS. Recall that the proposed estimator is found through a penalized regression:

$$\text{argmin}_{\zeta\in\mathcal{H}_K}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i-\langle\widehat{X}_{i,\mathrm{M}_\circ},\zeta\rangle_2)^2+\rho\big\|\zeta\big\|^2_{\mathcal{H}(K)}\right\}. \tag{2.6.1}$$

Decompose $\mathcal{H}(K)$ into two orthogonal subspaces, i.e., $\mathcal{H}_K=\mathcal{H}_0\oplus\mathcal{H}_1$, where $\mathcal{H}_0=\text{span}\{\omega_k(t)\}_{k=1}^p$ is finite-dimensional with orthogonal basis functions, and $\mathcal{H}_1$ is its orthogonal complement with kernel K_1. It is straightforward to verify (Wahba, 1990) that there exists $\mathbf{d}=(d_1,\ldots,d_p)^\top\in\mathbb{R}^p$ and $\mathbf{c}=(c_1,\ldots,c_n)^\top\in\mathbb{R}^n$ such that

$$\widehat{\zeta}_{\lambda,\mathrm{M}_\circ}(t)=\sum_{k=1}^{p}d_k\omega_k(t)+\sum_{i=1}^{n}c_iL_{K_1}(\widehat{X}_{i,\mathrm{M}_\circ})(t). \tag{2.6.2}$$

The following result shows how vectors $\mathbf{c}$ and $\mathbf{d}$ are computed.

Theorem 8. *The vectors* $\mathbf{d}$ *and* $\mathbf{c}$ *in* (2.6.2) *satisfy*

$$\begin{aligned}\Lambda\mathbf{c}+\Psi\mathbf{d}&=\mathbf{y},\\ \Psi^\top\mathbf{c}&=\mathbf{d},\end{aligned}$$

where

$$\begin{aligned}\Lambda&=n\rho I+\Sigma,\\ \Sigma_{ij}&=\iint_{[0,1]^2}\widehat{X}_{i,\mathrm{M}_\circ}(s)K_1(s,t)\widehat{X}_{j,\mathrm{M}_\circ}(t)\mathrm{d}s\mathrm{d}t,\quad 1\le i,j\le n,\\ \Psi_{ij}&=\int_0^1\widehat{X}_{i,\mathrm{M}_\circ}(t)\omega_j(t)\mathrm{d}t,\qquad\qquad 1\le i\le n,\,1\le j\le p.\end{aligned}$$

Then, he vectors $\mathbf{d}$ *and* $\mathbf{c}$ *are given by*

$$\begin{aligned}\mathbf{c}&=(\Lambda+\Psi^\top\Psi)^{-1}\mathbf{y},\\ \mathbf{d}&=\Psi^\top(\Lambda+\Psi^\top\Psi)^{-1}\mathbf{y}.\end{aligned}$$

Theorem 8 can be easily derived using the representer lemma for smoothing splines (Wahba, 1990) with a new penalized term in the $\|\cdot\|_{\mathcal{H}(K)}$ instead of $\|\cdot\|_{\mathcal{H}(K_1)}$. Using the $\mathcal{H}(K)$-norm as the penalization term avoids the numerical instability in solving for vectors $\mathbf{d}$ and $\mathbf{c}$ (page 13 in Wahba (1990)). We can now estimate the coefficients in (2.6.2) as long as the RKHS $\mathcal{H}(K)$ is known. For this purpose, we select the following RKHS as our estimation space:

$$\mathcal{H}(K) \;=\; \mathbb{W}_2[0,1] \;=\; \big\{\beta \,\big|\, \beta, \beta^{(1)} \text{ are absolutely continuous and } \beta^{(2)} \in \mathcal{L}_2[0,1]\big\},$$

where

$$\begin{aligned} K(s,t) \;&=\; \frac{1}{(2!)^2} B_2(s)B_2(t) - \frac{1}{4!} B_4(|s-t|), \\ B_2(t) \;&=\; t^2 - t + \frac{1}{6}, \\ B_4(t) \;&=\; t^4 - 2t^3 + t^2 - \frac{1}{30}, \end{aligned}$$

and the subspace $\mathcal{H}_0$ is span$\{1, t\}$. Then, the solution of (2.6.1) in the format of (2.6.2) can be written as

$$\widehat{\zeta}_{\lambda,\mathrm{M}_\circ}(t) \;=\; d_1 + d_2 t + \sum_{i=1}^{n} c_i \int_0^1 \widehat{X}_{i,\mathrm{M}_\circ}(s) K(t,s)\mathrm{d}s. \tag{2.6.3}$$

The penalty parameter ρ is selected by GCV as the minimizer of

$$\mathrm{GCV}(\rho) \;=\; \frac{\frac{1}{n}\big\|\widehat{\mathbf{y}} - \mathbf{y}\|^2}{\big(1 - \frac{1}{n}\mathrm{tr}(H(\rho))\big)^2},$$

where $\mathbf{y}$ and $\widehat{\mathbf{y}}$ are the true and estimated response values, respectively, and

$$H(\lambda) \;=\; (\Psi\Psi^\top + \Sigma)(n\rho I + \Sigma + \Psi\Psi^\top)^{-1}.$$

For both estimation approaches, the covariance of the functional data and its eigencomponents are estimated using `FPCA()` provided in the `R` package `fdapace`(Zhou et al., 2022). The Fraction-of-Variance (FVE) threshold used during the SVD of the fitted covariance function is 0.95, and the output time grid is a grid of 51 equally spaced points on $[0,1]$ including both end points.

Three different sample sizes $n = 50, 200, 500$ are considered in the simulation study. Each experiment was repeated 1000 times. To approximate the integral, we divided $[0, 1]$ into 50 equal width sub-intervals and used the trapezoidal rule. For each scenario, we consider 2 sampling schemes:

(a) randomly sample $5 \sim 10$ time points with measurement errors,

(b) sample all time points without measurement errors.

We report three quantities to measure each approach's relative accuracy normalized by the $\mathcal{L}_2$-norm of the true slope function, the variability of the estimation, and integrated mean squared error. Denote the average of the estimated slope functions by $\bar{\zeta}(t) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\zeta}_i(t)$, and three quantities are computed according to

(i) the relative bias in the $\mathcal{L}_2$-norm $\text{bias}_r = \|\zeta_0 - \bar{\zeta}\|_2/\|\zeta_0\|_2$,

(ii) the sample estimation variance $\text{var} = \frac{1}{n}\sum_{i=1}^{n}\int_0^1(\widehat{\zeta}_i(t) - \bar{\zeta}(t))^2\mathrm{d}t$,

(iii) the integrated mean squared error (IMSE) of the estimator $\text{imse} = \text{var} + \int_0^1(\bar{\zeta}(t) - \zeta_0(t))^2\mathrm{d}t$.

Table 2.1 presents the estimation accuracy under models (1)-(3) using both the normal equation approach (abbreviated as N.) and the reproducing kernel Hilbert space approach (abbreviated as R.) when measurement error is present and sampling plan (a) is used. Table 2.2 is organized in the same way except that the results are based on the dense sampling plan (b) for comparison purpose.

Table 2.1 shows, as expected, that a larger sample size leads to a smaller estimation error in both estimation methods. A similar pattern is also observed in Table 2.2, where both the estimation error and variability have smaller magnitude due to the dense sampling plan and no measurement errors. Furthermore, it is interesting that the RKHS approach works even when Assumption (3) is violated. Model (1) generates functional covariates using uniform random scores, and in this case equation (2.3.5) is not the true conditional mean but rather the best linear predictors. Models (2) and (3) are from elliptical families, and they satisfy Assumption 3 when no measurement error is present.

		$n = 50$			$n = 200$			$n = 500$		
model	method	bias_r	var	IMSE	bias_r	var	IMSE	bias_r	var	IMSE
1	N.	0.1719	1.1837	1.6954	0.1322	0.2776	0.5803	0.1280	0.1172	0.4009
	R.	0.0996	0.8253	0.9973	0.0567	0.1946	0.2503	0.0479	0.0713	0.1111
2	N.	0.1754	1.3934	1.9260	0.1360	0.3907	0.7109	0.1274	0.1613	0.4424
	R.	0.1054	1.0844	1.2766	0.0623	0.2515	0.3188	0.0477	0.0844	0.1238
3	N.	0.1834	2.2424	2.8252	0.1513	0.7914	1.1877	0.1361	0.3485	0.6692
	R.	0.1175	1.9404	2.1794	0.0722	0.6776	0.7679	0.0528	0.2826	0.3310

Table 2.1: The relative bias, variance and IMSE of the estimation of both methods under the sparse sampling plan (a).

The performance of the RKHS approach are indeed satisfactory. This demonstrates the robustness of this method. We also note that the RKHS approach tends to perform better in our experiments. In practice, the normal equation approach relies on delicate synchronization between the population covariance (its eigenfunctions) and the slope function, which is more affected by the covariance estimation and other hyperparameters choices during the estimation procedure.

		$n = 50$			$n = 200$			$n = 500$		
model	method	bias_r	var	IMSE	bias_r	var	IMSE	bias_r	var	IMSE
1	N.	0.1417	0.2085	0.5562	0.1336	0.0509	0.3601	0.1323	0.0213	0.3247
	R.	0.0390	0.2488	0.2751	0.0265	0.0681	0.0803	0.0199	0.0344	0.0413
2	N.	0.1428	0.2230	0.5763	0.1347	0.0536	0.3676	0.1331	0.0214	0.3281
	R.	0.0364	0.2622	0.2852	0.0272	0.0640	0.0768	0.0217	0.0326	0.0408
3	N.	0.1486	0.3791	0.7616	0.1359	0.1104	0.4302	0.1341	0.0490	0.3603
	R.	0.0363	0.2137	0.2365	0.0233	0.0491	0.0585	0.0192	0.0222	0.0286

Table 2.2: The relative bias, variance and IMSE of the estimation of both methods under the dense sampling plan (b).

2.6.2 Framingham Heart Study

We apply the proposed prediction method to the sparse functional data in the Framingham Heart Study (D'Agostino et al., 2001). The original data consist of the body mass index (BMI) of 5,209 subjects (2,336 men and 2,873 women) recruited at ages between 28 and 62, but we only have access to the data for 1213 individuals. The BMI is a risk factor for cardiovascular disease, and it is computed from the height (in meter) h and weight (in lbs) w of each subject as $703 \cdot w/h^2$. Our goal is to predict the BMI of an individual at an older age (65 to 70) using the BMI trajectory of this individual from age 35 to 60. To avoid complications due to death, we exclude subjects who died before 70. This results in 880 subjects. Since the BMI values were recorded longitudinally at different ages for different subjects, we have irregularly and sparsely recorded functional covariates with an average of 9 BMI measurements per subject. For the response, we use the average BMI recorded between age 65 to 70 to ensure that all subjects have a scalar response.

Both the normal equation and RKHS approaches described in the previous simulation section were used to estimate the slope function. For covariance estimation, the FVE threshold was set at 0.99. Figure 2.1 shows the slope function estimated from both methods. To avoid the boundary effects, we highlight the results (red and blue) that are at least one bandwidth away from the boundaries. The two approaches give slightly different but similar estimates that exhibit an increasing trend, which starts with near zero slopes before age 45 then a prominent increase in slopes in order ages. This suggests that more recent BMI values are more predictive of future BMI values. The r^2 value is 76.38% for the normal equation approach and 77.63% for the RKHS approach, indicating a good fit of the functional linear regression model.

To evaluate the performance of both methods in prediction, we randomly hold out 10% of the data (i.e., 80 subjects) for prediction and use the remaining data to estimate the slope function. The estimate is denoted as $\widehat{\zeta}_r$. The imputed functional covariates in the hold-out dataset are estimated using (2.3.6) and (2.3.7). We use the relative absolute error, defined as $\frac{1}{m}\sum_{j=1}^{m} |\widehat{y}_j - y_j|/|y_j|$, as the quality measure. Here the index j loops over the hold-out set, and there are $m = 80$ subjects in this set. The relative absolute

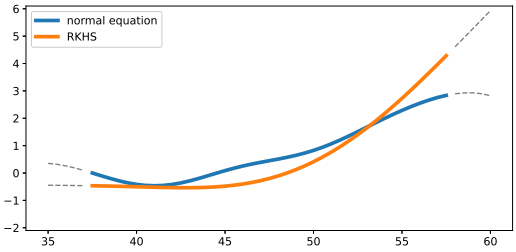


Figure 2.1: The slope function estimated from the normal equation approach and the RKHS approach. The x-axis represents the age.

error is 8.39% for the normal equation approach and 7.43% for the RKHS approach, indicating that both prediction approaches are effective with the RKHS approach slightly outperforms the normal equation approach.

Chapter 3

Deep Learning for Functional Data Analysis via Adaptive Bases

Despite their widespread success, the application of deep neural networks to functional data remains scarce today. The infinite dimensionality of functional data means standard learning algorithms can be applied only after appropriate dimension reduction, typically achieved via basis expansions. Currently, these bases are chosen a priori without the information for the task at hand and thus may not be effective for the designated task. We instead propose to adaptively learn these bases in an end-to-end fashion. We introduce neural networks that employ a new Basis Layer whose hidden units are each basis functions themselves implemented as a micro neural network. Our architecture learns to apply parsimonious dimension reduction to functional inputs that focuses only on information relevant to the target rather than irrelevant variation in the input function. Across numerous classification/regression tasks with functional data, our method empirically outperforms other types of neural networks, and we prove that our approach is statistically consistent with low generalization error.

3.1 Introduction

Deep learning has revolutionized data analysis and predictive modeling as its learned input representations capture more relevant aspects of a problem than representations based on manually selected features of the data. While the powerful capabilities of neural

networks (NN) have been clearly demonstrated for vector/image/text/audio/graph data, how to best adapt these models to functional data remains under-explored. Functional data are sample of random functions. The simplest examples are random curves defined over a (univariate) real-valued interval with one curve per individual subject in our dataset. Without loss of generality we assume that the interval is $[0, 1]$. The curve for one subject is thus a random function $X(t), t \in [0, 1]$, which can be viewed as a continuous stochastic process on $[0, 1]$. Extensively studied in the statistics literature (Ferraty and Vieu, 2006; Hsing and Eubank, 2015; Ramsay and Silverman, 2007; Wang et al., 2016), functional data appear frequently in scientific studies and daily life, such as in datasets of: air pollution, fMRI scans, growth curves, and sensors like wearable devices.

A fundamental property that distinguishes functional data from other data types is that they are intrinsically infinite dimensional and generated by smooth underlying processes. While high-dimensionality poses challenges in modeling and prediction, it brings benefits when data are generated from smooth or continuous functions. This is because the observed measurements at one location t_0 can inform us the values of $X(t)$ for t at nearby locations, thereby increasing the estimation efficiency. The smoothness assumption also makes functional data resilient to noise contamination as the magnitude of the noise can be estimated and statistical methods designed for functional data can accommodate noise in the observed data.

In practical applications, we are interested in using these continuous curves $X(t)$ to infer their relationship to some response variable Y, often to predict the response. More formally, we have a dataset $\{(X_i(t), Y_i)\}_{i=1}^n$ of i.i.d. samples of $X(t)$ and the associated Y, where $X(t)$ is a mean-square continuous process defined over $t \in [0, 1]$. For each subject i in our dataset, the observations of X_i serve as a functional covariate to predict scalar response variable Y_i, which may be either continuous or discrete. In many practical applications, the continuous process $X(t)$ is only be observed on a discrete time grid $\{t_1, \ldots, t_{J+1}\}$ in each observed X_i. Assuming there is an underlying map $\mathcal{T} : X(t) \mapsto Y$, our goal is to estimate $\mathcal{T}$ from the data (e.g. using a neural network).

A common pipeline for *functional data analysis* (FDA) is to summarize the informa-

tion contained in each function into a finite-dimensional vector and then carry out the analysis using existing models for the resulting multivariate data. Two popular dimension reduction approaches are Functional Principal Component Analysis (FPCA) (Besse and Ramsay, 1986; Li and Hsing, 2010; Rice and Silverman, 1991; Silverman, 1996; Yao et al., 2005a) and preselected basis expansions using, for example, B-splines (Cardot et al., 2003; Rice and Wu, 2001).

Existing approaches to deep learning for functional data rely on a straightforward pipeline that first applies classic functional data analysis methods and then a neural network in sequence. Rossi et al. (2002) handled functional data through discretization and functional parameterization of weight matrices, while Rossi et al. (2005) and Guss (2016) use basis function expansion to convert functional inputs into a vector form that can then be directly fed into to a standard neural network. Some universal approximation theory for functional neural networks has been established by Rossi et al. (2005) and Guss and Salakhutdinov (2019). However this two-stage fitting process in prior work is unable to fully leverage the representation-learning power and flexibility of deep learning.

In this paper, we propose to improve existing architectures by replacing these pre-specified choice of basis functions with adaptively learned bases that are implemented via micro neural networks (Lin et al., 2013) to which we backpropagate information regarding the response variable. A similar variant of our basic idea was briefly suggested by Rossi and Conan-Guez (2005), but their work never pursued the idea beyond a short comment stating it could be the possibile to implement the weight function of their functional neural network as an Multilayer Perceptron (MLP). Our design eliminates the need for a preprocessing step to convert random functions into vector inputs that otherwise typically requires a manually-prespecified choice of basis functions. Our architecture synchronizes the dimension reduction step and the nonlinear mapping step by adjusting the learned basis functions such that they only capture the information in $X(t)$ that is relevant to the output. Existing basis function representations instead seek to retain as much information about the input as possible, which may actually make supervised learning more difficult (Tishby and Zaslavsky, 2015).

Our main contribution is to propose an alternative neural architecture for *end-to-end* FDA that consists of a novel *Basis Layer* (BL) implemented via micro networks. We name the new network an *Adaptive Functional Neural Network* (AdaFNN). We study some theoretical properties of AdaFNN, establishing convergence and generalization error guarantees. Adding a BL into a neural network as feature extractors for functional inputs, the resulting model is empirically more accurate than existing methods and is simultaneously more parsimonious (meaning it requires less basis functions to model the random functions). Two types of regularizers are introduced to encourage basis orthogonality and basis sparsity. This regularization improves the resulting learned representations as well as the interpretability of the learned bases (especially when a small number of basis nodes are used). Moreover, our model can be trained end-to-end and thus composed with arbitrary differentiable operations without any alteration.

3.2 Related Work

3.2.1 Discretization of Functions

A straightforward application of neural networks to functional data is to treat the discretely observed functional values $\{X(t_1), \ldots, X(t_{J+1})\}$ as a high-dimensional vector and then input this vector into a neural network. This approach has been explored in Guss and Salakhutdinov (2019); Rossi and Conan-Guez (2005); Rossi et al. (2002, 2005). An alternative common approach is to treat the discretely sampled data from subjects as time series. However this approach has several disadvantages as it does not leverage the smoothness of the underlying data-generating process $X(t)$ and relies on additional strong assumptions such as stationarity. Lastly, most time-series methods are not designed for replicate observations (here we observe numerous draws of the underlying $X(t)$, one per subject). Thus we instead review the vector-based approach below. Since the $X(t)$ process is only observed at discrete time points $\{t_j\}_{j=1}^{J+1}$, one could use the vector $v_x = [X(t_1), \ldots, X(t_{J+1})]$ as the input of a standard neural network. Using the vector v_x, the original mapping $\mathcal{T}$ can be approximated by $\mathcal{T}_{\text{finite}} : v_x \mapsto Y$. Classical results imply that $\mathcal{T}_{\text{finite}}$ can be well approximated by a neural network with a sufficient number

of parameters. Furthermore, by increasing the partition resolution J, we can approximate $\mathcal{T}$ using $\mathcal{T}_{\text{finite}}$ with arbitrarily small error.

Drawbacks. In order to preserve critical information about the functional inputs, the discretization approach may require a high-dimensional vector, which hampers subsequent learning due to the curse of dimensionality. Furthermore, these discrete vector dimensions may fail to reflect the smoothness inherent to many functional covariates if they are contaminated by noise.

3.2.2 Basis Representation of Functions

To overcome the disadvantage of discretization, Rossi and Conan-Guez (2005) proposed to make use of the continuity of functional data and find a better finite-dimensional representation of a functional input before feeding it into a network. To be specific, let $\{\varphi_k(t)\}_{k=1}^K$ be a set of K continuous basis functions defined on $[0, 1]$. The task is to represent $X(t)$ using a vector $v_a = [a_1, \ldots, a_K]$ such that:

$$X(t) \approx \sum_{k=1}^{K} a_k \varphi_k(t). \tag{3.2.1}$$

Commonly used basis functions are Fourier basis functions, B-splines, or eigenfunctions obtained via spectral decomposition of $\text{cov}(X(s), X(t))$. After finding a basis expansion of $X(t)$, we can use the vector v_a instead of v_x as the input to a feedforward network. Normally the dimension K of v_a is much smaller than the dimension $J+1$ of the discretized data v_x, a clear advantage. Furthermore, the basis expansion in (3.2.1) automatically produces a smooth approximation of $X(t)$, which can reduce the noise contained in v_x.

Drawbacks. While the basis representation approach in (3.2.1) could recover the underlying smooth process of functional input, it does not take advantage of the key information contained in the response Y during its dimension reduction stage. Besides, both dimension reduction of functional inputs and parameterization of weight matrices (see FMLP on p.55, Rossi and Conan-Guez (2005)) require selection of basis functions. These bases are typically selected a priori and the number of bases needs to be chosen as well. We address these questions in this paper by proposing an adaptive approach to find the optimal bases that utilizes the information on Y and the specific learning task.

3.2.3 Micro Network Inside of a Network

Embedding smaller neural networks within a larger overall network architecture has been previously explored. For example, the Network in Network (NIN) model of Lin et al. (2013) replaces linear convolutions by a micro MLP. Empirically, the enhanced nonlinearity improves the model's ability to extract good features. Although NIN is conceptually related to our proposal, their operations differ in two ways. In our design, a basis node in a basis layer, which is also a micro MLP, is applied to the whole input. In contrast, a NIN micro network performs local convolution operations. Second, the micro MLP in NIN takes a small region of an image as its input and this process is slid across the whole image via convolution. Our basis micro network instead takes a fixed time point t as its input and this process operates on a full functional input $[X(t_1), \ldots, X(t_{J+1})]$ via numerical integration.

3.3 Methodology

To address the drawbacks of discretization and basis representation, we propose a novel neural network that adaptively learn the best basis functions for supervised learning tasks with functional inputs. Figure 3.1 shows the basic architecture of such a network (AdaFNN), and Algorithm 1 details the network computations used to produce predictions in a forward pass. After the initial basis layer, our network shares the same structure as a standard feedforward network. Each node in a BL outputs a scalar value computed as the inner product between the input function and the corresponding basis function at that node (Figure 3.2). Unlike handcrafted functions used in existing methods, we parameterize each basis function with a micro neural network that takes a scalar t as its input and outputs the value the basis function takes at t.

In the network depicted in Figure 3.1, the BL consists of d basis nodes (for some user-specified value of d). Each node represents the application of some basis function $\beta_i(t)$, for $i = 1, \ldots, d$ and $t \in [0, 1]$. The value output by each basis node, i.e., the *score* of $X(t)$ with respect to the basis function $\beta_i(t)$, is computed as:

$$c_i \;=\; \langle \beta_i, X \rangle \;=\; \int \beta_i(t) \cdot X(t)\, \mathrm{d}t. \tag{3.3.1}$$

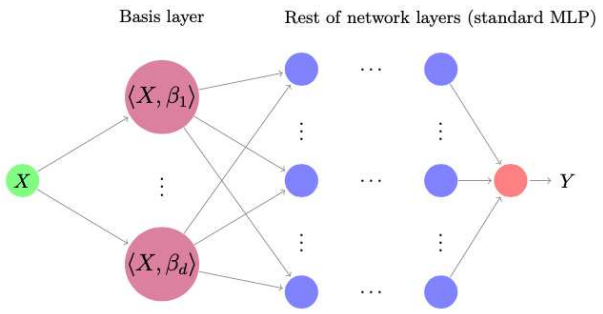


Figure 3.1: Neural network with our Basis Layer

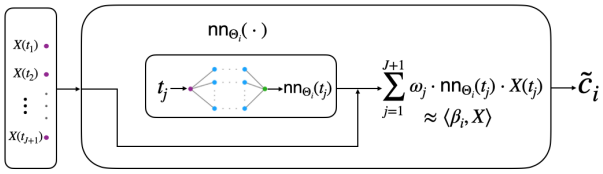


Figure 3.2: The i-th basis node in a Basis Layer.

The BL outputs form a vector $c = [c_1, \ldots, c_d] \in \mathbb{R}^d$. This vector is then fed through the rest of the AdaFNN network's layers, which after the BL are all standard fully-connected layers (as in a standard MLP, where Θ denotes the weights of all layers after the BL). The output of the overall network is $\widehat{Y} = \widehat{\mathcal{T}}(X) = \sigma_L\big(\cdots \sigma_1\big(W_1 c + b_1\big)\big)$, where $\sigma_1, \ldots, \sigma_L$ are the activation functions at each layer.

The bases β_i are used to project the functional input $X(t)$ to a vector representation c. While the form of these bases could in theory be selected a priori, such ad hoc selection violates the principle of end-to-end learning that drives deep learning's success. We would instead like to backpropagate information about Y through the network in order to adaptively identify optimal bases β_i. For this reason, we prefer the representation of input functions in (3.3.1) over (3.2.1), as the latter is less amenable to basis learning via backpropagation.

In the BL, each basis function $\beta_i(t)$ is itself parameterized by another neural network $\mathrm{nn}_{\Theta_i}(t) : t \mapsto \sigma_L^i\big(\cdots \sigma_1^i\big(W_1^i t + b_1^i\big)\big)$ with parameters Θ_i consisting of weights $\{W_l^i\}_{\ell=1}^L$ and biases $\{b_l^i\}_{\ell=1}^L$. In this work, these micro networks nn_{Θ_i} are simply MLPs whose only input are the values t_j at which the functional covariate $X(t)$ is measured (our micro

MLPs employ modern techniques such as skip connections, batch/layer normalization, and dropout). Note that these are the only locations at which we need to evaluate the learned basis function β_i in order to approximate $\langle \beta_i, X \rangle$. In practice, the integral in (3.3.1) must be approximated numerically, for example, using the rectangular or trapezoidal rule. Suppose that the domain $[0, 1]$ is equally discretized by the partition $\{t_j = (j-1)/J\}_{j=1}^{J+1}$. For each i, the score c_i is approximated by $\tilde{c}_i = \sum_{j=1}^{J+1} \omega_j \cdot \text{nn}_{\Theta_i}(t_j) \cdot X(t_j)$, where $\{\omega_j\}_{j=1}^{J+1}$ are weights used in a numerical integration algorithm. For instance, if trapezoidal rule is used in the network, then we would have $\omega_{J+1} = \omega_1 = 1/(2J)$ and $\omega_j = 1/J$ for $j = 2, \ldots, J$. Here equal spacing is not necessary. Our method works well as long as the grid is dense enough for the numerical integration to be accurate (see our provided code for a demonstration of AdaFNN with non-uniform spacing).

The loss of the network is calculated between a prediction $\widehat{Y}$ and the observed response Y, and is written as $\ell(\widehat{Y}, Y)$. Here we employ a standard loss function for prediction such as mean-squared-error for regression or cross-entropy for classification. The overall loss is the average loss across the whole sample (or a mini-batch of data for stochastic gradient training) and is denoted as $L(\{\Theta_i\}_{i=1}^d, \Theta) = \frac{1}{n} \sum_{i=1}^n \ell(\widehat{Y}_i, Y_i)$. The micro-networks and subsequent fully-connected network layers are all simultaneously trained (end-to-end) using this single objective $L(\{\Theta_i\}_{i=1}^d, \Theta)$ applied to the output predictions. In the next section, we also consider possible addition of orthogonality or sparsity regularization to this objective.

We make three observations about the proposed model. First, it can clearly be trained end-to-end with the BL. There is no need to select and even fine tune what type of basis functions should be used in representing input functions (or weight matrices). Second, since the parameters Θ_i are updated to minimize prediction loss, our learned basis functions are likely better suited for the desired task than handcrafted basis functions chosen without this information. The experiments in Section 3.5 show that our BL architecture can achieve higher accuracy than existing basis expansion methods while utilizing fewer basis functions (as each individual learned basis function can capture more predictive signal). Lastly, since each micro neural network is a composition of continuous functions (as

Algorithm 1 AdaFNN Forward Pass

Input: data $x = [X(t_1), \ldots, (X_{J+1})]$

Output: prediction $\widehat{y}$

Parameters: basis layer NN $\{\Theta_i\}_{i=1,\ldots,d}$, output NN Θ, integration weights $\omega_1, \ldots, \omega_{J+1}$, e.g. for trapezoid rule with equally-spaced t_j: $\omega_j = \frac{1}{J}$ if $2 \leq j \leq J$, else $= \frac{1}{2J}$

---

for $i = 1$ **to** d: $\tilde{c}_i \leftarrow \sum_{j=1}^{J+1} \omega_j \cdot \mathrm{nn}_{\Theta_i}(t_j) \cdot X(t_j)$

$\tilde{v}_c \leftarrow [\tilde{c}_1, \ldots, \tilde{c}_d] \in \mathbb{R}^d$

$\widehat{y} \leftarrow \mathrm{nn}_{\Theta}(\tilde{v}_c)$

return $\widehat{y}$

a standard MLP), all of the learned basis functions within our model will be continuous.

3.3.1 Regularization

Encouraging Basis Orthogonality. Without constraints, two BL nodes might learn similar basis functions and extract redundant information from the same functional input $X(t)$. To encourage different BL nodes to represent different (uncorrelated) information about the function, we can regularize them to be orthogonal. Recall that $L\big(\{\Theta_i\}_{i=1}^d, \Theta\big)$ is the loss function of a BL network. To encourage basis orthogonality, we introduce a regularization term which penalizes the cosine similarity between each pair of basis functions. The resulting loss optimized by the regularized network is

$$L_{\mathrm{perp}}\big(\{\Theta_i\}_{i=1}^d, \Theta\big) \;=\; L\big(\{\Theta_i\}_{i=1}^d, \Theta\big) + \lambda_1 \cdot \frac{1}{\binom{d}{2}} \sum_{j \neq j'} \frac{|\langle \mathrm{nn}_{\Theta_j}, \mathrm{nn}_{\Theta_{j'}} \rangle|}{\|\mathrm{nn}_{\Theta_j}\|_2 \|\mathrm{nn}_{\Theta_{j'}}\|_2},$$

where $\lambda_1 > 0$ controls the strength of the penalty. When a BL contains many nodes, enumerating all pairs becomes computationally expensive. Instead we randomly sample a few pairs at each mini-batch update and employ their average absolute cosine similarity as a stochastic estimate of the regularizer against which we optimize network parameters.

Encouraging Basis Sparsity. We can also regularize the shape of our learned basis functions. In domain selection problems (James et al., 2009; Wang et al., 2021; Zhou et al., 2013), the response is only related to the functional input over an (a priori unknown)

subset of its domain, i.e. $Y \perp\!\!\!\perp \{X(t')\}_{t' \notin \mathcal{I}} \mid \{X(t)\}_{t \in \mathcal{I}}$ for some $\mathcal{I} \subset [0,1]$. To encourage this desired property, it is sensible to learn a basis function whose value is zero outside of $\mathcal{I}$. While the number of nonzero values taken by the basis function is hard to optimize, we can penalize their L_1 norm as a tight convex relaxation of an L_0 norm to enforce basis sparsity. The resulted loss function is

$$L_{\text{sprs}}\big(\{\Theta_i\}_{i=1}^d, \Theta\big) \;=\; L\big(\{\Theta_i\}_{i=1}^d, \Theta\big) + \lambda_2 \cdot \frac{1}{s} \sum_{i \in \mathcal{S}} \int |\beta_i(t)| \, \mathrm{d}t,$$

where $\mathcal{S} \subseteq \{1, \ldots, d\}$ indicates which subset of basis functions we wish to sparsify, s is the number of elements in $\mathcal{S}$, and $\lambda_2 > 0$ controls the strength of the L_1 penalty. Even when we are not sure whether the domain selection assumption truly hold, learning sparse basis functions via this L_1 penalty can greatly improve the overall interpretability of our prediction model (Figure 3.6).

3.4 Theoretical Analysis

Here we discuss theoretical properties of the architecture proposed in the previous section. We provide a universal approximation theorem for this design and prove it achieves low generalization error under mild regularity conditions.

Let $\mathbb{C}([0,1])$ denote the space of continuous functions defined on the compact interval $[0,1]$. Assume that the underlying mapping $\mathcal{T} : X \mapsto Y$ is a composite of a finite-dimensional linear transformation and a subsequent non-linear transformation. We can write $\mathcal{T} = h \circ g$, where $g : \mathbb{C}([0,1]) \to \mathbb{R}^q$ is a linear continuous map, and $h : \mathbb{R}^q \to \mathbb{R}$ is a non-linear continuous map. By Riesz representation theorem, there exist square-integrable function γ_i with $i = 1, \ldots, q$ such that $g(X) = [\langle \gamma_1, X \rangle, \ldots, \langle \gamma_q, X \rangle]$. The approximate network parameterized by weights $\{\Theta_i\}_{i=1}^q$ and Θ is denoted as $\widehat{\mathcal{T}}$.

Theorem 9 (Consistency of the network)**.** *With the notations defined previously and following the conventions in the literature, we assume that:*

(i) the numerical integration at each basis node can be accurately evaluated as $J \to \infty$,

(ii) each network in $\widehat{\mathcal{T}}$ can have sufficient capacity.

Then for any $\epsilon > 0$, there exists a network $\widehat{\mathcal{T}}^*$ with weights $\{\Theta_i^*\}_{i=1}^q$ and Θ^* such that

$$\sup_{f \in \mathbb{C}([0,1]), \|f\|_2 \leq 1} |\widehat{\mathcal{T}}^*(f) - \mathcal{T}(f)| < \epsilon.$$

Hence, we have the following result: Let X be a continuous process defined on $[0,1]$. For any $\delta > 0$, there exists a network $\widehat{\mathcal{T}}^\star$ with weights $\{\Theta_i^\star\}_{i=1}^q$ and $\Theta^\star$ such that

$$\mathbb{P}\big(|\widehat{\mathcal{T}}^\star(X) - \mathcal{T}(X)| < \delta\big) > 1 - \delta.$$

Remark 1. Although Theorem 9 provides the consistency of the proposed architecture, it is not equivalent to identifying each true basis function consistently. The reason is that the individual basis functions are not identifiable since there are multiple ways to parameterize one map.

Remark 2. To make adequate predictions, the number of basis nodes d in AdaFNN should be sufficiently larger than the dimensionality q of $g(X)$, introduced in the assumptions of Theorem 1. In practice, we could also vary the choice of the number of basis nodes to find out which yields the best performance (lowest validation loss or fewest bases with low validation loss). Once the training is done, by investigating the learned bases, one can decide which ones seem to be relevant and use those as the basis functions to re-train a smaller subsequent network.

Next, we prove the proposed architecture can achieve small generalization error. Let $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be n i.i.d. copies of (X, Y), and there exist two constants $M_1, M_2 > 0$ such that both $\sup_{t \in [0,1]} |X(t)| \leq M_1$ and $|Y| \leq M_2$ hold almost surely. Let $\widehat{\mathcal{T}}_\Theta$ be a model proposed in Section 3.3 with its architecture fixed. In a slight abuse of notation, we use Θ to denote all the weights, including $\{\Theta_i\}_{i=1}^d$ and Θ, used in $\widehat{\mathcal{T}}_\Theta$. We use $\ell\big(\widehat{\mathcal{T}}_\Theta(X), Y\big)$ to denote the loss function. The population risk is defined as $r(\Theta) = \mathbb{E}\big[\ell\big(\widehat{\mathcal{T}}_\Theta(X), Y\big)\big]$, and the empirical risk is $r_n(\Theta) = \frac{1}{n}\sum_{i=1}^n \ell\big(\widehat{\mathcal{T}}_\Theta(X_i), Y_i\big)$. Usually the weights Θ are estimated via $\widehat{\Theta}$ produced by a random algorithm A based on the sample S. The generalization error is defined as $\big|\mathbb{E}_{S,A}[r(\widehat{\Theta}) - r_n(\widehat{\Theta})]\big|$. Suppose that we use a variant of stochastic gradient descent to train our model. At the t-th iteration ($1 \leq t \leq T$), the weights are updated with $\widehat{\Theta}_{t+1} = \widehat{\Theta}_t - \alpha_t \nabla_\Theta \ell\big(\widehat{\mathcal{T}}_\Theta(X_{i_t}), Y_{i_t}\big)\big|_{\Theta = \widehat{\Theta}_t}$, where the indices

i_t are randomly chosen (e.g., uniformly), and the learning rate α_t is monotonically non-increasing with $\alpha_t \leq c/t$ for some fixed constant $c > 0$. The following result is an application of Theorem 3.12 in Hardt et al. (2016).

Theorem 10 (Small generalization error)**.** *Assume that*

(i) the weight Θ is restricted on some compact region,

(ii) the loss function $\ell(\cdot, \cdot)$ and its gradient $\nabla\ell(\cdot, \cdot)$ are both Lipschitz.

Then for every pair of observation (X, Y), both $\ell\big(\widehat{\mathcal{T}}_\Theta(X), Y\big)$ and $\nabla_\Theta\ell\big(\widehat{\mathcal{T}}_\Theta(X), Y\big)$ are Lipschitz with respect to Θ. Hence, there exists some constant $c > 1$ such that

$$\big|\mathbb{E}_{S,A}[r(\widehat{\Theta}) - r_n(\widehat{\Theta})]\big| \lesssim \frac{T^{1-1/c}}{n}.$$

Remark 3. Assumption (i) in Theorem 10 is not restrictive since we can convert a constraint optimization problem to an equivalent penalized problem. In practice, we could minimize the regularized empirical risk, i.e., $r_n(\Theta) + \rho\|\text{vec}(\Theta)\|_2$, where $\text{vec}(\Theta)$ is the vectorized weights Θ, and $\rho > 0$ is a tuning parameter.

3.5 Experiments

Throughout, our experiments focus on methods that can handle general functional inputs, rather than approaches tailored for specific tasks like predicting future observations. The proposed AdaFNN network is compared to three baseline models: 'Raw data (#) + NN' (Rossi et al., 2002), 'B-spline (#) + NN' (Rossi et al., 2005), and 'FPCA_p + NN'(Rossi et al., 2005). Here the integer # denotes the input dimension of each network. The subscript p of 'FPCA' is the fraction of variation explained (FVE) used in selecting the number of principal components. The latter two baselines (B-spline and FPCA) have their roots in the field of functional data analysis (Cardot et al., 1999, 2003; Dou et al., 2012; Müller and Stadtmüller, 2005) and classification tasks (Chiou, 2012; Leng and Müller, 2006a; Müller, 2005; Song et al., 2008).

The first baseline simply discretizes the raw functional input as a vector that is fed into the network, so the dimension of the input vector is the number of points t at which

$X(t)$ has been observed. The other two baseline models consist of two steps. The first step involves transforming the functional input into a vector of scores from its B-spline or FPCA expansion (cf. (3.2.1)). The resulting vector representation is then fed as input into a network in the second step.

The number of bases used in the two baseline models is often determined by how well a function can be represented in the functional space spanned by these basis functions. For example, when using B-splines, we look at how well the selected spline functions can capture the trend of the raw data. A small number of B-splines may not be able to recover the trajectory very well. However, too many B-splines might overfit the functions. The choice of the number of principal components is based on the desired FVE by these principal components. Often, a sizeable FVE is expected, e.g., 90% to 99%. Similar to the selection of B-splines, we should not simply choose an FVE as large as possible. Large FVE may include (functional) principal components whose corresponding eigenvalues are very small, hence difficult to estimate. A good rule of thumb is to use the first few components that capture the bulk of the variation.

We write 'AdaFNN (λ_1,λ_2)' to indicate the level of regularization, where λ_1 (and λ_2) controls the degree of orthogonality (and L_1) regularization. AdaFNN was trained with 9 different combinations of the orthogonal regularization penalty $\lambda_1 \in \{0, 0.5, 1\}$ and L_1 regularization penalty $\lambda_2 \in \{0, 1, 2\}$. The performance of each configuration is reported for all simulations and real data tasks. Throughout we use * to indicate the λ_1, λ_2 values that performed best on the validation data, as these are the hyperparameter values that would be typically used in practice.

All models, including AdaFNN and the other NN baselines, employ the same architecture and training hyperparameters. A network with 3 hidden layers and 128 fully connected nodes per layer was used for Tasks 1-7 (real data) and our simulation studies. For Tasks 8 and 9 with small sample sizes, we used a smaller network with 2 hidden layers and 64 nodes and added dropout during training. All networks were trained up to 500 epochs (with 200-epoch early stopping patience) using mini-batches of size 128. AdaFNN merely uses 2 bases in simulation Cases 1 & 4, 3 bases in simulation Cases 2 & 3, and 4

bases in our applications to all 9 prediction tasks with real data. In contrast, we allowed the B-spline/FPCA baselines to either rely on a similar number or more basis functions since these models cannot optimize each of their bases (e.g. $\text{FPCA}_{0.99}$ often selected more than 10 principal components).

3.5.1 Simulation Studies

We demonstrate through simulations that baseline functional neural network models may miss relevant information but AdaFNN is able to capture the true signal while relying on fewer basis functions than the baselines. Four different simulation settings were considered, each is purposefully designed to illustrate a particular conceptual shortcoming of one of the baseline methods (with an extra fifth setting to highlight the utility of our proposed regularization).

We first describe the underlying data-generating process in each of the four settings. For $t \in [0, 1]$, define $\phi_1(t) = 1$ and $\phi_k(t) = \sqrt{2}\cos((k-1)\pi t)$, $k = 2, \ldots, 50$. Consider the process $X(t) = \sum_{k=1}^{50} c_k \phi_k(t)$, where $c_k = z_k r_k$, and r_k are i.i.d. uniform random variables on $[-\sqrt{3}, \sqrt{3}]$. The actual observations for $X_i(t)$ is the discrete data $\{X_i(t_j), j = 1, \ldots, 51\}$. We report the *mean squared prediction error* (MSE) achieved by each method on the test data.

Case 1: We set: $z_1 = 20, z_2 = z_3 = 5$, and $z_k = 1$ for $k \geq 4$. The response is $Y = c_3^2$, that is, $Y = (\langle X, \phi_3 \rangle)^2$, where the function ϕ_3 corresponds to the true predictive *signal*. This case is designed to show that a small FVE in FPCA may not suffice to capture the relevant signal.

Case 2: We set: $z_1 = z_3 = 5, z_5 = z_{10} = 3$, and $z_k = 1$ for other k. The response is $Y = c_5^2 = (\langle X, \phi_5 \rangle)^2$. Here the function ϕ_5 corresponds to the true predictive *signal* and is more complex than ϕ_3. The squared operation ensures a nonlinear relationship between the response and functional covariate. This case is designed to show that a small number of B-splines may not suffice to represent the $X(t)$ information relevant to the response.

Case 3: In Cases 1 & 2, both the response Y is free of noise and the functional input $X(t)$ is not contaminated by measurement errors. However, this setup is rarely realistic in practice. The goal here is to evaluate functional estimators in the presence of both

outcome noise and measurement errors in $X(t)$. We use the same model as Case 2 except that a mean zero Gaussian noise is added to the response, and the observation of $X(t)$ at each time point is perturbed by an additive mean zero Gaussian measurement error. The signal-to-noise ratios (SNR), defined as

$$\text{SNR}(X) = \frac{\sqrt{\int_0^1 (X(t))^2 \, \mathrm{d}t}}{\text{standard deviation of measurement error}}$$

for the functional input (due to $\mathbb{E}[X(t)] = 0$), is $\sqrt{10}$ to 1.

Case 4: This case studies how well AdaFNN captures multiple signals and its application in domain selection. The functional covariates $X_i(t)$ are generated similarly as in Cases 1 & 2, except that in Case 4: z_i are all taken to be 1. Two signals β_1 and β_2 are chosen as: $\beta_1(t) = (4 - 16t) \cdot 1\{0 \leq t \leq 1/4\}$ and $\beta_2(t) = (4 - 16|1/2 - t|) \cdot 1\{1/4 \leq t \leq 3/4\}$. The response is $Y = \langle \beta_2, X \rangle + (\langle \beta_1, X \rangle)^2$. Centered Gaussian noise is added to Y, and $X(t)$ is also contaminated by measurement error.

Case 5: The same setup as Case 4, but now with double the noise variance in Y (used to highlight our regularization).

Table **??** reports the MSEs for all methods. Under columns Cases 1 & 2, we see that AdaFNN exhibits the best performance in both settings. The performance of the baseline models is mixed. The top two principal components (with FVE at least 90%) are not able to detect any useful predictive signal in the data under Case 1, same with four B-splines bases in the simulation under Case 2. Thanks to its success in capturing the true basis function, AdaFNN performs strongly in both simulation settings. It is interesting to observe that each fitted basis function contains a fraction of the true signal function, and together they are able to recover it (Figures 3.3 and 3.4). That is, the true signal function can be represented as a linear combination of the fitted bases.

For Case 3, Table **??** (column 'Case 3') shows that all methods performed worse with noise added, but AdaFNN with orthogonality regularization remains superior to other methods. On the other hand, the L_1 regularizer is not helpful in Case 3. This is expected, because the true signal ϕ_5 does not have any zero region in its domain. Note that feeding the raw data as a vector into neural networks performs comparably to the two

Method	Case 1	Case 2	Case 3	Case 4
Raw data (51) + NN	0.015	0.038	0.275	0.334
B-spline (4) + NN	0.050	0.984	0.971	0.369
B-spline (15) + NN	0.013	0.019	0.206	0.251
$\text{FPCA}_{0.9}$ + NN	0.917	0.023	0.134	0.855
$\text{FPCA}_{0.99}$ + NN	0.003	0.036	0.239	0.667
AdaFNN (0.0, 0.0)	**0.001***	**0.003**	0.979	**0.193***
AdaFNN (0.0, 1.0)	0.995	0.007	0.978	0.982
AdaFNN (0.0, 2.0)	0.996	0.992	0.978	0.981
AdaFNN (0.5, 0.0)	0.004	0.005*	0.137*	0.571
AdaFNN (0.5, 1.0)	0.983	0.005	0.978	0.590
AdaFNN (0.5, 2.0)	0.134	0.008	0.978	0.981
AdaFNN (1.0, 0.0)	1.000	0.004	**0.127**	0.196
AdaFNN (1.0, 1.0)	0.009	0.006	0.974	0.606
AdaFNN (1.0, 2.0)	0.051	0.009	0.978	0.981

Table 3.1: Test-set MSE of predictions in simulation study. For each case, the asterisk indicates the best AdaFNN hyperparameters on the validation set, and the method with the best test MSE is marked in bold.

functional baseline models in the noiseless simulations. However, in noisy settings (Cases 3 & 4), this approach is inferior to the other two baseline models. This demonstrates the importance of exploiting functional properties whenever the input is a smooth function. The performance of the functional 'B-spline + NN' and 'FPCA + NN' approaches is mixed; each has its own advantage over the other in different scenarios.

For Case 4, Table **??** (column 'Case 4') shows that AdaFNN outperforms the other methods in this setting. At the same time, the proposed method also learns meaningful bases that correctly identify the relevant domain of interest (Figure 3.5). That on $[3/4, 1]$ both β_1 and β_2 are zero implies that the values of $X(t)$ over $[3/4, 1]$ have no effect on the response. None of the baseline methods is able to show this information, and thus AdaFNN is not only more accurate than existing models, but also more *interpretable*. In

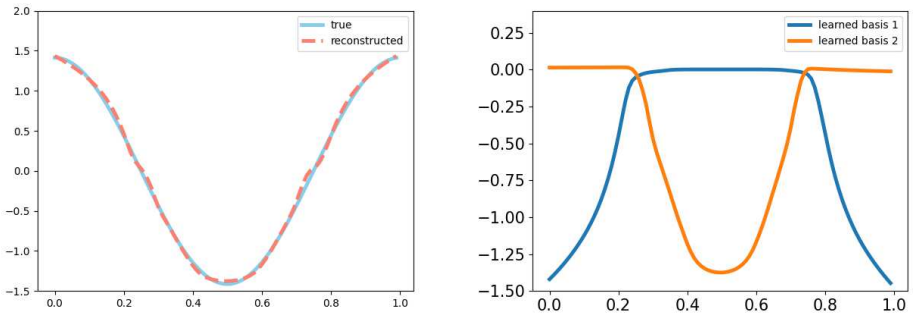


Figure 3.3: The left plot shows the true signal ϕ_3 (solid) and the reconstructed signal $\widehat{\phi}_3$ (dashed) from AdaFNN$(0, 0)$ (the regularization values with best validation MSE) in Case 1. The right plot shows each learned bases $\widehat{\beta}_1$ and $\widehat{\beta}_2$ from the same experiment. Note that: $\phi_3 \approx \widehat{\phi}_3 = \widehat{\beta}_2 - \widehat{\beta}_1$.

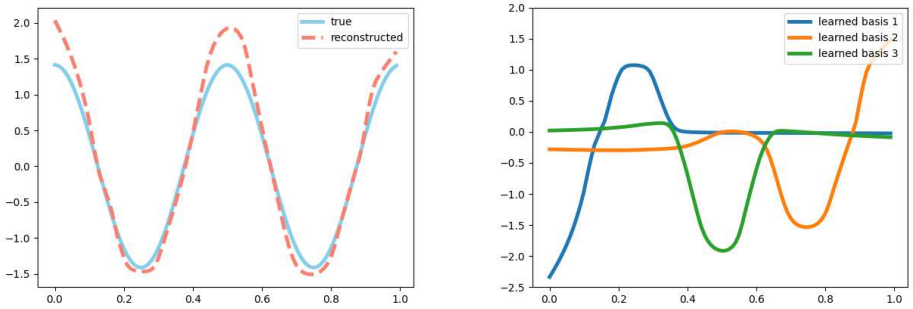


Figure 3.4: The left plot shows the true signal ϕ_5 (solid) and the reconstructed signal $\widehat{\phi}_5$ (dashed) from AdaFNN$(0.5, 0)$ (the regularization values with best validation MSE) in Case 2. The right plot shows each learned bases $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ from the same experiment. Note that: $\phi_5 \approx \widehat{\phi}_5 = \widehat{\beta}_3 - \widehat{\beta}_1 - \widehat{\beta}_2$.

summary, while some baseline methods perform well in certain simulation settings, none of these methods can achieve consistently strong performance like AdaFNN across all cases.

A final simulation, Case 5, demonstrates the utility of our proposed regularizationn. In this case, Table **??** shows that AdaFNN$(0.5, 0)$ and AdaFNN$(0, 0.1)$ clearly outperform AdaFNN without regularization, as well as all other methods (despite our regularized AdaFNN using **only 2 bases**, under which the other methods performed very poorly). Without any regularization, AdaFNN$(0, 0)$ learns **2 very similar bases** (left plot in

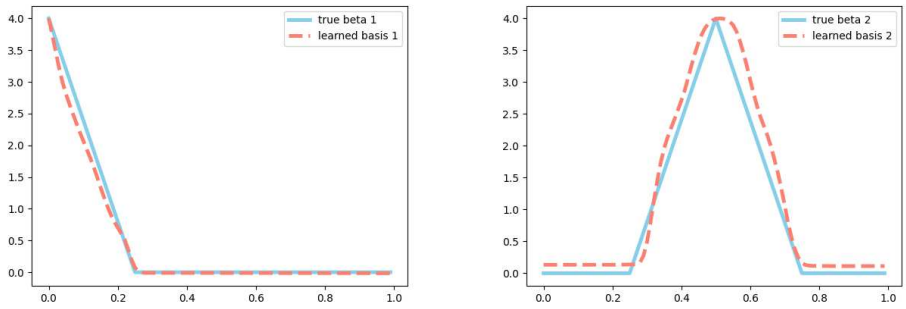


Figure 3.5: The left plot shows the true signal β_1 (solid) and a (scaled) learned signal $\widehat{\beta}_1$ (dashed) from AdaFNN$(0, 0)$ (the regularization values with best validation MSE) in Case 4. The right plot shows β_2 and a (scaled) $\widehat{\beta}_2$ from the same experiment.

Figure 3.6), while using either one of our proposed regularizers helps AdaFNN recover the true underlying bases (middle/right plots in Figure 3.6) and greatly improves its predictive performance.

METHOD	NO. BASES	MSE
RAW (51) + NN	51	0.339
B-SPLINE (4) + NN	4	0.382
B-SPLINE (15) + NN	15	0.257
$\text{FPCA}_{0.9}$ + NN	20	0.807
$\text{FPCA}_{0.99}$ NN	28	0.693
AdaFNN$(0.0, 0.0)$	2	0.598
AdaFNN$(0.5, 0.0)$	2	**0.231**
AdaFNN$(0.0, 0.1)$	2	**0.207**

Table 3.2: Test-set MSE of predictions in Case 5. AdaFNN with active regularization is highlighted in bold.

3.5.2 Application to Real Functional Datasets

Next, we evaluate the performance of AdaFNN and other neural FDA methods in nine different regression and classification tasks, using four datasets. In regression tasks, the

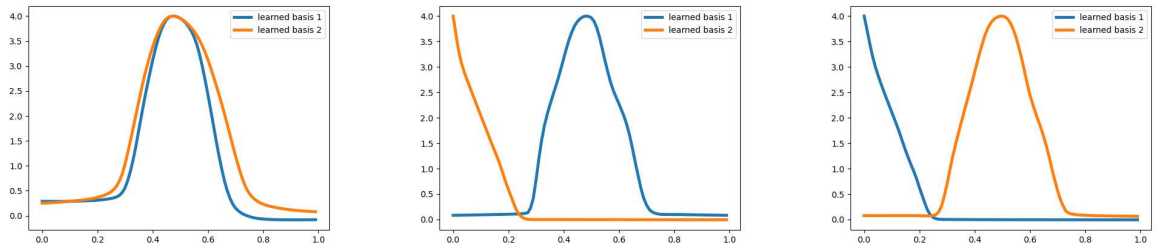


Figure 3.6: The (scaled) bases under simulation Case 5 learned by: AdaFNN$(0, 0)$ on the left, AdaFNN$(0.5, 0)$ in the middle, and AdaFNN$(0, 0.1)$ on the right.

Method	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9
Raw data (48) + NN	0.099	0.284	0.124	0.296	0.380	0.488	0.472	0.406	0.373
B-spline (15) + NN	0.094	0.306	0.137	0.326	0.335	**0.477**	0.429	0.413	0.387
$\text{FPCA}_{0.99}$ + NN	0.119	0.339	0.143	0.306	0.363	0.493	0.431	0.429	0.378
AdaFNN (0.0, 0.0)	**0.084***	0.290*	0.129*	0.311	0.365	**0.477**	**0.410***	0.377*	0.375
AdaFNN (0.0, 1.0)	0.094	0.276	0.126	0.327	0.561	0.479*	0.498	0.374	0.392
AdaFNN (0.0, 2.0)	0.097	0.276	0.129	0.324	0.596	0.481	0.473	0.381	0.445
AdaFNN (0.5, 0.0)	0.108	**0.260**	0.130	0.310*	0.380*	0.490	**0.410**	0.376	**0.368***
AdaFNN (0.5, 1.0)	0.089	0.279	0.126	0.324	0.616	0.486	0.494	**0.362**	0.413
AdaFNN (0.5, 2.0)	0.098	0.280	0.128	0.345	0.392	0.509	0.444	0.373	0.450
AdaFNN (1.0, 0.0)	**0.084**	0.288	**0.118**	**0.294**	**0.339**	0.485	0.413	0.378	0.406
AdaFNN (1.0, 1.0)	0.097	0.282	0.133	0.320	0.651	0.502	0.456	0.371	0.394
AdaFNN (1.0, 2.0)	0.092	0.279	0.127	0.326	0.371	0.510	0.414	0.374	0.416

Table 3.3: Comparing test-set performance of different methods' predictions on 9 functional datasets (MSE in regression, $1 - \text{AUC}$ in classification). For each dataset, the asterisk indicates which AdaFNN hyperparameters performed best on the validation set, and the best performing method on the test data is indicated in bold.

performance is again measured by MSE, while the performance is measured by the *area under the ROC curve* (ROC AUC) in classification tasks. Since our simulations show that B-spline (4) and $\text{FPCA}_{0.9}$ empirically underperform (Table **??**), we subsequently only consider the use of B-spline (15) and $\text{FPCA}_{0.99}$ on real data.

Electricity Data: Electricity consumption readings for 5567 London homes, where each household's electricity usage is recorded every half hour (UK Power Networks, 2015). The

functional covariate $X(t)$ is defined as the 48 measurements per household that constitute one day's electricity usage curve. Based on this $X(t)$, we consider four prediction tasks with different response variables:

1. Predict a household's total electricity consumption in the next week (week 2) based on its $X(t)$. [*regression*]
2. Predict a household's total electricity consumption in a later week (week 5) based on its $X(t)$. [*regression*]
3. Predict whether a household's morning (6am-12pm) electricity consumption in week 5 exceeds a certain threshold based on its $X(t)$. [*classification*]
4. Predict whether a household's consumption during the day (8am-17pm) exceeds night usage (17pm-12am) by a threshold in week 5 based on its $X(t)$. [*classification*]

The threshold values in tasks 3 and 4 are selected to ensure approximate class-balance in these classification problems. Because household consumption behavior is likely to change over a longer period, Tasks 2-4 are expected to be harder than Task 1. Results for these tasks are reported in columns Tasks 1-4 in Table **??**.

Wearable Device Data: This data consist of wearable device data from the National Health and Nutrition Examination Survey (NHANES) (NCHS, CDC 2020). Each subject in the study wore a device that continuously measures the intensity level of their physical activities within one week. The functional covariate $X(t)$ is the average activity levels every 30 minutes for one full day, resulting in a curve of 48 observations per subject. We examine whether physical activities are predictive of various health outcomes:

5. Predict whether a subject has diabetes. [*classification*]
6. Predict if subject feels chest pain. [*classification*]
7. Predict whether a subject experiences shortness of breath on stairs. [*classification*]

Tasks 6 and 7 aim at predicting a subject's cardiovascular health. Results for Tasks 5-7 are reported in column Task 5 to column Task 7 in Table **??**.

Mexfly and Medfly Data: The final two datasets pertain to Mexican fruit flies (Mexfly) (Carey et al., 2005) and Mediterranean fruit flies (Medfly) (Chiou et al., 2003), recording the number of eggs laid daily for each fly. Our task is to use early trajectories of egg-laying (daily number of eggs laid) to predict the *lifetime reproduction*, defined as the total number of eggs laid by the fly over its lifetime:

8. Predict lifetime reproduction of a Mexfly using its egg-laying curve $X(t)$ from day 1 to 30. [*regression*]
9. Predict lifetime reproduction of a Medfly using its egg-laying curve $X(t)$ from day 1 to 20. [*regression*]

The choice of the thresholds, day 20 and 30, is motivated by predicting lifetime reproduction based on early reproduction pattern in pre-peak period (peak usually occurs after 20 or 30 days depending on the species). Results are presented in Table **??** in columns Tasks 8 and 9.

Results. Empirically, AdaFNN performs better than all baseline methods in all 9 prediction tasks, demonstrating its advantage for diverse forms of real functional data spanning regression/classification problems. In contrast, none of the baseline methods consistently outperformed all other baselines across these tasks. Basis orthogonality and sparsity were used to improve learned representations and possibly get a better fit of the data (but like all regularization, the effectiveness of our proposed regularizers varies from dataset to dataset). Many of the best reported results are from AdaFNN with penalty $\lambda_1 > 0$, demonstrating that our orthogonal regularization technique improves the learned functional representations. While $\lambda_2 > 0$ only produces the most accurate AdaFNN model for one of the tasks, the L_1 penalty can remain useful for interpretability of the model. As with all regularizers, the optimal degree of regularization to employ also varies from dataset to dataset. By leveraging its superior representational capabilities, AdaFNN is also able to achieve superior accuracies with fewer bases than the B-spline + NN or FPCA + NN baselines.

3.6 Conclusion

This work presents a new approach to adapt representation learning techniques for functional data. Our proposed architecture does not require handcrafted bases to handle functional inputs, and learns the optimal bases for a particular dataset in an end-to-end manner. The Basis Layer compresses functional covariates in a linear fashion into a low-dimensional vector that reflects only those factors of variation most relevant to the response value. Traditional dimension reduction techniques like FPCA instead attempt to capture all variation in the functional input itself, regardless of its relationship to the response. There are many disadvantages to retaining global information about $X(t)$ rather than merely what is needed to infer Y, some of which are outlined in the *information bottleneck* principle of Tishby and Zaslavsky (2015).

Note that AdaFNN can be easily extended to vector-valued functional data, where either $t \in \mathbb{R}^p$ or $X(t) \in \mathbb{R}^p$ for $p > 1$. In the former case, each basis layer micro NN simply operates on vector-valued inputs t, while in the latter case, these micro NN would have larger output layers to produce vectors rather than scalars. Furthermore, our method can also be applied to data with both multiple functional covariates (can simply employ separate basis layers for each and pool their outputs) as well as auxiliary vector covariates in addition to $X(t)$ (can simply concatenate our basis layer output $\tilde{v}_c$ with these additional covariates before it is fed into the subsequent feedforward network).

As previously mentioned, AdaFNN is also directly applicable to non-uniform observations of $X(t)$, where the t_j are not equally spaced, although such cases require proper selection of the integration weights ω_j. However, successful application of AdaFNN to sparsely observed functional data, where the underlying process $X(t)$ is only observed at few locations t_j, remains nontrivial and likely requires improved numerical integration strategies as well as stronger inductive bias in the micro NN architecture Gunter et al. (2014). Nonetheless, we expect our adaptive Basis Layer will find broad applicability as a general-purpose representation learning tool for domains with functional data such as wearable devices, climatology, genomics, or neuroimaging.

Chapter 4

Bootstrap for Eigenvalues in High-Dimensions

In the context of principal components analysis (PCA), the bootstrap is commonly applied to solve a variety of inference problems, such as constructing confidence intervals for the eigenvalues of the population covariance matrix Σ. However, when the data are high-dimensional, there are relatively few theoretical guarantees that quantify the performance of the bootstrap. In this chapter we analyze how well the bootstrap can approximate the joint distribution of the leading eigenvalues of the sample covariance matrix $\widehat{\Sigma}$, and we establish non-asymptotic rates of approximation with respect to the multivariate Kolmogorov metric. Under certain assumptions, we show that the bootstrap can achieve the dimension-free rate of $\mathtt{r}(\Sigma)/\sqrt{n}$ up to logarithmic factors, where $\mathtt{r}(\Sigma)$ is the effective rank of Σ, and n is the sample size. From a methodological standpoint, our work also illustrates that applying a transformation to the eigenvalues of $\widehat{\Sigma}$ before bootstrapping is an important consideration in high-dimensional settings.

4.1 Introduction

Since the advent of the bootstrap itself (Diaconis and Efron, 1983), the bootstrap method has been widely applied in principal components analysis (PCA) and become part of standard practice in multivariate analysis (Davison and Hinkley, 1997; Jolliffe, 2002; Olive, 2017). It is well established in the literature that the bootstrap generally works in the

context of PCA with low-dimensional data (Beran and Srivastava, 1985; Eaton and Tyler, 1991). Furthermore, in cases where the bootstrap is known to encounter difficulties in low dimensions, such as in the case of tied population eigenvalues, various remedies have been proposed and analyzed (Beran and Srivastava, 1985; Dümbgen, 1993; Hall et al., 2009). Even though it has been widely applied to high-dimensional PCA problems (e.g., Fisher et al., 2016; Li and Ralph, 2019; Nguyen and Holmes, 2019; Stewart et al., 2019; Terry et al., 2018; Wagner, 2015; Webb-Vargas et al., 2017), the theory for bootstrap is relatively incomplete in this context. Although there is an extensive body of literature on distributional approximation for sample eigenvalues, their primary focus is establishing asymptotic properties. By and large, these results can be categorized into two parts, dealing with either classical asymptotics where p is held fixed as $n \to \infty$ (Anderson, 2003), or high-dimensional asymptotics where p/n converges to a positive constant as p and n diverge simultaneously (Bai and Silverstein, 2010). (The first case is essentially the low-dimensional scenario.) One fundamental limitation of both approaches is that the approximations expressed in asymptotics often do not reflect their practical performance. For example, it is hard to evaluate how accurate the confidence intervals constructed from asymptotic distributions can be on a collected data with a finite size. In addition, approximations based on analytical formulas are often tied to specific model assumptions, which can make it difficult to adapt such formulas outside of a given model.

Bootstrap methods can be a good solution to the second limitation, as they are data adaptive and usually used in a very flexible manner. In spite of that, the existing theoretical understanding of applying bootstrap to PCA tasks is still limited by the aforementioned problem, since the results are generally asymptotic (Beran and Srivastava, 1985; Eaton and Tyler, 1991; El Karoui and Purdom, 2019). Hence, the motivation of this chapter is two-fold. First, we want to explicitly quantify the accuracy of bootstrap approximation in terms of the sample size n and the effective rank of Σ. For example, our results can be used to quantify how close the coverage probabilities of bootstrap confidence intervals are to the nominal values. Another motivation is based on the fact that, until quite recently, most of the literature on bootstrap methods for PCA has been

limited to low-dimensional settings. It is of general interest to establish a more complete theoretical description of bootstrap methods for high-dimensional PCA—a point that was highlighted in a recent survey on this topic (Johnstone and Paul, 2018, §X.C).

In this chapter, we focus on approximating the joint distribution of leading eigenvalues of sample covariance matrices. Because the top eigenvalues are considered as dominant components in the spectral decomposition and often contain important summarizing information of the data in a broad range of applications, such as signal processing (Couillet and Debbah, 2011), and finance (Ruppert and Matteson, 2015). Accurately estimating them and understanding their fluctuation play a central role in related inference tasks and are crucial for domain-specific analysis. To better illustrate how sample eigenvalues and their uncertainty quantification can be applied in practice, we provide a real-data example based on stock market returns in Section 4.4.5 with the following two main focuses:

- *Selecting principal components.* A key step in applying PCA is to select a subset of principal components, often dominant ones, which are then used for tasks, such as summarizing data, reducing dimensions, and de-noising. Two commonly used approaches are (1) choosing a threshold (e.g., 95%) based on the proportion of explained variance $(\lambda_1(\widehat{\Sigma}) + \cdots + \lambda_k(\widehat{\Sigma}))/\operatorname{tr}(\widehat{\Sigma})$ and (2) the elbow rule based on eigengaps $\lambda_j(\widehat{\Sigma}) - \lambda_{j+1}(\widehat{\Sigma})$ in the scree plot of eigenvalues. The componentwise proportions $\lambda_j(\widehat{\Sigma})/\operatorname{tr}(\widehat{\Sigma})$ are also popular. Besides, the selection rules based on confidence intervals for the eigenvalues $\lambda_1(\Sigma), \ldots, \lambda_k(\Sigma)$ of the population covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ are also useful, as the construction of such intervals is directly linked to the distribution of the eigenvalues of $\widehat{\Sigma}$. It provides more insights on how accurate the selection can be based on the sample estimates, i.e., uncertainty quantification, which will be explained more below. For a general overview of selection rules, we refer to Jolliffe (2002).

- *Quantifying uncertainty.* In addition to PCA, the eigenvalues of a population covariance matrix plays a crucial role in understanding the performance of statistical methods for covariance estimation, regression, and classification (Dobriban and Wager, 2018; Hsu et al., 2014; Ledoit and Wolf, 2012). They also have domain-specific

meaning in applications ranging from finance to ecology (Chen et al., 2019; Fabozzi et al., 2007). For instance, finance practitioners uses principal components of a return vector of a number of cross-section stocks for risk analysis and portfolio selection Connor and Korajczyk (1993); Roll and Ross (1980). Hence, it is necessary to provide uncertainty quantification for sample eigenvalues to access estimation accuracy of these parameters, using approaches, such as evaluating variances and constructing confidence intervals–and again, this leads to the use of distributional approximation results for the sample eigenvalues.

To give an overall description of our contributions, consider the following setup. Let $X_1, \dots, X_n \in \mathbb{R}^p$ be centered i.i.d. observations with population covariance matrix $\Sigma = \mathbb{E}[X_1 X_1^\top]$. Also, let $\widehat{\Sigma} = \sum_{i=1}^n X_i X_i^\top / n$ denote the associated sample covariance matrix, and let $\widehat{\Sigma}^\star = \sum_{i=1}^n X_i^\star (X_i^\star)^\top / n$ be its bootstrap version, formed from random vectors $X_1^\star, \dots, X_n^\star$ that are sampled with replacement from the observations. In addition, let the eigenvalues of a symmetric matrix $A \in \mathbb{R}^{p \times p}$ be denoted as $\lambda_1(A) \geq \cdots \geq \lambda_p(A)$, and let $\boldsymbol{\lambda}_k(A) = (\lambda_1(A), \dots, \lambda_k(A))$ for a fixed integer $k < p$.

Following this notation, we study the finite-sample approximation to the distribution of leading k eigenvalues and establish non-asymptotic bounds on the multivariate Kolmogorov distance

$$\Delta_n = \sup_{t \in \mathbb{R}^k} \left| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)\big) \preceq t \Big) - \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star) - \boldsymbol{\lambda}_k(\widehat{\Sigma})\big) \preceq t \,\Big|\, X \Big) \right|,$$

where the relation $v \preceq w$ between two vectors $v, w \in \mathbb{R}^k$ means $v_j \leq w_j$ for all $j = 1, \dots, k$, and $\mathbb{P}(\cdot \,|X)$ refers to probability that is conditional on $X_1, \dots, X_n$. Under certain assumptions, our central result (Theorem 11) shows that the dimension-free bound

$$\Delta_n \;\leq\; \frac{C_n \, \mathtt{r}(\Sigma)}{\sqrt{n}} \tag{4.1.1}$$

holds with high probability, where $C_n > 0$ is a polylogarithmic function of n, and the quantity $\mathtt{r}(\Sigma)$ is the effective rank of Σ, defined by $\mathtt{r}(\Sigma) = \mathrm{tr}(\Sigma)/\lambda_1(\Sigma)$.

Several aspects of the bound (4.1.1) and the parameter $\mathtt{r}(\Sigma)$ are worth our attention. First, the effective rank always satisfies $1 \leq \mathtt{r}(\Sigma) \leq p$ whenever Σ is nonzero, and can be

interpreted as a proxy for the number of "dominant" principal components of Σ. Hence, even in very high-dimensional settings where $n \ll p$ and unbounded covariance scales where $\lambda_1(\Sigma) \to \infty$, the bound (4.1.1) shows that the bootstrap can perform well as long as the number of dominant components is not too large, which is precisely the situation where high-dimensional PCA is of greatest interest. Meanwhile, even in situations where $\mathtt{r}(\Sigma)$ is moderately large, e.g. $\mathtt{r}(\Sigma) \to \infty$ with $\mathtt{r}(\Sigma) = o(\sqrt{n})$, the bound (4.1.1) is still useful and can quantify the accuracy of the bootstrap. Indeed, in Section 4.4, our experiment results clearly demonstrate both points and confirm that the performance of the bootstrap is governed more by $\mathtt{r}(\Sigma)$ than the absolute dimension p, and that the bootstrap can still be accurate when $\mathtt{r}(\Sigma)$ is moderately large. More generally, it should also be mentioned that the theoretical role of effective rank in many other aspects of high-dimensional PCA has attracted considerable attention in recent years (e.g. Bunea and Xiao, 2015; Jung et al., 2018; Koltchinskii et al., 2020; Koltchinskii and Lounici, 2017; Lounici, 2014; Naumov et al., 2019).

Using a transformation prior to bootstrapping can be beneficial compared to directly bootstrapping the distribution of $\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)\big)$. It is a fundamental topic in the bootstrap literature and has been widely studied (e.g., Chernick, 2011; Davison and Hinkley, 1997; DiCiccio, 1984; DiCiccio and Efron, 1996; Konishi, 1991; Tibshirani, 1988). To be more specific, let h be a univariate scalar function, referred to as a transformation, and for any symmetric matrix $A \in \mathbb{R}^{p\times p}$, let $\boldsymbol{h}(\boldsymbol{\lambda}_k(A)) = (h(\lambda_1(A)), \ldots, h(\lambda_k(A)))$. Then, the distribution of $\sqrt{n}(\boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star)) - \boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma})))$ conditioning on the observations can be used to approximate the distribution of $\sqrt{n}(\boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma})) - \boldsymbol{h}(\boldsymbol{\lambda}_k(\Sigma)))$. Then the confidence intervals for $\lambda_j(\Sigma), 1 \le j \le k$ can be constructed by inverting the confidence intervals for $h(\lambda_j(\Sigma)), 1 \le j \le k$. (Extension discussion and algorithm details are provided in Sections 4.3 and 4.4.) A classical example of transformation is $h(x) = \log(x)$, which is known to be variance-stabilizing under certain conditions when $n \to \infty$ with p held fixed (Beran and Srivastava, 1985). With this in mind, a second contribution our analysis is an extended version of the bound (4.1.1) that can accommodate the use of certain transformations (see Theorem 12).

From a more methodological standpoint, our numerical experiments also shed new light on the role of transformations in bootstrap methods for high-dimensional PCA. Although we confirm that the classical logarithm transformation can be beneficial in low dimensions (small $\mathtt{r}(\Sigma)$, we show that it is less effective when $\mathtt{r}(\Sigma)$ is moderately large. Consequently, we explore some alternative transformations combined with partial standardization and data-adaptive parameter selection and provide numerical results demonstrating that there are opportunities to improve upon $h(x) = \log(x)$ in high dimensions. To put such empirical findings into perspective, we are not aware any prior work investigating how transformations can be used to enhance bootstrap methods in this context.

4.2 Related work

Quite recently, there has been an acceleration in the pace of research on bootstrap methods for high-dimensional sample covariance matrices, as evidenced in the papers El Karoui and Purdom (2019); Han et al. (2018); Johnstone and Paul (2018); Lopes et al. (2019, 2023); Naumov et al. (2019). Among these, the most relevant to our work is El Karoui and Purdom (2019), which examines both the successes and failures of the bootstrap in doing inference with the leading eigenvalues of $\widehat{\Sigma}$. In the negative direction, that paper focuses on a specialized model with $\lambda_1(\Sigma) > 1$ and $\lambda_2(\Sigma) = \cdots = \lambda_p(\Sigma) = 1$, which corresponds to a very large effective rank $\mathtt{r}(\Sigma) \asymp p$ that makes dimension reduction via PCA inherently difficult. In the positive direction, that paper deals with a different situation where Σ is assumed to have a near low-rank structure of the form

$$\Sigma = \begin{pmatrix} A & B \\ B^\top & C(\eta) \end{pmatrix}, \tag{4.2.1}$$

where A is of size $k \times k$ with $k \asymp 1$, and the diagonal blocks satisfy $\lambda_1(A) \asymp 1$, and $\lambda_1(C(\eta)) \lesssim n^{-\eta}$ for a fixed parameter $\eta > 1/2$. Working under an elliptical model, the paper (El Karoui and Purdom, 2019) shows that the bootstrap consistently approximates the distribution of $\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)\big)$ in an asymptotic framework where $p/n \lesssim 1$. In relation to our work, the most crucial distinction is that our results quantify the accuracy of the bootstrap with non-asymptotic *rates of approximation*. To illustrate the significance

of this, note that our bound (4.1.1) provides an explicit link between the size of $\mathtt{r}(\Sigma)$ and the accuracy bootstrap, whereas in an asymptotic setup, the effect of $\mathtt{r}(\Sigma)$ is hidden—because it "washes out in the limit". Our numerical experiments will also confirm that different sizes of $\mathtt{r}(\Sigma)$ can have an appreciable effect on the finite-sample accuracy of the bootstrap. In this way, our work indicates that the quantity $\mathtt{r}(\Sigma)/\sqrt{n}$ serves as a type of conceptual diagnostic for assessing the reliability of the bootstrap in high-dimensional PCA.

Beyond these points of contrast with El Karoui and Purdom (2019), there are several distinctions with regard to model assumptions. First, we work in a dimension-free setting where there are no restrictions on the size of p with respect to n. Second, the model based on (4.2.1) implicitly requires that $\lambda_j(\Sigma) \lesssim n^{-\eta}$ for all $j \geq k+1$, whereas this constraint on Σ is not used here. Third, it is straightforward to check that in the model based on (4.2.1), the condition $\eta > 1/2$ *implies* $\mathtt{r}(\Sigma) = o(\sqrt{n})$, which means that our bound (4.1.1) ensures bootstrap consistency in models that subsume the one based on (4.2.1). (As an example, if $p \asymp e^{m(n)}$ for some sequence of integers satisfying $m(n) = o(\sqrt{n})$ and if $\lambda_j(\Sigma) \asymp j^{-1}$, then the bound (4.1.1) implies bootstrap consistency, whereas this is not guaranteed by the previous result even when $p \asymp n$.)

Other works on bootstrap methods related to high-dimensional sample covariance matrices have dealt with models or statistics that are qualitatively different from those considered here. The papers Han et al. (2018); Lopes et al. (2023) look at bootstrapping the operator norm error $\sqrt{n}\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}}$, as well as variants of this statistic, such as $\sup_{u \in \mathcal{U}} \sqrt{n}|u^\top(\widehat{\Sigma} - \Sigma)u|/u^\top \Sigma u$, where $\mathcal{U}$ is a set of sparse vectors in the unit sphere of $\mathbb{R}^p$. In a different direction, the paper (Lopes et al., 2019) focuses on "linear spectral statistics" of the form $\sum_{j=1}^p f(\lambda_j(\widehat{\Sigma}))/p$, where $f : [0, \infty) \to \mathbb{R}$ is a smooth function. In that paper, it is shown that a type of parametric bootstrap procedure consistently approximates the distributions of such statistics when p/n converges to a positive limit. Lastly, the paper Naumov et al. (2019) deals with bootstrapping statistics related the eigenvectors of $\widehat{\Sigma}$.

Notation. For a random variable X and an integer $q \in \{1, 2\}$, define the ψ_q-Orlicz norm as $\|X\|_{\psi_q} = \inf\{t > 0 \mid \mathbb{E}[\exp(|X|^q/t^q)] \leq 2\}$. The random variable X is said to be

sub-exponential if $\|X\|_{\psi_1}$ is finite, and sub-Gaussian if $\|X\|_{\psi_2}$ is finite. In addition, for any $q \geq 1$, the L_q norm of X is defined as $\|X\|_q = (\mathbb{E}[|X|^q])^{1/q}$. For any vectors $u, v \in \mathbb{R}^p$, their inner product is $\langle u, v \rangle = \sum_{j=1}^p u_j v_j$. For any real numbers a and b, the expression $a \ll b$ is used in an informal sense to mean that b is much larger than a. Also, we use the notation $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. If $\{a_n\}$ and $\{b_n\}$ are two sequences on non-negative numbers, then the relation $a_n \lesssim b_n$ means that there is a positive constant c not depending on n such that $a_n \leq c\, b_n$ holds for all large n. When both of the conditions $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold, we write $a_n \asymp b_n$.

4.3 Main Results

We consider a sequence of models indexed by n, in which all parameters may depend on n except when stated otherwise. In particular, the dimension $p = p(n)$ is allowed to have arbitrary dependence on n. Likewise, if a parameter does not depend on n, then it is understood not to depend on p either. One of the few parameters that will be treated as fixed with respect to n is the positive integer $k < p$.

Assumption 6 (Data-generating model)**.**

(a). There is a non-zero positive semidefinite matrix $\Sigma \in \mathbb{R}^{p \times p}$, such that the ith observation is generated as $X_i = \Sigma^{1/2} Z_i$ for all $i = 1, \ldots, n$, where $Z_1, \ldots, Z_n \in \mathbb{R}^p$ are i.i.d. random vectors with $\mathbb{E}[Z_1] = 0$, and $\mathbb{E}[Z_1 Z_1^\top] = I_p$.

(b). The eigenvalues of Σ satisfy $\min_{1 \leq j \leq k} \big(\lambda_j(\Sigma) - \lambda_{j+1}(\Sigma)\big) \gtrsim \lambda_1(\Sigma)$.

(c). Let $u_j \in \mathbb{R}^p$ denote the jth eigenvector of Σ, and let $\Gamma \in \mathbb{R}^{k \times k}$ have entries given by $\Gamma_{jj'} = \mathbb{E}[(\langle u_j, Z_1 \rangle^2 - 1)(\langle u_{j'}, Z_1 \rangle^2 - 1)]$ for all $1 \leq j, j' \leq k$. Then, the matrix Γ satisfies $\lambda_k(\Gamma) \gtrsim 1$.

In connection with the model described by Assumption 6, our results will make reference to a moment parameter defined as $\beta_q = \max_{1 \leq j \leq p} \|\langle u_j, Z_1 \rangle^2\|_q$ for any $q \geq 1$.

Remarks. Regarding Assumption 6.(b), it ensures that there is some degree of separation between the leading eigenvalues of Σ. In less compact notation, the assumption states that there is a fixed constant $c > 0$ such that the inequality $\lambda_j(\Sigma) - \lambda_{j+1}(\Sigma) \geq c\lambda_1(\Sigma)$ holds for all $j = 1, \ldots, k$, and all large n. (There is no restriction on the size of c.) In general, a separation condition on the leading eigenvalues is unavoidable, because it is known both theoretically and empirically that the bootstrap can fail to approximate the distribution of $\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)\big)$ if the leading population eigenvalues are not distinct (Beran and Srivastava, 1987; Hall et al., 2009). In more technical terms, the source of this issue can be explained briefly as follows: If $\mathcal{S}^{p\times p}$ denotes the space of real symmetric $p \times p$ matrices, and if $\lambda_j(\cdot)$ is viewed as a functional from $\mathcal{S}^{p\times p}$ to $\mathbb{R}$, then $\lambda_j(\cdot)$ becomes non-differentiable at Σ in the case when $\lambda_j(\Sigma)$ is a repeated eigenvalue (i.e. with multiplicity larger than 1). In turn, this lack of smoothness makes it difficult for the bootstrap to approximate the distribution of $\sqrt{n}(\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma))$.

To interpret Assumption 6.(c), the matrix Γ serves a technical role as a surrogate for the correlation matrix of $\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)\big)$. Hence, the lower bound $\lambda_k(\Gamma) \gtrsim 1$ can be viewed as a type of non-degeneracy condition for the distribution of interest. The proposition below gives examples of well-established models in which Assumption 6.(c) holds. Namely, parts (i) and (ii) below respectively correspond to *Marčenko-Pastur models* and *elliptical models*. The latter case also illustrates that the entries of the vector Z_1 are not required to be independent.

Proposition 1. *(i) (Marčenko-Pastur case). Suppose that Assumption 6.(a) holds. In addition, suppose that the entries of Z_1 are independent, and there is a constant $\kappa > 1$ not depending on n such that $\min_{1\leq j\leq p} \mathbb{E}[Z_{1j}^4] \geq \kappa$. Then, Assumption 6.(c) holds.*
(ii) (Elliptical case). Let V be a random vector that is uniformly distributed on the unit sphere of $\mathbb{R}^p$, and let ξ be a non-negative scalar random variable independent of V that satisfies $\mathbb{E}[\xi^2] = p$ and $\mathbb{E}[\xi^4] < \infty$. Under these conditions, if Z_1 has the same distribution as ξV, then Assumption 6.(c) holds.

The proof of Proposition 1 is given in Section S1 of the supplementary material.

Bootstrap approximation. The following theorem is the central result of the paper, and quantifies the accuracy of the bootstrap when it is used to approximate the distribution of $\sqrt{n}(\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma))$.

Theorem 11. *Suppose that Assumption 6 holds and let $q = 5\log(kn)$. Then, there is a constant $c > 0$ not depending on n such that the event*

$$\sup_{t\in\mathbb{R}^k} \left| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)\big) \preceq t\Big) - \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star) - \boldsymbol{\lambda}_k(\widehat{\Sigma})\big) \preceq t \,\Big|\, X\Big) \right| \;\leq\; \frac{c\,\log(n)\,\beta_{3q}^3\,\mathtt{r}(\Sigma)}{\sqrt{n}} \tag{4.3.1}$$

holds with probability at least $1 - c/n$.

Remarks. The proof of Theorem 11 is given in Section S4 of the supplemenatary material. It is possible to provide a more concrete understanding of the bound (4.3.1) by looking at how the factors $\mathtt{r}(\Sigma)$ and β_{3q} behave in some well-known situations. For instance, consider the class of matrices Σ whose eigenvalues have a polynomial decay profile of the form $\lambda_j(\Sigma) \asymp j^{-\gamma}$, for some fixed constant $\gamma > 0$. This class offers a convenient point of reference, because it interpolates between models that have low-dimensional structure and those that do not. Specifically, the effective rank can be related to γ as

$$\mathtt{r}(\Sigma) \asymp \begin{cases} 1 & \text{if } \gamma > 1 \\ \log(p) & \text{if } \gamma = 1 \\ p^{1-\gamma} & \text{if } \gamma < 1. \end{cases}$$

With regard to the parameter β_{3q}, its dependence on q is simple to describe in some commonly considered cases. If the entries of Z_1 are i.i.d. and sub-Gaussian, then β_{3q} grows at most linearly in q, with $\beta_{3q} \lesssim q\|Z_{11}\|_{\psi_2}^2$. Alternatively, if the entries of Z_1 are i.i.d. and sub-exponential, then β_{3q} grows at most quadratically in q, with $\beta_{3q} \lesssim q^2\|Z_{11}\|_{\psi_1}^2$. (See Chapter 2 of Vershynin (2018) for further details.) Hence, a direct consequence of Theorem 11 in such cases is that bootstrap consistency holds when $\gamma > 1/2$, $p \asymp n$ and $\|Z_{11}\|_{\psi_1} \lesssim 1$. Likewise, when $\gamma > 1$, the bound in Theorem 11 nearly achieves the *parametric rate* $n^{-1/2}$ and is not influenced by the size of p at all. This conclusion also conforms with the numerical results that we present in Section 4.4.

From a more practical standpoint, it is possible to gauge the size of $\mathtt{r}(\Sigma)$ in an empirical way, by either estimating $\mathtt{r}(\Sigma)$ directly, or estimating upper bounds on it. Some

examples of upper bounds on $\mathtt{r}(\Sigma)$ for which straightforward estimation methods are known to be effective in high dimensions include $\mathrm{tr}(\Sigma)/\max_{1\leq j\leq p}\Sigma_{jj}$ and $\mathrm{tr}(\Sigma)^2/\|\Sigma\|_F^2$. (Although guarantees can be established for direct estimates of $\mathtt{r}(\Sigma)$ in high-dimensions, such results can involve a more complex set of considerations than the upper bounds just mentioned.)

Transformations. To briefly review the idea of transformations, they are often used to solve inference problems involving a parameter θ and an estimator $\widehat{\theta}$ for which the distribution of $(\widehat{\theta}-\theta)$ is difficult to approximate. In certain situations, this difficulty can be alleviated if there is a monotone function h for which the distribution of $(h(\widehat{\theta})-h(\theta))$ is easier to approximate. In turn, this allows for more accurate inference on the "transformed parameter" $h(\theta)$, and then the results can be inverted to do inference on θ. In light of this, our next result shows that the rates of bootstrap approximation established in Theorem 11 remain essentially unchanged when using the class of fractional power transformations from $[0,\infty)$ to $[0,\infty)$. This class will be denoted by $\mathcal{H}$, so that if $h\in\mathcal{H}$, then $h(x)=x^a$ for some $a\in(0,1]$.

Beyond the class of transformations just mentioned, the bootstrap can be combined with another type of transformation known as *partial standardization* (Lopes et al., 2020). Letting $h\in\mathcal{H}$ be a given function, and letting $\varsigma_j^2=\mathrm{var}\left(h(\lambda_j(\widehat{\Sigma}))\right)$ for each $j=1,\ldots,p$, this technique is well suited to bootstrapping "max statistics" of the form

$$M=\max_{1\leq j\leq k}\frac{h(\lambda_j(\widehat{\Sigma}))-h(\lambda_j(\Sigma))}{\varsigma_j^{\tau}}, \tag{4.3.2}$$

where $\tau\in[0,1]$ is a parameter that can be viewed as a degree of standardization. The ability to approximate the distribution of M is relevant to the construction of simultaneous confidence intervals for $\lambda_1(\Sigma),\ldots,\lambda_k(\Sigma)$. It also turns out that the choice of τ encodes a trade-off between the coverage accuracy and the width of such intervals, and that choosing an intermediate value $\tau\in(0,1)$ can offer benefits in relation to $\tau=0$ and $\tau=1$. This will be discussed in greater detail later in Section 4.4.

In order to state our extension of Theorem 11 in a way that handles both partial standardization and transformations $h\in\mathcal{H}$ in a unified way, we need to introduce a bit more

notation. First, when considering the bootstrap counterpart of a partially standardized statistic such as (4.3.2), the vector $\boldsymbol{\varsigma}_k^\tau = (\varsigma_1^\tau, \ldots, \varsigma_k^\tau)$ can be replaced with the estimate $\widehat{\boldsymbol{\varsigma}}_k^\tau = (\widehat{\varsigma}_1^\tau, \ldots, \widehat{\varsigma}_k^\tau)$, whose entries are defined by $\widehat{\varsigma}_j^2 = \mathrm{var}\big(h(\lambda_j(\widehat{\Sigma}^\star))\big|X\big)$ for all $j = 1, \ldots, p$. Second, the expression v/u involving vectors v and u denotes the vector obtained by entrywise division, $(v/u)_j = v_j/u_j$. (To handle the possibility zero denominators, events of the form $\{V/\widehat{\boldsymbol{\varsigma}}_k^\tau \preceq t\}$ are understood as $\{V \preceq t \odot \widehat{\boldsymbol{\varsigma}}_k^\tau\}$, where $V \in \mathbb{R}^k$ is random, $t \in \mathbb{R}^k$ is fixed, and $\odot$ is entrywise multiplication. Lemma S5.5 in the supplementary material also shows that such cases occur with negligible probability.) Lastly, recall that we write $\boldsymbol{h}(v) = (h(v_1), \ldots, h(v_k))$ for a k-dimensional vector v and transformation h.

Theorem 12. *Suppose that Assumption 6 holds. Fix a transformation $h \in \mathcal{H}$ and a constant $\tau \in [0,1]$ with respect to n, and let $q = 5\log(kn)$. Then, there is a constant $c > 0$ not depending on n, such that the event*

$$\sup_{t \in \mathbb{R}^k} \left| \mathbb{P}\Big(\frac{\boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma})) - \boldsymbol{h}(\boldsymbol{\lambda}_k(\Sigma))}{\boldsymbol{\varsigma}_k^\tau} \preceq t \Big) - \mathbb{P}\Big(\frac{\boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star)) - \boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma}))}{\widehat{\boldsymbol{\varsigma}}_k^\tau} \preceq t \,\Big|\, X \Big) \right| \;\leq\; \frac{c\,\log(n)\,\beta_{3q}^5\,\mathtt{r}(\Sigma)}{\sqrt{n}}$$

holds with probability at least $1 - c/n$.

Remarks. The proof of Theorem 12 is given in Section S5 of the supplementary material. To comment on the technical relationship between Theorems 11 and 12, it is important to call attention to the differences between asymptotic and non-asymptotic analysis. When using asymptotics, the process of showing that bootstrap consistency for $\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)\big)$ implies the same for $(\boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma})) - \boldsymbol{h}(\boldsymbol{\lambda}_k(\Sigma)))/\boldsymbol{\varsigma}_k^\tau$ can typically be handled with a brief argument, based on the delta method and the consistency of the estimate $\widehat{\boldsymbol{\varsigma}}_k^\tau$. However, when taking a non-asymptotic approach, this process is much more involved. For instance, it is necessary to establish fine-grained error bounds for $\widehat{\boldsymbol{\varsigma}}_k^2$, such as in showing that the uniform relative error $\|(\widehat{\boldsymbol{\varsigma}}_k^2 - \boldsymbol{\varsigma}_k^2)/\boldsymbol{\varsigma}_k^2\|_\infty$ is likely to be at most of order $n^{-1/2}\beta_{2q}^3\lambda_1(\Sigma)^2\mathtt{r}(\Sigma)$ up to logarithmic factors.

4.4 Numerical Results

In this section, we focus on the application of constructing simultaneous confidence intervals for $\lambda_1(\Sigma), \ldots, \lambda_k(\Sigma)$. This will be done in a variety of settings, corresponding to different values of n and p, as well as different values of effective rank, and different

choices of transformations. In a nutshell, there are two overarching conclusions to take away from the experiments: **(1)** In situations where $n \ll p$ and $\mathtt{r}(\Sigma) \asymp 1$, the bootstrap generally produces intervals with accurate coverage, which provides a confirmation of our theoretical results. **(2)** The classical log transformation mostly works well in low dimensions, but it can lead to coverage that is substantially below the nominal level when $\mathtt{r}(\Sigma)$ is moderately large. Nevertheless, we show that it is possible to find transformations that offer more reliable coverage in this challenging case. More generally, this indicates that alternative transformations are worth exploring in high-dimensional settings.

4.4.1 Simulation settings

The eigenvalues of the population covariance matrix Σ were chosen to have two different decay profiles:

(a) A polynomial decay profile $\lambda_j(\Sigma) = j^{-\gamma}$ for all $j = 1, \ldots, p$, with $\gamma \in \{0.7, 1.0, 1.3\}$.

(b) An exponential decay profile $\lambda_j(\Sigma) = \delta^j$ for all $j = 1, \ldots, p$, with $\delta \in \{0.7, 0.8, 0.9\}$.

As a clarification, it is important to note that the effective rank of Σ increases for larger values of δ, but decreases for larger values of γ. For the purposes of simulations, the choices (a) and (b) have the valuable property that the eigenvalues are parameterized in the same way for every choice of p, which facilitates the comparison of results across different dimensions. The matrix of eigenvectors for Σ was drawn uniformly from the set of $p \times p$ orthogonal matrices. The dimension p was taken from $\{10, 50, 100, 200\}$, and the sample size n ranged from 50 to 500. For each triple (n, p, γ) or (n, p, δ), the data $X_1, \ldots, X_n$ were generated in an i.i.d. manner with the following choices for the distribution of X_1:

(i) The vector $X_1 = \Sigma^{1/2}\xi V$ was generated with V being uniformly distributed on the unit sphere of $\mathbb{R}^p$, and ξ^2 being an exponential random variable independent of V with $\mathbb{E}[\xi^2] = p$.

(ii) The vector X_1 was generated from the Gaussian distribution $N(0, \Sigma)$.

For each parameter setting, we generated 1000 realizations of the dataset $X_1, \ldots, X_n$, and for each such realization, we generated $B := 1000$ sets of bootstrap samples of size n. When constructing simultaneous confidence intervals for $\lambda_1(\Sigma), \ldots, \lambda_k(\Sigma)$, the value of k was set to 5.

4.4.2 Bootstrap confidence intervals

For any $\alpha \in (0,1)$, we aim to construct approximate versions of ideal random intervals $\mathcal{I}_1, \ldots, \mathcal{I}_k$ that satisfy

$$\mathbb{P}\left(\bigcap_{j=1}^{k} \left\{ \lambda_j(\Sigma) \in \mathcal{I}_j \right\} \right) \;\geq\; 1 - \alpha. \tag{4.4.1}$$

To this end, consider the following max and min statistics, based on any choice of partial standardization parameter $\tau \in [0,1]$ and transformation h,

$$\begin{aligned} M &= \max_{1 \leq j \leq k} \frac{h(\lambda_j(\widehat{\Sigma})) - h(\lambda_j(\Sigma))}{\varsigma_j^\tau} \\ L &= \min_{1 \leq j \leq k} \frac{h(\lambda_j(\widehat{\Sigma})) - h(\lambda_j(\Sigma))}{\varsigma_j^\tau}. \end{aligned}$$

Letting $q_M(\alpha)$ and $q_L(\alpha)$ denote the respective α-quantiles of M and L for any $\alpha \in (0,1)$, it follows that the desired condition (4.4.1) holds if each interval $\mathcal{I}_j$ is defined as

$$\mathcal{I}_j = h^{-1}\left(\left[h(\lambda_j(\widehat{\Sigma})) - \varsigma_j^\tau q_M(1 - \tfrac{\alpha}{2}) \;,\; h(\lambda_j(\widehat{\Sigma})) - \varsigma_j^\tau q_L(\tfrac{\alpha}{2}) \right] \right), \tag{4.4.2}$$

with $h^{-1}([a,b])$ being understood as the preimage of $[a,b]$ under h.

To construct bootstrap intervals $\widehat{\mathcal{I}}_1, \ldots, \widehat{\mathcal{I}}_k$ based on (4.4.2), it is only necessary to replace $q_M(1-\frac{\alpha}{2})$, $q_L(\frac{\alpha}{2})$, and $\varsigma_1, \ldots, \varsigma_k$ with estimates. In detail, each ς_j is first estimated using the sample variance of B bootstrap replicates of the form $h(\lambda_j(\widehat{\Sigma}^\star))$. Next, the empirical $1-\frac{\alpha}{2}$ quantile of B bootstrap replicates of the form $M^\star = \max_{1 \leq j \leq k}[h(\lambda_j(\widehat{\Sigma}^\star)) - h(\lambda_j(\widehat{\Sigma}))]/\widehat{\varsigma}_j^\tau$ is taken as an estimate of $q_M(1-\frac{\alpha}{2})$, and similarly for $q_L(\frac{\alpha}{2})$.

Regarding the use of transformations, the following three options were included in the experiments:

- *log transformation*: $h(x) = \log(x)$ with $\tau = 0$.
- *standardization*: $h(x) = x$ with $\tau = 1$.

- *square-root transformation*: $h(x) = x^{1/2}$ with $\tau \in [0, 1]$ chosen data-adaptively.

In the case of the log transformation, the choice of $\tau = 0$ corresponds to the way that this transformation has been used in the classical literature (Beran and Srivastava, 1985), while in the case of standardization, the choice of $\tau = 1$ is definitional. For the square-root transformation, the use of a data-adaptive selection rule for $\tau \in [0, 1]$ is more nuanced, and can be informally explained in terms of the following ideas developed previously in (Lin et al., 2021; Lopes et al., 2020).

In essence, this choice can be understood in terms of a trade-off between two competing effects that occur in the extreme cases of $\tau = 1$ and $\tau = 0$. When using $\tau = 1$, the random variables $[\lambda_j(\widehat{\Sigma})^{1/2} - \lambda_j(\Sigma)^{1/2}]/\varsigma_j$ with $\varsigma_j^2 = \text{var}(\lambda_j(\widehat{\Sigma})^{1/2})$ and $j = 1, \ldots, k$ are on approximately "equal footing", which makes the behavior of the statistic M sensitive to their joint distribution (and likewise for L). By contrast, when $\tau = 0$ is used, the variables $[\lambda_j(\widehat{\Sigma})^{1/2} - \lambda_j(\Sigma)^{1/2}]$ will tend to be on different scales, and the variable on the largest scale, say j', will be the maximizer for M relatively often. In this situation, the statistic M is governed more strongly by the marginal distribution of $[\lambda_{j'}(\widehat{\Sigma})^{1/2} - \lambda_{j'}(\Sigma)^{1/2}]$. So, from this heuristic point of view, the choice of $\tau = 0$ can simplify the behavior of M relative to the case of $\tau = 1$, making the distribution of M easier to approximate. However, the choice of $\tau = 0$ also has the drawback that it can lead to simultaneous confidence intervals that are excessively wide, because the widths are no longer adapted to the different values $\varsigma_1, \ldots, \varsigma_k$ (since $\varsigma_1^0 = \cdots = \varsigma_k^0 = 1$).

To strike a balance between these competing effects, we used the following simple rule to select τ in the case of the square-root transformation. For a candidate value of τ, let $\widehat{\mathcal{I}}_1(\tau), \ldots, \widehat{\mathcal{I}}_k(\tau)$ denote the associated bootstrap intervals defined beneath equation (4.4.2) (so that the dependence on τ is explicit), and let $|\widehat{\mathcal{I}}_1(\tau)|, \ldots, |\widehat{\mathcal{I}}_k(\tau)|$ denote their widths. Also define $\widehat{\mu}(\tau) = \sum_{j=1}^k |\widehat{\mathcal{I}}_j(\tau)|/k$ and $\widehat{\sigma}(\tau)^2 = \sum_{i=1}^k (|\widehat{\mathcal{I}}_j(\tau)| - \widehat{\mu}(\tau))^2/k$. In this notation, we selected the value of τ that minimized $\widehat{\mu}(\tau) + \widehat{\sigma}(\tau)$ over the set of candidates $\{0.0, 0.1, \ldots, 0.9, 1.0\}$. Different variants of this type of criterion minimization rule have also been observed to be effective in other contexts (Lin et al., 2021; Lopes et al., 2020).

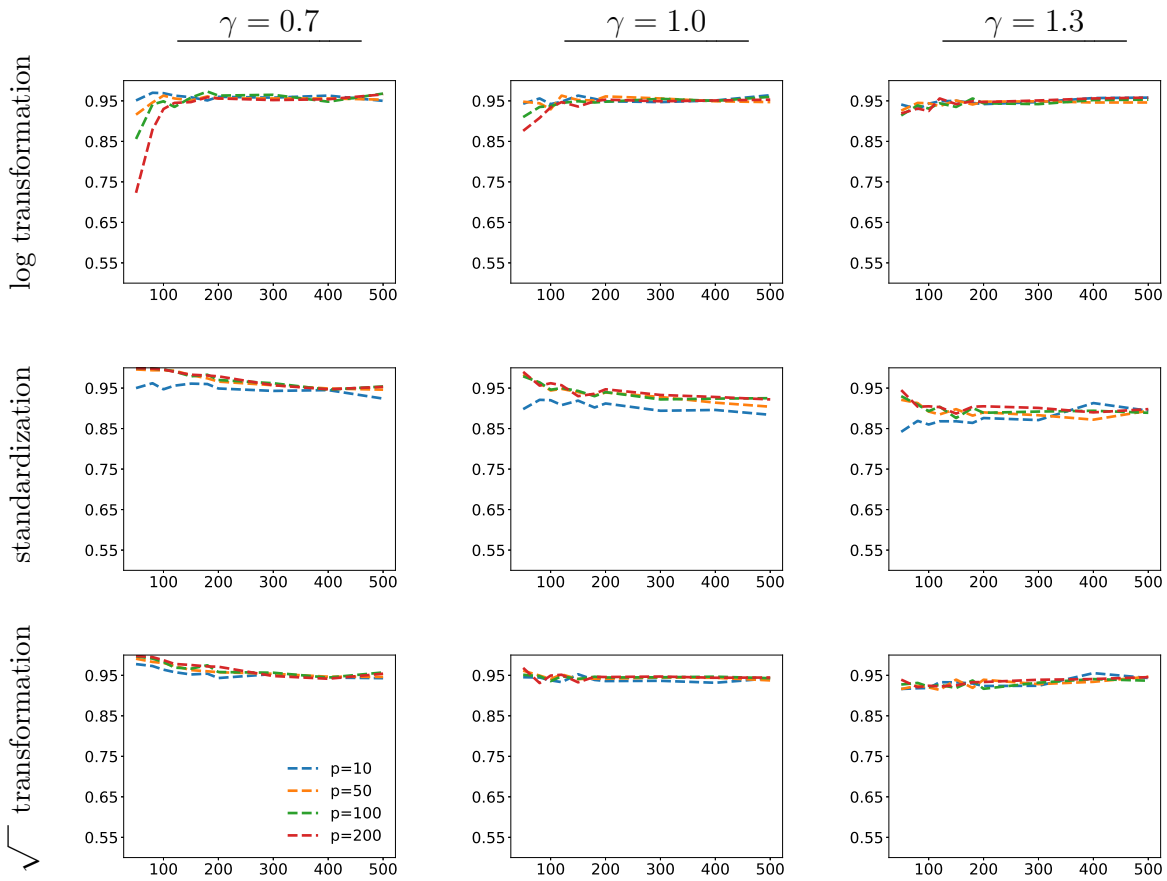


Figure 4.1: (Simultaneous coverage probability versus n in simulation model (i) with a polynomial decay profile). In each panel, the y-axis measures $\mathbb{P}(\cap_{j=1}^{5}\{\lambda_j(\Sigma) \in \widehat{\mathcal{I}}_j\})$ based on a nominal value of 95%, and the x-axis measures n. The colored curves correspond to the different values of p, indicated in the legend.

4.4.3 Discussion of coverage

Figure 4.1 contains nine panels displaying the results for the simultaneous coverage probability $\mathbb{P}(\cap_{j=1}^{5}\{\lambda_j(\Sigma) \in \widehat{\mathcal{I}}_j\})$, based on a nominal value of 95% (i.e. $\alpha = 0.05$) in the case of the simulation model (i) with a polynomial decay profile for the population eigenvalues. The figure summarizes a large amount of information, because it shows how the coverage depends on n, p, the eigenvalue decay parameter γ, and the three transformations described above. For each panel, the x-axis measures n, and the y-axis measures $\mathbb{P}(\cap_{j=1}^{5}\{\lambda_j(\Sigma) \in \widehat{\mathcal{I}}_j\})$. Results corresponding to the dimensions $p = 10, 50, 100, 200$ are plotted with colored curves that are labeled in the legend. The three rows of panels from

top to bottom correspond to the log transformation, ordinary standardization, and the square-root transformation. The three columns of panels from left to right correspond to the eigenvalue decay parameters $\gamma = 0.7, 1.0, 1.3$. In addition, Figure 4.2 displays analogous results for exponentially decaying population eigenvalues in model (i). Lastly, results for model (ii), as well as for a nominal value of 90% (instead of 95%), are provided in Section S8 of the supplementary material.

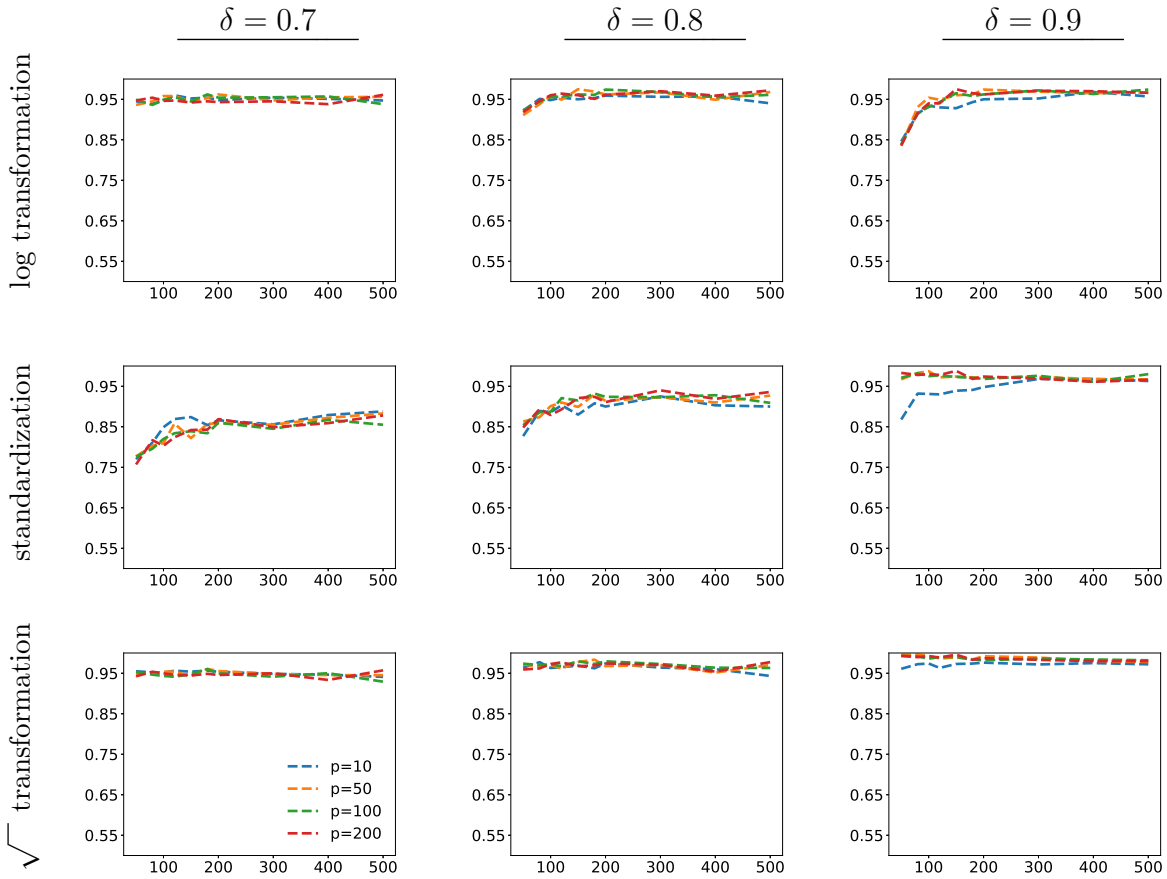
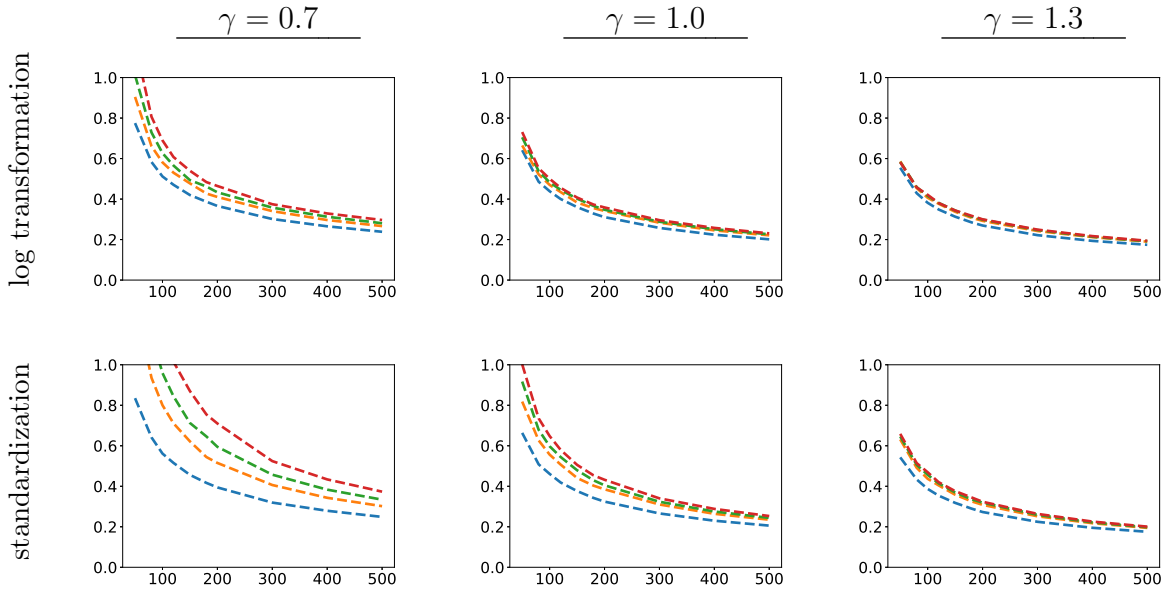


Figure 4.2: (Simultaneous coverage probability versus n in simulation model (i) with an exponential decay profile). The plotting scheme is the same as described in the caption of Figure 4.1, except that the three columns correspond to values of the eigenvalue decay parameter δ.

There are several notable patterns in Figures 4.1 to discuss. The first is that faster rates of decay tend to lead to better coverage accuracy—as anticipated by our theoretical results. In particular, when the eigenvalue decay parameter is set to $\gamma = 1.3$, the

coverage is rather accurate even when $n \ll p$. Furthermore, the accuracy is essentially unaffected by the dimension p in this situation, as indicated by the overlap of the four colored curves. On the other hand, as the decay parameter becomes smaller, the three transformations perform in different ways. For instance, when $\gamma = 0.7$, $p = 200$, and $n < 200$, the log transformation yields coverage that clearly falls short of the nominal level. By contrast, the standardization and square-root transformations tend to err more safely in the conservative direction when $\gamma = 0.7$. To give some indication of the difficulty of $\gamma = 0.7$, it should be noted that if γ were decreased slightly to 0.5 with $p \gtrsim n$, this would imply $\mathtt{r}(\Sigma)/\sqrt{n} \asymp \sqrt{p/n} \gtrsim 1$, in which case bootstrap consistency would not be guaranteed. When considering all three cases $\gamma = 0.7, 0.8, 0.9$ collectively, the square-root transformation seems to yield the best overall coverage results if conservative errors are viewed as preferable to anti-conservative ones.

Turning to the coverage results for exponential spectrum decay, the log and square-root transformations continue to follow the pattern that faster decay improves coverage accuracy. Also, the log transformation maintains its tendency to err in the anti-conservative direction, while the square-root transformation maintains its tendency to err in the conservative direction. Meanwhile, ordinary standardization yields larger errors in the anti-conservative direction than it did in the previous context.



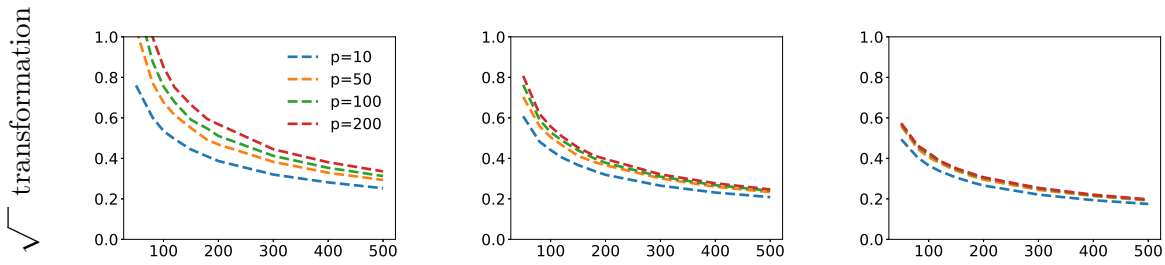


Figure 4.3: (Average width versus n in simulation model (i) with a polynomial decay profile). In each of the nine panels, the y-axis measures the average width $\mathbb{E}[|\widehat{\mathcal{I}}_1| + \cdots + |\widehat{\mathcal{I}}_5|]/5$, and the x-axis measures n. The colored curves correspond to the different values of $p = 10, 50, 100, 200$, indicated in the legend. The three rows and three columns correspond to labeled choices of transformations and values of the eigenvalue decay parameter γ.

4.4.4 Discussion of width

Beyond coverage probability, interval width is another important factor to consider when appraising confidence intervals. In Figures 4.3-C.4, the average width $\mathbb{E}[|\widehat{\mathcal{I}}_1| + \cdots + |\widehat{\mathcal{I}}_k|]/k$ is plotted on the y-axis as a function of the sample size n on the x-axis, with the underlying parameter settings being organized in the same manner as in Figures 4.1-C.2. (Corresponding results for settings based on model (ii) and a nominal value of 90% are presented in Section S8 of the supplementary material.) With regard to the three transformations, they produce intervals that have roughly similar widths across most parameter settings. However, at a more fine-grained level, the results in the case of polynomial spectrum decay show that the log transformation tends to yield slightly shorter widths than the square-root transformation, which in turn, tends to yield slightly shorter widths than ordinary standardization. In the case of exponential spectrum decay with $\delta = 0.9$, the same pattern is also apparent, while for smaller values of δ, there is not much difference among the transformations.

Aside from the transformations, there are two other general trends to notice. Within each of the 18 panels of Figures 4.3-C.4, there is a monotone relationship between width and the dimension p, with the width generally increasing as the dimension increases. Similarly, the width generally also increases as the effective rank $\mathtt{r}(\Sigma)$ increases.

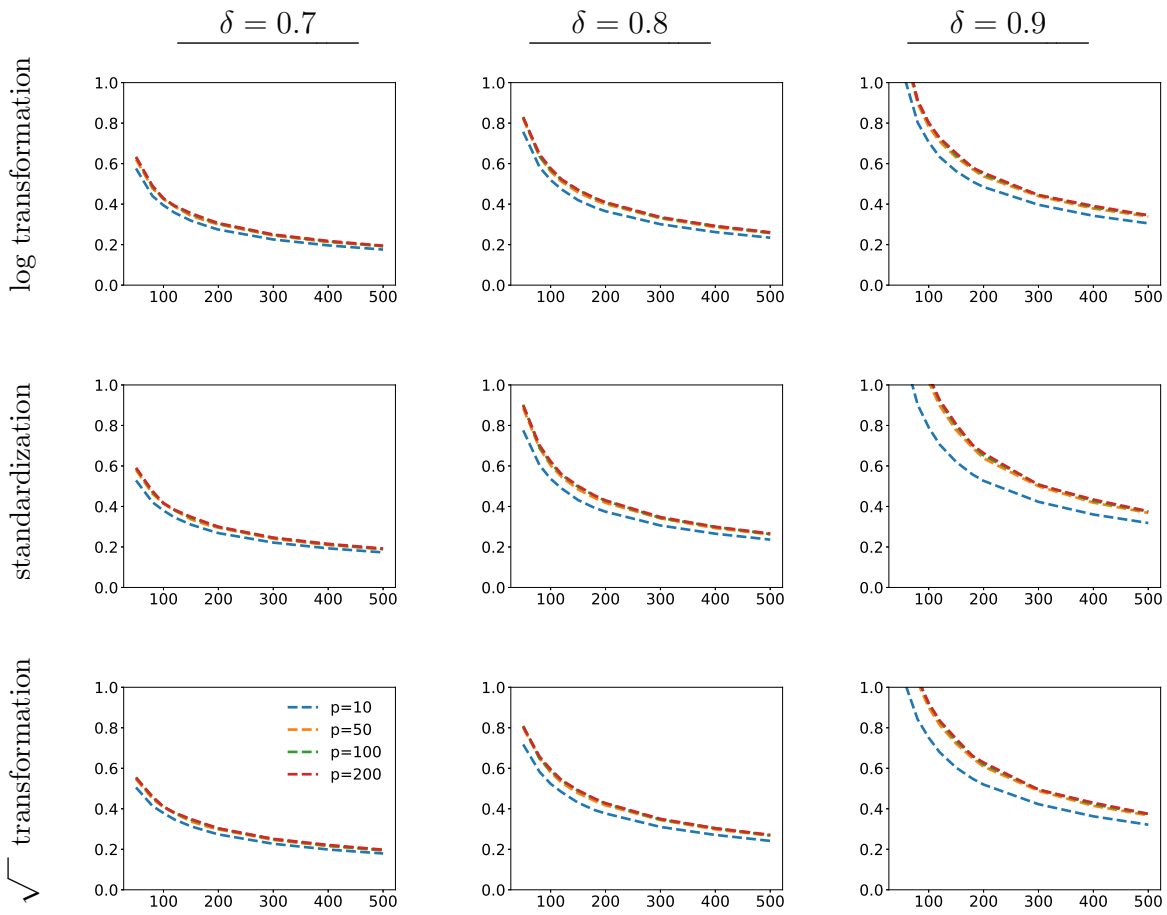


Figure 4.4: (Average width versus n in simulation model (i) with an exponential decay profile). The plotting scheme is the same as described in the caption of Figure 4.3, except that the three columns correspond to values of the eigenvalue decay parameter δ.

4.4.5 Illustration with stock market data

Within the context of finance, PCA is often applied to stock market return data for the purposes of risk analysis and portfolio selection (Fabozzi et al., 2007; Ruppert and Matteson, 2015). Here, we look at several high-dimensional datasets of stock market returns to illustrate how the bootstrap can be applied to do inference on parameters of interest in PCA.

Starting from one dataset of S&P 500 returns during the period February 2013 to December 2017 (Nugent, 2017), we isolated four distinct datasets in the following way. First, we ranked the 500 stocks based on their average monthly trading volume over the

stated time period. Second, we selected four subsets of the 500 stocks, corresponding to the top 50, 150, 200, and 300 members of the ranked list. Third, for each stock, we extracted its biweekly log returns over the time period, resulting in 118 log return values per stock. (The use of log returns rather than ordinary returns is a standard practice in finance (Ruppert and Matteson, 2015).) Altogether, this produced four data matrices of size $n \times p$ with the same number of rows $n = 118$, but differing numbers of columns $p = 50, 150, 200, 300$. In addition to being high-dimensional, these datasets also conform with our interest in settings that are well suited to PCA, since the empirical effective rank satisfies $\mathtt{r}(\widehat{\Sigma}) \leq 4$ for every dataset.

4.4.5.1 Inference on population eigenvalues

When PCA is used to analyze stock market returns, the leading eigenvectors and eigenvalues of the population covariance matrix Σ have special interpretations. Namely, the eigenvector corresponding to $\lambda_1(\Sigma)$ is often viewed as representing an overall "market portfolio", while subsequent eigenvectors represent "principal portfolios", which produce returns that are uncorrelated with the overall market return (Laloux et al., 2000). Also, the eigenvalues can be interpreted as the variances (or volatilities) of the returns associated with the principal portfolios. For this reason, the population eigenvalues are important for risk assessment, and so it is of interest to quantify the uncertainty in these unknown parameters.

For each of the four datasets described above, we applied the bootstrap method with square-root transformation from Section 4.4.2 to construct simultaneous confidence intervals for the leading ten eigenvalues $\lambda_1(\Sigma), \ldots, \lambda_{10}(\Sigma)$. The bootstrap intervals are plotted in Figure 4.5, based on a simultaneous coverage probability of 95%, with a black dot representing the sample eigenvalue $\lambda_j(\widehat{\Sigma})$ in the jth interval for $j = 1, \ldots 10$. Upon close inspection, it can be seen that $\lambda_j(\widehat{\Sigma})$ tends to sit slightly above the midpoint of the jth interval. This is encouraging, because it means that the bootstrap intervals are able to counteract the well-known phenomenon that the leading sample eigenvalues tend to be biased upwards in high-dimensional settings (Yao et al., 2015, Ch.11). In addition, as a way to gain extra empirical support for the bootstrap intervals, we carried out

the following exercise with estimates of $\lambda_1(\Sigma), \ldots, \lambda_{10}(\Sigma)$ computed via the method of QuEST (Ledoit and Wolf, 2015), which is designed for use in high-dimensional settings, and has been adopted frequently in the literature. Specifically, we verified that the QuEST estimate of $\lambda_j(\Sigma)$ was contained in the jth bootstrap interval for every $j = 1, \ldots, 10$ and $p = 50, 150, 200, 300$. Hence, this makes it more plausible that the bootstrap intervals also contain the population eigenvalues.

To comment further on the numerical results in Figure 4.5, first note that in every panel, the interval for $\lambda_1(\Sigma)$ is well separated from the intervals for $\lambda_2(\Sigma), \ldots, \lambda_{10}(\Sigma)$, while there is substantial overlap among the latter intervals. This type of situation occurs frequently when PCA is applied to stock market return data, and this is generally interpreted to mean that the overall behavior of the market has a much more dominant effect on returns than other types of economic factors (Laloux et al., 2000; Ruppert and Matteson, 2015). A second observation is that the bootstrap intervals can provide some additional insight into the relationship between $\lambda_2(\Sigma)$ and $\lambda_3(\Sigma)$. On one hand, a user who only looks at the sample eigenvalues $\lambda_2(\widehat{\Sigma})$ and $\lambda_3(\widehat{\Sigma})$ might be tempted to conclude that there is a clear difference between the population eigenvalues $\lambda_2(\Sigma)$ and $\lambda_3(\Sigma)$. On the other hand, a user who looks at the overlap of the second and third intervals would have more information to see that the difference between $\lambda_2(\Sigma)$ and $\lambda_3(\Sigma)$ might actually be negligible. Lastly, one more aspect of Figure 4.5 to mention is that the relative positions of the ten intervals stay approximately the same for each of the four dimensions $p = 50, 150, 200, 300$. Given that the empirical effective rank satisfies $\mathtt{r}(\widehat{\Sigma}) \leq 4 \ll p$ for every dataset, this makes sense from the standpoint of our theoretical results, which indicate that the bootstrap should be relatively insensitive to the ambient dimension compared to the effective rank.

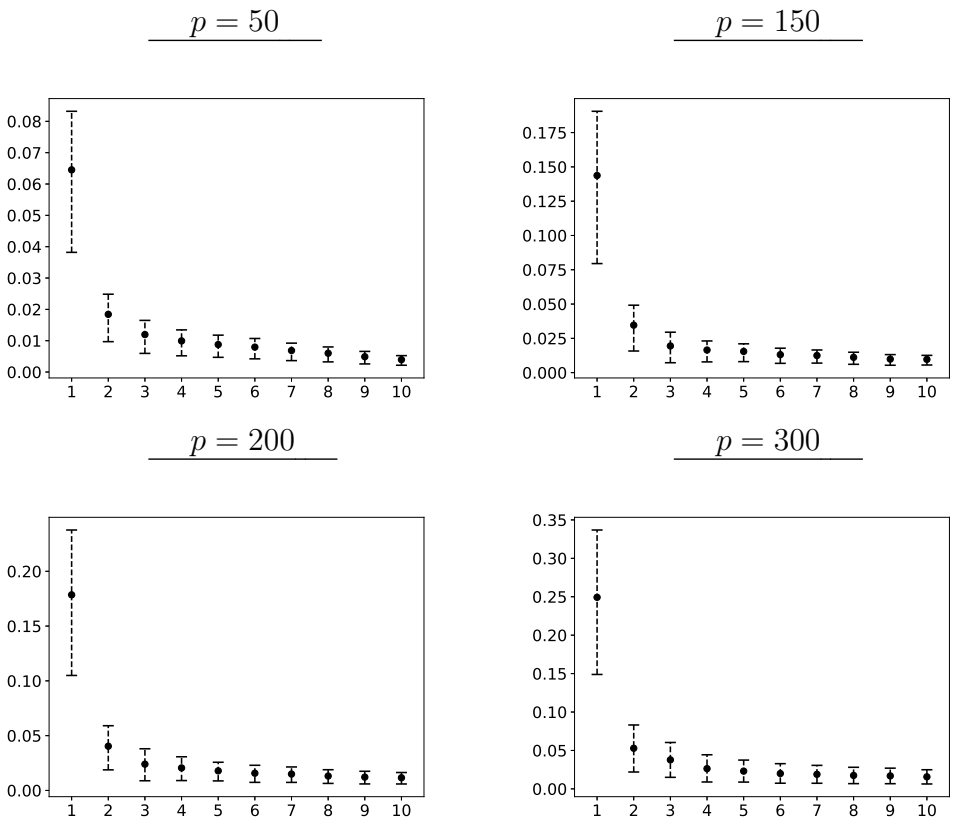


Figure 4.5: (Simultaneous confidence intervals for $\lambda_1(\Sigma), \ldots, \lambda_{10}(\Sigma)$.) In each panel, the y-axis corresponds to the magnitude of eigenvalues, and the x-axis corresponds to the index $j = 1, \ldots, 10$. The intervals are based on a simultaneous coverage probability of 95%, and the black dots represent the sample eigenvalues $\lambda_j(\widehat{\Sigma})$ for each j.

4.4.5.2 Inference on proportions of explained variance

The proportions of explained variance, denoted $\pi_j(\Sigma) = \sum_{i=1}^{j} \lambda_i(\Sigma)/\operatorname{tr}(\Sigma)$ for $j = 1, \ldots, p$, often play a decisive role in applications of PCA, since they form of the basis of standard decision rules for selecting an appropriate number of components. To complement our previous example dealing with the population eigenvalues $\lambda_1(\Sigma), \ldots, \lambda_{10}(\Sigma)$, this subsection looks instead at inference with simultaneous confidence intervals for the parameters $\pi_1(\Sigma), \ldots, \pi_{10}(\Sigma)$. As before, the bootstrap method with square-root transformation from Section 4.4.2 was used to construct the intervals based on a simultaneous coverage probability of 95%. The results are given in Figure 4.6, with black dots showing the locations of the empirical proportions $\pi_j(\widehat{\Sigma})$ within the jth interval.

Due to the fact that the proportions $\pi_1(\Sigma), \ldots, \pi_p(\Sigma)$ are unknown, one of the most widely used rules for selecting the number of components is to choose the smallest number

k for which $\pi_k(\widehat{\Sigma})$ exceeds a given threshold. Although this rule may be appropriate when dealing with low-dimensional data, it is known in the literature that this rule can be unreliable in high-dimensional settings, because it tends to select too few components (Ledoit and Wolf, 2015). One way of avoiding this pitfall is to consider the following simple modification, based on simultaneous bootstrap confidence intervals for the proportions: If $\widehat{l}_j$ denotes the lower endpoint of the jth interval, then the bootstrap-based rule selects the smallest number k for which $\widehat{l}_k$ exceeds the threshold. Consequently, if the intervals perform properly with a simultaneous coverage probability of 95%, then the bootstrap-based rule will select a sufficient number of components with at least 95% probability.

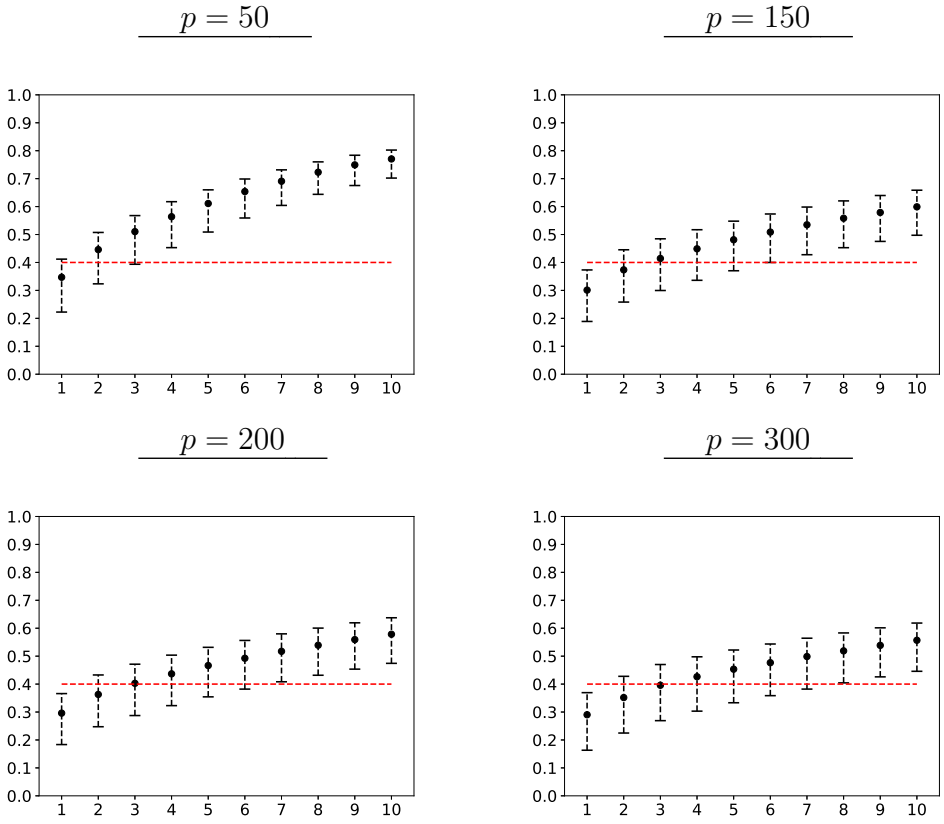


Figure 4.6: (Simultaneous confidence intervals for the proportions of explained variance $\pi_1(\Sigma), \ldots, \pi_{10}(\Sigma)$.) In each panel, the y-axis corresponds to the magnitude of proportions, and the x-axis corresponds to the index $j = 1, \ldots, 10$. The intervals are based on a simultaneous coverage probability of 95%, and the black dots represent the empirical proportions $\pi_j(\widehat{\Sigma})$ for each j.

To illustrate the difference between the two rules, a red line corresponding to a particular threshold of 0.4 has been drawn in each panel of Figure 4.6. (The value of 0.4

has no special importance, and is used only for ease of presentation.) In each of the four cases $p = 50, 150, 200, 300$, the original rule selects the respective values $k = 2, 3, 3, 4$. By contrast, the bootstrap-based rule selects the respective values $k = 4, 6, 7, 8$, and hence, it clearly counteracts the problem of selecting too few components. Moreover, from a financial standpoint, it makes sense that extra components are needed as p increases, because as a wider variety of stocks are included in the data, there is greater opportunity for the returns to be influenced by economic factors that are not captured by the previously leading components.

Appendix A

Supplementary Material for Chapter 2

Organization. We prove the major theorems in the first 7 sections. Intermediate results on truncation and its estimation are presented in Propositions A.8 and A.9. Supporting lemmas are put in Section A.10. Notations follow from those defined in the main text. We also note that the symbols $c, C, \ldots$ are constants not depending on the sample size n throughout the proofs. Their dependence on the parameters can change from line to line. Whenever the domain of a function is clear, we suppress its argument. For example, we use $\tilde{X}$ to denote $\tilde{X}(t)$ when there is no ambiguity.

A.1 Proof of Theorem 1

To show that model (2.2.2) and model (2.2.3) are equivalent, we need to show the following:

1. $\tilde{X}_i(t)$ has mean $\mu(t)$.
2. The new noise e_1 has mean 0 and finite variance.
3. $\tilde{X}_i(t)$ and e_i are uncorrelated.

It is straightforward to verify that $\mathbb{E}[\tilde{X}_i(t)] = \mu(t)$ and $\mathbb{E}[e_i] = 0$. Next, we have

$$\begin{aligned} \mathrm{var}(e_i) &= \mathbb{E}\big[\langle \zeta_0, X_i - \tilde{X}_i \rangle_2 + \epsilon_i\big)^2\big] \\ &= \mathbb{E}\big[\langle \zeta_0, X_i - \tilde{X}_i \rangle_2^2\big] + \sigma^2 + 2\mathbb{E}\big[\epsilon_i \langle \zeta_0, X_i - \tilde{X}_i \rangle_2\big]. \end{aligned}$$

Assumption (2) implies that the first and the third terms are finite. For (c), it suffices to show that $X_i(t) - \tilde{X}_i(t)$ and $\tilde{X}_i(t)$ are uncorrelated. By the discussion in Chapter 11, Section 2 in van der Vaart (1998), for every $t, s \in [0,1]$, we have

$$\mathbb{E}[(X_i(t) - \tilde{X}_i(t))\tilde{X}_i(s)] \;=\; 0.$$

Combining $\mathbb{E}[X_i(t)] = \mathbb{E}[\tilde{X}_i(t)] = 0$, we have, for $s, t \in [0,1]$,

$$\mathrm{cov}(X(t) - \tilde{X}(t), \tilde{X}(s)) \;=\; \mathbb{E}[(X(t) - \tilde{X}(t))\tilde{X}(s)] \;=\; 0.$$

□

A.2 Proof of Theorem 2

Recall that $\widehat{\zeta}_{\mathrm{M}} = \sum_{k=1}^{\mathrm{M}} \widehat{z}_k \widehat{\phi}_k$ and $\zeta_0 = \sum_{j=k}^{\mathrm{M}} z_k \phi_k$. The triangle inequality implies

$$\begin{aligned}
\|\widehat{\zeta}_{\mathrm{M}} - \zeta_0\|_2^2 \;&\leq\; 2\sum_{k=1}^{\mathrm{M}} (\widehat{z}_k - z_k)^2 + 2\left\| \sum_{k=1}^{\mathrm{M}} z_k \widehat{\phi}_k - \zeta_0 \right\|_2 \\
&\leq\; 4\sum_{k=1}^{\mathrm{M}} (\widehat{z}_k - \check{z}_k)^2 + 4\sum_{k=1}^{K} (\check{z}_k - z_k)^2 + 2\left\| \sum_{k=1}^{\mathrm{M}} z_k \widehat{\phi}_k - \zeta_0 \right\|_2^2,
\end{aligned}$$

where $\check{z}_k$ is defined through $g_k = z_k \lambda_k = \widehat{\lambda}_k \check{z}_k$. Lemmas A.10.1, A.10.2, and A.10.3 control each term on the right hand side. Substituting these results into the inequality yields the stated outcome. □

A.3 Proof of Theorem 3

At a high level, the proof for the estimators of $\zeta_0(t)$ consists of three steps. In the first step, we try to solve a surrogate problem, i.e, minimizing the loss $\ell(\zeta; \{\tilde{X}_i\}_{i=1}^n)$, where $\{\tilde{X}_i\}_{i=1}^n$ are the theoretically imputed functions. In the second step, we aim at solving $\min_{\zeta \in \mathcal{H}(K)} \ell(\zeta; \{\tilde{X}_{i,\mathrm{M}_\circ}\}_{i=1}^n)$, where $\tilde{X}_{i,\mathrm{M}_\circ}(t)$ is the truncated $\mathrm{M}_\circ$-dimensional approximation of $\tilde{X}_i(t)$.

We then evaluate the distance between minimizers of $\ell(\zeta; \{\tilde{X}_i\}_{i=1}^n)$ and $\ell(\zeta; \{\tilde{X}_{i,\mathrm{M}_\circ}\}_{i=1}^n)$. Since $\tilde{X}_{i,\mathrm{M}_\circ}$ is not observed, we estimate it in the third step by $\widehat{X}_{i,\mathrm{M}_\circ}$ and then solve $\min_{\zeta \in \mathcal{H}(K)} \ell(\zeta; \{\widehat{X}_{i,\mathrm{M}_\circ}\}_{i=1}^n)$. The third step is concluded by showing the distance between

the minimizers of $\ell(\zeta; \{\tilde{X}_{i,\mathrm{M}_\circ}\}_{i=1}^n)$ and $\ell(\zeta; \{\widehat{X}_{i,\mathrm{M}_\circ}\}_{i=1}^n)$ is small. It will become obvious later that the distance in the last step dominates all other errors in the first two steps.

Recall the estimator from (2.4.6) is $\widehat{\zeta}_\rho$. The estimation error can be broken into three pieces:

$$\|\widehat{\zeta}_{\rho,\mathrm{M}_\circ} - \zeta_0\|_2 \ \leq \ \mathrm{I} + \mathrm{II} + \mathrm{III},$$

where I, II, and III are defined as:

$$\mathrm{I} \ = \ \|\tilde{\zeta}_\rho - \zeta_0\|_2, \tag{A.3.1}$$

$$\mathrm{II} \ = \ \|\widehat{\zeta}_{\rho,\mathrm{M}_\circ} - \tilde{\zeta}_{\rho,\mathrm{M}_\circ}\|_2, \tag{A.3.2}$$

$$\mathrm{III} \ = \ \|\tilde{\zeta}_{\rho,\mathrm{M}_\circ} - \tilde{\zeta}_\rho\|_2, \tag{A.3.3}$$

and $\tilde{\zeta}_\rho$ and $\tilde{\zeta}_{\rho,\mathrm{M}_\circ}$ are

$$\tilde{\zeta}_\rho(t) \ = \ \mathrm{argmin}_{\zeta\in\mathcal{H}(K)}\left\{\frac{1}{n}\sum_{i=1}^n \left(Y_i - \langle\zeta, \tilde{X}_i\rangle_2\right)^2 + \rho\left\|\zeta\right\|_{\mathcal{H}(K)}^2\right\}, \tag{A.3.4}$$

$$\tilde{\zeta}_{\rho,\mathrm{M}_\circ}(t) \ = \ \mathrm{argmin}_{\zeta\in\mathcal{H}(K)}\left\{\frac{1}{n}\sum_{i=1}^n \left(Y_i - \langle\zeta, \widehat{X}_{i,\mathrm{M}_\circ}\rangle_2\right)^2 + \rho\left\|\zeta\right\|_{\mathcal{H}(K)}^2\right\}. \tag{A.3.5}$$

The difference between (A.3.5) and (2.4.6) is that $\widehat{X}_{i,\mathrm{M}_\circ}(t)$ is replaced by its theoretical counterpart $\tilde{X}_{i,\mathrm{M}_\circ}(t)$ in (2.4.6). We complete the proof in the subsequent paragraphs by showing that the three terms on the right hand side of (A.3) are small. □

Proof of term I

First, we introduce a few notations to help present the proof. The sample covariance of $\{\tilde{X}_i(t)\}_{i=1}^n$ is

$$\tilde{\Gamma}_n(s,t) = \frac{1}{n}\sum_{i=1}^n \tilde{X}_i(s)\tilde{X}_i(t).$$

Recall that the kernel R satisfies

$$L_R(f) \ = \ L_{K^{1/2}}(L_{\tilde{\Gamma}}(L_{K^{1/2}}(f))), \quad f \in L_2[0,1].$$

Based on $\tilde{\Gamma}(s,t)$, the empirical estimate R_n of R satisfies

$$L_{R_n}(f) \ = \ L_{K^{1/2}}(L_{\tilde{\Gamma}_n}(L_{K^{1/2}}(f))), \quad f \in L_2[0,1].$$

Also recall that $\mathcal{H}(K) = L_{K^{1/2}}(\mathcal{L}_2[0,1])$. So there exist unique functions $\mathrm{f}_0, \mathrm{f}_\rho \in \mathcal{L}_2[0,1]/\ker(L_{K^{1/2}})$ such that $\zeta_0 = L_{K^{1/2}}(\mathrm{f}_0)$ and $\tilde{\zeta}_\rho = L_{K^{1/2}}(\mathrm{f}_\rho)$. The optimization problem (A.3.4) is equivalent to

$$\mathrm{f}_\rho \;=\; \mathrm{argmin}_{f\in\mathcal{L}_2[0,1]}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \langle \tilde{X}_i, L_{K^{1/2}}(f)\rangle_2\right)^2 + \rho\,\|f\|_2^2\right\}.$$

Some algebra gives us the solution

$$\mathrm{f}_\rho \;=\; (L_{R_n} + \rho\mathbf{1})^{-1}(L_{R_n}(\mathrm{f}_0) + h_n),$$

where $h_n = \frac{1}{n}\sum_{i=1}^{n} e_i L_{K^{1/2}}(\tilde{X}_i)$. We write the population counterpart of f_ρ as

$$\mathrm{f}_\rho^* \;=\; (L_R + \rho\mathbf{1})^{-1} L_R(\mathrm{f}_0).$$

It is straightforward to verify the following inequality

$$\left\|\tilde{\zeta}_\rho - \zeta_0\right\|_2^2 \;=\; \left\|L_{K^{1/2}}(\mathrm{f}_\rho - \mathrm{f}_0)\right\|_2^2 \;\le\; c\left\|\mathrm{f}_\rho - \mathrm{f}_0\right\|_2^2. \tag{A.3.6}$$

The term $\|\mathrm{f}_\rho - \mathrm{f}_0\|_2$ on the left hand side can be further bounded by

$$\left\|\mathrm{f}_\rho - \mathrm{f}_0\right\|_2 \;\le\; \left\|\mathrm{f}_\rho - \mathrm{f}_\rho^*\right\|_2 + \left\|\mathrm{f}_\rho^* - \mathrm{f}_0\right\|_2. \tag{A.3.7}$$

To bound the first term, we apply the triangle inequality again and obtain

$$\left\|\mathrm{f}_\rho - \mathrm{f}_\rho^*\right\|_2 \;\le\; \mathrm{E}_1 + \mathrm{E}_2, \tag{A.3.8}$$

where two terms on the right hand side are defined as

$$\begin{aligned}
\mathrm{E}_1 \;&=\; \left\|\left((L_{R_n} + \rho\mathbf{1})^{-1} - (L_R + \rho\mathbf{1})^{-1}\right)(L_{R_n}(\mathrm{f}_0) + h_n)\right\|_2,\\
\mathrm{E}_2 \;&=\; \left\|(L_R + \rho\mathbf{1})^{-1}\left(L_{R_n}(\mathrm{f}_0) + h_n - L_R(\mathrm{f}_0)\right)\right\|_2.
\end{aligned}$$

For the term E_1, we have

$$\begin{aligned}
\mathrm{E}_1 \;&\le\; \left\|(L_{R_n} + \rho\mathbf{1})^{-1} - (L_R + \rho\mathbf{1})^{-1}\right\|_{\mathrm{op}}\left\|L_{R_n}(\mathrm{f}_0) + h_n\right\|_2,\\
&\le\; \rho^{-2}\|R_n - R\|_2\left\|L_{R_n}(\mathrm{f}_0) + h_n\right\|_2,\\
&\le\; \rho^{-2}\|\tilde{\Gamma}_n - \tilde{\Gamma}\|_2\left\|L_{R_n}(\mathrm{f}_0) + h_n\right\|_2,
\end{aligned} \tag{A.3.9}$$

where the second last line follows from Lemma A.10.5, and

$$\mathrm{E}_2 \;\leq\; \rho^{-1}\left(\left\|L_{R_n}(\mathrm{f}_0) - L_R(\mathrm{f}_0)\right\|_2 + \left\|h_n\right\|_2\right), \tag{A.3.10}$$

Applying large sample results for functional data (e.g., Theorems 7.7.2, 7.7.6, 8.1.1, and 8.1.2 in Hsing and Eubank (2015)), we obtain

$$\begin{aligned}
\|\tilde{\Gamma}_n - \Gamma\|_2 &= O_p(n^{-1/2}),\\
L_{R_n}(\mathrm{f}_0) - L_R(\mathrm{f}_0) &= O_p(n^{-1/2}),\\
h_n &= O_p(n^{-1/2}).
\end{aligned}$$

Hence, by substituting these results into (A.3.9) and (A.3.10), we obtain

$$\mathrm{E}_1 + \mathrm{E}_2 \;=\; O_p\big(\rho^{-2}n^{-1/2}\big),$$

and (A.3.8) implies that

$$\|\mathrm{f}_\rho - \mathrm{f}_\rho^*\|_2 \;=\; O_p\big(\rho^{-2}n^{-1/2}\big). \tag{A.3.11}$$

The second term is easily handled by Lemma A.1

$$\left\|\mathrm{f}_\rho^* - \mathrm{f}_0\right\|_2 \;=\; O(\rho^\nu). \tag{A.3.12}$$

Substituting (A.3.12) and (A.3.11) into (A.3.6) yields

$$\|\mathrm{f}_\rho - \mathrm{f}_0\|_2 \;=\; O_p(\rho^{-2}n^{-1/2} + \rho^\nu),$$

which completes the proof. □

Remark. The same rate also holds for the RKHS-norm of $\left\|\tilde{\zeta}_\rho - \zeta_0\right\|_{\mathcal{H}(K)}$ due to the inequality $\left\|\tilde{\zeta}_\rho - \zeta_0\right\|_{\mathcal{H}(K)} \leq c_0\|\tilde{\mathrm{f}}_\rho - \mathrm{f}_0\|_2$, where $c_0 = \sup_{\|h\|_2=1}\|L_{K^{1/2}}(h)\|_{\mathcal{H}(K)}$.

Proof of terms II and III

Both terms II and III can be handled together using Lemma A.10.6, which provides the $\mathcal{L}_2$ distance of the RKHS estimators based on the imputed process and an approximation of it. For example, if we know how to control the difference $\|\tilde{X}_i - \tilde{X}_{i,\mathrm{M}_\circ}\|_2$, we can measure

the asymptotic order of $\|\tilde{\zeta}_\rho - \tilde{\zeta}_{\rho,\mathrm{M}_\circ}\|_2$. Likewise, knowing $\|\tilde{X}_{i,\mathrm{M}_\circ} - \widehat{X}_{i,\mathrm{M}_\circ}\|_2$ yields the asymptotic order of $\|\tilde{\zeta}_{\rho,\mathrm{M}_\circ} - \widehat{\zeta}_{\rho,\mathrm{M}_\circ}\|_2$. Hence, combining Propositions 1 and 2, we obtain

$$\begin{aligned} \mathrm{II} &= O_p\big(\rho^{-2}{\mathrm{M}_\circ}^{\gamma+1}(\log(n)/n)^{1/3}\big), \\ \mathrm{III} &= O_p\big(\rho^{-2}{\mathrm{M}_\circ}^{-(\alpha_1-1)/2}\big). \end{aligned}$$

□

A.4 Proof of Theorem 4

Recall that X^* denotes a new functional trajectory, and $\tilde{X}^*$ is the conditional expectation conditioning on its observations. We use the proposed imputation method to approximate $\tilde{X}^*$ by a truncated estimate $\widehat{X}^*_{\mathrm{M}_\circ}$. Then the expected prediction error is

$$\begin{aligned} \mathbb{E}\big[\big(\langle\zeta_0, \tilde{X}^*\rangle_2 - \langle\zeta_0, \widehat{X}^*_{\mathrm{M}_\circ}\rangle_2\big)^2\big] &= \mathbb{E}\big[\langle\zeta_0, \tilde{X}^* - \widehat{X}^*_{\mathrm{M}_\circ}\rangle_2^2\big] \\ &\leq \big\|\zeta_0\big\|_2^2\,\mathbb{E}\big[\big\|\tilde{X}^* - \widehat{X}^*_{\mathrm{M}_\circ}\big\|_2^2\big] \\ &\leq c\,\big(\mathbb{E}\big[\big\|\tilde{X}^* - \tilde{X}^*_{\mathrm{M}_\circ}\|_2^2\big] + \mathbb{E}\big[\big\|\tilde{X}^*_{\mathrm{M}_\circ} - \widehat{X}^*_{\mathrm{M}_\circ}\big\|_2^2\big]\big), \\ &= O_p\big(\,{\mathrm{M}_\circ}^{-\alpha_1+1} + {\mathrm{M}_\circ}^{2(\gamma+1)}\big(\log(n)/n\big)^{2/3}\,\big), \end{aligned}$$

where the last line has used Propositions 1 and 2. Taking $\mathrm{M}_\circ = (n/\log(n))^{2/(3\alpha_1+6\gamma+3)}$ yields the stated result. □

A.5 Proof of Theorem 5

By the triangle inequality, it is straightforward to verify

$$\begin{aligned} \mathbb{E}\big[\big(\langle\zeta_0, \tilde{X}^*\rangle_2 - \langle\widehat{\zeta}, \widehat{X}^*_{\mathrm{M}_\circ}\rangle_2\,\big)^2\big] &\leq c\,\big(\mathbb{E}\big[\langle\zeta_0 - \widehat{\zeta}, \tilde{X}^*\rangle_2^2\big] + \mathbb{E}\big[\langle\widehat{\zeta}, \tilde{X}^* - \widehat{X}^*_{\mathrm{M}_\circ}\rangle_2^2\big]\big) \\ &= O_p\big(\|\widehat{\zeta} - \zeta_0\|_2^2\big) + O_p\big(\|\tilde{X}^* - \widehat{X}^*_{\mathrm{M}_\circ}\|_2^2\big) \\ &= O_p\big(\,\delta^2 + {\mathrm{M}_\circ}^{-\alpha_1+1} + {\mathrm{M}_\circ}^{2(\gamma+1)}\big(\log(n)/n\big)^{2/3}\,\big), \end{aligned}$$

where the last line has used Proposition 1 and 2. □

A.6 Proof of Theorem 6

The proof follows from substituting the result of Theorem 2 into δ and taking $\mathrm{M} = \mathrm{M}_\circ = (n/\log(n))^{2/(6\alpha_1+6\gamma+3\min\{\alpha_1,2\beta_1\}))}$. Note that the choice of $\mathrm{M}_\circ$ here is different from its choice in Theorem 2. □

A.7 Proof of Theorem 7

Let $X^*(t)$ be a new process independent of the original sample and $\tilde{X}^*(t)$ be its theoretical imputation. The estimated truncated process consisting of the first $\mathrm{M}_\circ$ scores is denoted as $\widehat{X}^*_{\mathrm{M}_\circ}(t)$. Since the estimated slope function $\widehat{\zeta}_{\rho,\mathrm{M}_\circ}$ does not depend on $X^*(t)$, the prediction error is

$$\mathbb{E}\big[\big(\langle \widehat{X}^*_{\mathrm{M}_\circ}, \widehat{\zeta}_{\rho,\mathrm{M}_\circ}\rangle_2 - \langle \tilde{X}^*, \zeta_0\rangle_2\big)^2\big] \;\le\; 2\,\big(\mathrm{T}_1 + \mathrm{T}_2 + \mathrm{T}_3\big),$$

where

$$\mathrm{T}_1 \;=\; \mathbb{E}\big[\langle \widehat{X}^*_{\mathrm{M}_\circ} - \tilde{X}^*, \widehat{\zeta}_{\rho,\mathrm{M}_\circ}\rangle_2^2\big], \tag{A.7.1}$$

$$\mathrm{T}_2 \;=\; \mathbb{E}\big[\langle \tilde{X}^*, \widehat{\zeta}_{\rho,\mathrm{M}_\circ} - \tilde{\zeta}_\rho\rangle_2^2\big], \tag{A.7.2}$$

$$\mathrm{T}_3 \;=\; \mathbb{E}\big[\langle \tilde{X}^*, \tilde{\zeta}_\rho - \zeta_0\rangle_2^2\big]. \tag{A.7.3}$$

Denote $r_n = (\log(n)/n)^{2(\alpha_1-1)/(3(\alpha_1+2\gamma+1))}$. The first term T_1 has a convergence rate $O_p(r_n)$ due to the triangle inequality and Proposition 1 and 2. The proof of (A.3.2) and (A.3.3) in Theorem 3 implies that the second term T_2 has a convergence rate $O_p(\rho^{-4}r_n)$. It remains to prove a bound for the third term T_3.

To this end, we further break down T_3 according to

$$\begin{aligned}\mathbb{E}\big[\langle \tilde{X}^*, \tilde{\zeta}_\rho - \zeta_0\rangle_2^2\big] \;&=\; \big\|L_{R^{1/2}}\mathrm{f}_\rho - L_{R^{1/2}}\mathrm{f}_0\big\|_2^2 \\ &\le\; 2\big(\big\|L_{R^{1/2}}(\mathrm{f}_\rho - \mathrm{f}^*_\rho)\big\|_2^2 + \big\|L_{R^{1/2}}(\mathrm{f}^*_\rho - \mathrm{f}_0)\big\|_2^2\big),\end{aligned} \tag{A.7.4}$$

where $L_{R^{1/2}}$ satisfies $L_R(f) = L_{R^{1/2}}(L_{R^{1/2}}(f))$. Next, we will provide a bound for the

second term. It is straightforward to verify

$$\begin{aligned}\left\|L_{R^{1/2}}(\mathrm{f}_\rho^* - \mathrm{f}_0)\right\|_2^2 &= \left\|L_{R^{1/2}}\Big(\sum_{k=1}^{\infty}\Big(\frac{\theta_k}{\rho+\theta_k}f_k - f_k\Big)\varphi_k\Big)\right\|_2^2 \\ &= \sum_{k=1}^{\infty}\frac{\rho^2\,\theta_k\,f_k^2}{(\rho+\theta_k)^2}.\end{aligned}$$

Let $q' = 2/(2+(2\beta_2+\alpha_2-1)/\alpha_2)$ such that $(K_\rho'+1)^{-\alpha_2} < \rho^{q'} \le (K_\rho')^{-\alpha_2}$. Then following a similar argument in the proof of Lemma A.10.4, we have

$$\begin{aligned}\sum_{k=1}^{\infty}\frac{\rho^2\,\theta_k\,f_k^2}{(\rho+\theta_k)^2} &\lesssim \rho^{2-2q'}\int_1^{K_\rho'} x^{-(2\beta_2+\alpha_2)}\mathrm{d}x + \int_{K_\rho'+1}^{\infty} x^{-(2\beta_2+\alpha_2)}\mathrm{d}x \\ &\lesssim \rho^{2-2q'} + \rho^{q'(2\beta_2+\alpha_2-1)/\alpha_2} \qquad \text{(A.7.5)} \\ &= \rho^{2\nu_\circ},\end{aligned}$$

where we define $\nu_\circ = \frac{(2\beta_2+\alpha_2-1)/\alpha_2}{2+(2\beta_2+\alpha_2-1)/\alpha_2}$. Thus we conclude that

$$\left\|L_{R^{1/2}}(\mathrm{f}_\rho^* - \mathrm{f}_0)\right\|_2^2 = O(\rho^{2\nu_\circ}). \qquad \text{(A.7.6)}$$

Now we turn to the first term in (A.7.4), whose proof is more involved. Consider the following breakdown:

$$\begin{aligned}L_{R^{1/2}}(\mathrm{f}_\rho^* - \mathrm{f}_\rho) = {}& L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}L_{R^{1/2}}L_{R^{1/2}}(\mathrm{f}_\rho^* - \mathrm{f}_0) && \text{(A.7.7)} \\ &+ L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}(L_{R_n} - L_R)(\mathrm{f}_\rho^* - \mathrm{f}_0) && \text{(A.7.8)} \\ &+ \rho L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}\mathrm{f}_\rho^* && \text{(A.7.9)} \\ &- L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}h_n && \text{(A.7.10)} \\ &+ L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}(L_{R_n} - L_R)(\mathrm{f}_\rho^* - \mathrm{f}_\rho). && \text{(A.7.11)}\end{aligned}$$

For term (A.7.7), we have

$$\begin{aligned}
&\big\|L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}L_{R^{1/2}}L_{R^{1/2}}(\mathrm{f}_\rho^*-\mathrm{f}_0)\big\|_2^2\\
&\leq\ \big\|L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}L_{R^{1/2}}\big\|_{\mathrm{op}}^2\big\|L_{R^{1/2}}(\mathrm{f}_\rho^*-\mathrm{f}_0)\big\|_2^2\\
&\leq\ \sup_{k\geq 1}\frac{\theta_k}{\theta_k+\rho}\cdot\big\|L_{R^{1/2}}(\mathrm{f}_\rho^*-\mathrm{f}_0)\big\|_2^2\\
&=\ O(\rho^{2\nu_\circ}).
\end{aligned}$$

Turning to the term (A.7.8)

$$\begin{aligned}
&\big\|L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}(L_{R_n}-L_R)(\mathrm{f}_\rho^*-\mathrm{f}_0)\big\|_2^2\\
&\leq\ \big\|L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}\big\|_{\mathrm{op}}^2\big\|L_{R_n}-L_R\big\|_{\mathrm{op}}^2\big\|\mathrm{f}_\rho^*-\mathrm{f}_0\big\|_2^2\\
&\leq\ \left(\sup_{k\geq 1}\frac{\sqrt{\theta_k}}{\theta_k+\rho}\right)^2\cdot\|\tilde{\Gamma}-\Gamma\|_2^2\cdot\rho^{2\nu} \qquad \text{(A.7.12)}\\
&=\ O_p\big(n^{-1}\rho^{2\nu-1}\big),
\end{aligned}$$

where in the second last line, we have used $\|\mathrm{f}_\rho^*-\mathrm{f}_0\|_2=O(\rho^\nu)$ from Lemma A.10.4 and the relation

$$\big\|L_{R^{1/2}}(L_R+\rho\mathbf{1})^{-1}\big\|_{\mathrm{op}}\ =\ \sup_{k\geq 1}\frac{1}{\sqrt{\theta_k}+\rho/\sqrt{\theta_k}}\ \leq\ \frac{1}{2\sqrt{\rho}}, \qquad \text{(A.7.13)}$$

and the last line follows from $\|\tilde{\Gamma}-\Gamma\|_2=O_p(n^{-1/2})$ (Theorems 7.7.6 and 8.1.2 in Hsing and Eubank (2015)).

For the term (A.7.9), we apply the argument in (A.7.5) and obtain

$$\big\|\rho L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}\mathrm{f}_\rho^*\big\|_2^2\ =\ \sum_{k=1}^{\infty}\frac{\rho^2\,\theta_k^3\,f_k^2}{(\theta_k+\rho)^4}\ \lesssim\ \rho^{2\nu'},$$

where the last line defines $\nu'=\frac{(2\beta_2+3\alpha_2-1)/\alpha_2}{4+(2\beta_2+3\alpha_2-1)/\alpha_2}$. Note that $\nu'>\nu_\circ$.

To handle the term (A.7.10), we re-use (A.7.13) and $h_n=O_p(n^{-1/2})$ and obtain

$$\big\|L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}h_n\big\|_{\mathcal{L}_2}^2\ =\ O_p(n^{-1}\rho^{-1}).$$

For the term (A.7.11), using (A.3.11) and (A.7.12) we conclude that

$$\begin{aligned}
&\big\|L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}(L_{R_n}-L_R)(\mathrm{f}^*_\rho-\mathrm{f}_\rho)\big\|_2^2\\
\leq\ &\big\|L_{R^{1/2}}(L_R+\rho\,\mathbf{1})^{-1}(L_{R_n}-L_R)\big\|_{\mathrm{op}}^2\big\|\mathrm{f}^*_\rho-\mathrm{f}_\rho\big\|_2^2\\
=\ &O_p(n^{-2}\rho^{-5}).
\end{aligned}$$

Combining all these five results yields a bound for the second term in (A.7.4):

$$\big\|L_{R^{1/2}}(\mathrm{f}^*_\rho-\mathrm{f}_\rho)\big\|_2^2\ =\ O_p(\rho^{2\nu_\circ}+\rho^{-1}n^{-1}).$$

Lastly, taking $\rho=r_n^{1/(4+2\nu_\circ)}$ in (A.7.1), (A.7.2), and (A.7.3) completes the proof. □

A.8 Proof of Proposition 1

Recall $\tilde{A}_{ik}=\mathbb{E}[A_{ik}|\mathbf{W}_i(\mathbf{T}_i)]$. Jensen's inequality implies

$$\begin{aligned}
\mathbb{E}\big[\big\|\tilde{X}_i(t)-\tilde{X}_{i,\mathrm{M}_\circ}(t)\big\|_2\big]\ &=\ \mathbb{E}\bigg[\bigg(\sum_{k=\mathrm{M}_\circ+1}^{\infty}\tilde{A}_{ik}^2\bigg)^{1/2}\bigg]\\
&\leq\ \bigg(\sum_{k=\mathrm{M}_\circ+1}^{\infty}\mathbb{E}[\tilde{A}_{ik}^2]\bigg)^{1/2}\\
&=\ c\bigg(\sum_{k=\mathrm{M}_\circ+1}^{\infty}k^{-\alpha_1}\bigg)^{1/2}\\
&\leq\ c\,{\mathrm{M}_\circ}^{-(\alpha_1-1)/2}.
\end{aligned}$$

An application of Markov inequality leads to the stated result. □

A.9 Proof of Proposition 2

The local linear smoothing method is used to estimate the mean function $\mu(t)$ and covariance function $\Gamma(s,t)$. Zhang and Wang (2016) assume that the number of observations per subject N_i is fixed, while in this paper N_i are random. With a slight modification of the proof in of Corollaries 5.2 and 5.4 in Zhang and Wang (2016), we can obtain

$$\begin{aligned}
\sup_{t\in[0,1]}|\widehat{\mu}(t)-\mu(t)|\ &=\ O\bigg(h_\mu^2+\sqrt{\frac{\log(n)}{nh_\mu\mathbb{E}[N_1]}}\bigg)\quad a.s.,\\
\sup_{s,t\in[0,1]}|\widehat{\Gamma}(s,t)-\Gamma(s,t)|\ &=\ O\bigg(h_\mu^2+\sqrt{\frac{\log(n)}{nh_\mu\mathbb{E}[N_1]}}+h_\Gamma^2+\sqrt{\frac{\log(n)}{nh_\Gamma^2\mathbb{E}[N_1]}}\bigg)\quad a.s.,
\end{aligned}$$

where h_μ, h_Γ are bandwidth choices for estimating mean and covariance. Using these results, we can improve the rates in Corollary 1 and Theorem 2 of Yao et al. (2005a) with the choices of $h_\mu = (\log(n)/n)^{1/5}$ and $h_\Gamma = (\log(n)/n)^{1/6}$. Hence we obtain

$$\sup_{t\in[0,1]} |\widehat{\mu}(t) - \mu(t)| \;=\; O\big((\log(n)/n)^{2/5}\big) \quad a.s., \tag{A.9.1}$$

$$\sup_{s,t\in[0,1]} \big|\widehat{\Gamma}(s,t) - \Gamma(s,t)\big| \;=\; O\big((\log(n)/n)^{1/3}\big) \quad a.s., \tag{A.9.2}$$

$$|\widehat{\sigma}_\eta^2 - \sigma_\eta^2| \;=\; O_p\big((\log(n)/n)^{1/3}\big). \tag{A.9.3}$$

By Assumption (A2.3) and applying Theorem 5.1.8 of Hsing and Eubank (2015), we obtain

$$\|\widehat{\phi}_k - \phi_k\|_2 \;=\; O_p\big(k^\gamma\,(\log(n)/n)^{1/3}\big). \tag{A.9.4}$$

With (A.9.1), (A.9.2) and (A.9.4) in hand, we have

$$\big|\widehat{A}_{ik} - \tilde{A}_{ik}\big| \;=\; O_p\big(\big\|\widehat{\Sigma}^{-1}_{\mathbf{W}_i(\mathbf{T}_i)} - \Sigma^{-1}_{\mathbf{W}_i(\mathbf{T}_i)}\big\|_{\text{op}} + \big\|\widehat{\lambda}_k\widehat{\boldsymbol{\phi}}_{ik} - \lambda_k\boldsymbol{\phi}_{ik}\big\|_2\big).$$

For the first term on the right hand side, we apply Lemma A.3 in Facer and Müller (2003) and obtain

$$\big\|\widehat{\Sigma}^{-1}_{\mathbf{W}_i(\mathbf{T}_i)} - \Sigma^{-1}_{\mathbf{W}_i(\mathbf{T}_i)}\big\|_{\text{op}} \;\le\; c\,\big\|\Sigma^{-1}_{\mathbf{W}_i(\mathbf{T}_i)}\big\|^2_{\text{op}}\big\|\widehat{\Sigma}_{\mathbf{W}_i(\mathbf{T}_i)} - \Sigma_{\mathbf{W}_i(\mathbf{T}_i)}\big\|_{\text{op}}.$$

Since $\|\Sigma^{-1}_{\mathbf{W}_i(\mathbf{T}_i)}\|_{\text{op}} \le \sigma_\eta^{-2}$, we can further obtain

$$\begin{aligned}
&\big\|\widehat{\Sigma}^{-1}_{\mathbf{W}_i(\mathbf{T}_i)} - \Sigma^{-1}_{\mathbf{W}_i(\mathbf{T}_i)}\big\|_{\text{op}} \\
&= O_p\Big(\big\|\widehat{\Sigma}_{\mathbf{W}_i(\mathbf{T}_i)} - \Sigma_{\mathbf{W}_i(\mathbf{T}_i)}\big\|_{\text{op}}\Big) \\
&= O_p\Big(\sup_{1\le j_1,j_2\le N_i}\Big|[\widehat{\Sigma}_{\mathbf{W}_i(\mathbf{T}_i)}]_{j_1j_2} - [\Sigma_{\mathbf{W}_i(\mathbf{T}_i)}]_{j_1j_2}\Big|\Big) \\
&= O_p\Big(\big|\widehat{\sigma}_\eta^2 - \sigma_\eta^2\big| + \sup_{s,t\in[0,1]}\big|\widehat{\Gamma}(s,t) - \Gamma(s,t)\big|\Big) \\
&= O_p\big((\log(n)/n)^{1/3}\big),
\end{aligned} \tag{A.9.5}$$

where the last line has used (A.9.1) and (A.9.3).

For the second term, notice that

$$\big\|\widehat{\lambda}_k\widehat{\boldsymbol{\phi}}_{ik} - \lambda_k\boldsymbol{\phi}_{ik}\big\|_2 \;\le\; \sqrt{N_i}\sup_{1\le j\le N_i}\big|\widehat{\lambda}_k\widehat{\phi}_k(T_{ij}) - \lambda_k\phi_k(T_{ij})\big|,$$

and

$$\lambda_k\phi_k(t) \;= \int_0^1 \Gamma(s,t)\phi_k(s)\mathrm{d}s.$$

The triangle inequality implies

$$\begin{aligned}
&\sup_{1\le j\le N_i}\big|\widehat{\lambda}_k\widehat{\phi}_k(T_{ij}) - \lambda_k\phi_k(T_{ij})\big| \\
&= \sup_{1\le j\le N_i}\Big|\int_0^1 \widehat{\Gamma}(s,t_{ij})\widehat{\phi}_k(s)\mathrm{d}s - \int_0^1 \Gamma(s,t_{ij})\phi_k(s)\mathrm{d}s\Big| \\
&\le \sup_{t\in[0,1]}\Big|\int_0^1 (\widehat{\Gamma}(s,t)-\Gamma(s,t))\widehat{\phi}_k(s)\mathrm{d}s\Big| + \sup_{t\in[0,1]}\int_0^1 \Gamma(s,t)\big|\widehat{\phi}_k(s)-\phi_k(s)\big|\mathrm{d}s \\
&\le c\Big(\sup_{t\in[0,1]}\Big((\widehat{\Gamma}(s,t)-\Gamma(s,t))^2\Big)^{1/2} + \sup_{s,t\in[0,1]}|\Gamma(s,t)|\big\|\widehat{\phi}_k(t)-\phi_k(t)\big\|_2\Big) \\
&\le c\Big(\sup_{s,t\in[0,1]}\big|\widehat{\Gamma}(s,t)-\Gamma(s,t)\big| + \|\widehat{\phi}_k-\phi_k\|_2\Big).
\end{aligned}$$

Thus,

$$\big\|\widehat{\lambda}_k\widehat{\boldsymbol{\phi}}_{ik} - \lambda_k\boldsymbol{\phi}_{ik}\big\|_2 \;=\; O_p\big(k^\gamma(\log(n)/n)^{1/3}\big). \tag{A.9.6}$$

Substituting (A.9.5), (A.9.6) into (A.9) and using (A.9.4), we have

$$\begin{aligned}
\big\|\widehat{X}_{i,\mathrm{M}_\circ}(t) - \tilde{X}_{i,\mathrm{M}_\circ}(t)\big\|_2 \;&=\; \sum_{k=1}^{\mathrm{M}_\circ}|\widehat{A}_{ik}-\tilde{A}_{ik}|O_p(1) + \sum_{k=1}^{\mathrm{M}_\circ}|\tilde{A}_{ik}|\|\widehat{\phi}_k-\phi_k\|_2 \\
&=\; O_p\Big(\sum_{k=1}^{\mathrm{M}_\circ}k^\gamma\,(\log(n)/n)^{1/3}\Big) \\
&=\; O_p\big(\mathrm{M}_\circ{}^{\gamma+1}(\log(n)/n)^{1/3}\big).
\end{aligned}$$

□

A.10 Supporting Lemmas for Theorems

A.10.1 Lemmas for Theorem 2

As a preliminary step, we argue that the convergence rate of the local linear estimator of $g(t)$ is a 1-D rate. The proof of Theorem 1 in Yao et al. (2005a) establishes the convergence

rate of $\widehat{g}(t)$, where Y_{ij} is replaced by V_iU_{ij} with

$$\begin{aligned} V_i &= Y_i + e_i, \\ U_{ij} &= X_i(T_{ij}) + \varepsilon_{ij}. \end{aligned}$$

The substitution does not change the local linear estimator shown in (29). This is because the corresponding Nadaraya-Watson estimator, when Y_{ij} is replaced by V_iU_{ij}, satisfies

$$\mathbb{E}\Big[\frac{\sum_i \sum_j w_{ij} V_i U_{ij}/\mathbb{E}[N]}{\sum_i \sum_j w_{ij}/\mathbb{E}[N]}\Big] = \mathbb{E}[V_iU_{ij}] = \mathbb{E}[Y_iX_i(T_{ij})].$$

Lemma A.10.1. *Suppose the Assumptions in Theorem 2 hold. Then*

$$\sum_{k=1}^{\mathrm{M}} (\widehat{z}_k - \check{z}_k)^2 = O_p\big(\mathrm{M}^{2\alpha_1+1}\, r_1^2 + \mathrm{M}^{2(\alpha_1+\gamma)+1}\, r_2^2\big). \tag{A.10.1}$$

Proof. Recall that $\widehat{z}_k = \widehat{g}_k/\widehat{\lambda}_k$ and $g_k = z_k\lambda_k = \widehat{\lambda}_k\check{z}_k$. We have the following equality,

$$\widehat{z}_k = \check{z}_k + \widehat{\lambda}_j^{-1}(\mathrm{T}_{1k} + \mathrm{T}_{2k} + \mathrm{T}_{3k}), \tag{A.10.2}$$

where T_{1k}, T_{2k}, and T_{3k} are defined as

$$\begin{aligned} \mathrm{T}_{1k} &= \langle \widehat{g} - g, \phi_j \rangle_2, \\ \mathrm{T}_{2k} &= \langle g, \widehat{\phi}_k - \phi_k \rangle_2, \\ \mathrm{T}_{3k} &= \langle \widehat{g} - g, \widehat{\phi}_k - \phi_k \rangle_2. \end{aligned}$$

Combining the relation $\sup_{1\le k} |\widehat{\lambda}_k - \lambda_k| \le \|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}}$, the rate (A.9.2) in the proof of Proposition 2 implies

$$\max_{1\le k\le \mathrm{M}} |\widehat{\lambda}_k - \lambda_k| = O\big((\log(n)/n)^{1/3}\big) \quad a.s..$$

So there exists $n(\mathrm{M}) \ge 1$ such that for all $n \ge n(\mathrm{M})$,

$$\max_{1\le k\le \mathrm{M}} |\widehat{\lambda}_k - \lambda_k| \le \lambda_{\mathrm{M}}/2 \quad a.s.,$$

and thus $\widehat{\lambda}_k \geq \lambda_k/2$ *a.s.* for all $n \geq n(\mathrm{M})$. Substituting (A.10.2) into (A.10.1) yields

$$\begin{aligned}
\sum_{k=1}^{\mathrm{M}} (\widehat{z}_k - \check{z}_k)^2 \;&\leq\; 3 \sum_{k=1}^{\mathrm{M}} \widehat{\lambda}_k^{-2} \big(\mathrm{T}_{1k}^2 + \mathrm{T}_{2k}^2 + \mathrm{T}_{3k}^2 \big) \\
&\leq\; 12 \sum_{k=1}^{\mathrm{M}} \lambda_k^{-2} \big(\mathrm{T}_{1k}^2 + \mathrm{T}_{2k}^2 + \mathrm{T}_{3k}^2 \big) \\
&\leq\; 12 \sum_{k=1}^{\mathrm{M}} \lambda_k^{-2} \big(\mathrm{T}_{1k}^2 + \mathrm{T}_{2k}^2 \big) + 12 \|\widehat{g} - g\|_2^2 \sum_{k=1}^{\mathrm{M}} \lambda_k^{-2} \|\widehat{\phi}_k - \phi_k\|_2^2, \qquad \text{(A.10.3)}
\end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality. The proof is complete once we derive a bound for each term on the right hand side of (A.10.3).

First, we apply Cauchy-Schwarz inequality again to T_{1k},

$$|\mathrm{T}_{1k}| \;\leq\; \|\widehat{g} - g\|_2,$$

and thus obtain

$$|\mathrm{T}_{1k}|^2 \;=\; O_p\big(r_1^2\big). \qquad \text{(A.10.4)}$$

For the term T_{2k}, using the bound

$$\|\widehat{\phi}_k - \phi_k\|_2 \;\leq\; c\, \delta_k^{-1} \|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}}, \qquad \text{(A.10.5)}$$

and the perturbation theory (Hsing and Eubank, 2015, Chapter 5), we have

$$\begin{aligned}
|\mathrm{T}_{2k}| \;&\leq\; c \|\widehat{\phi}_k - \phi_k\|_2 \qquad \text{(A.10.6)} \\
&\leq\; c\, \delta_k^{-1} \|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}}.
\end{aligned}$$

Hence,

$$\sum_{k=1}^{\mathrm{M}} \lambda_k^{-2} \mathrm{T}_{2k}^2 \;\leq\; c \sum_{k=1}^{\mathrm{M}} j^{2\gamma + 2\alpha_1} \|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}}^2,$$

and

$$\sum_{k=1}^{\mathrm{M}} \lambda_k^{-2} \mathrm{T}_{2k}^2 \;=\; O_p\big(\mathrm{M}^{2(\alpha_1+\gamma)+1}\, r_2^2\big). \qquad \text{(A.10.7)}$$

Lastly, note that due to (A.10.6), $\sum_{k=1}^{\mathrm{M}} \lambda_k^{-2} \|\widehat{\phi}_k - \phi_k\|_2^2$ has the same convergence order as $\sum_{k=1}^{\mathrm{M}} \lambda_k^{-2} \mathrm{T}_{2k}^2$. Hence, combining (A.10.4), (A.10.7), and (A.10.5) yields the stated outcome. □

Lemma A.10.2. *Suppose the assumptions in Theorem 2 hold. Then*

$$\sum_{k=1}^{\mathrm{M}}(\widehat{z}_k - z_k)^2 \;=\; O_p\big(\mathrm{M}^{2(\alpha_1-\beta_1)+1}\, r_2^2\big). \tag{A.10.8}$$

Proof. Recall that $g_k = z_k\lambda_k = \check{z}_k\widehat{\lambda}_k$. Then we have the following equality

$$\begin{aligned}\check{z}_k - z_k \;&=\; g_k/\widehat{\lambda}_k - g_k/\lambda_k \\ &=\; \big(\widehat{\lambda}_k\lambda_k\big)^{-1}(\lambda_k - \widehat{\lambda}_k)g_k.\end{aligned}$$

Using this equal relation and Wely's inequality, we can further derive

$$\begin{aligned}\sum_{k=1}^{\mathrm{M}}(\check{z}_k - z_k)^2 \;&\le\; c\sum_{k=1}^{\mathrm{M}}\lambda_k^{-2}z_k^2(\widehat{\lambda}_k - \lambda_k)^2 \\ &\le\; c\,\mathrm{M}^{2(\alpha_1-\beta_1)+1}\,\|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}}^2. \end{aligned}\tag{A.10.9}$$

Substituting $\|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}} = O_p(r_2)$ into (A.10.9) finishes the proof. □

Lemma A.10.3. *Suppose the assumptions in Theorem 2 hold. Then*

$$\Big\|\sum_{k=1}^{\mathrm{M}} z_j\widehat{\phi}_j - \zeta_0\Big\|_2^2 \;=\; O_p\big(\mathrm{M}^{2(\gamma-\beta_1)+1}\, r_2^2 + \mathrm{M}^{-2\beta_1+1}\big). \tag{A.10.10}$$

Proof. The triangle inequality implies

$$\begin{aligned}\Big\|\sum_{k=1}^{\mathrm{M}} z_k\widehat{\phi}_k - \zeta_0\Big\|_2^2 \;&\le\; 2\sum_{k=1}^{\mathrm{M}} z_k^2\|\widehat{\phi}_k - \phi_k\|_2^2 + 2\sum_{k=\mathrm{M}+1}^{\infty} z_k^2 \\ &\le\; c\,\|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}}^2\sum_{k=1}^{\mathrm{M}} z_k^2\delta_k^{-2} + 2\sum_{k=\mathrm{M}+1}^{\infty} z_k^2. \end{aligned}\tag{A.10.11}$$

For the two summations, we have

$$\sum_{k=1}^{\mathrm{M}} z_k^2\delta_k^{-2} \;=\; \sum_{k=1}^{\mathrm{M}} k^{2(\gamma-\beta_1)} \;=\; O\big(\mathrm{M}^{2(\gamma-\beta_1)+1}\big), \tag{A.10.12}$$

and

$$\sum_{k=\mathrm{M}+1}^{\infty} z_k^2 \;=\; \sum_{k=\mathrm{M}+1} k^{-2\beta_1} \;=\; O\big(\mathrm{M}^{-2\beta_1+1}\big). \tag{A.10.13}$$

Substituting (A.10.12) and (A.10.13) into (A.10.11) completes the proof. □

A.10.2 Lemmas for Theorem 3

Lemma A.10.4. *Suppose that the Assumptions in Theorem 3 hold. There exists a constant* $\nu = \frac{(2\beta_2-1)/\alpha_2}{2+(2\beta_2-1)/\alpha_2} \in (0,1)$ *such that*

$$\|\mathrm{f}_\rho^* - \mathrm{f}_0\|_2 \;=\; O(\lambda^\nu).$$

Proof. It is straightforward to verify

$$\begin{aligned}\|\mathrm{f}_\rho^* - \mathrm{f}_0\|_2^2 \;&=\; \sum_{k=1}^{\infty}\left(\frac{\theta_k f_k}{\rho+\theta_k} - f_k\right)^2 \\ &=\; \sum_{k=1}^{\infty}\frac{\rho^2 f_k^2}{(\rho+f_k)^2}.\end{aligned}$$

Recall that $|f_k| \le ck^{-\beta_2}$ and $\theta_k \asymp k^{-\alpha_2}$. Some algebra gives us

$$\sum_{k=1}^{\infty}\frac{\rho^2 f_k^2}{(\rho+\theta_k)^2} \;\lesssim\; \sum_{k=1}^{\infty}\frac{\rho^2 k^{-2\beta_2}}{(\rho+k^{-\alpha_2})^2}.$$

Without loss of generality, we may assume that $0<\rho<1$. For a fixed ρ, there exists an integer K_ρ such that

$$(K_\rho+1)^{-\alpha_2} \;<\; \rho^q \;\le\; K_\rho^{-\alpha_2},$$

where $q = 2/(2+(2\beta_2-1)/\alpha_2) \in (0,1)$. Due to this inequality, we have

$$\begin{aligned}\sum_{k=1}^{\infty}\frac{\rho^2 k^{-2\beta_2}}{(\rho+k^{-\alpha_2})^2} \;&\le\; \sum_{k=1}^{K_\rho}\frac{\rho^2 k^{-2\beta_2}}{(\rho+\rho^q)^2} + \sum_{k=K_\rho+1}^{\infty} k^{-2\beta_2} \\ &\lesssim\; \rho^{2-2q}\int_1^{K_\rho} x^{-2\beta_2}\mathrm{d}x + \int_{K_\rho+1}^{\infty} x^{-2\beta_2}\mathrm{d}x \\ &\lesssim\; \rho^{2-2q} + \rho^{q(2\beta_2-1)/\alpha_2} \\ &=\; \rho^{2\nu},\end{aligned}$$

which completes the proof. □

Lemma A.10.5. *Let* A, B *be two self-adjoint and compact linear operators. Define* $U = A+\lambda\mathbf{1}$ *and* $V = B+\lambda\mathbf{1}$, $\lambda>0$. *Assume that*

$$\|A-B\|_{\mathrm{op}} \;=\; \Delta,$$

and $\Delta \to 0$. *Then we have*

$$\|U^{-1} - V^{-1}\|_{\text{op}} \;\leq\; \lambda^{-2}\Delta.$$

Proof. Let $M = 2\max\{\|U\|_{\text{op}}, \|V\|_{\text{op}}\}$. Define two scaled operators, $U_0 = U/M$ and $V_0 = V/M$. It is straightforward to verify that $\|U_0\|_{\text{op}}, \|V_0\|_{\text{op}} \in (0,1)$ and

$$\|U^{-1} - V^{-1}\|_{\text{op}} \;=\; M^{-1}\|U_0^{-1} - V_0^{-1}\|_{\text{op}}. \tag{A.10.14}$$

It is sufficient to find a bound for $\|U_0^{-1} - V_0^{-1}\|_{\text{op}}$ in order to bound the right hand side of (A.10.14). Note that we also have $\|\mathbf{1} - U_0\|_{\text{op}}, \|\mathbf{1} - V_0\|_{\text{op}} \in (0,1)$. To verify this, take $v \in \mathcal{L}_2[0,1]$ with $\|v\|_2 = 1$. Let $\{(\lambda_k, \psi_k(t))\}_{k=1}^{\infty}$ be the eigen-pairs of U with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda$. Write $v = \sum_{k=1}^{\infty} v_k\psi_k$ where $v_k = \langle v, \psi_k\rangle$. Then we can verify

$$\begin{aligned}
\|\mathbf{1} - U_0\|_{\text{op}} \;&=\; \sup_{\|v\|_2=1} \|(\mathbf{1} - M^{-1}U)v\|_2, \\
&=\; \sup_{\|v\|_2=1} \Big\| \sum_{k=1}^{\infty} \Big(v_k - \frac{1}{M}\lambda_k v_k\Big)\psi_k \Big\|_2 \\
&=\; \sup_{\|v\|_2=1} \Big(\sum_{k=1}^{\infty} v_k^2 \Big(1 - \frac{\lambda_k}{M}\Big)^2 \Big)^{1/2} \\
&=\; 1 - \frac{\lambda}{M} < 1.
\end{aligned}$$

The same line of argument applies to $\|\mathbf{1} - V_0\|_{\text{op}}$.

Expanding U_0^{-1} and V_0^{-1} using Neumann series yields

$$U_0^{-1} \;=\; \sum_{k=0}^{\infty}(\mathbf{1} - U_0)^k, \quad V_0^{-1} \;=\; \sum_{k=0}^{\infty}(\mathbf{1} - V_0)^k.$$

The triangle inequality gives us

$$\|U_0^{-1} - V_0^{-1}\|_{\text{op}} \;\leq\; \sum_{k=1}^{\infty} \big\|(\mathbf{1} - U_0)^k - (\mathbf{1} - V_0)^k\big\|_{\text{op}}. \tag{A.10.15}$$

Denote

$$M_0 \;=\; \|\mathbf{1} - U_0\|_{\text{op}} \;=\; \|\mathbf{1} - V_0\|_{\text{op}} \;=\; 1 - \frac{\lambda}{M}.$$

For each $k \geq 1$ we have

$$\begin{aligned}
\left\|(\mathbf{1}-U_0)^k-(\mathbf{1}-V_0)^k\right\|_{\text{op}} &= \left\|(U_0-V_0)\sum_{i=1}^{k}(\mathbf{1}-U_0)^{k-i}(\mathbf{1}-V_0)^{i-1}\right\|_{\text{op}} \\
&\leq \|U_0-V_0\|_{\text{op}}\sum_{i=1}^{k}M_0^{(k-i)+(i-1)} \\
&= \|U_0-V_0\|_{\text{op}}\,kM_0^{k-1}. \qquad \text{(A.10.16)}
\end{aligned}$$

Now we are ready to bound (A.10.14) using (A.10.15) and (A.10.16),

$$\begin{aligned}
\|U^{-1}-V^{-1}\|_{\text{op}} &= M^{-1}\|U_0^{-1}-V_0^{-1}\|_{\text{op}} \\
&\leq M^{-1}\|U_0-V_0\|_{\text{op}}\sum_{k=1}^{\infty}kM_0^{k-1} \\
&= M^{-1}\|U_0-V_0\|_{\text{op}}\Big(\sum_{k=1}^{\infty}x^k\Big)'\Big|_{x=M_0} \\
&= (1-M_0)^{-2}M^{-2}\|A-B\|_{\text{op}} \\
&= (M-(M-\lambda))^{-2}\|A-B\|_{\text{op}} \\
&= \lambda^{-2}\|A-B\|_{\text{op}},
\end{aligned}$$

which completes the proof. □

Lemma A.10.6. *Let $\{z_{1i}(t)\}_{i=1}^n$ and $\{z_{2i}(t)\}_{i=1}^n$ be samples of two processes. Assume that for each $i=1,\ldots,n$, z_{2i} is an approximation for z_{1i}, and the approximation error is uniform across all i,*

$$\left\|z_{1i}-z_{2i}\right\|_2 = O_p(\Delta).$$

Let $\zeta_j, j=1,2$ be the minimizers of the penalized regressions

$$\zeta_j = \mathit{argmin}_{\zeta\in\mathcal{H}(K)}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\langle\zeta,z_{ji}\rangle_2\right)^2+\rho\left\|\zeta\right\|_{\mathcal{H}(K)}^2\right\}.$$

Then

$$\|\zeta_1-\zeta_2\|_2 = O_p(\rho^{-2}\Delta).$$

Proof. The original minization task is equivalent to finding the minimizer of the following quadratic function

$$\mathscr{L}_j(f) \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \langle z_{ji}, L_{K^{1/2}}(f)\rangle_{\mathcal{L}_2}\big)^2 + \rho\,\|f\|_{\mathcal{L}_2}^2, \quad j=1,2,$$

whose solution is

$$\widehat{f_j} \;=\; \big(\mathbb{C}_{n,j} + \rho\mathbf{1}\big)^{-1} L_{K^{1/2}}\bigg\{\frac{1}{n}\sum_{i=1}^{n} Y_i z_{ji}\bigg\},$$

where

$$\mathbb{C}_{n,j}(f) \;=\; L_{K^{1/2}}(L_{C_{n,j}}(L_{K^{1/2}}(f))),$$

and

$$C_{n,j}(t,s) = \frac{1}{n}\sum_{i=1}^{n} z_{1i}(t) z_{1i}(s),$$

is the sample covariance. Then the solutions to the original problems are $\widehat{\zeta_j} = L_{K^{1/2}}(\widehat{f_j})$. Given that for each $i = 1,\ldots,n$, two processes z_{1i} and z_{2i} are close, the goal is to compare how close $\widehat{f_1}$ and $\widehat{f_2}$ are, which further implies how close $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are.

By the triangle inequality, we have

$$\big\|\widehat{f_1} - \widehat{f_2}\big\|_{\mathcal{L}_2} \;\le\; \mathrm{I} + \mathrm{II},$$

where the two terms on the right side are defined according to

$$\mathrm{I} \;=\; \bigg\|(\mathbb{C}_{n,1} + \rho\mathbf{1})^{-1} L_{K^{1/2}}\bigg\{\frac{1}{n}\sum_{i=1}^{n} Y_i(z_{1i} - z_{2i})\bigg\}\bigg\|_2, \tag{A.10.17}$$

$$\mathrm{II} \;=\; \bigg\|\big((\mathbb{C}_{n,1} + \rho\mathbf{1})^{-1} - (\mathbb{C}_{n,2} + \lambda\mathbf{1})^{-1}\big) L_{K^{1/2}}\bigg\{\frac{1}{n}\sum_{i=1}^{n} Y_i z_{2i}\bigg\}\bigg\|_2. \tag{A.10.18}$$

To handle (A.10.17), we have

$$\begin{aligned}
\text{slowromancapi@} \;&\le\; \big\|(\mathbb{C}_{n,1} + \rho\mathbf{1})^{-1}\big\|_{\mathrm{op}} \bigg\| L_{K^{1/2}}\bigg\{\frac{1}{n}\sum_{i=1}^{n} Y_i(z_{1i} - z_{2i})\bigg\}\bigg\|_2 \\
&\le\; c\,\rho^{-1}\cdot\frac{1}{n}\sum_{i=1}^{n}|Y_i|\big\|z_{1i} - z_{2i}\big\|_2 \\
&=\; O_p(\rho^{-1}\Delta).
\end{aligned}$$

For (A.10.18), we can verify that

$$\left\| L_{K^{1/2}} \left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i z_{2i} \right\} \right\|_{\mathcal{L}_2} \;\leq\; \left\| L_{K^{1/2}} \right\|_{\text{op}} \cdot \frac{1}{n} \sum_{i=1}^{n} |Y_i| \left\| z_{2i} \right\|_2 \;=\; O_p(1).$$

So it suffices to find the rate of $\|(\mathbb{C}_{n,1}+\rho\mathbf{1})^{-1}-(\mathbb{C}_{n,2}+\rho\mathbf{1})^{-1}\|_{\text{op}}$. Note that both $\mathbb{C}_{n,1}+\rho\mathbf{1}$ and $\mathbb{C}_{n,2}+\rho\mathbf{1}$ are invertible due to the addition of $\rho\mathbf{1}$. Moreover, we have

$$\begin{aligned} \left\| (\mathbb{C}_{n,1} + \rho\mathbf{1}) - (\mathbb{C}_{n,2} + \rho\mathbf{1}_n) \right\|_{\text{op}} \;&=\; \left\| L_{K^{1/2}} \circ (L_{C_{n,1}} - L_{C_{n,2}}) \circ L_{K^{1/2}} \right\|_{\text{op}} \\ &\leq\; c \left\| C_{n,1} - C_{n,2} \right\|_2, \end{aligned}$$

and

$$\begin{aligned} & \left\| C_{n,1} - C_{n,2} \right\|_2 \\ =\; & \left\| \frac{1}{n} \sum_{i=1}^{n} z_{1i}(t) z_{1i}(s) - \frac{1}{n} \sum_{i=1}^{n} z_{2i}(t) z_{2i}(s) \right\|_2 \\ \leq\; & \left\| \frac{1}{n} \sum_{i=1}^{n} z_{1i}(t)(z_{1i}(s) - z_{2i}(s)) \right\|_2 + \left\| \frac{1}{n} \sum_{i=1}^{n} (z_{2i}(t) - z_{1i}(s)) z_{2i}(s) \right\|_2 \qquad \text{(A.10.19)} \\ \leq\; & \frac{1}{n} \sum_{i=1}^{n} (\left\| z_{1i} \right\|_2 + \left\| z_{2i} \right\|_2) \left\| z_{1i} - z_{2i} \right\|_2 \qquad \text{(A.10.20)} \\ =\; & O_p(\Delta). \end{aligned}$$

Hence, we have

$$\left\| (\mathbb{C}_{n,1} + \rho\mathbf{1}) - (\mathbb{C}_{n,2} + \rho\mathbf{1}) \right\|_{\text{op}} \;=\; O_p(\Delta).$$

Then Lemma A.10.5 implies that

$$\left\| (\mathbb{C}_{n,1} + \rho\mathbf{1})^{-1} - (\mathbb{C}_{n,2} + \rho\mathbf{1})^{-1} \right\|_{\text{op}} \;=\; O_p(\rho^{-2}\Delta).$$

Combining bounds on (A.10.17) and (A.10.18) completes the proof. □

A.11 Discussion on Assumption 3

We propose to impute the process $X_i(t)$ based on the short vector $\mathbf{W}_i(\mathbf{T}_i)$ collected from the subjects, i.e., $\tilde{X}_i(t) = \mathbb{E}[X_i(t)|\mathbf{W}_i(\mathbf{T}_i)]$. By and large this strategy still works, but its

theoretical justification needs more discussion. This section is devoted to closing this gap. We will first introduce elliptical distributions and then discuss how this class of random elements fit in our theory.

Elliptical distributions

For a comprehensive survey on this topic, see, e.g., Frahm (2004). An introduction to elliptical random process can be found in Bali and Boente (2009) and Boente et al. (2014).

Let $\mathbf{X} \in \mathbb{R}^d$ be a d-dimensional random vector. Cambanis et al. (1981) showed that $\mathbf{X}$ is an elliptical random vector $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if and only if $\mathbf{X} \overset{\mathcal{D}}{=} \boldsymbol{\mu} + \mathcal{R}\boldsymbol{\Lambda}\mathbf{U}^{(k)}$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is a non-random constant, $\mathcal{R}$ is a non-negative random variable, $\boldsymbol{\Lambda}$ is a $d \times k$ non-random matrix with rank $k(\leq d)$, and $\mathbf{U}^{(k)}$ is a k-dimensional random vector independent of $\mathcal{R}$ and uniformly distributed on the sphere $\mathcal{S}^{k-1}$. With this definition, we have

$$\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu} \quad \text{and} \quad \operatorname{cov}(\mathbf{X}) = \frac{\mathbb{E}[\mathcal{R}^2]}{d}\boldsymbol{\Sigma}, \tag{A.11.1}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$ has rank k. Note that both the multivariate normal and multivariate t distributions are elliptical distributions.

Next we derive the conditional distribution of an elliptical distribution. Let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top$ and $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$ with $\mathbf{X}_1$ the vector of the first k coordinates of $\mathbf{X}$ ($k < d$) such that $\operatorname{cov}(\mathbf{X}_1)$ is not singular. Let

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{11}$ is a $k \times k$ matrix and $\boldsymbol{\Sigma}_{12}$ is a $k \times (d-k)$ matrix. Then we have (Cambanis et al., 1981),

$$\mathbb{E}[\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1] = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1). \tag{A.11.2}$$

Although $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{21}$ are unobservable, using (A.11.1) we can re-write (A.11.2) as

$$\begin{aligned} \mathbb{E}[\mathbf{X}_2|\mathbf{X}_1 = \mathbf{x}_1] &= \boldsymbol{\mu}_2 + (\mathbb{E}[\mathcal{R}^2]\boldsymbol{\Sigma}_{21}/d)(\mathbb{E}[\mathcal{R}^2]\boldsymbol{\Sigma}_{11}/d)^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \\ &= \boldsymbol{\mu}_2 + \operatorname{cov}(\mathbf{X}_2, \mathbf{X}_1)\operatorname{cov}(\mathbf{X}_1)^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1). \end{aligned} \tag{A.11.3}$$

In the case where $\boldsymbol{\mu}_2 = \mathbf{0}$, this mirrors the conditional expectation relation (2.3.5). We provide two examples that satisfy the linearity Assumption (3).

Example.

(E1) $X_i(t)$ is a Gaussian process, and the measurement errors $\{\eta_{ij}\}_{j=1}^{N_i}$ are i.i.d. independent Gaussian random variables.

(E2) The scores $\{A_{ik}\}_{k=1}^{\infty}$ and measurement errors $\{\eta_{ij}\}_{j=1}^{N_i}$ are jointly elliptical.

Note that multivariate Gaussian is a special case of elliptical distributions. So it is straightforward to verify that the conditional expectation of scores under (E1) and (E2) satisfy the relation (2.3.5) and thus Assumption 3. Furthermore, (E2) is intriguing. Because it is not implied by elliptical $X_i(t)$ and i.i.d. elliptical measurement errors $\{\eta_{ij}\}_{j=1}^{N_i}$ as $(A_{ik}, \mathbf{W}_i(\mathbf{T}_i))$ need not be jointly elliptical under these assumptions. This is very different from the Gaussian case in (E1), where Gaussian $X_i(t)$ and i.i.d. Gaussian measurement errors $\{\eta_{ij}\}_{j=1}^{N_i}$ imply that $(A_{ik}, \mathbf{W}_i(\mathbf{T}_i))$ are jointly Gaussian, hence Assumption 3 is satisfied. The lack of a similar result for elliptical distributions is best understood through the fact that the sum of two independent t-distributed random variables no longer follows a t-distribution.

Alternative assumption on conditional scores

The joint elliptical assumption is not innocent. It does not hold if the measurement errors η_{ij} are independent across j, as commonly assumed in the literature. To accommodate independent measurement errors, we can make the following alternative assumption to replace Assumption (A3′) by

(A3′) The linear relationship

$$\mathbb{E}[A_{ik}|\mathbf{X}_i(\mathbf{T}_i)] = a_{ik} + b_{ik}^{\top}\mathbf{X}_i(\mathbf{T}_i)$$

holds, and the measurement errors $\{\eta_{ij}\}_{j=1}^{N_i}$ are i.i.d. across both i and j.

However, a very different proof for the consistency of estimating $\zeta_0(t)$ is needed with such an assumption, as Proposition 2 does not hold under Assumption (A3′).

Below we provide a short argument to show why $\tilde{X}_{i,\mathrm{M}_\circ}(t)$ can not be consistently estimated under Assumption (A3′).

To begin, it is straightforward to verify

$$\tilde{A}_{ik} \;=\; \mathbb{E}[A_{ik}|\mathbf{X}_i(\mathbf{T}_i)] \;=\; \gamma_k \boldsymbol{\phi}_{ik}^\top \boldsymbol{\Sigma}^{-1}_{\mathbf{X}_i(\mathbf{T}_i)} \mathbf{X}_i(\mathbf{T}_i).$$

The quantities $\gamma_k, \boldsymbol{\phi}_{ik}, \Sigma_{\mathbf{X}_i(\mathbf{T}_i)}, \boldsymbol{\mu}_i(\mathbf{T}_i)$ can be estimated, but $\mathbf{X}_i(\mathbf{T}_i)$ is unobservable. So we can only use $\mathbf{W}_i(\mathcal{T}_i)$ and obtain the score estimate

$$\tilde{A}^\star_{ik} \;=\; \gamma_k \boldsymbol{\phi}_{ik}^\top \boldsymbol{\Sigma}^{-1}_{\mathbf{X}_i(\mathbf{T}_i)} \mathbf{W}_i(\mathbf{T}_i). \tag{A.11.4}$$

Due to measurement errors, the difference $d_{ik} = \tilde{A}_{ik} - \tilde{A}^\star_{ik}$ does not vanish as $n \to \infty$. Thus $\tilde{A}_{ik}$ cannot be consistently approximated by $\tilde{A}^\star_{ik}$. Hence, Proposition 2 no longer holds. Luckily, one can show that collectively the impact of measurement error is asymptotically negligible for estimation in the RKHS approach. (Note that the normal equation approach directly targets population quantities, thus it is not restricted by this assumption.)

An important difference between $\tilde{A}_{ik}$ in (2.3.5) and $\tilde{A}^\star_{ik}$ is the covariance matrix in the conditional scores. Due to measurement error, the smallest eigenvalue of $\boldsymbol{\Sigma}_{\mathbf{W}_i(\mathcal{T}_i)}$ in $\tilde{A}_{ik}$ is at least σ_η^2, which is not guaranteed in $\boldsymbol{\Sigma}_{\mathcal{X}_i(\mathcal{T}_i)}$. So we need another restriction on $\boldsymbol{\Sigma}_{\mathcal{X}_i(\mathcal{T}_i)}$, that is, $\mathbb{P}(\lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{X}_i(\mathcal{T}_i)}) \geq c) = 1$ for some fixed constant c, which prevents $\boldsymbol{\Sigma}^{-1}_{\mathcal{X}_i(\mathcal{T}_i)}$ from blowing up.

Now we turn to the estimation part. Let $\widehat{A}^\star_{ik} = \widehat{\gamma}_k \widehat{\boldsymbol{\phi}}_{ik}^\top \widehat{\boldsymbol{\Sigma}}^{-1}_{\mathbf{X}_i(\mathbf{T}_i)} \mathbf{W}_i(\mathbf{T}_i)$ be an estimate for $\tilde{A}^\star_{ik}$. There is a non-negligible error $d_{ik} = \tilde{A}_{ik} - \tilde{A}^\star_{ik}$ in estimating $\tilde{A}_{ik}$ by $\widehat{A}^\star_{ik}$ that would not go to zero as the sample size n increases. Furthermore, we can find out their difference between $\widehat{X}^\star_{i,\mathrm{M}_\circ}(t) = \sum_{k=1}^{\mathrm{M}_\circ} \widehat{A}^\star_{ik} \widehat{\phi}_k(t)$ and $\tilde{X}_{i,\mathrm{M}_\circ}(t)$, that is,

$$\widehat{X}^\star_{i,\mathrm{M}_\circ}(t) - \tilde{X}_{i,\mathrm{M}_\circ}(t) \;=\; \sum_{k=1}^{\mathrm{M}_\circ} (\widehat{A}^\star_{ik} - \tilde{A}^\star_{ik}) \widehat{\phi}_k(t) + \sum_{k=1}^{\mathrm{M}_\circ} d_{ik} \phi_k(t) - \sum_{k=1}^{\mathrm{M}_\circ} 2 d_{ik} (\widehat{\phi}_k(t) - \phi_k(t)).$$

If we substitute $z_{1i}(t)$ and $z_{2i}(t)$ with $\tilde{X}_{i,\mathrm{M}_\circ}(t)$ and $\widehat{X}^\star_{i,\mathrm{M}_\circ}(t)$ respectively, the first term in

(A.10.19) becomes

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{i,\mathrm{M}_\circ}(t)\big(\tilde{X}_{i,\mathrm{M}_\circ}(t)-\widehat{X}^{\star}_{i,\mathrm{M}_\circ}(t)\big) \;=\; & -\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{i,\mathrm{M}_\circ}(t)\sum_{k=1}^{\mathrm{M}_\circ}(\widehat{A}^{\star}_{ik}-\tilde{A}^{\star}_{ik})\widehat{\phi}_k(t)\\ & -\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{i,\mathrm{M}_\circ}(t)\sum_{k=1}^{\mathrm{M}_\circ}d_{ik}\phi_k(t)\\ & +\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{i,\mathrm{M}_\circ}(t)\sum_{k=1}^{\mathrm{M}_\circ}2d_{ik}(\widehat{\phi}_k(t)-\phi_k(t)).\end{aligned}$$

The first term and the third term retain the same convergence rates as $\widehat{A}^{\star}_{ik}$ (consistent for $\tilde{A}^{\star}_{ik}$) and $\widehat{\phi}_k(t)$ (consistent for $\phi_k(t)$) respectively. Note that for each $i = 1, \cdots, n$, $\tilde{X}_{i,\mathrm{M}_\circ}(t)\sum_{k=1}^{\mathrm{M}_\circ} d_{ik}\phi_k(t)$ has mean zero and variance at most of order $O(\mathrm{M}_\circ)$, and they are independently and identically distributed. Hence the second term is of order at most $O_p(\mathrm{M}_\circ \cdot n^{-1/2})$, which is negligible compared to the asymptotic orders of the first and third terms. The second term in (A.10.19) can be handled using the same argument. This effectively shows that the asymptotic convergence rate of the RKHS estimator stays the same. As a last comment, we note that although it is possible to retain the consistency of estimator with Assumption (A3′), the prediction results are no longer valid, since the imputed processes cannot be consistently estimated due to (A.11.4).

Appendix B

Supplementary Material for Chapter 3

B.1 Proof of Theorem 1

Proof. Let $\tau_i : \mathbb{R}^q \to \mathbb{R}$ be a projection such that $\tau_i(v) = v_i$ for $v = (v_1, \ldots, v_q) \in \mathbb{R}^q$. It is straightforward to verify that τ_is are linear and continuous and $v = \big(\tau_1(v), \ldots, \tau_q(v)\big)$. Using this property, we can write $g = \big(\tau_1 \circ g, \ldots, \tau_q \circ g\big)$ with each $\tau_i \circ g : \mathbb{C}([0,1]) \to \mathbb{R}$ being a continuous linear functional. We further denote $c_i = \tau_i \circ g(f)$ and write $v_c = (c_1, \ldots, c_q)^\top$. By Riesz representation theorem, for each $i = 1, \ldots, q$, there exists $b_i \in \mathcal{L}_2([0,1])$ such that $\tau_i \circ g(f) = \langle f, b_i \rangle$ for every $f \in \mathbb{C}([0,1])$ and $\|\tau_i \circ g\|_{\text{op}} = \sup_{\|f\|=1} |\langle f, b_i \rangle| = \|b_i\|_2$. Since $\|f\|_2 \leq 1$, we can check that $\|v_c\|_2 \leq K$ where $K = \sqrt{q} \max_i \|b_i\|_2$.

Classical universal approximation results (Cybenko, 1989; Funahashi, 1989; Hornik, 1991; Stinchcombe, 1999) imply that for any $\epsilon > 0$, there exists a network with weights Θ^* such that $\sup_{\|v\|_2 \leq 2K} |\text{nn}_{\Theta^*}(v) - h(v)| < \epsilon/2$. Since h is uniformly continuous on the compact set $D = \{v \in \mathbb{R}^q \,|\, \|v\|_2 \leq 2K\}$, there exists $0 < \rho_\epsilon < K$ such that for any $v_1, v_2 \in D$, $\|v_1 - v_2\|_2 < \rho_\epsilon$ implies $|h(v_1) - h(v_2)| < \epsilon/2$.

On the other hand, it is well known that $\mathbb{C}([0,1])$ is dense in $\mathcal{L}_2([0,1])$. For each $i = 1, \ldots, q$, there exists $\tilde{b}_i \in \mathbb{C}([0,1])$ such that $\|\tilde{b}_i - b_i\|_2 < \rho_\epsilon/(2\sqrt{q})$. Classical universal approximation results imply that there exists a set of networks, each of which has weights Θ_i^*, such that $\sup_{t \in [0,1]} |\text{nn}_{\Theta_i^*}(t) - \tilde{b}_i(t)| < \rho_\epsilon/(2\sqrt{q})$. Denote $\tilde{c}_i = \langle \text{nn}_{\Theta_i^*}, f \rangle$ and write

$\tilde{v}_c = (\tilde{c}_1, \ldots, \tilde{c}_q)^\top$. Then, we have

$$|\tilde{c}_i - c_i| \leq |\langle \mathrm{nn}_{\Theta_i^*} - \tilde{b}_i, f\rangle| + |\langle \tilde{b}_i - b_i, f\rangle| < \rho_\epsilon/\sqrt{q}$$

and thus $\|\tilde{v}_c - v_c\|_2 < \rho_\epsilon$.

Therefore, by linking these steps together, we have

$$\begin{aligned}
\sup_{\substack{f\in\mathbb{C}([0,1])\\ \|f\|_2\leq 1}} |\widehat{\mathcal{T}}^*(f) - \mathcal{T}(f)| &\leq \sup_{\substack{\|\tilde{v}_c - v_c\|_2<\rho_\epsilon\\ v_c,\tilde{v}_c\in D}} |\mathrm{nn}_{\Theta^*}(\tilde{v}_c) - h(v_c)| \\
&\leq \sup_{v\in D} |\mathrm{nn}_{\Theta^*}(v) - h(v)| + \sup_{\substack{\|v_1 - v_2\|_2<\rho_\epsilon\\ v_1,v_2\in D}} |h(v_1) - h(v_2)| \\
&< \epsilon/2 + \epsilon/2 = \epsilon.
\end{aligned}$$

For the second part of the claim, first note that X is a continuous random process defined on a compact interval. Its norm $\|X\|_2$ is a random variable on $\mathbb{R}$ and thus $\|X\|_2$ is stochastically bounded. In other words, for any $\delta > 0$, there exists a constant $M_\delta > 0$ such that $\mathbb{P}(\|X\|_2 \leq M_\delta) > 1-\delta$. Then, it is easy to verify that by using $\|X\|_2 \leq M_\delta$ in replacement of $\|f\|_2 \leq 1$ in the arguments above, we can obtain $\sup_{f\in\mathbb{C}([0,1]),\|f\|_2\leq M_\delta} |\widehat{\mathcal{T}}^\star(f) - \mathcal{T}(f)| < \delta$ for a suitable $\widehat{\mathcal{T}}^\star$. □

B.2 Proof of Theorem 2

Proof. Without loss of generality, we may assume that the model $\widehat{\mathcal{T}}_\Theta$ has one basis node in its BL and one hidden layer (with m nodes) in the subsequent network. Suppose that a numerical integration algorithm with weights $\{\omega_j\}_{j=1}^{J+1}$ and grid $\{t_j\}_{j=1}^{J+1}$ is used. Then, the model $\widehat{\mathcal{T}}_\Theta$ can be expressed as $\widehat{\mathcal{T}}_\Theta(X) = \sigma\big(\sum_{l=1}^m \theta_{l1}\sigma\big(\theta_{l2}\sum_{j=1}^{J+1}\omega_j X(t_j)\cdot \mathrm{nn}_{\tilde{\Theta}}(t_j) + \theta_{l3}\big) + \theta_4\big)$, where σ is a nonlinear Lipschitz activation function (e.g., sigmoid or ReLU), and $\tilde{\Theta}$ denotes the parameters of the basis function network $\mathrm{nn}_{\tilde{\Theta}}$. Note that here Θ denotes the collection of $\tilde{\Theta}$, θ_{l1}s, θ_{l2}s, θ_{l3}s, and θ_4.

It is straightforward to verify that both $\widehat{\mathcal{T}}_\Theta$ and $\nabla_\Theta \widehat{\mathcal{T}}_\Theta$ are bounded Lipschitz functions in Θ with bounded Lipschitz constants due to the condition $\sup_{t\in[0,1]} |X(t)| \leq M_1$ and Assumption (i). We may also assume that the the loss function $\ell(\cdot,\cdot)$ is non-negative,

since we can always shift the bounded loss function by a positive constant so that its minimum value is non-negative. Using the condition $|Y| \leq M_2$, Assumption (ii), chain rule, and the fact that a composition of two Lipschitz functions is also a Lipschitz function, we conclude that $\ell\big(\widehat{\mathcal{T}}_\Theta(X), Y\big)$ and $\nabla_\Theta \ell\big(\widehat{\mathcal{T}}_\Theta(X), Y\big)$ are also Lipschitz in Θ with bounded Lipschitz constants. Therefore, the second part of the theorem follows directly from Theorem 3.12 in Hardt et al. (2016) by checking its conditions. □

B.3 Experiment Details

Table **??** summarizes the number of bases used in each of the four simulation settings and each of the nine tasks in the data experiments.

Method	Case 1	Case 2	Case 3	Case 4	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9
Raw data + NN	51	51	51	51	48	48	48	48	48	48	48	30	20
	0.015	0.038	0.275	0.334	0.099	0.284	0.124	0.296	0.380	0.488	0.472	0.406	0.373
B-spline (4) + NN	4	4	4	4	4	4	4	4	4	4	4	4	4
	0.050	0.984	0.971	0.369	-	-	-	-	-	-	-	-	-
B-spline (15) + NN	15	15	15	15	15	15	15	15	15	15	15	15	15
	0.013	0.019	0.206	0.251	0.094	0.306	0.137	0.326	0.335	**0.477**	0.429	0.413	0.387
$\text{FPCA}_{0.9}$ + NN	2	3	4	20	5	5	10	11	1	1	1	4	3
	0.917	0.023	0.134	0.855	-	-	-	-	-	-	-	-	-
$\text{FPCA}_{0.99}$ + NN	4	19	20	28	17	17	24	24	5	4	2	12	7
	0.003	0.036	0.239	0.667	0.119	0.339	0.143	0.306	0.363	0.493	0.431	0.429	0.378
AdaFNN	2	3	3	2	4	4	4	4	4	4	4	4	4
	0.001	**0.003**	**0.127**	**0.193**	**0.084**	**0.260**	**0.118**	**0.294**	**0.339**	**0.477**	**0.410**	**0.362**	**0.368**

Table B.1: Number of bases used for the four simulation settings and nine tasks on datasets. The error rate of each method (already in the main paper) is also reported right below the method. The smallest error is marked in bold.

B.3.1 Model description

Preprocessing before training. For all tasks, the functional input is standardized entry-wise using function `StandardScaler` in `Python` package `sklearn.preprocessing`. The response is also standardized using the same function in all regression tasks.

Basis Layer in AdaFNN. As each basis node in the Basis Layer is implemented as a micro network, the scale of its output can vary with the initialization of the micro network weights. To stabilize the numerical integration at a basis node, we can normalize

the output of the micro networks, i.e., scale the output of each basis node such that its numerical integral is equal to 1. Another benefit of this normalization is that we do not have to normalize the learned basis functions again when applying the orthogonality/sparsity regularization.

B-spline scores. The B-spline scores used to represent functional inputs as a baseline method are computed using the spline functions (e.g., `smooth.spline`) in `R`, and the coefficients are computed individually for each curve.

FPCA scores. We use the function `FPCA` in the `R` package `fdapace` (Zhou et al., 2022) to compute functional principal component scores. Principal components are estimated and selected based on the training dataset first and then used to compute the principal component scores of curves in the test dataset.

B.3.2 Data description

Simulated Datasets: Simulation datasets are generated based on the following model

$$X(t) = \sum_{k=1}^{50} c_k \phi_k(t), \quad t \in [0, 1],$$

where terms on the right hand are defined as:

1. $\phi_1(t) = 1$ and $\phi_k(t) = \sqrt{2}\cos((k-1)\pi t)$, $k = 2, \ldots, 50$;
2. $c_k = z_k r_k$, and r_k are i.i.d. uniform random variables on $[-\sqrt{3}, \sqrt{3}]$.

Four simulation cases correspond to different configurations of z_ks:

1. In Case 1, $z_1 = 20, z_2 = z_3 = 5$, and $z_k = 1$ for $k \geq 4$. The response is $y = (\langle \phi_3, X \rangle)^2$;
2. In Case 2, $z_1 = z_3 = 5, z_5 = z_{10} = 3$, and $z_k = 1$ for other k. The response is $y = (\langle \phi_5, X \rangle)^2$.
3. Case 3 has the same configurations as Case 2. The observed response is

$$\tilde{y} = y + \epsilon = (\langle \phi_5, X \rangle)^2 + \epsilon, \quad \epsilon \sim N(0, 3/10).$$

For each time point t_j, the observed $X(t_j)$ is

$$\tilde{X}(t_j) = X(t_j) + \eta_j, \quad \eta_j \overset{\text{i.i.d.}}{\sim} N(0, 114/10).$$

4. In Case 4, $z_k = 1$ for all k. The response is $y = \langle \beta_2, X \rangle + (\langle \beta_1, X \rangle)^2$, where

$$\beta_1(t) = (4 - 16t) \cdot 1\{0 \leq t \leq 1/4\}$$

and

$$\beta_2(t) = (4 - 16|1/2 - t|) \cdot 1\{1/4 \leq t \leq 3/4\}.$$

The observed response is

$$\tilde{y} = y + \epsilon, \quad \epsilon \sim N(0, 1/10).$$

For each time point t_j, the observed $X(t_j)$ is

$$\tilde{X}(t_j) = X(t_j) + \eta_j, \quad \eta_j \overset{\text{i.i.d.}}{\sim} N(0, 5).$$

5. Case 5 has the same setup as Case 4, but with double the noise variance in Y.

In each case, 4000 curves are generated, among which 3200 are used for training while the rest 800 are left for testing. During training, 20% of the training data, i.e., 640 curves, are held out as a validation set.

Case 1 is designed to illustrate the weakness of the FPCA+NN approach, where the first two principle components explain at least 90% of the variability of the functional input X and are selected, but the true signal ϕ_3, which explains a very small fraction of the variability of X, is in the later principal components.

Case 2 is designed to illustrate the weakness of the B-spline+NN approach; here the true signal ϕ_5 in the functional covariate cannot be well represented by a small set of, e.g., four, B-splines.

Case 3 is designed to illustrate that AdaFNN is able to learn meaning basis functions and does better than other baseline methods in the appearance of both measurement error and noise.

Case 4 is designed to illustrate that AdaFNN can be used to select relevant domains and achieve smaller prediction error.

Case 5 is designed to illustrate the effectiveness of the regularizers.

Real Datasets:

Electricity Data (UK Power Networks, 2015). The study contains electricity consumption readings for 5567 London households which participated in the Low Carbon London project from November 2011 to February 2014. Every half hour, there is a recording of the total electricity usage during the past half hour, so the total number of daily observations is 48. The observation periods vary among households. So, we select a period of 5 weeks in which the number of households in the project is the highest. This results in 5503 households. Since daily consumption is noisy, we take the average daily consumption during the first week period (to produce a smoother curve) as our functional covariate $X(t)$ for all prediction tasks. The training and test split is $4:1$, and 20% of the training data are held out for validation during the training process.

NHANES Data (NCHS, CDC 2020). The study contains wearable device readings of physical intensity of 7742 subjects during a one week period. According to the data documentation, the device was the ActiGraph AM-7164 (formerly the CSA/MTI AM-7164), manufactured by ActiGraph located in Ft. Walton Beach, FL. It is programmed to detect and record the magnitude of acceleration or "intensity" of movement. The original intensity readings were summed over 1-minute epoch. To construct the functional input, we further aggregate the readings over 30 minute intervals, and this results in 48 observations per day. We take the average daily intensity curve for the whole week as the functional input $X(t)$. The prediction tasks here is to classify the health conditions of individual subjects using their physical intensity curve $X(t)$. Since not everyone reported their health conditions and there are invalid and incomplete information in some subjects, we ended up with 6555 subjects for Task 5, 2416 for Task 6, and 2412 for Task 7. The training and test split is taken to be $4:1$, and 20% of training data are held out for validation during the training process.

Mexfly (Carey et al., 2005) and Medfly(Chiou et al., 2003). These two datasets are very similar, so we describe them together. The Mexfly data contain 1072 subjects and the Medfly data consist of 1000 subjects. For both data, the number of eggs laid daily was recorded for individual flies in the study until death. It is of biological interest to

explore how early reproduction shapes lifetime reproduction, defined as the total number of eggs laid in a lifetime, which is perhaps the single most important biological question from the evolution point of view. We thus aim to predict the lifetime reproduction of a fly using its early reproduction trajectory $X(t)$, which is the number of eggs laid daily from birth till a specified day M shortly before peak reproduction period. This day M is 20 for the Medflies and 30 for the Mexflies. Next, we remove flies that died before day M and the resulted sample size is 872 for Mexflies (Task 8) and 870 for Medflies (Task 9). The training and testing split $4:1$ is the same as before.

Appendix C

Supplementary Material for Chapter 4

Organization. In Appendices S4 and S5, we prove Theorems 11 and 12 respectively. The building blocks for these results are Theorems 13 and 14, which are presented in Appendices S2 and S3 respectively. Technical lemmas are given in Appendix S6. Proposition 1 from the main text is proved in Appendix S1. Background results are stated in Appendix S7. Appendix S8 provides additional plots of simulation results. Appendix 4.4.5 presents real-data examples based on stock market returns. Appendix S9 describes the computational cost of implementing the bootstrap. Lastly, Appendix S10 provides sensitivity analysis with regard to Assumption 1(b).

Conventions. Throughout the proofs of Theorems 11 and 12, we may assume without loss of generality that $n \geq 3$, and that for any constant $\epsilon \in (0, 1)$ fixed with respect to n, the following inequality holds

$$\frac{\log(n)\, \beta_{3q}^3\, \mathtt{r}(\Sigma)}{n^{1/2}} \;\leq\; \epsilon, \tag{C.0.1}$$

where $q = 5\log(kn)$. (If $n < 3$ or if (C.0.1) does not hold, then the constant c in the statements of Theorems 11 and 12 may be taken as $c = 3 \vee \frac{1}{\epsilon}$, which makes the results trivially true.) In addition, we will frequently re-use the symbol c to denote a constant that does not depend on n, and we will allow its value to vary with each appearance.

Notation. The spectral decomposition for Σ will be written as

$$\Sigma = U\Lambda U^\top, \tag{C.0.2}$$

where $\Lambda = \mathrm{diag}(\lambda_1(\Sigma), \ldots, \lambda_p(\Sigma)) \in \mathbb{R}^{p\times p}$, and the jth column of $U \in \mathbb{R}^{p\times p}$ is the jth eigenvector of Σ. For a vector $v \in \mathbb{R}^p$, the ℓ_∞ and ℓ_2-norms are denoted as $\|v\|_\infty = \max_{1\le j\le p} |v_j|$ and $\|v\|_2 = (\sum_{j=1}^p v_j^2)^{1/2}$. For a $p_1 \times p_2$ matrix M, the following three norms will be used: $\|M\|_{\mathrm{op}} = \sup_{\|v\|_2=1} \|Mx\|_2$, $\|M\|_1 = \max_{1\le j\le p_2} \sum_{i=1}^{p_1} |M_{ij}|$, and $\|M\|_\infty = \max_{1\le i\le p_1} \sum_{j=1}^{p_2} |M_{ij}|$. If A is a symmetric $p \times p$ matrix and $j \in \{1, \ldots, p\}$, then $\Lambda_j(A)$ denotes the $j \times j$ diagonal matrix formed by the largest j eigenvalues of A,

$$\Lambda_j(A) = \mathrm{diag}\big(\lambda_1(A), \ldots, \lambda_j(A)\big).$$

Also, for each $j = 1, \ldots, p$, let $d_j(A) = A_{jj}$ be the jth diagonal element of A, and let $\boldsymbol{d}_j(A)$ be the vector of the first j diagonal entries

$$\boldsymbol{d}_j(A) = (d_1(A), \ldots, d_j(A)).$$

If B is another symmetric $p \times p$ matrix, then the relation $B \succcurlyeq A$ means that $B - A$ is positive semidefinite. The symbol 1_j denotes the j-dimensional all-ones vector. For a univariate scalar function h with first derivative h', define $\boldsymbol{h}'(v) = (h'(v_1), \ldots, h'(v_p))$. Lastly, for two vectors u and v, the symbol $u \odot v$ denotes the vector obtained from entrywise multiplication, $(u \odot v)_j = u_j v_j$.

S1 Proof of Proposition 1

Proof of Proposition 1(i)**.** For each $l = 1, \ldots, p$, let $\kappa_l = \mathbb{E}[Z_{1l}^4]$. Also recall that in part (i) of the proposition, the entries of Z_1 are assumed to be independent. For any pair of indices (j, j') satisfying $1 \le j, j' \le k$, it follows from the equation (9.8.6) in Bai and Silverstein (2010) that the corresponding entry of $\Gamma \in \mathbb{R}^{k\times k}$ is

$$\begin{aligned}
\Gamma_{jj'} &= \mathbb{E}\big[\big(Z_1^\top (u_j u_j^\top) Z_1 - \mathrm{tr}(u_j u_j^\top)\big)\big(Z_1^\top (u_{j'} u_{j'}^\top) Z_1 - \mathrm{tr}(u_{j'} u_{j'}^\top)\big)\big] \\
&= 2 \cdot 1\{j = j'\} + \sum_{l=1}^p (\kappa_l - 3) u_{jl}^2 u_{j'l}^2.
\end{aligned}$$

If we let H denote the $p \times k$ matrix whose jth column is equal to $(u_{j1}^2, \ldots, u_{jp}^2)$, then the previous entrywise expression for Γ can be written in matrix form as

$$\Gamma \;=\; 2I_k + H^\top (D - 3I_p) H,$$

where $D = \mathrm{diag}(\kappa_1, \ldots, \kappa_p)$. Now recall the assumption that there is a constant $\kappa > 1$ not depending on n such that $\min_{1 \le l \le p} \kappa_l \ge \kappa$. If we consider the case when $\kappa \ge 3$, then it is clear that $D - 3I_p \succcurlyeq 0$, which implies $\Gamma \;\succcurlyeq\; 2I_k$, and hence $\lambda_k(\Gamma) \ge 2$.

On the other hand, in the case when $\kappa < 3$, we have

$$\begin{aligned} \Gamma \;&\succcurlyeq\; 2I_k - (3-\kappa) H^\top H \\ &\succcurlyeq \big(2 - (3-\kappa)\|H\|_{\mathrm{op}}^2\big) I_k. \end{aligned} \tag{S1.1}$$

With regard to the quantity $\|H\|_{\mathrm{op}}^2$, observe that

$$\begin{aligned} \|H\|_{\mathrm{op}}^2 \;&\le\; \|H\|_1 \|H\|_\infty \\ &=\; \max_{1 \le j \le k} \|u_j\|_2^2 \cdot \max_{1 \le i \le p} \sum_{j=1}^k u_{ji}^2 \\ &\le\; 1. \end{aligned}$$

Combining this with (S1.1) gives $\Gamma \;\succcurlyeq\; (2 - (3-\kappa))I_k$, and hence $\lambda_k(\Gamma) \ge \kappa - 1$, which completes the proof. □

Proof of Proposition 1(ii)**.** For $1 \le j, j' \le k$, it follows from Lemma A.1 in Hu et al. (2019) that

$$\begin{aligned} \Gamma_{jj'} \;&=\; \mathrm{cov}\Big(Z_1^\top (u_j u_j^\top) Z_1 \,,\, Z_1^\top (u_{j'} u_{j'}^\top) Z_1 \Big) \\ &=\; \frac{\mathbb{E}[\xi^4]}{p(p+2)} \big(1 + 2\,\mathrm{tr}(u_j u_j^\top u_{j'} u_{j'}^\top)\big) - 1 \\ &=\; 2 \cdot 1\{j = j'\} \cdot \frac{\mathbb{E}[\xi^4]}{p(p+2)} + \frac{\mathbb{E}[\xi^4]}{p(p+2)} - 1. \end{aligned}$$

When written in matrix form, this is equivalent to

$$\Gamma \;= a\, I_k + b\, 1_k 1_k^\top$$

where $a = \frac{2\,\mathbb{E}[\xi^4]}{p(p+2)}$ and $b = \frac{\mathbb{E}[\xi^4]}{p(p+2)} - 1$. Consequently, one eigenvalue of Γ is equal to $a + kb$, and the rest are equal to a. Furthermore, due to the basic inequality $\mathbb{E}[\xi^4] \geq (\mathbb{E}[\xi^2])^2 = p^2$, and the fact that $p \geq 2$, we have the lower bounds

$$a \;\geq\; \tfrac{2}{1+2/p} \geq\; 1$$

and

$$\begin{aligned} a + kb \;&\geq\; \tfrac{2}{1+2/p} + k\Big(\tfrac{1}{1+2/p} - 1\Big) \\ &=\; \tfrac{2(1-k/p)}{1+2/p} \\ &\geq\; 1 - \tfrac{k}{p} \\ &\gtrsim\; 1, \end{aligned}$$

which completes the proof. □

S2 Gaussian Approximation

Theorem 13 (Gaussian approximation)**.** *Suppose that the conditions of Theorem 11 hold and let* $\zeta \sim N(0, \Lambda_k \Gamma \Lambda_k)$*. Then,*

$$\sup_{t \in \mathbb{R}^k} \Big| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)\big) \preceq t \Big) - \mathbb{P}\Big(\zeta \preceq t \Big) \Big| \;\lesssim\; \frac{\log(n)\, \beta_{2q}^3\, \mathtt{r}(\Sigma)}{n^{1/2}}. \tag{S2.1}$$

Proof. For each $i = 1, \ldots, n$, let $\tilde{X}_i = \Lambda^{1/2} U^\top Z_i$, and let the associated sample covariance matrix be denoted as

$$\tilde{\Sigma} \;=\; \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i^\top.$$

This implies $\tilde{\Sigma} = U^\top \widehat{\Sigma} U$, and so the eigenvalues of $\widehat{\Sigma}$ may be equivalently written as $\lambda_j(\widehat{\Sigma}) = \lambda_j(\tilde{\Sigma})$ for every $j = 1, \ldots, p$. For future reference, it will also be helpful to note that $\mathbb{E}[\tilde{\Sigma}] = \Lambda$.

To partition $\tilde{\Sigma}$ into suitable blocks, we will write

$$\tilde{\Sigma} \;=\; \begin{pmatrix} \tilde{\Sigma}[1,1] & \tilde{\Sigma}[1,2] \\ \tilde{\Sigma}[1,2]^\top & \tilde{\Sigma}[2,2] \end{pmatrix}, \tag{S2.2}$$

where the matrix $\tilde{\Sigma}[1,1]$ is of size $k \times k$, and the matrix $\tilde{\Sigma}[2,2]$ is of size $(p-k) \times (p-k)$.

Based on the notation above, the desired result can be broken down as follows:

$$\sup_{t \in \mathbb{R}^k} \left| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)\big) \preceq t \Big) - \mathbb{P}\Big(\zeta \preceq t \Big) \right| \;\leq\; \mathrm{I} + \mathrm{II}$$

where the two terms on the right are defined as

$$\mathrm{I} \;=\; \sup_{t \in \mathbb{R}^k} \left| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{\lambda}_k(\tilde{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)\big) \preceq t \Big) - \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{d}_k(\tilde{\Sigma}[1,1]) - \boldsymbol{\lambda}_k(\Sigma)\big) \preceq t \Big) \right|, \quad \text{(S2.3)}$$

$$\mathrm{II} \;=\; \sup_{t \in \mathbb{R}^k} \left| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{d}_k(\tilde{\Sigma}[1,1]) - \boldsymbol{\lambda}_k(\Sigma)\big) \preceq t \Big) - \mathbb{P}\Big(\zeta \preceq t \Big) \right|. \quad \text{(S2.4)}$$

For each of these terms, Lemmas S2.4, and S2.3 respectively give the following bounds

$$\mathrm{I} \;\lesssim\; \frac{\log(n)\, \beta_{2q}^3\, \mathtt{r}(\Sigma)}{n^{1/2}}$$

$$\mathrm{II} \;\lesssim\; \frac{\beta_4^3}{n^{1/2}},$$

completing the proof. □

S2.1 Lemmas for Gaussian approximation

Lemma S2.1. *Suppose that the conditions of Theorem 13 hold. Then, there is a constant $c > 0$ not depending on n such that the event*

$$\left\| \sqrt{n}\Big(\boldsymbol{\lambda}_k(\tilde{\Sigma}) - \boldsymbol{\lambda}_k(\tilde{\Sigma}[1,1]) \Big) \right\|_\infty \;\leq\; \frac{c\, \log(n)\, \beta_{2q}\, \lambda_1(\Sigma)\, \mathtt{r}(\Sigma)}{n^{1/2}}$$

holds with probability at least $1 - \frac{c}{n^4}$.

Proof. Recall the partition (S2.2) of the matrix $\tilde{\Sigma}$ in the proof of Theorem 13. By Wielandt's inequality (Lemma S7.2), the following inequality holds when the event $\{\lambda_k(\tilde{\Sigma}[1,1]) > \lambda_1(\tilde{\Sigma}[2,2])\}$ occurs,

$$\max_{1 \leq j \leq k} \left| \lambda_j(\tilde{\Sigma}) - \lambda_j(\tilde{\Sigma}[1,1]) \right| \;\leq\; \frac{\|\tilde{\Sigma}[1,2]\|_{\mathrm{op}}^2}{\lambda_k(\tilde{\Sigma}[1,1]) - \lambda_1(\tilde{\Sigma}[2,2])}. \quad \text{(S2.5)}$$

We will derive a high-probability upper bound for the right side of (S2.5) by separately handling the numerator and denominator.

To control the numerator in the bound (S2.5), we may apply Lemma S6.1 to conclude that

$$\big\| \|\tilde{\Sigma}[1,2]\|_{\mathrm{op}}^2 \big\|_q \;\lesssim\; \frac{q\,\beta_{2q}\,\lambda_1(\Sigma)\;\mathrm{tr}(\Sigma)}{n^{1-3/(2q)}}.$$

Then, for any $t > 0$, Chebyshev's inequality yields the tail bound

$$\mathbb{P}\big(\|\tilde{\Sigma}[1,2]\|_{\mathrm{op}}^2 \geq e\,t\big) \;\leq\; \frac{e^{-q}\;\big\| \|\tilde{\Sigma}[1,2]\|_{\mathrm{op}}^2 \big\|_q^q}{t^q}.$$

Recalling the choice $q = 5\log(kn)$ and taking $t = \frac{c}{n}\log(n)\beta_{2q}\lambda_1(\Sigma)\,\mathrm{tr}(\Sigma)$ for a sufficiently large constant c, it follows that the event

$$\|\tilde{\Sigma}[1,2]\|_{\mathrm{op}}^2 \;\leq\; \frac{c\,\log(n)\,\beta_{2q}\,\lambda_1(\Sigma)^2\,\mathtt{r}(\Sigma)}{n} \tag{S2.6}$$

holds with probability at least $1 - \frac{1}{n^4}$.

To handle the denominator in the bound (S2.5), let $\Pi \in \mathbb{R}^{k\times p}$ denote the matrix whose ith row is the ith standard basis vector in $\mathbb{R}^p$. Then, Weyl's inequality implies

$$\begin{aligned}\max_{1\leq j\leq k} |\lambda_j(\tilde{\Sigma}[1,1]) - \lambda_j(\Sigma)| \;&\leq\; \big\|\Pi(\tilde{\Sigma} - \Lambda)\Pi^\top\big\|_{\mathrm{op}} \\ &\leq\; \|\tilde{\Sigma} - \Lambda\|_{\mathrm{op}}.\end{aligned}$$

Similarly, we have

$$\max_{k+1\leq j\leq p} |\lambda_{j-k}(\tilde{\Sigma}[2,2]) - \lambda_j(\Sigma)| \;\leq\; \|\tilde{\Sigma} - \Lambda\|_{\mathrm{op}}.$$

Combining the last two steps yields the following bound for some positive constant c_1 not depending on n,

$$\begin{aligned}\lambda_k(\tilde{\Sigma}[1,1]) - \lambda_1(\tilde{\Sigma}[2,2]) \;&=\; \big(\lambda_k(\tilde{\Sigma}[1,1]) - \lambda_k(\Sigma)\big) + \big(\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)\big) - \big(\lambda_1(\tilde{\Sigma}[2,2]) - \lambda_{k+1}(\Sigma)\big) \\ &\geq\; c_1\lambda_1(\Sigma) - 2\|\tilde{\Sigma} - \Lambda\|_{\mathrm{op}}, \end{aligned} \tag{S2.7}$$

where Assumption 6.(b) has been used in the last line.

To complete the proof, it suffices to show that $\|\tilde{\Sigma} - \Lambda\|_{\mathrm{op}} \leq \frac{c_1}{4}\lambda_1(\Sigma)$ holds with high probability. This may be accomplished using Lemma S7.3, which gives the following bound,

$$\left(\mathbb{E}\|\tilde{\Sigma}-\Lambda\|_{\mathrm{op}}^{q}\right)^{1/q} \;\lesssim\; \left(\frac{\sqrt{q}\left(\mathbb{E}\|\tilde{X}_1\|_2^{2q}\right)^{\frac{1}{2q}}\left\|\mathbb{E}[\tilde{X}_1\tilde{X}_1^{\top}]\right\|_{\mathrm{op}}^{1/2}}{n^{1/2-3/(2q)}}\right) \bigvee \left(\frac{q\left(\mathbb{E}\|\tilde{X}_1\|_2^{2q}\right)^{1/q}}{n^{1-3/q}}\right). \quad \text{(S2.8)}$$

This bound can be simplified by noting that $\mathbb{E}[\tilde{X}_1\tilde{X}_1^{\top}] = \Lambda$ and

$$\begin{aligned}\left(\mathbb{E}\|\tilde{X}_1\|_2^{2q}\right)^{1/q} &= \left\|\|\tilde{X}_1\|_2^2\right\|_q \\ &= \left\|\sum_{j=1}^{p}\lambda_j\langle u_j, Z_1\rangle^2\right\|_q \qquad \text{(S2.9)} \\ &\leq \operatorname{tr}(\Sigma)\,\beta_q.\end{aligned}$$

Therefore, the bound (S2.8) reduces to

$$\begin{aligned}\left(\mathbb{E}\|\tilde{\Sigma}-\Lambda\|_{\mathrm{op}}^{q}\right)^{1/q} &\lesssim \sqrt{\frac{q\,\beta_q\,\lambda_1(\Sigma)^2\,\mathtt{r}(\Sigma)}{n^{1-3/q}}} \bigvee \frac{q\,\beta_q\,\lambda_1(\Sigma)\,\mathtt{r}(\Sigma)}{n^{1-3/q}} \\ &\lesssim \sqrt{\frac{q\,\beta_q\,\lambda_1(\Sigma)^2\,\mathtt{r}(\Sigma)}{n^{1-3/q}}}, \qquad \text{(S2.10)}\end{aligned}$$

where the second line has used the condition (C.0.1).

To apply the previous bound, Chebyshev's inequality implies that for any $t > 0$,

$$\mathbb{P}\big(\|\tilde{\Sigma}-\Lambda\|_{\mathrm{op}} \geq e\;t\big) \;\leq\; \frac{e^{-q}\;\mathbb{E}\|\tilde{\Sigma}-\Lambda\|_{\mathrm{op}}^{q}}{t^q}.$$

Recalling the choice $q = 5\log(kn)$ and taking $t = \frac{c\lambda_1(\Sigma)}{\sqrt{n}}\sqrt{\log(n)\beta_q\mathtt{r}(\Sigma)}$ for a sufficiently large constant c not depending on n, we conclude that the event

$$\|\tilde{\Sigma}-\Lambda\|_{\mathrm{op}} \;\leq\; c\lambda_1(\Sigma)\sqrt{\frac{\log(n)\,\beta_q\,\mathtt{r}(\Sigma)}{n}} \qquad \text{(S2.11)}$$

holds with probability at least $1-\frac{1}{n^4}$. Furthermore, if we apply this bound to (S2.7) and use the condition (C.0.1), then the bound

$$\lambda_k(\tilde{\Sigma}[1,1]) - \lambda_1(\tilde{\Sigma}[2,2]) \;\geq\; \frac{c_1}{2}\lambda_1(\Sigma) \qquad \text{(S2.12)}$$

holds with probability at least $1-\frac{1}{n^4}$.

The stated result is obtained by combining the bounds (S2.5), (S2.6), and (S2.12). $\square$

The next result shows that the random vector $\sqrt{n}\big(\boldsymbol{\lambda}_k(\tilde{\Sigma}[1,1]) - \boldsymbol{\lambda}_k(\Sigma)\big)$ is well approximated by $\sqrt{n}\big(\boldsymbol{d}_k(\tilde{\Sigma}[1,1]) - \boldsymbol{\lambda}_k(\Sigma)\big)$ in an entrywise sense.

Lemma S2.2. *Suppose that the conditions of Theorem 13 hold. Then, there is a constant $c > 0$ not depending on n such that the event*

$$\Big\|\sqrt{n}\Big(\boldsymbol{\lambda}_k(\tilde{\Sigma}[1,1]) - \boldsymbol{d}_k(\tilde{\Sigma}[1,1])\Big)\Big\|_\infty \;\leq\; \frac{c\,\log(n)\,\beta_{2q}\,\lambda_1(\Sigma)\,\mathtt{r}(\Sigma)}{n^{1/2}}$$

holds with probability at least $1 - \frac{c}{n^4}$.

Proof. We will use an iterative argument. As an initial step, first partition the matrix $\tilde{\Sigma}[1,1]$ as

$$\tilde{\Sigma}[1,1] = \begin{pmatrix} d_1(\tilde{\Sigma}[1,1]) & V_1 \\ V_1^\top & D_1 \end{pmatrix},$$

where $d_1(\tilde{\Sigma}[1,1])$ is a scalar, and D_1 is a $(k-1)\times(k-1)$ matrix. To lighten the notational burden of subscripts when handling matrices of different sizes, we will write $\boldsymbol{\lambda}(A) = (\lambda_1(A), \ldots, \lambda_r(A))$, as well as $\boldsymbol{d}(A) = (A_{11}, \ldots, A_{rr})$ for any symmetric matrix $A \in \mathbb{R}^{r\times r}$ and integer $r \geq 1$. By the triangle inequality, we have

$$\begin{aligned}\Big\|\boldsymbol{\lambda}(\tilde{\Sigma}[1,1]) - \boldsymbol{d}(\tilde{\Sigma}[1,1])\Big\|_\infty \;&\leq\; \Big\|\boldsymbol{\lambda}(\tilde{\Sigma}[1,1]) - \begin{bmatrix} d_1(\tilde{\Sigma}[1,1]) \\ \boldsymbol{\lambda}(D_1) \end{bmatrix}\Big\|_\infty + \Big\|\begin{bmatrix} d_1(\tilde{\Sigma}[1,1]) \\ \boldsymbol{\lambda}(D_1) \end{bmatrix} - \boldsymbol{d}(\tilde{\Sigma}[1,1])\Big\|_\infty, \\ &=\; \mathrm{T} + \mathrm{T}'\end{aligned} \tag{S2.13}$$

where we have defined the random variables T and T′ through the last line. Later on, we will show that the event

$$\mathrm{T} \;\leq\; \frac{c\,\log(n)\,\beta_{2q}\,\mathrm{tr}(\Sigma)}{n} \tag{S2.14}$$

holds with probability at least $1 - \frac{2}{kn^4}$.

Next, to handle T', we partition the matrix D_1 as

$$D_1 = \begin{pmatrix} d_1(D_1) & V_2 \\ V_2^\top & D_2 \end{pmatrix},$$

where $d_1(D_1) = d_2(\tilde{\Sigma}[1,1])$ is a scalar, and D_2 is a $(k-2) \times (k-2)$ matrix. Proceeding in a similar manner to (S2.13), and noting that $d_1(\tilde{\Sigma}[1,1])$ is the same as the first entry of $\boldsymbol{d}(\tilde{\Sigma}[1,1])$, we have

$$\begin{aligned} \mathrm{T}' &\leq \left\| \boldsymbol{\lambda}(D_1) - \begin{bmatrix} d_1(D_1) \\ \boldsymbol{\lambda}(D_2) \end{bmatrix} \right\|_\infty + \left\| \begin{bmatrix} d_1(D_1) \\ \boldsymbol{\lambda}(D_2) \end{bmatrix} - \boldsymbol{d}(D_1) \right\|_\infty, \\ &= \mathrm{T}'' + \mathrm{T}''', \end{aligned} \tag{S2.15}$$

where the random variables T'' and T''' have been defined through the last line. The argument that will be used to prove (S2.14) can also be used to show that the event

$$\mathrm{T}'' \leq \frac{c \log(n) \beta_{2q} \operatorname{tr}(\Sigma)}{n}$$

holds with probability at least $1 - \frac{c}{kn^4}$. Likewise, we can combine $k-1$ iterations of this process by summing the bounds on $\mathrm{T}, \mathrm{T}'', \mathrm{T}'''', \ldots$ appearing at each iteration.

To complete the proof, it remains to validate the claim (S2.14). Let $D_0 = \tilde{\Sigma}[1,1]$, and observe that

$$\mathrm{T} = |\lambda_1(D_0) - d_1(D_0)| \vee \max_{2 \leq j \leq k} |\lambda_j(D_0) - \lambda_{j-1}(D_1)|. \tag{S2.16}$$

Wielandt's inequality (Lemma S7.2) gives the following bounds when the event $\{d_1(D_0) > \lambda_1(D_1)\}$ occurs,

$$|\lambda_1(D_0) - d_1(D_0)| \leq \frac{\|V_1\|_{\mathrm{op}}^2}{d_1(D_0) - \lambda_1(D_1)}$$

and

$$|\lambda_j(D_0) - \lambda_{j-1}(D_1)| \leq \frac{\|V_1\|_{\mathrm{op}}^2}{d_1(D_0) - \lambda_1(D_1)},$$

for all $j = 2, \ldots, k$. So, in light of the formula for T given in (S2.16), we conclude that the bound

$$\mathrm{T} \leq \frac{\|V_1\|_{\mathrm{op}}^2}{d_1(D_0) - \lambda_1(D_1)}. \tag{S2.17}$$

holds whenever the event $\{d_1(D_0) > \lambda_1(D_1)\}$ holds. By using the argument at (S2.12) from the proof of Lemma S2.1, it can be shown that the denominator in the previous bound satisfies

$$d_1(D_0) - \lambda_1(D_1) \geq c\lambda_1(\Sigma)$$

with probability at least $1 - \frac{1}{kn^4}$ for some constant $c > 0$ not depending on n. Next, in order to derive an upper bound on $\|V_1\|_{\mathrm{op}}$, note that for any value of k, the matrix $V_1^\top$ is contained in the submatrix of $\tilde{\Sigma}$ indexed by $\{2, \ldots, p\} \times \{1\}$. Hence, $\|V_1\|_{\mathrm{op}}$ is upper bounded by the operator norm of that submatrix, which is the same as $\tilde{\Sigma}[1,2]^\top$ in the particular case when $k = 1$. Due to this observation, it follows from Lemma S6.1 that there is a constant $c > 0$ not depending on n, such that for any choice of k, the bound

$$\|V_1\|_{\mathrm{op}}^2 \leq \frac{c\, q\, \beta_{2q}\, \lambda_1(\Sigma)\, \mathrm{tr}(\Sigma)}{n} \tag{S2.18}$$

holds with probability at least $1 - \frac{1}{kn^4}$, where we continue to use $q = 5\log(kn)$. Thus, combining the last few steps establishes the claim (S2.14).

To comment on how this argument can be applied iteratively to T'' and its successors, the previous reasoning involving Weilandt's inequality shows that the bound

$$\mathrm{T}'' \;\leq\; \frac{\|V_2\|_{\mathrm{op}}^2}{d_1(D_1) - \lambda_1(D_2)}$$

holds whenever the event $\{d_1(D_1) > \lambda_1(D_2)\}$ holds. In turn, a lower bound on the denominator $d_1(D_1) - \lambda_1(D_2)$ that is proportional to $\lambda_1(\Sigma)$ can be established in the same manner as for $d_1(D_0) - \lambda_1(D_1)$. Meanwhile, the numerator can be handled by noting that for any value of k, the matrix $V_2^\top$ is contained in the submatrix of $\tilde{\Sigma}$ indexed by $\{3, \ldots, p\} \times \{1, 2\}$. The latter matrix is the same as $\tilde{\Sigma}[1,2]^\top$ in the particular case of $k = 2$, and consequently, Lemma S6.1 can be used to establish a bound on $\|V_2\|_{\mathrm{op}}^2$ that is of the same form as (S2.18). This completes the proof. □

The next lemma provides a Gaussian approximation result for $\sqrt{n}\big(\boldsymbol{d}_k(\tilde{\Sigma}[1,1]) - \boldsymbol{\lambda}_k(\Sigma)\big)$.

Lemma S2.3. *Suppose that the conditions of Theorem 13 hold, and let* II *be as defined in* (S2.4). *Then,*

$$\text{II} \;\lesssim\; \frac{\beta_4^3}{n^{1/2}}.$$

Proof. For each $i = 1, \ldots, n$, let the random vector $W_i \in \mathbb{R}^k$ have its jth entry defined as $W_{ij} = \langle u_j, Z_i\rangle^2 - 1$, where u_j is the jth eigenvector of Σ. Letting $Y_i = \Lambda_k W_i$, we have $\sqrt{n}\big(\boldsymbol{d}_k(\tilde{\Sigma}[1,1]) - \boldsymbol{\lambda}_k(\Sigma)\big) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$. Also, note that the covariance matrix of Y_i is given by $\mathbb{E}[Y_i Y_i^\top] = \Lambda_k \Gamma \Lambda_k$, with Γ as defined in Assumption 6.(c).

Applying Bentkus' multivariate Berry-Esseen theorem (Lemma S7.5) yields

$$\sup_{t\in\mathbb{R}^k}\left|\mathbb{P}\Big(\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n Y_i \preceq t\Big) - \mathbb{P}\big(\zeta \preceq t\big)\right| \;\lesssim\; \frac{\mathbb{E}\big\|(\Lambda_k\Gamma\Lambda_k)^{-1/2}Y_1\big\|_2^3}{n^{1/2}}. \tag{S2.19}$$

The proof is complete once we derive a bound on $\mathbb{E}\big\|(\Lambda_k\Gamma\Lambda_k)^{-1/2}Y_1\big\|_2^3$. Due to Assumption 6.(c), we have

$$\big\|(\Lambda_k\Gamma\Lambda_k)^{-1/2}Y_1\big\|_2^2 \;=\; W_1^\top\Gamma^{-1}W_1 \;\le\; c\|W_1\|_2^2\,,$$

almost surely for some constant $c > 0$ not depending on n. In turn, Lyapunov's inequality implies

$$\begin{aligned}
\mathbb{E}\big\|(\Lambda_k\Gamma\Lambda_k)^{-1/2}Y_1\big\|_2^3 \;&\lesssim\; \mathbb{E}\big[(W_1^\top W_1)^2\big]^{3/4} \\
&=\; \Big\|\sum_{j=1}^k \big(\langle u_j, Z_1\rangle^2 - 1\big)^2\Big\|_2^{3/2} \\
&\le\; \Big(\sum_{j=1}^k \big\|\langle u_j, Z_1\rangle^2 - 1\big)^2\big\|_2\Big)^{3/2} \\
&\lesssim\; \beta_4^3.
\end{aligned}$$

Substituting this bound into (S2.19) completes the proof. □

Lemma S2.4. *Suppose that the conditions of Theorem 13 hold, and let* I *be as defined in* (S2.3). *Then,*

$$\text{I} \;\lesssim\; \frac{\log(n)\,\beta_{2q}^3\,\mathtt{r}(\Sigma)}{n^{1/2}}. \tag{S2.20}$$

Proof. For any $\epsilon > 0$ and $t \in \mathbb{R}^k$, define the two events

$$\mathcal{A}(t) = \Big\{ \sqrt{n}\big(\boldsymbol{d}_k(\tilde{\Sigma}[1,1]) - \boldsymbol{\lambda}_k(\Sigma)\big) \preceq t \Big\},$$

$$\mathcal{B}(\epsilon) = \Big\{ \sqrt{n}\big\|\boldsymbol{\lambda}_k(\tilde{\Sigma}) - \boldsymbol{d}_k(\tilde{\Sigma}[1,1])\big\|_\infty \geq \epsilon \Big\}.$$

It follows from Lemma S7.6 that

$$\mathrm{I} \;\leq\; \mathbb{P}\big(\mathcal{B}(\epsilon)\big) \;+\; \sup_{t\in\mathbb{R}^k} \Big|\mathbb{P}\big(\mathcal{A}(t+\epsilon\,1_k)\big) - \mathbb{P}\big(\mathcal{A}(t-\epsilon\,1_k)\big)\Big|. \tag{S2.21}$$

The first term $\mathbb{P}(\mathcal{B}(\epsilon))$ can be handled by Lemmas S2.1 and S2.2, which show that there is a constant $c > 0$ not depending on n such that if $\epsilon = \frac{c}{\sqrt{n}}\,\log(n)\,\beta_{2q}\,\lambda_1(\Sigma)\,\mathtt{r}(\Sigma)$, then

$$\mathbb{P}(\mathcal{B}(\epsilon)) \;\lesssim\; \tfrac{1}{n}.$$

Next, the anti-concentration term in (S2.21) can be bounded through an approximation involving the Gaussian vector $\zeta \sim N(0, \Lambda_k \Gamma \Lambda_k)$,

$$\begin{aligned}\sup_{t\in\mathbb{R}^k} \Big|\mathbb{P}\big(\mathcal{A}(t+\epsilon 1_k)\big) - \mathbb{P}\big(\mathcal{A}(t-\epsilon 1_k)\big)\Big| \;&\leq\; 2\,\mathrm{II} \;+\; \sup_{t\in\mathbb{R}^k} \Big|\mathbb{P}\Big(\zeta \preceq t+\epsilon 1_k\Big) - \mathbb{P}\big(\zeta \preceq t-\epsilon 1_k\big)\Big| \\ &=\; 2\,\mathrm{II} \;+\; \mathrm{J}(\epsilon),\end{aligned}$$

where the quantity $\mathrm{J}(\epsilon)$ is defined by the last line, and $\mathrm{II} \lesssim \beta_4^3/n^{1/2}$ holds by Lemma S2.3. To handle $\mathrm{J}(\epsilon)$, we need the following lower bound

$$\min_{1\leq j\leq k} \mathrm{var}(\zeta_j) \;=\; \min_{1\leq j\leq k} \lambda_j(\Sigma)^2\,\Gamma_{jj} \;\gtrsim\; \lambda_k(\Sigma)^2,$$

where we have used Assumption 6.(c). Based on this lower bound, Nazarov's inequality (Lemma S7.4) yields

$$\begin{aligned}\mathrm{J}(\epsilon) \;&\lesssim\; \frac{\epsilon}{\lambda_k(\Sigma)} \\ &\lesssim\; \frac{\log(n)\,\beta_{2q}\,\mathtt{r}(\Sigma)\,\lambda_1(\Sigma)/\lambda_k(\Sigma)}{n^{1/2}}.\end{aligned} \tag{S2.22}$$

Combining the previous bounds and noting that Assumption 6.(b) implies $\lambda_1(\Sigma)/\lambda_k(\Sigma) \lesssim 1$, we obtain the stated result. □

S3 Bootstrap Approximation

To introduce some notation, first note that the bootstrapped vectors $X_i^\star$ can be represented theoretically as $X_i^\star = \Sigma^{1/2} Z_i^\star$ for each $i = 1, \ldots, n$, where $Z_1^\star, \ldots, Z_n^\star$ are sampled with replacement from $Z_1, \ldots, Z_n$. Also, let $\widehat{\Sigma}^\star = \frac{1}{n}\sum_{i=1}^n X_i^\star (X_i^\star)^\top$, and let the diagonal matrix of the largest k sample eigenvalues be denoted as

$$\widehat{\Lambda}_k = \operatorname{diag}\big(\lambda_1(\widehat{\Sigma}), \ldots, \lambda_k(\widehat{\Sigma})\big).$$

For each $i = 1, \ldots, n$, recall the vector $W_i \in \mathbb{R}^k$ whose ith entry is defined as $W_{ij} = \langle u_j, Z_i\rangle^2 - 1$, where u_j is the jth eigenvector of Σ. Also let $\bar{W} = \frac{1}{n}\sum_{i=1}^n W_i$. In light of the fact that $\Gamma = \mathbb{E}[W_1 W_1^\top]$ with $\mathbb{E}[W_1] = 0$, the empirical counterpart of Γ is defined as

$$\widehat{\Gamma} \;=\; \frac{1}{n}\sum_{i=1}^n (W_i - \bar{W})(W_i - \bar{W})^\top. \tag{S3.1}$$

Lastly, recall that $\mathbb{P}(\cdot|X)$ and $\mathbb{E}[\cdot|X]$ refer to probability and expectation that are conditional on $X_1, \ldots, X_n$.

Theorem 14 (Bootstrap approximation)**.** *Suppose that Assumption 6 holds, and let $q = 5\log(kn)$. Also, let $\xi \in \mathbb{R}^k$ be a random vector that is conditionally distributed as $N\big(0, \Lambda_k \widehat{\Gamma} \Lambda_k\big)$, given the observations $X_1, \ldots, X_n$. Then, there is a constant $c > 0$ not depending on n such that the event*

$$\sup_{t\in\mathbb{R}^k} \Big| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star) - \boldsymbol{\lambda}_k(\widehat{\Sigma})\big) \preceq t \,\Big|\, X\Big) - \mathbb{P}\Big(\xi \preceq t \,\Big|\, X\Big)\Big| \;\le\; \frac{c\,\log(n)\,\beta_{3q}^3\,\mathtt{r}(\Sigma)}{n^{1/2}} \tag{S3.2}$$

holds with probability at least $1 - \frac{c}{n}$.

Proof. For each $i = 1, \ldots, n$, define

$$\tilde{X}_i^\star = \Lambda^{1/2} U^\top Z_i^\star,$$

as well as the matrix

$$\tilde{\Sigma}^\star = \frac{1}{n}\sum_{i=1}^n (\Lambda^{1/2} U^\top Z_i^\star)(\Lambda^{1/2} U^\top Z_i^\star)^\top.$$

Based on this definition, we have $\tilde{\Sigma}^\star = U^\top \widehat{\Sigma}^\star U$, and hence

$$\lambda_j(\tilde{\Sigma}^\star) = \lambda_j(\widehat{\Sigma}^\star)$$

for all $j = 1, \ldots, p$. By analogy with the proof of Theorem 13, we partition $\tilde{\Sigma}^\star$ as

$$\tilde{\Sigma}^\star = \begin{pmatrix} \tilde{\Sigma}^\star[1,1] & \tilde{\Sigma}^\star[1,2] \\ \tilde{\Sigma}^\star[1,2]^\top & \tilde{\Sigma}^\star[2,2] \end{pmatrix}, \tag{S3.3}$$

where the matrix $\tilde{\Sigma}^\star[1,1]$ is of size $k \times k$, and the matrix $\tilde{\Sigma}^\star[2,2]$ is of size $(p-k) \times (p-k)$.

To proceed, consider the bound

$$\sup_{t \in \mathbb{R}^k} \Big| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star) - \boldsymbol{\lambda}_k(\widehat{\Sigma})\big) \preceq t \,\Big|\, X \Big) - \mathbb{P}\big(\xi \preceq t \,\big|\, X\big) \Big| \;\leq\; \widehat{\mathrm{I}} \;+\; \widehat{\mathrm{II}}$$

where the terms on the right are defined as

$$\widehat{\mathrm{I}} \;=\; \sup_{t \in \mathbb{R}^k} \Big| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{\lambda}_k(\tilde{\Sigma}^\star) - \boldsymbol{\lambda}_k(\widehat{\Sigma})\big) \preceq t \,\Big|\, X \Big) - \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{d}_k(\tilde{\Sigma}^\star[1,1]) - \boldsymbol{d}_k(\tilde{\Sigma}[1,1])\big) \preceq t \,\Big|\, X \Big) \Big|, \tag{S3.4}$$

$$\widehat{\mathrm{II}} \;=\; \sup_{t \in \mathbb{R}^k} \Big| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{d}_k(\tilde{\Sigma}^\star[1,1]) - \boldsymbol{d}_k(\tilde{\Sigma}[1,1])\big) \preceq t \,\Big|\, X \Big) - \mathbb{P}\Big(\xi \preceq t \,\Big|\, X\Big) \Big|. \tag{S3.5}$$

Lemmas S3.4 and S3.3 ensure that there is a constant $c > 0$ not depending on n such that the following events

$$\widehat{\mathrm{I}} \;\leq\; \frac{c\, \log(n)\, \beta_{3q}^3\, \mathtt{r}(\Sigma)}{n^{1/2}} \tag{S3.6}$$

$$\widehat{\mathrm{II}} \;\leq\; \frac{c\, \beta_{3q}^3}{n^{1/2}}. \tag{S3.7}$$

each hold with probability $1 - \frac{c}{n}$. Combining these bounds gives the stated result. □

S3.1 Lemmas for bootstrap approximation

Lemma S3.1. *Suppose that the conditions of Theorem 14 hold. Then, there is a constant $c > 0$ not depending on n such that the event*

$$\mathbb{P}\bigg(\Big\| \sqrt{n}\Big(\boldsymbol{\lambda}_k(\tilde{\Sigma}^\star) - \boldsymbol{\lambda}_k(\tilde{\Sigma}^\star[1,1])\Big) \Big\|_\infty \;\geq\; \frac{c\, \log(n) \beta_{2q}\, \lambda_1(\Sigma)\, \mathtt{r}(\Sigma)}{n^{1/2}} \,\bigg|\, X \bigg) \;\leq\; \frac{c}{n^4} \tag{S3.8}$$

holds with probability at least $1 - \frac{c}{n}$.

Proof. Let $\pi(X)$ denote the conditional probability on the left side of (S3.8). By Markov's inequality, we have

$$\mathbb{P}(\pi(X) \geq \tfrac{1}{n^4}) \;\leq\; n^4 \mathbb{E}[\pi(X)],$$

and so it is sufficient to show that there is a constant $c > 0$ not depending on n such that $\mathbb{E}[\pi(X)] \leq c/n^5$. In other words, it is enough to show that the event

$$\Big\| \sqrt{n} \Big(\boldsymbol{\lambda}_k(\tilde{\Sigma}^\star) - \boldsymbol{\lambda}_k(\tilde{\Sigma}^\star[1,1]) \Big) \Big\|_\infty \;\leq\; \frac{c\, \log(n) \beta_{2q}\, \lambda_1(\Sigma)\, \mathtt{r}(\Sigma)}{n^{1/2}}$$

holds with probability at least $1 - c/n^5$.

Whenever the event $\big\{ \lambda_k(\tilde{\Sigma}^\star[1,1]) > \lambda_1(\tilde{\Sigma}^\star[2,2]) \big\}$ holds, Wielandt's inequality (Lemma S7.2) implies

$$\max_{1 \leq j \leq k} \big| \lambda_j(\tilde{\Sigma}^\star) - \lambda_j(\tilde{\Sigma}^\star[1,1]) \big| \;\leq\; \frac{\|\tilde{\Sigma}^\star[1,2]\|_{\mathrm{op}}^2}{\lambda_k(\tilde{\Sigma}^\star[1,1]) - \lambda_1(\tilde{\Sigma}^\star[2,2])}. \tag{S3.9}$$

The denominator in (S3.9) can be controlled by analogy with the proof of Lemma S2.1: It follows from Weyl's inequality and Assumption 6.(b) that

$$\begin{aligned} \lambda_k(\tilde{\Sigma}^\star[1,1]) - \lambda_1(\tilde{\Sigma}^\star[2,2]) \;&\geq\; \big(\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma) \big) - 2\|\tilde{\Sigma}^\star - \Lambda\|_{\mathrm{op}} \\ &\geq c_1 \lambda_1(\Sigma) - 2\|\tilde{\Sigma}^\star - \Lambda\|_{\mathrm{op}}, \end{aligned} \tag{S3.10}$$

for some constant $c_1 > 0$ not depending on n. So, controlling the denominator in (S3.9) amounts to showing that the random variable $\|\tilde{\Sigma}^\star - \Lambda\|_{\mathrm{op}}$ is small with high probability. We will proceed by upper bounding $\|\tilde{\Sigma}^\star - \tilde{\Sigma}\|_{\mathrm{op}} + \|\tilde{\Sigma} - \Lambda\|_{\mathrm{op}}$ with high probability. Applying Lemma S6.3 with the choice $q = 5\log(kn)$, it follows from a simple marginalization argument that the event

$$\|\tilde{\Sigma}^\star - \tilde{\Sigma}\|_{\mathrm{op}} \;\leq\; c\lambda_1(\Sigma) \sqrt{\frac{\log(n)\, \beta_q\, \mathtt{r}(\Sigma)}{n}} \tag{S3.11}$$

holds with probability at least $1 - \frac{c}{n^5}$. Next, recalling the bound (S2.11) in Lemma S2.1, there is a constant $c > 0$ not depending on n, such that the event

$$\|\tilde{\Sigma} - \Lambda\|_{\mathrm{op}} \;\leq\; c\lambda_1(\Sigma) \sqrt{\frac{\log(n)\, \beta_q\, \mathtt{r}(\Sigma)}{n}} \tag{S3.12}$$

holds with probability at least $1 - \frac{c}{n^5}$. Combining the last two bounds with (S3.10) and the condition (C.0.1) implies that the event $\{\lambda_k(\tilde{\Sigma}^\star[1,1]) - \lambda_1(\tilde{\Sigma}^*[2,2]) > \frac{c_1}{2}\lambda_1(\Sigma)\}$ holds with probability at least $1 - c/n^5$, where $q = 5\log(kn)$.

Lastly, to address the numerator in (S3.9), it follows from Lemma S6.2 and a simple marginalization argument that the event

$$\|\tilde{\Sigma}^\star[1,2]\|_{\text{op}}^2 \;\leq\; \frac{c\,\log(n)\,\beta_{2q}\,\lambda_1(\Sigma)\;\text{tr}(\Sigma)}{n} \tag{S3.13}$$

holds with probability at least $1 - \frac{c}{n^5}$. This completes the proof. □

Lemma S3.2. *Suppose that the conditions of Theorem 14 hold. Then, there is a constant $c > 0$ not depending on n such that the event*

$$\mathbb{P}\bigg(\Big\|\sqrt{n}\Big(\boldsymbol{\lambda}_k(\tilde{\Sigma}^\star[1,1]) - \boldsymbol{d}_k(\tilde{\Sigma}^\star[1,1])\Big)\Big\|_\infty \geq \frac{c\,\log(n)\,\beta_{2q}\,\lambda_1(\Sigma)\,\mathtt{r}(\Sigma)}{n^{1/2}}\,\bigg|X\bigg) \;\leq\; \frac{c}{n^4} \tag{S3.14}$$

holds with probability at least $1 - \frac{c}{n}$.

Proof. As in the proof of Lemma S2.2, we will lighten the use of subscripts by writing $\boldsymbol{\lambda}(A) = (\lambda_1(A), \ldots, \lambda_r(A))$ and $\boldsymbol{d}(A) = (A_{11}, \ldots, A_{rr})$ for any symmetric matrix $A \in \mathbb{R}^{r\times r}$ and integer $r \geq 1$. By the same reasoning used at the beginning of the proof of Lemma S3.1, it is sufficient to show that the event

$$\Big\|\sqrt{n}\Big(\boldsymbol{\lambda}(\tilde{\Sigma}^\star[1,1]) - \boldsymbol{d}(\tilde{\Sigma}^\star[1,1])\Big)\Big\|_\infty \;\leq\; \frac{c\,\log(n)\,\beta_{2q}\,\lambda_1(\Sigma)\,\mathtt{r}(\Sigma)}{n^{1/2}}$$

holds with probability at least $1 - \frac{c}{n^5}$. Overall, the current proof is similar to that of Lemma S2.2. Let $D_0^\star = \tilde{\Sigma}^\star[1,1]$, and for each $r = 1, \ldots, k-1$, partition the matrix $\tilde{\Sigma}^\star[1,1]$ recursively as

$$D_{r-1}^\star \;=\; \begin{pmatrix} d_1(D_{r-1}^\star) & V_r^\star \\ (V_r^\star)^\top & D_r^\star \end{pmatrix},$$

where $d_1(D_{r-1}^\star)$ is a scalar, and $D_r^\star$ is of size $(k-r)\times(k-r)$. Based on the proof of Lemma S2.2, it suffices to show that there is a constant $c > 0$ not depending on n, such that for any $r = 1, \ldots, k$, the event

$$\Big\|\boldsymbol{\lambda}(D_{r-1}^\star) - \begin{bmatrix} d_1(D_{r-1}^\star) \\ \boldsymbol{\lambda}(D_r^\star) \end{bmatrix}\Big\|_\infty \;\leq\; \frac{c\,\log(n)\,\beta_{2q}\,\lambda_1(\Sigma)\,\mathtt{r}(\Sigma)}{n}$$

holds with probability at least $1 - \frac{c}{kn^5}$.

Using the reasoning that led to (S2.17) in the proof of Lemma S2.2, Wielandt's inequality (Lemma S7.2) implies that if the event $\{d_1(D^\star_{r-1}) > \lambda_1(D^\star_r)\}$ holds, then the following event also holds

$$\left\| \boldsymbol{\lambda}(D^\star_{r-1}) - \begin{bmatrix} d_1(D^\star_{r-1}) \\ \boldsymbol{\lambda}(D^\star_r) \end{bmatrix} \right\|_\infty \;\leq\; \frac{\|V^\star_r\|^2_{\text{op}}}{d_1(D^\star_{r-1}) - \lambda_1(D^\star_r)}. \tag{S3.15}$$

The numerator in this bound (S3.15) can be controlled with Lemma S6.2 and a simple marginalization argument, which imply that under the choice $q = 5\log(kn)$, there is a constant $c > 0$ not depending on n such that the event

$$\|V^\star_r\|^2_{\text{op}} \;\leq\; \frac{c\,\log(n)\,\beta_{2q}\,\lambda_1(\Sigma)\;\text{tr}(\Sigma)}{n}$$

holds with probability at least $1 - \frac{1}{kn^5}$.

To control the denominator in the bound (S3.15), Weyl's inequality and the reasoning in the proof of Lemma S2.2 based on Assumption 6.(b) imply

$$d_1(D^\star_{r-1}) - \lambda_1(D^\star_r) \;\geq\; c_2\lambda_1(\Sigma) - 2\|\tilde{\Sigma}^\star - \Lambda\|_{\text{op}},$$

for some constant $c_2 > 0$ not depending on n. Next, by combining the bound (S2.11) and Lemma S6.3, it follows that the event

$$\|\tilde{\Sigma}^\star - \Lambda\|_{\text{op}} \;\leq\; c\lambda_1(\Sigma)\sqrt{\frac{\log(n)\,\beta_q\,\mathtt{r}(\Sigma)}{n}} \tag{S3.16}$$

holds with probability at least $1 - \frac{c}{kn^5}$. Therefore, condition (C.0.1) implies that the lower bound

$$d_1(D^\star_{r-1}) - \lambda_1(D^\star_r) \;\geq\; \tfrac{c_2}{2}\lambda_1(\Sigma)$$

holds with probability at least $1 - \frac{1}{kn^5}$. This completes the proof.

□

Lemma S3.3. *Suppose that the conditions of Theorem 14 hold, and let* $\widehat{\mathrm{II}}$ *be as defined in* (S3.5)*. Then, there is a constant* $c > 0$ *not depending on* n *such that the event*

$$\widehat{\mathrm{II}} \;\leq\; \frac{c\,\beta_{3q}^3}{n^{1/2}}$$

holds with probability at least $1 - \frac{c}{n}$*.*

Proof. Let $W_1, \ldots, W_n$ and $\bar{W}$ be as defined at the beginning of Appedix S3. Also, define the vector $W_i^\star \in \mathbb{R}^k$ with jth entry $W_{ij}^\star = \langle u_j, Z_i^\star\rangle^2 - 1$, and define $Y_i^\star = \Lambda_k(W_i^\star - \bar{W})$. These definitions give the relation

$$\sqrt{n}\big(\boldsymbol{d}_k(\tilde{\Sigma}^\star[1,1]) - \boldsymbol{d}_k(\tilde{\Sigma}[1,1])\big) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i^\star.$$

Also note that $\mathbb{E}[Y_i^\star|X] = 0$, and

$$\mathbb{E}\big[Y_i^\star (Y_i^\star)^\top \big| X\big] \;=\; \Lambda_k \widehat{\Gamma} \Lambda_k.$$

Due to Bentkus' multivariate Berry-Esseen theorem (Lemma S7.5), we have

$$\sup_{t\in\mathbb{R}^k}\left|\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i^\star \preceq t \;\middle|\; X\right) - \mathbb{P}\big(\xi \preceq t \mid X\big)\right| \leq \frac{c\;\mathbb{E}\big[\|(\Lambda_k\widehat{\Gamma}\Lambda_k)^{-1/2}Y_1^\star\|_2^3 \,\big|\, X\big]}{n^{1/2}}. \quad \text{(S3.17)}$$

Applying Lemma S6.5 with $q = 5\log(kn)$ and using Chebyshev's inequality, there is a constant $c > 0$ not depending on n such that the event

$$\mathbb{E}\big[\|(\Lambda_k\widehat{\Gamma}\Lambda_k)^{-1/2}Y_1^\star\|_2^3 \,\big|\, X\big] \;\leq\; c\,\beta_{3q}^3$$

holds with probability at least $1 - \frac{c}{n}$. □

Lemma S3.4. *Suppose that the conditions of Theorem 14 hold, and let* $\widehat{\mathrm{I}}$ *be as defined in* (S3.4)*. Then, there is a constant* $c > 0$ *not depending on* n *such that the event*

$$\widehat{\mathrm{I}} \;\leq\; \frac{c\,\log(n)\,\beta_{3q}^3\,\mathtt{r}(\Sigma)}{n^{1/2}} \quad \text{(S3.18)}$$

holds with probability at least $1 - \frac{c}{n}$*.*

Proof. In a similar manner to the proof of Lemma S2.4, the left side of (S3.18) can be bounded by the sum $2\widehat{\mathrm{I}\mathrm{I}} + \widehat{\mathrm{G}}_1(\epsilon) + 2\widehat{\mathrm{G}}_2(\epsilon) + 2\widehat{\mathrm{G}}_3$, with the last three terms defined for any $\epsilon > 0$ according to

$$\begin{aligned}
\widehat{\mathrm{G}}_1(\epsilon) &= \mathbb{P}\Big(\Big\| \sqrt{n}\Big(\boldsymbol{\lambda}_k(\tilde{\Sigma}^\star) - \boldsymbol{\lambda}_k(\widehat{\Sigma}) \Big) \; - \; \sqrt{n}\Big(\boldsymbol{d}_k(\tilde{\Sigma}^\star[1,1]) - \boldsymbol{d}_k(\tilde{\Sigma}[1,1]) \Big) \Big\|_\infty \geq \epsilon \; \Big| \; X \Big), \\
\mathrm{G}_2(\epsilon) &= \sup_{t \in \mathbb{R}^k} \Big| \mathbb{P}\Big(\zeta \preceq t + \epsilon 1_k \Big) - \mathbb{P}\Big(\zeta \preceq t - \epsilon 1_k \Big) \Big|, \\
\widehat{\mathrm{G}}_3 &= \sup_{t \in \mathbb{R}^k} \Big| \mathbb{P}\big(\zeta \preceq t \big) - \mathbb{P}\big(\xi \preceq t \,\big|\, X \big) \Big|.
\end{aligned}$$

First, recall from Lemma S3.3 that there is a constant $c > 0$ not depending on n such that $\widehat{\mathrm{I}\mathrm{I}} \leq c\beta_{3q}^3 / n^{1/2}$ holds with probability at least $1 - c/n$.

Next, when handling the terms $\widehat{\mathrm{G}}_1(\epsilon)$, $\mathrm{G}_2(\epsilon)$, and $\widehat{\mathrm{G}}_3$ below, we will take ϵ to be of the form $\epsilon = \frac{c}{\sqrt{n}} \log(n) \beta_{2q} \lambda_1(\Sigma) \mathtt{r}(\Sigma)$. Using the choice $q = 5\log(kn)$, Lemmas S2.1, S2.2, S3.1 and S3.2 imply that there is a constant $c > 0$ not depending on n such that the event

$$\widehat{\mathrm{G}}_1(\epsilon) \; \leq \; \frac{c}{n}$$

holds with probability at least $1 - \frac{c}{n}$. With regard to $\mathrm{G}_2(\epsilon)$, observe that it is deterministic and equal to $\mathrm{J}(\epsilon)$ in the proof of Lemma S2.4. Therefore, the bound (S2.22) gives

$$\mathrm{G}_2(\epsilon) \; \lesssim \frac{\log(n)\, \beta_{2q}\, \mathtt{r}(\Sigma)}{n^{1/2}}.$$

Lastly, the bound (S4.1) (to be established in Appendix S4) implies that the event

$$\widehat{\mathrm{G}}_3 \leq \; \frac{c\, \log(n)\, \beta_{2q}^2}{n^{1/2}}$$

holds with probability at least $1 - \frac{c}{n}$. Combining the last several bounds yields the stated result. □

S4 Proof of Theorem 11

By comparing the bounds (S2.1) and (S3.2) in Theorems 13 and 14, it is enough to show that there is a constant $c > 0$ not depending on n such that the event

$$\sup_{t \in \mathbb{R}^k} \Big| \mathbb{P}\big(\xi \preceq t \,\big|\, X \big) - \mathbb{P}\big(\zeta \preceq t \big) \Big| \; \leq \; \frac{c\, \log(n)\, \beta_{2q}^2}{n^{1/2}} \tag{S4.1}$$

holds with probability at least $1 - \frac{c}{n}$. To this end, define the three matrices

$$\begin{aligned} C_k &= \Lambda_k \Gamma \Lambda_k, \\ \widehat{C}_k &= \Lambda_k \widehat{\Gamma} \Lambda_k, \\ \widehat{B}_k &= C_k^{-1/2} \widehat{C}_k C_k^{-1/2} - I_k. \end{aligned}$$

By Lemma S7.7, there is a constant $c > 0$ not depending on n such that the bound

$$\sup_{t \in \mathbb{R}^k} \Big| \mathbb{P}\big(\xi \preceq t \,\big|\, X\big) - \mathbb{P}\big(\zeta \preceq t\big) \Big| \;\leq\; c \, \|\widehat{B}_k\|_{\mathrm{op}}$$

holds almost surely. Furthermore, due to Assumptions 6.(b) and 6.(c), the following bounds also hold almost surely,

$$\begin{aligned} \|\widehat{B}_k\|_{\mathrm{op}} &\leq \big\|C_k^{-1/2}\big\|_{\mathrm{op}}^2 \|\widehat{C}_k - C_k\|_{\mathrm{op}} \\ &\leq c\big(\tfrac{\lambda_1(\Sigma)}{\lambda_k(\Sigma)}\big)^2 \big\|\widehat{\Gamma} - \Gamma\big\|_{\mathrm{op}} \\ &\leq c \, \big\|\widehat{\Gamma} - \Gamma\big\|_{\mathrm{op}}. \end{aligned}$$

Next, Lemma S6.4 implies that the event

$$\|\widehat{\Gamma} - \Gamma\|_{\mathrm{op}} \;\leq\; \frac{c\, q\, \beta_{2q}^2}{n^{1/2}}$$

holds with probability at least $1 - e^{-q}$, and in light of the stated choice of $q = 5 \log(kn)$, the proof is complete. □

S5 Proof of Theorem 12

A simple rescaling argument can be used to show that the left side of the bound in Theorem 12 is the same as

$$\sup_{t \in \mathbb{R}^k} \left| \mathbb{P}\left(\boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma})/\lambda_1(\Sigma)) - \boldsymbol{h}(\boldsymbol{\lambda}_k(\Sigma)/\lambda_1(\Sigma)) \preceq t \right) - \mathbb{P}\left(\frac{\boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star)/\lambda_1(\Sigma)) - \boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma})/\lambda_1(\Sigma))}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau} \preceq t \,\middle|\, X \right) \right|, \tag{S5.1}$$

where we note that the quantity $(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau$ is invariant to rescaling of the observations. Hence, without loss of generality, we may assume that $\lambda_1(\Sigma) = 1$ in the remainder of this appendix. Consequently, Assumption 6.(b) implies there is a positive constant $c_\circ$ not

depending on n such that $\lambda_j(\Sigma) \in [c_\circ, 1]$ holds for all $j = 1, \dots, k$. (These points will sometimes be used without being explicitly mentioned in this appendix.) The proof is completed by combining Lemmas S5.1 and S5.2 given below. □

S5.1 Lemmas for bootstrap with transformations

Lemma S5.1. *Suppose that the conditions of Theorem 12 hold.*

(i) Let $\zeta \in \mathbb{R}^k$ be a random vector that is distributed as $N(0, \Lambda_k \Gamma \Lambda_k)$. Then,

$$\sup_{t \in \mathbb{R}^k} \Big| \mathbb{P}\Big(\sqrt{n}\big(\boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma})) - \boldsymbol{h}(\boldsymbol{\lambda}_k(\Sigma))\big) \preceq t \Big) - \mathbb{P}\Big(\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma)) \odot \zeta \preceq t \Big) \Big| \lesssim \frac{\log(n)\, \beta_{2q}^3\, \mathtt{r}(\Sigma)}{n^{1/2}}.$$

(ii) Let $\xi \in \mathbb{R}^k$ be a random vector that is conditionally distributed as $N\big(0, \Lambda_k \widehat{\Gamma} \Lambda_k\big)$ given the observations $X_1, \dots, X_n$. Then, there is a constant $c > 0$ not depending on n such that the event

$$\sup_{t \in \mathbb{R}^k} \left| \mathbb{P}\left(\frac{\sqrt{n}\big(\boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star)) - \boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma}))\big)}{(\widehat{\boldsymbol{\varsigma}}_k / \boldsymbol{\varsigma}_k)^\tau} \preceq t \,\middle|\, X \right) - \mathbb{P}\left(\frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma})) \odot \xi}{(\widehat{\boldsymbol{\varsigma}}_k / \boldsymbol{\varsigma}_k)^\tau} \preceq t \,\middle|\, X \right) \right| \leq \frac{c \log(n)\, \beta_{3q}^3\, \mathtt{r}(\Sigma)}{n^{1/2}}$$

holds with probability at least $1 - \frac{c}{n}$.

Proof. Part (i). Applying a Taylor expansion for each $j = 1, \dots, k$, we obtain

$$\begin{aligned} \sqrt{n}\big(h\big(\lambda_j(\widehat{\Sigma})\big) - h\big(\lambda_j(\Sigma)\big)\big) &= \sqrt{n}\, h'\big(\lambda_j(\Sigma)\big)\, \big(\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big) \\ &\quad + \sqrt{n}\big(\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big)^2 \int_0^1 (1-t)\, h''\big(\lambda_j(\Sigma) + t(\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma))\big) \mathrm{d}t. \end{aligned}$$

Let S, T, V, and R be random vectors in $\mathbb{R}^k$ whose entries are defined by

$$\begin{aligned} S_j &= \sqrt{n}\big(h\big(\lambda_j(\widehat{\Sigma})\big) - h\big(\lambda_j(\Sigma)\big)\big), \\ T_j &= h'\big(\lambda_j(\Sigma)\big)\zeta_j, \\ V_j &= \sqrt{n}\, h'\big(\lambda_j(\Sigma)\big)\, \big(\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big), \\ R_j &= \sqrt{n}\big(\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big)^2 \int_0^1 (1-t)\, h''\big(\lambda_j(\Sigma) + t(\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma))\big) \mathrm{d}t. \end{aligned}$$

Lemma S7.8 implies that for any $r > 0$,

$$\sup_{t \in \mathbb{R}^k} \big| \mathbb{P}(S \preceq t) - \mathbb{P}(T \preceq t) \big| \lesssim \sup_{t \in \mathbb{R}^k} \big| \mathbb{P}(V \preceq t) - \mathbb{P}(T \preceq t) \big| + \frac{r}{\min_{1 \leq j \leq k} \sqrt{\mathrm{var}(T_j)}} + \mathbb{P}(\|R\|_\infty \geq r). \tag{S5.2}$$

Theorem 13 gives a bound on the first term,

$$\sup_{t\in\mathbb{R}^k}\Big|\mathbb{P}\big(V \preceq t\big) - \mathbb{P}\big(T \preceq t\big)\Big| \;\lesssim\; \frac{\log(n)\,\beta_{2q}^3\,\mathtt{r}(\Sigma)}{n^{1/2}}. \tag{S5.3}$$

Next, recall from the discussion at the beginning of this appendix that we may assume there is a constant $c_\circ > 0$ not depending on n such that $\lambda_j(\Sigma) \in [c_\circ, 1]$ holds for all $j = 1, \ldots, k$. In turn, combining this with Assumption 6.(c) implies

$$\begin{aligned}\operatorname{var}(T_j) \;&=\; \lambda_j(\Sigma)^2 \cdot \Gamma_{jj} \cdot h'\big(\lambda_j(\Sigma)\big)^2 \\ &\gtrsim\; 1. \end{aligned} \tag{S5.4}$$

It remains to find a bound for the last term in (S5.2) and to choose a suitable value for r. For each $j = 1, \ldots, k$, define the event

$$\mathcal{A}_j = \Big\{\big|\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big| \le \frac{c_\circ}{2}\Big\}.$$

Then,

$$\mathbb{P}\big(|R_j| \ge r\big) \;\le\; \mathbb{P}\big(\{|R_j| \ge r\} \cap \mathcal{A}_j\big) + \mathbb{P}\big(\mathcal{A}_j^c\big).$$

Using the choice $q = 5\log(kn)$ and Weyl's inequality, the bound (S2.10) implies that

$$\begin{aligned}\max_{1\le j\le k}\big\|\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big\|_q \;&\le\; \big\|\|\tilde{\Sigma} - \Lambda\|_{\mathrm{op}}\big\|_q \\ &\lesssim\; \frac{\lambda_1(\Sigma)\sqrt{\log(n)\beta_q\mathtt{r}(\Sigma)}}{n^{1/2}} \\ &\le\; \frac{c_\circ}{2e}, \end{aligned} \tag{S5.5}$$

where the condition (C.0.1) has been used in the last step, and the rationale for the constant $\frac{c_\circ}{2e}$ will be seen in the next step. By Chebyshev's inequality, we have

$$\max_{1\le j\le k}\mathbb{P}(\mathcal{A}_j^c) \;\le\; \max_{1\le j\le k}\frac{\big\|\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big\|_q^q}{(c_\circ/2)^q} \;\le\; \frac{(c_\circ/(2e))^q}{(c_\circ/2)^q} \;=\; e^{-q} \;\le\; \frac{1}{kn}. \tag{S5.6}$$

In order to handle $\mathbb{P}\big(\{|R_j| \ge r\} \cap \mathcal{A}_j\big)$, first let C denote the following supremum

$$C = \sup\Big\{|h''(x)|\;\Big|\; c_\circ/2 \le x \le 1 + c_\circ/2\Big\}, \tag{S5.7}$$

which is finite and does not depend on n. Taking $r = \frac{c}{\sqrt{n}}\lambda_1(\Sigma)^2 \log(kn)\beta_q \mathtt{r}(\Sigma)$ for a sufficiently large constant c, we have

$$\begin{aligned}
\mathbb{P}\big(\{|R_j| \geq r\} \cap \mathcal{A}_j\big) \;&\leq\; \mathbb{P}\Big(\sqrt{n}\big(\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big)^2 \cdot \tfrac{C}{2} \geq r\Big) \\
&\leq\; \mathbb{P}\Big(\big|\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big| \geq \frac{\lambda_1(\Sigma)}{n^{1/2}}\sqrt{(2c/C)\log(kn)\beta_q \mathtt{r}(\Sigma)}\Big) \\
&\leq\; \tfrac{1}{kn}, \qquad \text{(S5.8)}
\end{aligned}$$

where the previous step uses (S5.5). Combining the last several steps shows that $\mathbb{P}(|R_j| \geq r) \;\leq\; \frac{2}{kn}$ holds for all $j = 1, \ldots, k$ and so a union bound gives

$$\mathbb{P}(\|R\|_\infty \geq r) \;\lesssim\; \tfrac{1}{n}. \qquad \text{(S5.9)}$$

Substituting the bounds (S5.3), (S5.4), and (S5.9) into (S5.2) proves the statement (i).

Part (ii). First note that by rescaling, it is enough to show that the event

$$\sup_{t\in\mathbb{R}^k}\left|\mathbb{P}\left(\sqrt{n}\big(\boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star)) - \boldsymbol{h}(\boldsymbol{\lambda}_k(\widehat{\Sigma}))\big) \preceq t \,\middle|\, X\right) - \mathbb{P}\left(\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))\odot\xi \preceq t \,\middle|\, X\right)\right| \;\leq\; \frac{c\,\log(n)\,\beta_{3q}^3\,\mathtt{r}(\Sigma)}{n^{1/2}}$$

holds with probability at least $1 - c/n$. Similar to (S5.2), the left side can be decomposed as

$$\begin{aligned}
&\sup_{t\in\mathbb{R}^k}\Big|\mathbb{P}\big(S^\star \preceq t \,|\, X\big) - \mathbb{P}\big(T^\star \preceq t \,|\, X\big)\Big| \\
&\qquad\lesssim\; \sup_{t\in\mathbb{R}^k}\Big|\mathbb{P}\big(V^\star \preceq t \,|\, X\big) - \mathbb{P}\big(T^\star \preceq t \,|\, X\big)\Big| + \frac{r}{\min_{1\leq j\leq k}\sqrt{\mathrm{var}(T_j^\star \,|\, X)}} \;+\; \mathbb{P}(\|R^\star\|_\infty \geq r \,|\, X), \qquad \text{(S5.10)}
\end{aligned}$$

where $S^\star$, $T^\star$, $R^\star$, $V^\star$ are vectors in $\mathbb{R}^k$ defined as

$$\begin{aligned}
S_j^\star \;&=\; \sqrt{n}\big(h\big(\lambda_j(\widehat{\Sigma}^\star)\big) - h\big(\lambda_j(\widehat{\Sigma})\big)\big), \\
T_j^\star \;&=\; h'\big(\lambda_j(\widehat{\Sigma})\big)\,\xi_j, \\
V_j^\star \;&=\; \sqrt{n}\,h'\big(\lambda_j(\widehat{\Sigma})\big)\,\big(\lambda_j(\widehat{\Sigma}^\star) - \lambda_j(\widehat{\Sigma})\big), \\
R_j^\star \;&=\; \sqrt{n}\big(\lambda_j(\widehat{\Sigma}^\star) - \lambda_j(\widehat{\Sigma})\big)^2 \int_0^1 (1-t)\,h''\big(\lambda_j(\widehat{\Sigma}) + t(\lambda_j(\widehat{\Sigma}^\star) - \lambda_j(\widehat{\Sigma}))\big)\mathrm{d}t.
\end{aligned}$$

The first term of the bound (S5.10) is handled by Theorem 14. For the middle term, first note that $\mathrm{var}(T_j^\star \,|\, X) = \lambda_j(\Sigma)^2 \cdot \widehat{\Gamma}_{jj} \cdot h'(\lambda_j(\widehat{\Sigma}))^2$. Using the condition (C.0.1), it follows from (S2.11) and Lemma S6.4 that for each $j = 1, \ldots, k$, the events

$$\lambda_j(\widehat{\Sigma}) \geq c\, \lambda_j(\Sigma),$$

$$\widehat{\Gamma}_{jj} \geq c\, \Gamma_{jj},$$

$$h'(\lambda_j(\widehat{\Sigma})) \geq c\, h'(\lambda_j(\Sigma))$$

each hold with probability at least $1 - \frac{c}{kn}$. Hence, Assumptions 6.(b) and (c) imply that the event

$$\min_{1 \leq j \leq k} \mathrm{var}(T_j^\star \,|\, X) \geq c$$

holds with probability at least $1 - \frac{c}{n}$. Next, define the event

$$\mathcal{A}_j^\star = \Big\{ \big|\lambda_j(\widehat{\Sigma}^\star) - \lambda_j(\widehat{\Sigma})\big| \leq \frac{c_\circ}{2} \Big\}$$

with the constant $c_\circ$ having the same definition as in Part (i). For the last term on the right side of (S5.10), we claim that the event

$$\mathbb{P}(\|R^\star\|_\infty \geq r \,|\, X) \;\leq\; \frac{c}{n}$$

holds with probability at least $1 - \frac{c}{n}$ when r is appropriately chosen. This can be established with a union bound

$$\mathbb{P}(\|R^\star\|_\infty \geq r \,|\, X) \;\leq\; \sum_{j=1}^{k} \mathbb{P}\big(\big\{ |R_j^\star| \geq r \big\} \cap \mathcal{A}_j^\star \,\big|\, X \big) + \mathbb{P}\big((\mathcal{A}_j^\star)^c \,|\, X \big). \tag{S5.11}$$

Analogously to (S5.6), we can apply Lemma S6.3 with $q = 5\log(kn)$ and the condition (C.0.1) to conclude that for each $j = 1, \ldots, k$ the event

$$\max_{1 \leq j \leq k} \mathbb{P}\big((\mathcal{A}_j^\star)^c \,|\, X \big) \;\leq\; \frac{c}{kn}$$

holds with probability at least $1 - \frac{c}{kn}$. For the first term on the right side of (S5.11), we may use an argument similar to the one leading up to (S5.8) with $r = \frac{c}{\sqrt{n}} \lambda_1(\Sigma)^2 \log(kn) \beta_q \mathtt{r}(\Sigma)$ to show that for each $j = 1, \ldots, k$ the event

$$\mathbb{P}\big(\{ |R_j^\star| \geq r \} \cap \mathcal{A}_j^\star \,\big|\, X \big) \;\leq\; \frac{c}{kn}$$

holds with probability at least $1 - \frac{c}{kn}$. Combining the last few bounds completes the proof. $\square$

Lemma S5.2. *Suppose the conditions of Theorem 12 hold. Then, there is a constant $c > 0$ not depending on n such that the event*

$$\sup_{t\in\mathbb{R}^k}\left|\mathbb{P}\Big(\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\odot\zeta \preceq t\Big) - \mathbb{P}\bigg(\frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))\odot\xi}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau} \preceq t \,\bigg|\, X\bigg)\right| \;\leq\; \frac{c\,\log(n)\,\beta_{2q}^5\,\mathtt{r}(\Sigma)}{n^{1/2}} \tag{S5.12}$$

holds with probability at least $1 - \frac{c}{n}$.

Proof. Define the matrices

$$\begin{aligned}
\mathsf{C}_k &= \operatorname{diag}\big(\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\big)\Big(\Lambda_k\Gamma\Lambda_k\Big)\operatorname{diag}\big(\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\big)\\
\widehat{\mathsf{C}}_k &= \operatorname{diag}\bigg(\frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau}\bigg)\Big(\Lambda_k\widehat{\Gamma}\Lambda_k\Big)\operatorname{diag}\bigg(\frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau}\bigg)\\
\widehat{\mathsf{B}}_k &= \mathsf{C}_k^{-1/2}\widehat{\mathsf{C}}_k\mathsf{C}_k^{-1/2} - I_k.
\end{aligned}$$

By Lemma S7.7, the following bound holds almost surely

$$\sup_{t\in\mathbb{R}^k}\left|\mathbb{P}\bigg(\frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))\odot\xi}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau} \preceq t \,\bigg|\, X\bigg) - \mathbb{P}\Big(\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\odot\zeta \preceq t\Big)\right| \;\leq\; c\,\|\widehat{\mathsf{B}}_k\|_{\mathrm{op}} \tag{S5.13}$$

for some constant $c > 0$ not depending on n. Using several applications of the triangle inequality, the following bound holds almost surely,

$$\begin{aligned}
\|\widehat{\mathsf{B}}_k\|_{\mathrm{op}} &\leq \big\|\mathsf{C}_k^{-1/2}\big\|_{\mathrm{op}}^2\|\mathsf{C}_k - \widehat{\mathsf{C}}_k\|_{\mathrm{op}}\\
&\leq \big\|\mathsf{C}_k^{-1/2}\big\|_{\mathrm{op}}^2\cdot\Big\|\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\Big\|_\infty\cdot\Big\|\Lambda_k\Gamma\Lambda_k\Big\|_{\mathrm{op}}\cdot\bigg\|\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma)) - \frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau}\bigg\|_\infty \qquad \text{(S5.14)}\\
&\quad+\quad \big\|\mathsf{C}_k^{-1/2}\big\|_{\mathrm{op}}^2\cdot\Big\|\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\Big\|_\infty\cdot\Big\|\Lambda_k(\Gamma-\widehat{\Gamma})\Lambda_k\Big\|_{\mathrm{op}}\cdot\bigg\|\frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau}\bigg\|_\infty\\
&\quad+\quad \big\|\mathsf{C}_k^{-1/2}\big\|_{\mathrm{op}}^2\cdot\bigg\|\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma)) - \frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau}\bigg\|_\infty\cdot\Big\|\Lambda_k\widehat{\Gamma}\Lambda_k\Big\|_{\mathrm{op}}\cdot\bigg\|\frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau}\bigg\|_\infty.
\end{aligned}$$

The leading factor satisfies

$$\big\|\mathsf{C}_k^{-1/2}\big\|_{\mathrm{op}}^2 \;\lesssim\; 1$$

because we may assume that there is a constant $c_\circ > 0$ not depending on n such that $\lambda_j(\Sigma) \in [c_\circ, 1]$ holds for all $j = 1, \ldots, k$ (as discussed at the beginning of this appendix),

and also because $\lambda_k(\Gamma) \gtrsim 1$ by Assumption 6.(c). Also, the quantity $\|\widehat{\Gamma} - \Gamma\|_{\text{op}}$ is handled by Lemma S6.4, which shows there is a constant $c > 0$ not depending on n such that the event

$$\|\widehat{\Gamma} - \Gamma\|_{\text{op}} \;\leq\; \frac{c\,\log(n)\,\beta_{2q}^2}{n^{1/2}} \tag{S5.15}$$

holds with probability at least $1 - \frac{c}{n}$. Combining this with the condition (C.0.1) and the bound $\|\Gamma\|_{\text{op}} \lesssim \beta_2^2$, it follows that

$$\|\Lambda_k \widehat{\Gamma} \Lambda_k\|_{\text{op}} \;\leq\; \|\Lambda_k \Gamma \Lambda_k\|_{\text{op}} + \|\Lambda_k(\widehat{\Gamma} - \Gamma)\Lambda_k\|_{\text{op}} \;\leq\; c\,\beta_2^2$$

holds with probability at least $1 - \frac{c}{n}$. To handle the remaining quantities in the bound (S5.14), note that the triangle inequality yields

$$\left\| \frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^{\tau}} - \boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma)) \right\|_{\infty} \leq \left\|(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^{-\tau}\right\|_{\infty} \left\|\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma})) - \boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\right\|_{\infty} + \left\|\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\right\|_{\infty} \left\|(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^{-\tau} - 1_k\right\|_{\infty}. \tag{S5.16}$$

Using Lemmas S5.3, S5.4, and S5.5, as well as some elementary inequalities, the following bounds hold with probability at least $1 - c/n$ for any $\tau \in [0, 1]$,

$$\begin{aligned} \left\|(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^{-\tau} - 1_k\right\|_{\infty} \;&\leq\; \left\|1_k - \boldsymbol{\varsigma}_k/\widehat{\boldsymbol{\varsigma}}_k\right\|_{\infty} \\ &\leq\; \left\| \tfrac{\widehat{\boldsymbol{\varsigma}}_k^2 - \boldsymbol{\varsigma}_k^2}{\widehat{\boldsymbol{\varsigma}}_k \boldsymbol{\varsigma}_k} \right\|_{\infty} \\ &\leq\; \frac{c\,\log(n)\,\beta_{2q}^3\,\mathtt{r}(\Sigma)}{n^{1/2}}. \end{aligned} \tag{S5.17}$$

Furthermore, using the condition (C.0.1), this implies that the event

$$\left\|(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^{-\tau}\right\|_{\infty} \leq c \tag{S5.18}$$

also holds with probability at least $1 - c/n$. Next, we derive upper bounds for $\left\|\boldsymbol{h}'(\lambda_j(\widehat{\Sigma}))\right\|_{\infty}$ and $\left\|\boldsymbol{h}'(\lambda_j(\widehat{\Sigma})) - \boldsymbol{h}'(\lambda_j(\Sigma))\right\|_{\infty}$. Using (S2.11) in the proof of Lemma S2.1, it follows that the bounds

$$\begin{aligned} \left\|\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma})) - \boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\right\|_{\infty} \;&\leq\; c\,\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \\ &\leq\; c\lambda_1(\Sigma)\sqrt{\frac{\log(n)\,\beta_q\,\mathtt{r}(\Sigma)}{n}} \end{aligned} \tag{S5.19}$$

hold with probability at least $1 - \frac{c}{n}$. Substituting bounds (S5.17), (S5.18), and (S5.19), into (S5.16), it follows that the event

$$\left\| \frac{\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))}{(\widehat{\boldsymbol{\varsigma}}_k/\boldsymbol{\varsigma}_k)^\tau} - \boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma)) \right\|_\infty \;\leq\; \frac{c\,\log(n)\,\beta_{2q}^3\,\mathtt{r}(\Sigma)}{n^{1/2}} \tag{S5.20}$$

holds with probability at least $1 - \frac{c}{n}$. Lastly, observe that similar reasoning implies that the event

$$\left\|\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma}))\right\|_\infty \;\leq\; \left\|\boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\right\|_\infty + \left\|\boldsymbol{h}'(\boldsymbol{\lambda}_k(\widehat{\Sigma})) - \boldsymbol{h}'(\boldsymbol{\lambda}_k(\Sigma))\right\|_\infty \;\leq\; c \tag{S5.21}$$

holds with probability at least $1 - \frac{c}{n}$. By combining the last several bounds with (S5.14), the proof is complete. □

Lemma S5.3. *Suppose that the conditions of Theorem 12 hold. Then,*

$$\max_{1\leq j\leq k} \left|\mathbb{E}[\lambda_j(\widehat{\Sigma})] - \lambda_j(\Sigma)\right| \;\lesssim\; \frac{\log(n)\,\beta_{2q}\,\mathrm{tr}(\Sigma)}{n} \tag{i}$$

and

$$\max_{1\leq j\leq k} \left|\,\mathrm{var}\left(\lambda_j(\widehat{\Sigma})\right) - \lambda_j(\Sigma)^2\,\Gamma_{jj}/n\right| \;\lesssim\; \frac{\log(n)\,\beta_{2q}^2\,\lambda_1(\Sigma)\,\mathrm{tr}(\Sigma)}{n^{3/2}}. \tag{ii}$$

Also, there is a constant $c > 0$ not depending on n such that the events

$$\max_{1\leq j\leq k} \left|\mathbb{E}[\lambda_j(\widehat{\Sigma}^\star)|X] - \lambda_j(\widehat{\Sigma})\right| \;\leq\; \frac{c\,\log(n)\,\beta_{2q}\,\mathrm{tr}(\Sigma)}{n} \tag{iii}$$

and

$$\max_{1\leq j\leq k} \left|\,\mathrm{var}\left(\lambda_j(\widehat{\Sigma}^\star)\middle|X\right) - \lambda_j(\Sigma)^2\,\widehat{\Gamma}_{jj}/n\right| \;\leq\; \frac{c\,\log(n)\,\beta_{2q}^2\,\lambda_1(\Sigma)\,\mathrm{tr}(\Sigma)}{n^{3/2}} \tag{iv}$$

each hold with probability at least $1 - \frac{c}{n}$.

Proof. Part (i): Recalling the choice $q = 5\log(kn)$ from the statement of Theorem 12, it follows from Lemmas S2.1 and S2.2 that there is a constant c not depending on n such that the event

$$\max_{1\leq j\leq k} \left|\lambda_j(\widehat{\Sigma}) - d_j(\tilde{\Sigma}[1,1])\right| \;\leq\; \frac{c\,\log(n)\,\beta_{2q}\,\mathrm{tr}(\Sigma)}{n} \tag{S5.22}$$

holds with probability at least $1 - \frac{c}{n^4}$. To simplify presentation, let δ be a number of the form $\delta = \frac{c}{n}\log(n)\beta_{2q}\,\mathrm{tr}(\Sigma)$ and define the event

$$\mathcal{E}_j = \left\{|\lambda_j(\widehat{\Sigma}) - d_j(\tilde{\Sigma}[1,1])| \leq \delta\right\}$$

for each $j = 1, \ldots, k$. Also, as a temporary short hand, let U_j and V_j denote the random variables

$$U_j = \lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma) \qquad \text{and} \qquad V_j = d_j(\tilde{\Sigma}[1,1]) - \lambda_j(\Sigma).$$

Noting that $\mathbb{E}[d_j(\tilde{\Sigma}[1,1])] = \lambda_j(\Sigma)$, we have

$$\begin{aligned} \big|\mathbb{E}[\lambda_j(\widehat{\Sigma})] - \lambda_j(\Sigma)\big| &= |\mathbb{E}[U_j] - \mathbb{E}[V_j]| \\ &\leq \|U_j - V_j\|_2 \\ &\leq \big\||U_j - V_j| \cdot 1\{\mathcal{E}_j\}\big\|_2 + \big\||U_j - V_j| \cdot 1\{\mathcal{E}_j^c\}\big\|_2 \\ &\lesssim \delta + \|\lambda_1(\tilde{\Sigma})\|_4 \left(\mathbb{P}(\mathcal{E}_j^c)\right)^{1/4}, \end{aligned} \tag{S5.23}$$

where the fact $d_j(\tilde{\Sigma}[1,1]) = d_j(\tilde{\Sigma}) \leq \lambda_1(\tilde{\Sigma}) = \lambda_1(\widehat{\Sigma})$ has been used in the last step. Due to (S5.22), we have $\mathbb{P}(\mathcal{E}_j^c) \leq c/n^4$, and so it is adequate to derive a simple upper bound on $\|\lambda_1(\tilde{\Sigma})\|_4$ using (S2.10) and condition (C.0.1) as follows

$$\begin{aligned} \|\lambda_1(\tilde{\Sigma})\|_4 &\leq \lambda_1(\Sigma) + \big\|\|\tilde{\Sigma} - \Lambda\|_{\text{op}}\big\|_q, \\ &\leq c\,\lambda_1(\Sigma). \end{aligned} \tag{S5.24}$$

Hence, using the last bound in (S5.23) completes the proof of Part (i).

Part (ii): First note that

$$\begin{aligned} \big|\operatorname{var}\big(\lambda_j(\widehat{\Sigma})\big) - \lambda_j(\Sigma)^2\,\Gamma_{jj}/n\big| &= |\operatorname{var}(U_j) - \operatorname{var}(V_j)| \\ &\leq 2\,(\|U_j\|_2 + \|V_j\|_2)\,\|U_j - V_j\|_2. \end{aligned} \tag{S5.25}$$

Also note that the intermediate steps in Part (i) give

$$\|U_j - V_j\|_2 \lesssim \frac{\log(n)\,\beta_{2q}\,\operatorname{tr}(\Sigma)}{n}.$$

From the proof of Lemma S2.3, we have the identity

$$V_j = d_j(\tilde{\Sigma}[1,1]) - \lambda_j(\Sigma) = \frac{1}{n}\sum_{i=1}^{n}(\Lambda_k W_i)_j,$$

which leads to

$$\|V_j\|_2^2 \;=\; \Big\|\frac{1}{n}\sum_{i=1}^{n}(\Lambda_k W_i)_j\Big\|_2^2 \;\lesssim\; \frac{\beta_2^2\,\lambda_1(\Sigma)^2}{n}.$$

Also, we can bound $\|U_j\|_2^2$ as follows

$$\begin{aligned}\|U_j\|_2^2 \;&\le\; 2\|V_j\|_2^2 + 2\|U_j - V_j\|_2^2 \\ &\lesssim\; \frac{\beta_2^2\,\lambda_1(\Sigma)^2}{n} + \left(\frac{\log(n)\beta_{2q}\,\mathrm{tr}(\Sigma)}{\sqrt{n}}\right)^2 \frac{1}{n} \\ &\lesssim\; \frac{\beta_2^2\,\lambda_1(\Sigma)^2}{n}, \end{aligned} \tag{S5.26}$$

where the third step uses the condition (C.0.1). Combining the last several bounds completes the proof of Part (ii).

Part (iii): The proof is similar to that of Part (i). Lemmas S3.1 and S3.2 imply that the event

$$\mathbb{P}\bigg(\Big\|\sqrt{n}\Big(\boldsymbol{\lambda}_k(\tilde{\Sigma}^\star) - \boldsymbol{d}_k(\tilde{\Sigma}^\star[1,1])\Big)\Big\|_\infty \;\ge\; \frac{c\,\log(n)\beta_{2q}\,\mathrm{tr}(\Sigma)}{n^{1/2}}\,\bigg|\,X\bigg) \;\le\; \frac{c}{n^4} \tag{S5.27}$$

holds with probability at least $1 - \frac{c}{n}$. Letting δ have the same form as in Part (i), define the event

$$\mathcal{E}_j^\star = \Big\{|\lambda_j(\widehat{\Sigma}^\star) - d_j(\tilde{\Sigma}^\star[1,1])| \le \delta\Big\}$$

for each $j = 1, \ldots, k$. Also, define

$$U_j^\star = \lambda_j(\widehat{\Sigma}^\star) - d_j(\tilde{\Sigma}[1,1]) \quad \text{and} \quad V_j^\star = d_j(\tilde{\Sigma}^\star[1,1]) - d_j(\tilde{\Sigma}[1,1])$$

as the bootstrap counterparts of U_j and V_j. To proceed, note that $\mathbb{E}[d_j(\tilde{\Sigma}^\star[1,1])|X] = d_j(\tilde{\Sigma}[1,1])$, and so

$$\big|\mathbb{E}[\lambda_j(\widehat{\Sigma}^\star)|X] - \lambda_j(\widehat{\Sigma})\big| \;\le\; |\mathbb{E}[U_j^\star|X] - \mathbb{E}[V_j^\star|X]| + |d_j(\tilde{\Sigma}[1,1]) - \lambda_j(\widehat{\Sigma})|. \tag{S5.28}$$

The first term on the right side can be handled similarly to (S5.23),

$$\begin{aligned}|\mathbb{E}[U_j^\star|X] - \mathbb{E}[V_j^\star|X]| \;&\le\; \mathbb{E}\big[(U_j^\star - V_j^\star)^2 1\{\mathcal{E}_j^\star\}\big|X\big]^{1/2} \;+\; \mathbb{E}\big[(U_j^\star - V_j^\star)^2 1\{(\mathcal{E}_j^\star)^c\}\big|X\big]^{1/2} \\ &\le\; c\Big(\delta + \big(\mathbb{E}\big[\lambda_1^4(\tilde{\Sigma}^\star)\,\big|\,X\big]\big)^{1/4}\big(\mathbb{P}(\cup_{j=1}^k(\mathcal{E}_j^\star)^c|X)\big)^{1/4}\Big),\end{aligned}$$

where the fact $d_j(\tilde{\Sigma}^\star[1,1]) = d_j(\tilde{\Sigma}^\star) \le \lambda_1(\tilde{\Sigma}^\star)$ has been used in the last step. Due to (S5.27), we have $\mathbb{P}(\cup_{j=1}^k (\mathcal{E}_j^\star)^c | X) \le c/n^4$ with probability at least $1 - c/n$. Furthermore, it follows from Lemma S6.3 and (S2.11) that the event

$$\begin{aligned}
\big(\mathbb{E}\big[\lambda_1(\tilde{\Sigma}^\star)^4 \,\big|\, X\big]\big)^{1/4} &\le \lambda_1(\Sigma) + \|\tilde{\Sigma} - \Sigma\|_{\mathrm{op}} + \big(\mathbb{E}\big[\|\tilde{\Sigma}^\star - \tilde{\Sigma}\|_{\mathrm{op}}^q \,\big|\, X\big]\big)^{1/q} \\
&\le \lambda_1(\Sigma) + c\,\lambda_1(\Sigma)\sqrt{\frac{\log(n)\beta_q \mathtt{r}(\Sigma)}{n}} \qquad \text{(S5.29)} \\
&\le c\,\lambda_1(\Sigma)
\end{aligned}$$

holds with probability at least $1 - \frac{c}{n}$, where the last line has used condition (C.0.1). Thus, the event

$$\max_{1\le j\le k} |\mathbb{E}[U_j^\star|X] - \mathbb{E}[V_j^\star|X]| \;\le\; \frac{c\,\log(n)\,\beta_{2q}\,\mathrm{tr}(\Sigma)}{n}$$

holds with probability at least $1 - c/n$. For the second term on the right side of (S5.28), note that (S5.22) implies that the event

$$\max_{1\le j\le k} |d_j(\tilde{\Sigma}[1,1]) - \lambda_j(\widehat{\Sigma})| \;\le\; \frac{c\,\log(n)\,\beta_{2q}\,\mathrm{tr}(\Sigma)}{n}$$

holds with probability at least $1 - c/n$. Applying the last several steps into (S5.28) completes the proof of Part (*iii*).

Part (*iv*): The proof is essentially analogous to that of Part (*ii*), and so the details are omitted. □

Lemma S5.4. *Suppose that the conditions of Theorem 12 hold. Then,*

$$\max_{1\le j\le k} \big|\,\mathrm{var}\big(h(\lambda_j(\widehat{\Sigma}))\big) - h'(\lambda_j(\Sigma))^2\,\mathrm{var}\big(\lambda_j(\widehat{\Sigma})\big)\big| \;\lesssim\; \frac{\log(n)\,\beta_{2q}^2\,\lambda_1(\Sigma)\;\mathrm{tr}(\Sigma)}{n^{3/2}}. \qquad (i)$$

In addition, there is a constant $c > 0$ not depending on n such that the events

$$\max_{1\le j\le k} \big|\,\mathrm{var}\big(h(\lambda_j(\widehat{\Sigma}^\star))\big|X\big) - h'(\lambda_j(\widehat{\Sigma}))^2\,\mathrm{var}\big(\lambda_j(\widehat{\Sigma}^\star)\big|X\big)\big| \;\le\; \frac{c\,\log(n)\,\beta_{2q}^2\,\lambda_1(\Sigma)\;\mathrm{tr}(\Sigma)}{n^{3/2}} \quad (ii)$$

and

$$\max_{1\le j\le k} \big|\,\mathrm{var}\big(h(\lambda_j(\widehat{\Sigma}^\star))\big|X\big) - \mathrm{var}\big(h(\lambda_j(\widehat{\Sigma}))\big)\big| \;\le\; \frac{c\,\log(n)\,\beta_{2q}^3\,\lambda_1(\Sigma)\;\mathrm{tr}(\Sigma)}{n^{3/2}} \qquad (iii)$$

each hold with probability at least $1 - \frac{c}{n}$.

Proof. Part (i): For each $j = 1, \ldots, k$ define the random variables

$$S_j = h(\lambda_j(\widehat{\Sigma})) - h(\lambda_j(\Sigma)) \quad \text{and} \quad T_j = h'(\lambda_j(\Sigma)) \big(\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big). \tag{S5.30}$$

In this notation, we have

$$\begin{aligned} \max_{1 \leq j \leq k} \big| \operatorname{var}\big(h(\lambda_j(\widehat{\Sigma}))\big) - h'(\lambda_j(\Sigma))^2 \operatorname{var}\big(\lambda_j(\widehat{\Sigma})\big) \big| &= \max_{1 \leq j \leq k} \big| \operatorname{var}(S_j) - \operatorname{var}(T_j) \big| \\ &\leq \max_{1 \leq j \leq k} 2(\|S_j\|_2 + \|T_j\|_2)\|S_j - T_j\|_2. \end{aligned} \tag{S5.31}$$

As a way of handling $\|S_j - T_j\|_2$, first note that the argument used to establish (S5.9) can also be used to show that the event

$$\max_{1 \leq j \leq k} |S_j - T_j| \leq \frac{c \log(n)\, \beta_{2q} \operatorname{tr}(\Sigma)}{n} \tag{S5.32}$$

holds with probability at least $1 - \frac{c}{n^4}$. Using the bound (S5.32) and an argument analogous to the one leading up to (S5.23), we obtain

$$\|S_j - T_j\|_2 \lesssim \frac{\log(n) \beta_{2q} \operatorname{tr}(\Sigma)}{n} + \frac{\|S_j\|_4 + \|T_j\|_4}{n}. \tag{S5.33}$$

With regard to $\|T_j\|_4$ we use (S5.24) to obtain the following conservative but adequate bound,

$$\begin{aligned} \|T_j\|_4 &\lesssim \|\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\|_4 \\ &\leq \|\lambda_j(\widehat{\Sigma})\|_4 + \lambda_j(\Sigma) \\ &\lesssim \lambda_1(\Sigma). \end{aligned} \tag{S5.34}$$

To handle $\|S_j\|_4$, first note that the concavity of h implies that $S_j \leq T_j$ almost surely, and so

$$0 \leq h(\lambda_j(\widehat{\Sigma})) \leq T_j + h(\lambda_j(\Sigma)).$$

In turn, this yields

$$\begin{aligned}\|S_j\|_4 &= \|h(\lambda_j(\widehat{\Sigma})) - h(\lambda_j(\Sigma))\|_4 \\ &\leq \ \|h(\lambda_j(\widehat{\Sigma}))\|_4 \ + \ h(\lambda_j(\Sigma)) \\ &\leq \|T_j\|_4 \ + \ 2h(\lambda_j(\Sigma)) \\ &\lesssim \ \lambda_1(\Sigma). \end{aligned} \tag{S5.35}$$

So, substituting (S5.34) and (S5.35) into (S5.33) gives

$$\|S_j - T_j\|_2 \ \lesssim \ \frac{\log(n)\,\beta_{2q}\,\mathrm{tr}(\Sigma)}{n}. \tag{S5.36}$$

Now we turn to bounding $\|T_j\|_2$ and $\|S_j\|_2$ in (S5.31). Using Lemma S5.3 and the condition (C.0.1) we have

$$\begin{aligned}\|T_j\|_2 \ &\lesssim \ \sqrt{\mathrm{var}\left(\lambda_j(\widehat{\Sigma})\right)} + \left|\mathbb{E}[\lambda_j(\widehat{\Sigma})] - \lambda_j(\Sigma)\right| \\ &\lesssim \ \frac{\lambda_j(\Sigma)\,\beta_2}{\sqrt{n}} + \Big(\frac{\log(n)\,\beta_{2q}\,\mathtt{r}(\Sigma)}{\sqrt{n}}\Big)\frac{\lambda_1(\Sigma)}{\sqrt{n}} \\ &\lesssim \ \frac{\lambda_1(\Sigma)\,\beta_2}{\sqrt{n}}\end{aligned}$$

Likewise, we may bound $\|S_j\|_2$ as

$$\begin{aligned}\|S_j\|_2 \ &\leq \ \|T_j\|_2 + \|S_j - T_j\|_2 \\ &\lesssim \ \frac{\lambda_1(\Sigma)\,\beta_2}{\sqrt{n}} + \Big(\frac{\log(n)\,\beta_{2q}\,\mathtt{r}(\Sigma)}{\sqrt{n}}\Big)\frac{\lambda_1(\Sigma)}{\sqrt{n}} \\ &\lesssim \ \frac{\lambda_1(\Sigma)\,\beta_2}{\sqrt{n}},\end{aligned}$$

which completes the proof of Part (i).

Part (ii): Define random vectors $S^\star, T^\star \in \mathbb{R}^k$ with entries given by

$$S_j^\star = h(\lambda_j(\widehat{\Sigma}^\star)) - h(\lambda_j(\widehat{\Sigma})) \quad \text{and} \quad T_j^\star = h'(\lambda_j(\widehat{\Sigma}))\big(\lambda_j(\widehat{\Sigma}^\star) - \lambda_j(\widehat{\Sigma})\big).$$

As an initial step, it can be shown that there is a constant $c > 0$ not depending on n such that the event

$$\mathbb{P}\bigg(\|S^\star - T^\star\|_\infty \;\geq\; \frac{c \log(n) \beta_{2q} \operatorname{tr}(\Sigma)}{n} \,\bigg|\, X \bigg) \;\leq\; \frac{c}{n^4} \tag{S5.37}$$

holds with probability at least $1 - \frac{c}{n}$. Verifying this is similar to the proof of Lemma S3.1, and can be handled by showing that

$$\|S^\star - T^\star\|_\infty \;\leq\; \frac{c \, \log(n) \, \beta_{2q} \, \operatorname{tr}(\Sigma)}{n}$$

holds with probability at least $1 - \frac{c}{n^5}$. In turn, this can be shown using the condition (C.0.1) and the entrywise Taylor expansion

$$S_j^\star - T_j^\star = \big(\lambda_j(\widehat{\Sigma}^\star) - \lambda_j(\widehat{\Sigma})\big)^2 \int_0^1 (1-t) \, h''\big(\lambda_j(\widehat{\Sigma}) + t(\lambda_j(\widehat{\Sigma}^\star) - \lambda_j(\widehat{\Sigma}))\big) \mathrm{d}t.$$

along with (S3.11) and (S3.12).

To proceed with the rest of the proof, we can bound the left side of (ii) in a manner that is similar to (S5.31),

$$\max_{1 \leq j \leq k} \big| \operatorname{var}(S_j^\star | X) - \operatorname{var}(T_j^\star | X) \big| \leq \max_{1 \leq j \leq k} 2\big((\mathbb{E}[(S_j^\star)^2 | X])^{1/2} + (\mathbb{E}[(T_j^\star)^2 | X])^{1/2} \big) \big(\mathbb{E}\big[(S_j^\star - T_j^\star)^2 \,|\, X \big] \big)^{1/2}. \tag{S5.38}$$

It follows from an argument analogous to (S5.23) and an application of (S5.37) that the event

$$\max_{1 \leq j \leq k} \big(\mathbb{E}\big[(S_j^\star - T_j^\star)^2 \,|\, X \big] \big)^{1/2} \;\leq\; \frac{c}{n} \bigg(\log(n) \, \beta_{2q} \operatorname{tr}(\Sigma) + \max_{1 \leq j \leq k} \Big((\mathbb{E}[(S_j^\star)^4 | X])^{1/4} + (\mathbb{E}[(T_j^\star)^4 | X])^{1/4} \Big) \bigg)$$

holds with probability at least $1 - c/n$. Bounds on the conditional fourth moments can also be derived similarly to the way that the bounds (S5.34) and (S5.35) were. Namely, by using (S2.11) and (S5.29), it can be shown that the events

$$\max_{1 \leq j \leq k} (\mathbb{E}[(T_j^\star)^4 | X])^{1/4} \;\leq\; c \, \lambda_1(\Sigma)$$

and

$$\max_{1 \leq j \leq k} (\mathbb{E}[(S_j^\star)^4 | X])^{1/4} \;\leq\; c \, \lambda_1(\Sigma)$$

each hold with probability at least $1 - \frac{c}{n}$. Hence, the event

$$\max_{1 \le j \le k} \big(\mathbb{E}\big[(S_j^\star - T_j^\star)^2 \,\big|\, X \big] \big)^{1/2} \;\le\; \frac{c \, \log(n) \, \beta_{2q} \, \mathrm{tr}(\Sigma)}{n}$$

holds with probability at least $1 - \frac{c}{n}$.

To address the conditional L_2 norms of $T_j^\star$ and $S_j^\star$ in (S5.38), first note that

$$(\mathbb{E}[(T_j^\star)^2 | X])^{1/2} \;\le\; h'(\lambda_j(\Sigma)) \big(\mathrm{var}(\lambda_j(\widehat{\Sigma}^\star) | X)^{1/2} + \big| \mathbb{E}[\lambda_j(\widehat{\Sigma}^\star) | X] - \lambda_j(\widehat{\Sigma}) \big| \big),$$

In turn, Lemmas S5.3 and S6.4 as well as the condition (C.0.1) imply that the event

$$\begin{aligned} \max_{1 \le j \le k} (\mathbb{E}[(T_j^\star)^2 | X])^{1/2} \;&\le\; \frac{c \, \lambda_j(\Sigma) \, \beta_2}{\sqrt{n}} + \Big(\frac{\log(n) \beta_{2q} \mathtt{r}(\Sigma)}{\sqrt{n}} \Big) \frac{c \, \lambda_1(\Sigma)}{\sqrt{n}} \\ &\le\; \frac{c \, \lambda_1(\Sigma) \, \beta_2}{\sqrt{n}} \end{aligned}$$

holds with probability at least $1 - \frac{c}{n}$. Furthermore, this implies that the event

$$\begin{aligned} \max_{1 \le j \le k} (\mathbb{E}[(S_j^\star)^2 | X])^{1/2} \;&\le\; (\mathbb{E}[(T_j^\star)^2 | X])^{1/2} + \big(\mathbb{E}\big[(S_j^\star - T_j^\star)^2 \,\big|\, X \big] \big)^{1/2} \\ &\le\; \frac{c \, \lambda_1(\Sigma) \, \beta_2}{\sqrt{n}} + \Big(\frac{\log(n) \beta_{2q} \mathtt{r}(\Sigma)}{\sqrt{n}} \Big) \frac{c \, \lambda_1(\Sigma)}{\sqrt{n}} \\ &\le\; \frac{c \, \lambda_1(\Sigma) \, \beta_2}{\sqrt{n}} \end{aligned}$$

holds with probability at least $1 - \frac{c}{n}$. The proof is completed by combining the last several steps with (S5.38).

Part (*iii*): Using several applications of the triangle inequality, we have

$$\begin{aligned} \big| \mathrm{var}\big(h(\lambda_j(\widehat{\Sigma}^\star)) \big| X \big) - \mathrm{var}\big(h(\lambda_j(\widehat{\Sigma})) \big| \;\le\; & \big| \mathrm{var}\big(h(\lambda_j(\widehat{\Sigma}^\star)) \big| X \big) - h'(\lambda_j(\widehat{\Sigma}))^2 \, \mathrm{var}\big(\lambda_j(\widehat{\Sigma}^\star) \big| X \big) \big| \\ & + \; \big| \mathrm{var}\big(h(\lambda_j(\widehat{\Sigma})) - h'(\lambda_j(\Sigma))^2 \, \mathrm{var}\big(\lambda_j(\widehat{\Sigma}) \big) \big| \\ & + \; \big| h'(\lambda_j(\widehat{\Sigma}))^2 \, \mathrm{var}\big(\lambda_j(\widehat{\Sigma}^\star) | X \big) - h'(\lambda_j(\Sigma))^2 \, \mathrm{var}\big(\lambda_j(\widehat{\Sigma}) \big) \big| \end{aligned}$$

The first and second terms on the right side have been handled by Parts (*ii*) and (*i*) respectively. To handle the third term, we may use the triangle inequality to obtain

$$\big| h'(\lambda_j(\widehat{\Sigma}))^2 \, \mathrm{var}\big(\lambda_j(\widehat{\Sigma}^\star) \big| X \big) - h'(\lambda_j(\Sigma))^2 \, \mathrm{var}\big(\lambda_j(\widehat{\Sigma}) \big) \big| \;\le\; \mathrm{M}_j + \mathrm{M}_j',$$

where the two terms on the right are defined as

$$\begin{aligned} \mathrm{M}_j &= h'(\lambda_j(\widehat{\Sigma}))^2 \cdot \big| \operatorname{var}\big(\lambda_j(\widehat{\Sigma}^\star)\big|X\big) - \operatorname{var}\big(\lambda_j(\widehat{\Sigma})\big)\big| \\ \mathrm{M}_j' &= \operatorname{var}\big(\lambda_j(\widehat{\Sigma})\big) \cdot \big|h'(\lambda_j(\widehat{\Sigma}))^2 - h'(\lambda_j(\Sigma))^2\big| \end{aligned}$$

Using (S2.11) and the mean value theorem, it can be shown that the event

$$\max_{1\le j\le k} \big|h'(\lambda_j(\widehat{\Sigma}))^2 - h'(\lambda_j(\Sigma))^2\big| \;\le\; c\sqrt{\frac{\log(n)\,\beta_q\,\mathtt{r}(\Sigma)}{n}}$$

holds with probability at least $1-\frac{c}{n}$. Also, using Lemma S5.3 and the condition (C.0.1), it follows that

$$\max_{1\le j\le k} \operatorname{var}\big(\lambda_j(\widehat{\Sigma})\big) \;\lesssim\; \frac{\lambda_1(\Sigma)^2\,\beta_2^2}{n}.$$

Hence, the event

$$\max_{1\le j\le k} \mathrm{M}_j' \;\le\; \frac{c\,\lambda_1(\Sigma)^2\,\log(n)\,\beta_{2q}^{5/2}\,\mathtt{r}(\Sigma)}{n^{3/2}} \tag{S5.39}$$

holds with probability at least $1-\frac{c}{n}$.

Regarding the quantity M_j, observe that the bound

$$\begin{aligned} \big|\operatorname{var}(\lambda_j(\widehat{\Sigma}^\star)|X) - \operatorname{var}(\lambda_j(\widehat{\Sigma}))\big| \;\le\; & \big|\operatorname{var}(\lambda_j(\widehat{\Sigma}^\star)|X) - \lambda_j(\Sigma)^2\widehat{\Gamma}_{jj}/n\big| \\ + \;& \big|\operatorname{var}(\lambda_j(\widehat{\Sigma})) - \lambda_j(\Sigma)^2\Gamma_{jj}/n\big| \\ + \;& \frac{\lambda_j(\Sigma)^2\,\|\widehat{\Gamma}-\Gamma\|_{\mathrm{op}}}{n} \end{aligned}$$

holds almost surely. Consequently, it follows from Lemmas S5.3 and S6.4, that the event

$$\max_{1\le j\le k} \mathrm{M}_j \;\le\; \frac{c\,\log(n)\,\beta_{2q}^2\,\lambda_1(\Sigma)\,\operatorname{tr}(\Sigma)}{n^{3/2}} \tag{S5.40}$$

holds with probability at least $1-\frac{c}{n}$. Combining (S5.39) and (S5.40) completes the proof of Part (*iii*). □

Lemma S5.5. *Suppose that the conditions of Theorem 12 hold. Then,*

$$\min_{1\le j\le k} \operatorname{var}\big(h(\lambda_j(\widehat{\Sigma}))\big) \;\gtrsim\; \tfrac{\lambda_1(\Sigma)^2}{n},$$

and there is a constant $c \ge 1$ not depending on n such that the event

$$\min_{1\le j\le k} \operatorname{var}\big(h(\lambda_j(\widehat{\Sigma}^\star))\big|X\big) \;\ge\; \tfrac{\lambda_1(\Sigma)^2}{cn}$$

holds with probability at least $1 - \frac{c}{n}$.

Proof. For any $j = 1, \dots, k$, and any $t > 0$, Chebyshev's inequality and the triangle inequality give

$$\begin{aligned}\operatorname{var}\big(h(\lambda_j(\widehat{\Sigma}))\big) \;&\ge\; t^2\,\mathbb{P}\Big(\big|h(\lambda_j(\widehat{\Sigma})) - \mathbb{E}\big[h(\lambda_j(\widehat{\Sigma}))\big]\big| \ge t\Big)\\ &\ge t^2\,\mathbb{P}\Big(\sqrt{n}\big|h(\lambda_j(\widehat{\Sigma})) - h(\lambda_j(\Sigma))\big| \ge \sqrt{n}\big(t + \big|\mathbb{E}\big[h(\lambda_j(\widehat{\Sigma}))\big] - h(\lambda_j(\Sigma))\big|\big)\Big).\end{aligned} \tag{S5.41}$$

Next, let S_j and T_j be as defined in (S5.30). Applying Part (i) of Lemma S5.3 along with (S5.36) shows that the bounds

$$\begin{aligned}\big|\mathbb{E}\big[h(\lambda_j(\widehat{\Sigma}))\big] - h(\lambda_j(\Sigma))\big| \;&\le\; |\mathbb{E}[T_j]| + |\mathbb{E}[S_j] - \mathbb{E}[T_j]|\\ &\lesssim\; |h'(\lambda_j(\Sigma))|\,\big|\mathbb{E}\big[\lambda_j(\widehat{\Sigma}) - \lambda_j(\Sigma)\big]\big| + \|S_j - T_j\|_2\\ &\lesssim \frac{\log(n)\,\beta_{2q}\,\operatorname{tr}(\Sigma)}{n},\end{aligned}$$

hold for every $j = 1, \dots, k$. Hence, by taking $t = \lambda_1(\Sigma)/\sqrt{n}$ in (S5.41), and using Lemma S5.1 with the condition (C.0.1), there is a constant $c > 0$ not depending on n such that

$$\begin{aligned}\operatorname{var}\big(h(\lambda_j(\widehat{\Sigma}))\big) \;&\ge\; t^2\,\mathbb{P}\Big(\sqrt{n}\big|h(\lambda_j(\widehat{\Sigma})) - h(\lambda_j(\Sigma))\big| \ge c\Big)\\ &\gtrsim\; \frac{\lambda_1(\Sigma)^2}{n}\left\{\mathbb{P}\Big(h'(\lambda_j(\Sigma))\,\zeta_j \ge c\Big) - \frac{\log(n)\,\beta_{2q}^3\,\mathtt{r}(\Sigma)}{n^{1/2}}\right\}\\ &\gtrsim\; \frac{\lambda_1(\Sigma)^2}{n},\end{aligned}$$

for every $j = 1, \dots, k$. Lastly, the proof for the corresponding lower bound on $\operatorname{var}(h(\lambda_j(\widehat{\Sigma}^\star))|X)$ is analogous and so the details are omitted. □

S6 Proof of Technical Lemmas

Lemma S6.1. *Suppose that Assumption 6 holds and let $q \geq 5\log(kn)$. Then,*

$$\big\| \|\tilde{\Sigma}[1,2]\|_{\mathrm{op}}^2 \big\|_q \;\lesssim\; \frac{q\,\beta_{2q}\,\lambda_1(\Sigma)\,\mathrm{tr}(\Sigma)}{n^{1-3/(2q)}}.$$

Proof. We will need two auxiliary matrices to extract $\tilde{\Sigma}[1,2]$ from $\tilde{\Sigma}$. Define matrices $\Pi_1 \in \mathbb{R}^{k\times p}$ and $\Pi_2 \in \mathbb{R}^{p\times(p-k)}$ according to

$$\Pi_1 \;=\; \begin{bmatrix} I_k & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \Pi_2 \;=\; \begin{bmatrix} \mathbf{0} \\ I_{p-k} \end{bmatrix},$$

which allow us to write $\tilde{\Sigma}[1,2] = \Pi_1\tilde{\Sigma}\Pi_2$. Next, let $\mathbb{B}^k$ and $\mathbb{B}^{p-k}$ denote the unit ℓ_2-balls in $\mathbb{R}^k$ and $\mathbb{R}^{p-k}$, and let $\mathbb{T} = \mathbb{B}^k \times \mathbb{B}^{p-k}$. With this notation in hand, it follows that

$$\begin{aligned} \|\tilde{\Sigma}[1,2]\|_{\mathrm{op}} \;&=\; \sup_{(u,v)\in\mathbb{T}} u^\top \Pi_1\tilde{\Sigma}\Pi_2 v \\ &=\; \sup_{(u,v)\in\mathbb{T}} (U\Lambda^{1/2}\Pi_1^\top u)^\top \bigg(\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top\bigg)(U\Lambda^{1/2}\Pi_2 v). \end{aligned} \tag{S6.1}$$

Observe that the vectors $U\Lambda^{1/2}\Pi_1^\top u$ and $U\Lambda^{1/2}\Pi_2 v$ both lie in the ellipsoid $\mathcal{E} := U\Lambda^{1/2}(\mathbb{B}^p)$, and are orthogonal. Therefore,

$$\|\tilde{\Sigma}[1,2]\|_{\mathrm{op}} \;\leq\; \sup_{\substack{(\mathsf{u},\mathsf{v})\in\mathcal{E}\times\mathcal{E} \\ \langle \mathsf{u},\mathsf{v}\rangle = 0}} \frac{1}{n}\sum_{i=1}^n \langle \mathsf{u}, Z_i\rangle\langle \mathsf{v}, Z_i\rangle. \tag{S6.2}$$

It will be convenient to write the summands in terms of the matrix $\mathcal{Q}(\mathsf{u},\mathsf{v}) := \frac{1}{2}(\mathsf{u}\mathsf{v}^\top + \mathsf{v}\mathsf{u}^\top)$, namely

$$\langle \mathsf{u}, Z_i\rangle\langle \mathsf{v}, Z_i\rangle = Z_i^\top \mathcal{Q}(\mathsf{u},\mathsf{v}) Z_i.$$

Under this definition, it can be checked that if u and v are any pair of orthogonal vectors, then $\mathbb{E}[Z_i^\top \mathcal{Q}(\mathsf{u},\mathsf{v})Z_i] = 0$. Hence, if we subtract $\mathbb{E}[Z_i^\top \mathcal{Q}(\mathsf{u},\mathsf{v})Z_i]$ from the ith term in (S6.2) for each $i = 1,\ldots,n$, and subsequently drop the constraint $\langle \mathsf{u},\mathsf{v}\rangle = 0$ from the supremum, then we obtain the bound

$$\|\tilde{\Sigma}[1,2]\|_{\mathrm{op}} \;\leq\; \sup_{(\mathsf{u},\mathsf{v})\in\mathcal{E}\times\mathcal{E}} \bigg(\frac{1}{n}\sum_{i=1}^n Z_i^\top \mathcal{Q}(\mathsf{u},\mathsf{v})Z_i - \mathbb{E}[Z_i^\top \mathcal{Q}(\mathsf{u},\mathsf{v})Z_i]\bigg). \tag{S6.3}$$

Next, for a given pair of vectors $\mathsf{u}, \mathsf{v} \in \mathcal{E}$, define two associated vectors

$$\begin{aligned} \mathsf{w} &= \mathsf{u}/2 + \mathsf{v}/2, \\ \tilde{\mathsf{w}} &= \mathsf{u}/2 - \mathsf{v}/2, \end{aligned}$$

which satisfy the algebraic relation

$$\mathcal{Q}(\mathsf{u}, \mathsf{v}) = \mathsf{w}\mathsf{w}^\top - \tilde{\mathsf{w}}\tilde{\mathsf{w}}^\top.$$

To proceed, define the matrix $A \in \mathbb{R}^{p \times 2p}$ as the column concatenation $A = \Big(U\Lambda^{1/2}, U\Lambda^{1/2}\Big)$, and note that both w and $\tilde{\mathsf{w}}$ lie in the ellipsoid $\mathcal{E}' := A(\mathbb{B}^{2p})$. As a result, if we write $\xi_i = A^\top Z_i$, then we have

$$\begin{aligned} \|\tilde{\Sigma}[1,2]\|_{\text{op}} &\leq \sup_{(\mathsf{w},\tilde{\mathsf{w}}) \in \mathcal{E}' \times \mathcal{E}'} \left(\left| \frac{1}{n} \sum_{i=1}^n \langle Z_i, \mathsf{w} \rangle^2 - \mathbb{E}[\langle Z_i, \mathsf{w} \rangle^2] \right| + \left| \frac{1}{n} \sum_{i=1}^n \langle Z_i, \tilde{\mathsf{w}} \rangle^2 - \mathbb{E}[\langle Z_i, \tilde{\mathsf{w}} \rangle^2] \right| \right) \\ &\leq \sup_{w \in \mathbb{B}^{2p}} \left| \frac{2}{n} \sum_{i=1}^n \langle A^\top Z_i, w \rangle^2 - \mathbb{E}[\langle A^\top Z_i, w \rangle^2] \right| \\ &= \left\| \frac{2}{n} \sum_{i=1}^n \xi_i \xi_i^\top - \mathbb{E}[\xi_1 \xi_1^\top] \right\|_{\text{op}}. \end{aligned} \tag{S6.4}$$

Next, Lemma S7.3 implies that

$$\left\| \left\| \frac{1}{n} \sum_{i=1}^n \xi_i \xi_i^\top - \mathbb{E}[\xi_i \xi_i^\top] \right\|_{\text{op}} \right\|_q \leq c \left(\sqrt{\frac{q}{n^{1-3/q}}} \, \|\mathbb{E}[\xi_1 \xi_1^\top]\|_{\text{op}}^{1/2} \left(\mathbb{E}\|\xi_1\|_2^{2q} \right)^{\frac{1}{2q}} \right) \bigvee \left(\frac{q}{n^{1-3/q}} \left(\mathbb{E}\|\xi_1\|_2^{2q} \right)^{1/q} \right). \tag{S6.5}$$

We can further compute

$$\begin{aligned} \big\|\mathbb{E}[\xi_1 \xi_1^\top]\big\|_{\text{op}} &= \big\|A^\top \mathbb{E}[Z_1 Z_1^\top] A\big\|_{\text{op}} \\ &= \|A\|_{\text{op}}^2 \\ &\lesssim \lambda_1(\Sigma), \end{aligned} \tag{S6.6}$$

and

$$\begin{aligned} \left(\mathbb{E}\|\xi_1\|_2^{2q} \right)^{1/q} &= 2 \big\| Z_1^\top U \Lambda U^\top Z_1 \big\|_q \\ &= 2 \left\| \sum_{j=1}^p \lambda_j(\Sigma) \langle u_j, Z_1 \rangle^2 \right\|_q \\ &\lesssim \operatorname{tr}(\Sigma) \beta_q. \end{aligned} \tag{S6.7}$$

To finish, we use the relation $\big\|\|\tilde{\Sigma}_{12}\|_{\rm op}^2\big\|_q = \big\|\|\tilde{\Sigma}_{12}\|_{\rm op}\big\|_{2q}^2$. Specifically, the previous two bounds can be substituted into (S6.4) and (S6.5) while replacing q with $2q$ and using the condition (C.0.1). □

Lemma S6.2. *Suppose that Assumption 6 holds and let $q \geq 5\log(kn)$. Then, there is a constant $c > 0$ not depending on n such that the event*

$$\Big(\mathbb{E}\big[\|\tilde{\Sigma}^\star[1,2]\|_{\rm op}^{2q} \,\big|\, X\big]\Big)^{1/q} \;\leq\; \frac{c\,q\,\beta_{2q}\,\lambda_1(\Sigma)\;\mathrm{tr}(\Sigma)}{n^{1-3/(2q)}}$$

holds with probability at least $1 - ce^{-q}$.

Proof. We will follow the same notation that was used in the proof of Lemma S6.1. Repeating the argument from that proof up to (S6.3) and using $2q$ in place of q we have

$$\Big(\mathbb{E}\big[\|\tilde{\Sigma}^\star[1,2]\|_{\rm op}^{2q}\big|X\big]\Big)^{\frac{1}{2q}} \;\leq\; \left(\mathbb{E}\bigg[\bigg(\sup_{(\mathsf{u},\mathsf{v})\in\mathcal{E}\times\mathcal{E}}\frac{1}{n}\sum_{i=1}^n (Z_i^\star)^\top\mathcal{Q}(\mathsf{u},\mathsf{v})Z_i^\star - \mathbb{E}[Z_i^\top\mathcal{Q}(\mathsf{u},\mathsf{v})Z_i]\bigg)^{2q}\bigg|X\bigg]\right)^{\frac{1}{2q}}$$

Next, observe that

$$\mathbb{E}\big[(Z_1^\star)^\top\mathcal{Q}(\mathsf{u},\mathsf{v})Z_1^\star \,\big|\, X\big] \;=\; \frac{1}{n}\sum_{i=1}^n Z_i^\top\mathcal{Q}(\mathsf{u},\mathsf{v})Z_i$$

and so the triangle inequality for the conditional L_{2q} norm gives

$$\begin{aligned}\Big(\mathbb{E}\big[\|\tilde{\Sigma}^\star[1,2]\|_{\rm op}^{2q}\big|X\big]\Big)^{\frac{1}{2q}} \;\leq\; &\left(\mathbb{E}\bigg[\bigg(\sup_{(\mathsf{u},\mathsf{v})\in\mathcal{E}\times\mathcal{E}}\frac{1}{n}\sum_{i=1}^n (Z_i^\star)^\top\mathcal{Q}(\mathsf{u},\mathsf{v})Z_i^\star - \mathbb{E}\big[(Z_i^\star)^\top\mathcal{Q}(\mathsf{u},\mathsf{v})Z_i^\star \,\big|\, X\big]\bigg)^{2q}\bigg|X\bigg]\right)^{\frac{1}{2q}} \\ &+ \sup_{(\mathsf{u},\mathsf{v})\in\mathcal{E}\times\mathcal{E}}\bigg(\frac{1}{n}\sum_{i=1}^n Z_i^\top\mathcal{Q}(\mathsf{u},\mathsf{v})Z_i - \mathbb{E}[Z_i^\top\mathcal{Q}(\mathsf{u},\mathsf{v})Z_i]\bigg).\end{aligned} \tag{S6.8}$$

With regard to the second term in the last bound, the proof of Lemma S6.1 shows (via Chebyshev's inequality and condition (C.0.1)) that the event

$$\sup_{(\mathsf{u},\mathsf{v})\in\mathcal{E}\times\mathcal{E}}\bigg(\frac{1}{n}\sum_{i=1}^n Z_i^\top\mathcal{Q}(\mathsf{u},\mathsf{v})Z_i - \mathbb{E}[Z_1^\top\mathcal{Q}(\mathsf{u},\mathsf{v})Z_1]\bigg) \;\leq\; c\sqrt{\frac{q\beta_q\lambda_1(\Sigma)\,\mathrm{tr}(\Sigma)}{n^{1-3/q}}}. \tag{S6.9}$$

holds with probability at least $1 - ce^{-q}$, for some constant $c > 0$ not depending on n. Therefore, the proof of the current lemma is complete once we derive a similar bound for the first term on

the right side of (S6.8).

Let $\xi_i^\star = A^\top Z_i^\star$, with A as defined in the proof of Lemma S6.1. Then, by following the argument leading up to (S6.4) and applying Lemma S7.3, there is a constant $c > 0$ not depending on n such that

$$\begin{aligned}
&\Big(\mathbb{E}\Big[\Big(\sup_{(\mathsf{u},\mathsf{v})\in\mathcal{E}\times\mathcal{E}}\frac{1}{n}\sum_{i=1}^n (Z_i^\star)^\top \mathcal{Q}(\mathsf{u},\mathsf{v})Z_i^\star - \mathbb{E}\big[(Z_1^\star)^\top \mathcal{Q}(\mathsf{u},\mathsf{v})Z_1^\star \,\big|\, X\big]\Big)^{2q}\,\Big|\, X\Big]\Big)^{\frac{1}{2q}} \\
&\qquad \le\; c\left(\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i=1}^n \xi_i^\star(\xi_i^\star)^\top - \mathbb{E}\big[\xi_1^\star(\xi_1^\star)^\top \,\big|\, X\big]\Big\|_{\mathrm{op}}^{2q}\,\Big|\, X\Big]\right)^{\frac{1}{2q}} \\
&\qquad \le\; c\left(\sqrt{\frac{2q}{n^{1-\frac{3}{2q}}}}\cdot\big(\big\|\mathbb{E}[\xi_1^\star(\xi_1^\star)^\top \,|\, X]\big\|_{\mathrm{op}}\big)^{1/2}\cdot\big(\mathbb{E}\big[\|\xi_1^\star\|_2^{4q}\,\big|\, X\big]\big)^{\frac{1}{4q}}\right)\bigvee\left(\frac{q}{n^{1-\frac{3}{2q}}}\cdot\big(\mathbb{E}\big[\|\xi_1^\star\|_2^{4q}\,\big|\, X\big]\big)^{\frac{1}{2q}}\right).
\end{aligned} \tag{S6.10}$$

Next, the triangle inequality implies

$$\begin{aligned}
\big\|\mathbb{E}\big[\xi_1^\star(\xi_1^\star)^\top \,\big|\, X\big]\big\|_{\mathrm{op}} \;&=\; \Big\|\frac{1}{n}\sum_{i=1}^n (A^\top Z_i)(A^\top Z_i)^\top\Big\|_{\mathrm{op}} \\
&\le\; \Big\|\frac{1}{n}\sum_{i=1}^n (A^\top Z_i)(A^\top Z_i)^\top - A^\top A\Big\|_{\mathrm{op}} + 4\lambda_1(\Sigma),
\end{aligned}$$

where we have used $\|A^\top A\|_{\mathrm{op}} \le 4\lambda_1(\Sigma)$ and $\mathbb{E}[(A^\top Z_1)(A^\top Z_1)^\top] = A^\top A$. By applying Lemma S7.3 to the first term (along with the condition (C.0.1) and Chebyshev's inequality), it follows that the event

$$\big\|\mathbb{E}\big[\xi_1^\star(\xi_1^\star)^\top \,\big|\, X\big]\big\|_{\mathrm{op}} \le c\lambda_1(\Sigma) \tag{S6.11}$$

holds with probability at least $1 - e^{-q}$, for some constant $c > 0$ not depending on n.

It remains to develop an upper bound for $\big(\mathbb{E}\big[\|\xi_1^\star\|_2^{4q}\,\big|\, X\big]\big)^{\frac{1}{2q}}$. According to the definition of $\xi_i^\star$, we have

$$\begin{aligned}
\|\xi_1^\star\|_2^2 \;&=\; (Z_1^\star)^\top AA^\top Z_1^\star \\
&=\; 2(Z_1^\star)^\top \Sigma Z_1^\star,
\end{aligned}$$

and so

$$\begin{aligned}\big(\mathbb{E}\big[\|\xi_1^\star\|_2^{4q}\,\big|\,X\big]\big)^{\frac{1}{2q}} &= 2\bigg(\frac{1}{n}\sum_{i=1}^n (Z_i^\top \Sigma Z_i)^{2q}\bigg)^{\frac{1}{2q}} \\ &= 2S,\end{aligned}$$

where the non-negative random variable S is defined by the last line. For any $t > 0$, Chebyshev's inequality implies

$$\begin{aligned}\mathbb{P}(S \geq e\,t) &\leq \frac{e^{-2q}\,\mathbb{E}[S^{2q}]}{t^{2q}} \\ &= \frac{e^{-2q}\,\|Z_1^\top \Sigma Z_1\|_{2q}^{2q}}{t^{2q}} \\ &= e^{-2q}\left(\frac{1}{t}\bigg\|\sum_{j=1}^p \lambda_j(\Sigma)\langle u_j, Z_1\rangle^2\bigg\|_{2q}\right)^{2q} \\ &\leq e^{-2q}\left(\frac{\beta_{2q}\,\mathrm{tr}(\Sigma)}{t}\right)^{2q}.\end{aligned} \tag{S6.12}$$

Combining the last few steps and using the choice $t = \beta_{2q}\,\mathrm{tr}(\Sigma)$, it follows that there is a constant $c > 0$ not depending on n such that the event

$$\big(\mathbb{E}\big[\|\xi_1^\star\|_2^{4q}\,\big|\,X\big]\big)^{\frac{1}{2q}} \leq c\,\beta_{2q}\;\mathrm{tr}(\Sigma) \tag{S6.13}$$

holds with probability at least $1 - e^{-2q}$. Hence, we may substitute (S6.11) and (S6.13) into (S6.10), and use the condition (C.0.1) to conclude that the event

$$\bigg(\mathbb{E}\bigg[\bigg(\sup_{(\mathsf{u},\mathsf{v})\in\mathcal{E}\times\mathcal{E}}\frac{1}{n}\sum_{i=1}^n (Z_i^\star)^\top \mathcal{Q}(\mathsf{u},\mathsf{v})Z_i^\star - \mathbb{E}\big[(Z_1^\star)^\top \mathcal{Q}(\mathsf{u},\mathsf{v})Z_1^\star\,\big|\,X\big]\bigg)^{2q}\,\bigg|\,X\bigg]\bigg)^{\frac{1}{2q}} \leq c\sqrt{\frac{2q\,\beta_{2q}\,\lambda_1(\Sigma)\;\mathrm{tr}(\Sigma)}{n^{1-\frac{3}{2q}}}} \tag{S6.14}$$

holds with probability at least $1 - ce^{-q}$. Finally, the proof is completed by combining (S6.14) and (S6.9) with (S6.8). □

Lemma S6.3. *Suppose that Assumption 6 holds and let $q \geq 5\log(kn)$. Then, there is a constant $c > 0$ not depending on n, such that the event*

$$\big(\mathbb{E}\big[\|\tilde{\Sigma}^\star - \tilde{\Sigma}\|_{\mathrm{op}}^q\,\big|\,X\big]\big)^{1/q} \leq c\sqrt{\frac{q\,\beta_q\,\lambda_1(\Sigma)\;\mathrm{tr}(\Sigma)}{n^{1-3/q}}}$$

holds with probability at least $1 - ce^{-q}$.

Proof. Letting $\xi_1^\star = \Lambda^{1/2} U^\top Z_1^\star$ and using Lemma S7.3, we have

$$\big(\mathbb{E}\big[\|\tilde{\Sigma}^\star - \tilde{\Sigma}\|_{\text{op}}^q \,\big|\, X\big]\big)^{1/q} \;\leq\; c\cdot\bigg(\sqrt{\frac{q}{n^{1-3/q}}}\,\|\tilde{\Sigma}\|_{\text{op}}^{1/2}\,\big(\mathbb{E}\big[\|\xi_1^\star\|_2^{2q}\,\big|\,X\big]\big)^{\frac{1}{2q}}\bigg) \bigvee \bigg(\frac{q}{n^{1-3/q}}\,\big(\mathbb{E}\big[\|\xi_1^\star\|_2^{2q}\,\big|\,X\big]\big)^{\frac{1}{q}}\bigg). \tag{S6.15}$$

Using (S2.10) in the proof of Lemma S2.1, as well as the condition (C.0.1), it follows that the bound

$$\begin{aligned}
\|\tilde{\Sigma}\|_{\text{op}} \;&\leq\; \|\tilde{\Sigma} - \Lambda\|_{\text{op}} + \|\Lambda\|_{\text{op}} \\
&\leq\; c\,\lambda_1(\Sigma)\sqrt{\frac{q\,\beta_q\,\mathtt{r}(\Sigma)}{n^{1-3/q}}} + \lambda_1(\Sigma) \\
&\leq\; c\,\lambda_1(\Sigma)
\end{aligned} \tag{S6.16}$$

holds with probability at least $1 - e^{-q}$. Also, recall that the argument leading up to (S6.13) implies that the event

$$\big(\mathbb{E}\big[\|\xi_1^\star\|_2^{2q}\,\big|\,X\big]\big)^{\frac{1}{2q}} \;\leq\; c\sqrt{\beta_q\,\text{tr}(\Sigma)} \tag{S6.17}$$

holds with probability at least $1 - e^{-q}$, for some constant $c > 0$ not depending on n. Combining (S6.16) and (S6.17) with (S6.15) and the condition (C.0.1) completes the proof. □

Lemma S6.4. *Suppose that Assumption 6 holds and let* $q \geq 5\log(kn)$. *Then, there is a constant* $c > 0$ *not depending on* n *such that the bound*

$$\|\widehat{\Gamma} - \Gamma\|_{\text{op}} \;\leq\; \frac{c\,q\,\beta_{2q}^2}{n^{1/2}} \tag{S6.18}$$

holds with probability at least $1 - e^{-q}$.

Proof. Let $W_1, \ldots, W_n$ and $\bar{W}$ be as defined at the beginning of Appendix S3. Note that $\Gamma = \mathbb{E}[W_1 W_1^\top]$, and that the definition of $\widehat{\Gamma}$ in (S3.1) gives

$$\widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^n W_i W_i^\top - \bar{W}\bar{W}^\top.$$

We may apply Lemma S7.3 to obtain

$$\begin{aligned}\left(\mathbb{E}\left\|\widehat{\Gamma}-\Gamma\right\|_{\mathrm{op}}^{q}\right)^{1/q} &\leq \left(\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}W_iW_i^{\top}-\mathbb{E}[W_1W_1^{\top}]\right\|_{\mathrm{op}}^{q}\right)^{1/q} + \left(\mathbb{E}\|\bar{W}\bar{W}^{\top}\|_{\mathrm{op}}^{q}\right)^{1/q} \\ &\lesssim \left(\sqrt{\frac{q}{n^{1-3/q}}}\left(\|\mathbb{E}[W_1W_1^{\top}]\|_{\mathrm{op}}\right)^{1/2}\left(\mathbb{E}\|W_1\|_2^{2q}\right)^{1/(2q)}\right)\bigvee\left(\frac{q}{n^{1-3/q}}\left(\mathbb{E}\|W_1\|_2^{2q}\right)^{1/q}\right) + \left(\mathbb{E}\|\bar{W}\|_2^{2q}\right)^{1/q}.\end{aligned} \tag{S6.19}$$

It is straightforward to verify that

$$\|\mathbb{E}[W_1W_1^{\top}]\|_{\mathrm{op}} \leq \mathrm{tr}(\Gamma) \lesssim \beta_2^2.$$

Also, we have

$$\begin{aligned}\left(\mathbb{E}\|W_1\|_2^{2q}\right)^{1/q} &= \left\|\sum_{j=1}^{k}\left(\langle u_j,Z_1\rangle^2-1\right)^2\right\|_q \\ &\leq \sum_{j=1}^{k}\left\|\langle u_j,Z_1\rangle^2-1\right\|_{2q}^2 \\ &\lesssim \beta_{2q}^2.\end{aligned}$$

To handle the term involving $\bar{W}$, observe that the previous bound and Lemma S7.1 lead to

$$\begin{aligned}\left(\mathbb{E}\|\bar{W}\|_2^{2q}\right)^{\frac{1}{2q}} &\lesssim q\left(\left(\mathbb{E}\|\bar{W}\|_2^2\right)^{1/2}+\left(\textstyle\sum_{i=1}^{n}\mathbb{E}\|\frac{1}{n}W_i\|_2^{2q}\right)^{\frac{1}{2q}}\right) \\ &\lesssim q\left(\frac{\left(\mathbb{E}\|W_1\|_2^2\right)^{1/2}}{n^{1/2}}+\frac{\left(\mathbb{E}\|W_1\|_2^{2q}\right)^{\frac{1}{2q}}}{n^{1-\frac{1}{2q}}}\right) \\ &\lesssim q\left(\frac{\beta_2}{n^{1/2}}+\frac{\beta_{2q}}{n^{1-\frac{1}{2q}}}\right) \\ &\lesssim \frac{q\,\beta_{2q}}{n^{1/2}}\end{aligned} \tag{S6.20}$$

Using condition (C.0.1), we may substitute the last few bounds into (S6.19) to obtain

$$\left(\mathbb{E}\left\|\widehat{\Gamma}-\Gamma\right\|_{\mathrm{op}}^{q}\right)^{1/q} \lesssim \frac{q\,\beta_{2q}^2}{n^{1/2}} \tag{S6.21}$$

Combining this result with Chebyshev's inequality completes the proof.

□

For the next lemma, let $Y_1^\star$ be as defined at the beginning of the proof of Lemma S3.3.

Lemma S6.5. *Suppose that Assumption 6 holds and let $q \geq 5\log(kn)$. Then, there is a constant $c > 0$ not depending on n, such that the event*

$$\mathbb{E}\Big[\big\|(\Lambda_k\widehat{\Gamma}\Lambda_k)^{-1/2}Y_1^\star\big\|_2^3\,\Big|\,X\Big] \;\leq\; c\,\beta_{3q}^3.$$

holds with probability at least $1 - ce^{-q}$.

Proof. Note that

$$\begin{aligned}
\big\|(\Lambda_k\widehat{\Gamma}\Lambda_k)^{-1/2}Y_1^\star\big\|_2^2 \;&=\; (W_1^\star - \bar{W})^\top\widehat{\Gamma}^{-1}(W_1^\star - \bar{W}) \\
&\leq\; \frac{(W_1^\star - \bar{W})^\top(W_1^\star - \bar{W})}{\lambda_k(\widehat{\Gamma})} \\
&\leq\; \frac{2}{\lambda_k(\widehat{\Gamma})}\bigg(\sum_{j=1}^k\big(\langle u_j, Z_1^\star\rangle^2 - 1\big)^2 + \bar{W}^\top\bar{W}\bigg).
\end{aligned}$$

This implies

$$\begin{aligned}
\mathbb{E}\Big[\big\|(\Lambda_k\widehat{\Gamma}\Lambda_k)^{-1/2}Y_1^\star\big\|_2^3\Big|X\Big] \;&\leq\; \frac{c}{\lambda_k(\widehat{\Gamma})^{3/2}}\bigg(\frac{1}{n}\sum_{i=1}^n\Big(\sum_{j=1}^k\big(\langle u_j, Z_i\rangle^2 - 1\big)^2\Big)^{3/2} + (\bar{W}^\top\bar{W})^{3/2}\bigg) \\
&=\; \frac{c\,\big(T + (\bar{W}^\top\bar{W})^{3/2}\big)}{\lambda_k(\widehat{\Gamma})^{3/2}},
\end{aligned} \tag{S6.22}$$

where the first line has used the convexity of the function $x \mapsto x^{3/2}$, and the non-negative random variable T is defined in the second line. By the triangle inequality for the L^q norm, we have

$$\begin{aligned}
\|T\|_q \;&\leq\; \bigg(\Big\|\sum_{j=1}^k\big(\langle u_j, Z_1\rangle^2 - 1\big)^2\Big\|_{3q/2}^{3q/2}\bigg)^{1/q} \\
&\lesssim\; \max_{1\leq j\leq k}\Big\|\langle u_j, Z_1\rangle^2 - 1\Big\|_{3q}^3 \\
&\lesssim\; \beta_{3q}^3.
\end{aligned} \tag{S6.23}$$

So, by Chebyshev's inequality, there is a constant $c > 0$ not depending on n such that the event

$$T \;\leq\; c\,\beta_{3q}^3$$

holds with probability at least $1 - e^{-q}$. The bound (S6.20) in the proof of Lemma S6.4 and the condition (C.0.1) also imply that there is a constant $c > 0$ not depending on n such that the bound

$$(\bar{W}^\top \bar{W})^{3/2} \;\leq\; \Big(\frac{c\,q\,\beta_{2q}}{n^{1/2}}\Big)^{3/2} \;\leq\; c$$

holds with probability at least $1 - e^{-q}$.

Now we turn to showing that $\lambda_k\big(\widehat{\Gamma}\big)$ is greater than a positive constant with high probability. Due to Weyl's inequality we have $\lambda_k(\widehat{\Gamma}) \;\geq\; \lambda_k(\Gamma) - \|\widehat{\Gamma} - \Gamma\|_{\text{op}}$. Using Lemma S6.4, Assumption 6.(c), and the condition (C.0.1), it follows that there is a constant $c > 0$ not depending on n such that the bound

$$\lambda_k(\widehat{\Gamma}) \geq c, \tag{S6.24}$$

holds with probability at least $1 - e^{-q}$. Combining the bounds (S6.23) and (S6.24) with (S6.22) completes the proof. □

S7 Background Results

Lemma S7.1 (Theorem 1 in Talagrand (1989))**.** *Let $X_1, \ldots, X_n$ be independent centered random elements of a Banach space with norm $\|\cdot\|$. Then, there is an absolute constant $c > 0$ such that the following inequality holds for any $q \geq 1$,*

$$\Big(\mathbb{E}\Big\|\sum_{i=1}^n X_i\Big\|^q\Big)^{1/q} \;\leq\; \frac{cq}{1+\log(q)}\Big(\mathbb{E}\Big\|\sum_{i=1}^n X_i\Big\| + \Big(\mathbb{E}\max_{1\leq i\leq n}\|X_i\|^q\Big)^{1/q}\Big).$$

Lemma S7.2 (Weilandt's inequality Eaton and Tyler (1991); Wielandt and Meyer (1967))**.** *Consider a real symmetric $p \times p$ matrix*

$$A = \begin{pmatrix} B & C \\ C^\top & D \end{pmatrix},$$

where B is $k \times k$ and D is $(p-k) \times (p-k)$. If $\lambda_k(B) > \lambda_1(D)$, then

$$0 \leq \lambda_j(A) - \lambda_j(B) \leq \frac{\lambda_1(CC^\top)}{\lambda_j(B) - \lambda_1(D)}, \quad j = 1, \ldots, k$$

and

$$0 \leq \lambda_{p-k-i}(D) - \lambda_{p-i}(A) \leq \frac{\lambda_1(CC^\top)}{\lambda_k(B) - \lambda_{p-k-i}(D)}, \quad i = 0, \ldots, p-k-1.$$

Lemma S7.3 (Proposition 1 in Lopes et al. (2023))**.** *Let $\xi, \ldots, \xi_n \in \mathbb{R}^p$ be i.i.d. random vectors, let $q \geq 3$, and define the quantity*

$$r(q) = q \cdot \frac{\left(\mathbb{E}\|\xi_1\|_2^{2q}\right)^{1/q}}{\left\|\mathbb{E}[\xi_1 \xi_1^\top]\right\|_{\text{op}}}.$$

Then, there is an absolute constant $c > 0$ such that

$$\left(\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \xi_i \xi_i^\top - \mathbb{E}[\xi_1 \xi_1^\top]\right\|_{\text{op}}^q\right)^{1/q} \;\leq\; c \cdot \left\|\mathbb{E}[\xi_1 \xi_1^\top]\right\|_{\text{op}} \cdot \left(\sqrt{\frac{r(q)}{n^{1-3/q}}} \bigvee \frac{r(q)}{n^{1-3/q}}\right)$$

The following anti-concentration lemma originates from Nazarov (2003), and was further elucidated in (Chernozhukov et al., 2017b, Theorem 1).

Lemma S7.4 (Nazarov's inequality)**.** *Let $Y = (Y_1, \ldots, Y_p)$ be a centered Gaussian random vector in $\mathbb{R}^p$ and suppose that the parameter $\underline{\sigma}^2 = \min_{1 \leq j \leq p} \mathbb{E}[Y_j^2]$ is positive. Then for every $y \in \mathbb{R}^p$ and $\delta > 0$,*

$$\mathbb{P}(Y \preceq y + \delta 1_k) - \mathbb{P}(Y \preceq y) \;\leq\; \frac{\delta}{\underline{\sigma}}\big(\sqrt{2\log(p)} + 2\big).$$

The following lemma is Bentkus' multivariate Berry-Esseen theorem.

Lemma S7.5 (Theorem 1.1 in (Bentkus, 2003))**.** *Let $V_1, \ldots, V_n$ be i.i.d. random vectors $\mathbb{R}^d$, with zero mean and identity covariance matrix. Furthermore, let ζ be a standard Gaussian vector in $\mathbb{R}^d$, and let $\mathcal{A}$ denote the collection of all Borel convex subsets of $\mathbb{R}^d$. Then, there is an absolute constant $c > 0$ such that*

$$\sup_{A \in \mathcal{A}} \left|\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n V_i \in A\Big) - \mathbb{P}(\zeta \in A)\right| \;\leq\; \frac{c \cdot d^{1/4} \cdot \mathbb{E}[\|V_1\|_2^3]}{n^{1/2}}.$$

For the statement of Lemma S7.6 below, we need to introduce a bit of notation. For any $r > 0$ and set $A \subset \mathbb{R}^p$, define the outer r-neighborhood as $A^r = \big\{x \in \mathbb{R}^p \,|\, d(x, A) \leq r\big\}$, where $d(x, A) = \inf\{\|x - y\| \,|\, y \in A\}$, and $\|\cdot\|$ is any norm on $\mathbb{R}^p$. The corresponding inner r-neighborhood may be defined as $A^{-r} = \big\{x \in A \,|\, B(x, r) \subset A\big\}$, where $B(x, r) = \{y \in \mathbb{R}^p | \|x - y\| \leq r\}$.

Lemma S7.6 (Lemma 7.3 in (Lopes, 2022))**.** *Let $\|\cdot\|$ be any norm on $\mathbb{R}^p$, and let $\zeta, \xi \in \mathbb{R}^p$ be any two random vectors. Then, the following inequality holds for any Borel set $A \subset \mathbb{R}^p$, and any $r > 0$,*

$$|\mathbb{P}(\zeta \in A) - \mathbb{P}(\xi \in A)| \;\leq\; \mathbb{P}\big(\xi \in (A^r \setminus A^{-r})\big) \;+\; \mathbb{P}\big(\|\zeta - \xi\| \geq r\big).$$

The following lemma is a consequence of Pinsker's inequality and the proof of Lemma A.7 in the paper Spokoiny and Zhilova (2015).

Lemma S7.7 (Spokoiny and Zhilova (2015))**.** *Let ζ and $\tilde{\zeta}$ be centered Gaussian vectors in $\mathbb{R}^k$ with respective covariance matrices C and $\tilde{C}$. Also, suppose that C is invertible, and let $B = C^{-1/2}\tilde{C}C^{-1/2} - I_k$. Then, there is an absolute constant $c > 0$ such that*

$$\sup_{t \in \mathbb{R}^k} \Big|\mathbb{P}(\zeta \preceq t) - \mathbb{P}(\tilde{\zeta} \preceq t)\Big| \;\leq\; c\sqrt{k}\|B\|_{\mathrm{op}}.$$

The last background lemma follows from the proof of (Lopes et al., 2020, Lemma D.3), Lemma S7.6, and Nazarov's inequality (Lemma S7.4).

Lemma S7.8. *Let U, V, and R be random vectors in $\mathbb{R}^k$ that satisfy $U = V + R$. Also, let $W = (W_1, \ldots, W_k)$ be a centered Gaussian random vector in $\mathbb{R}^k$ and suppose that the parameter $\underline{\sigma}^2 = \min_{1 \leq j \leq k} \mathbb{E}[W_j^2]$ is positive. Then there is an absolute constant $c > 0$, such that the following bound holds for any $r > 0$,*

$$\sup_{t \in \mathbb{R}^k} \big|\mathbb{P}(U \preceq t) - \mathbb{P}(W \preceq t)\big| \;\leq\; 3 \cdot \sup_{t \in \mathbb{R}^k} \big|\mathbb{P}(V \preceq t) - \mathbb{P}(W \preceq t)\big| + \frac{cr\sqrt{\log(k)}}{\underline{\sigma}} + \mathbb{P}(\|R\|_\infty \geq r).$$

S8 Additional Numerical Results

Nominal value of 95% in model (ii). The following four figures are presented in the same manner as in the main text for a 95% nominal value, except that they are based on simulation model (ii).

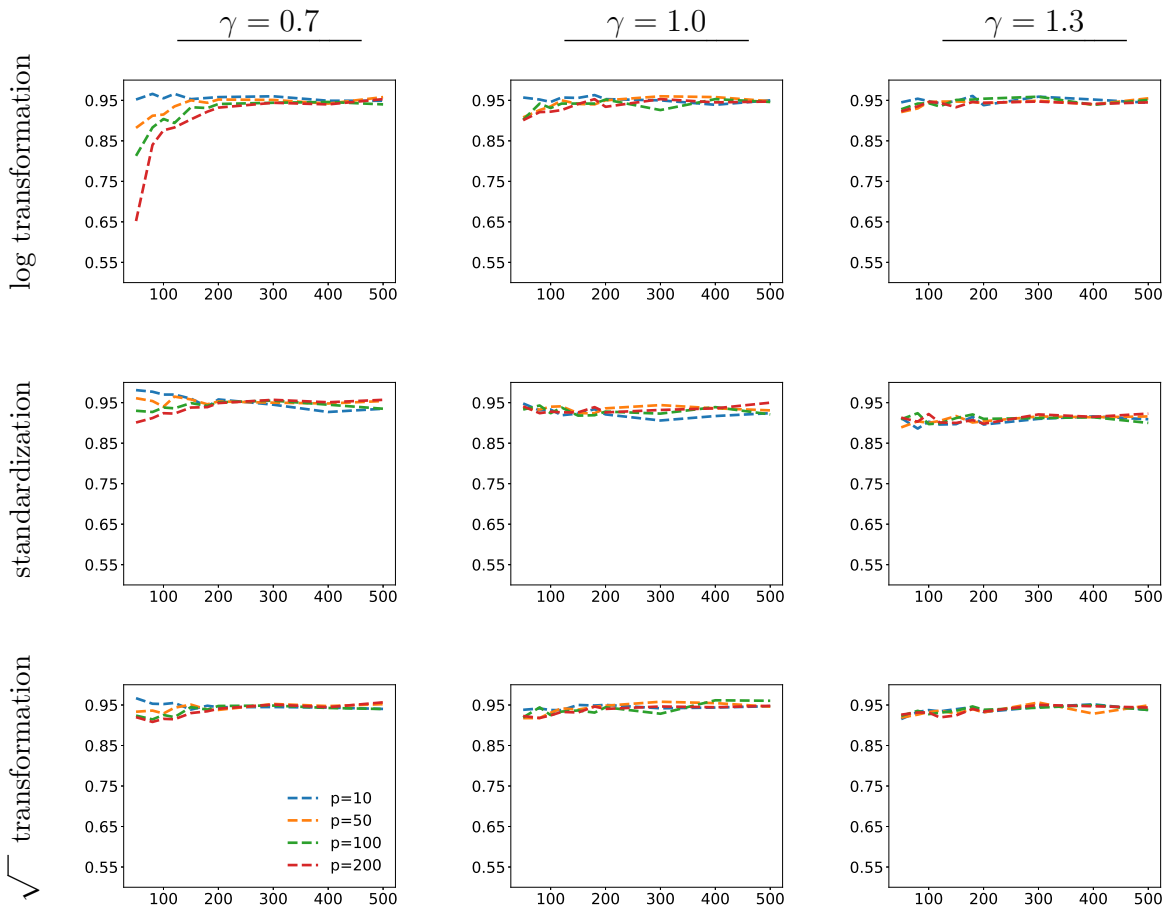


Figure C.1: (Simultaneous coverage probability versus n in simulation model (ii) with a polynomial decay profile). The plotting scheme is the same as described in the caption of Figure 4.1 in the main text.

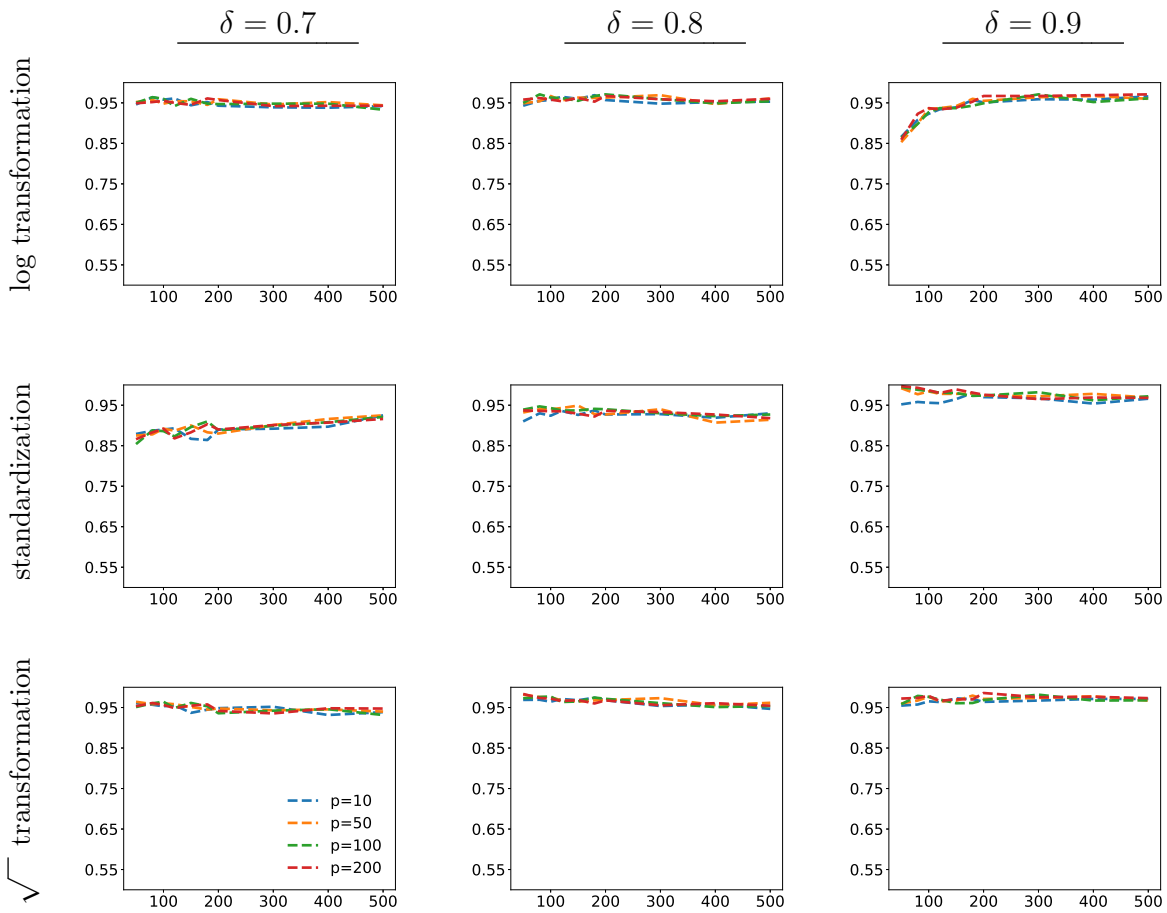


Figure C.2: (Simultaneous coverage probability versus n in simulation model (ii) with an exponential decay profile). The plotting scheme is the same as described in the caption of Figure 4.2 in the main text.

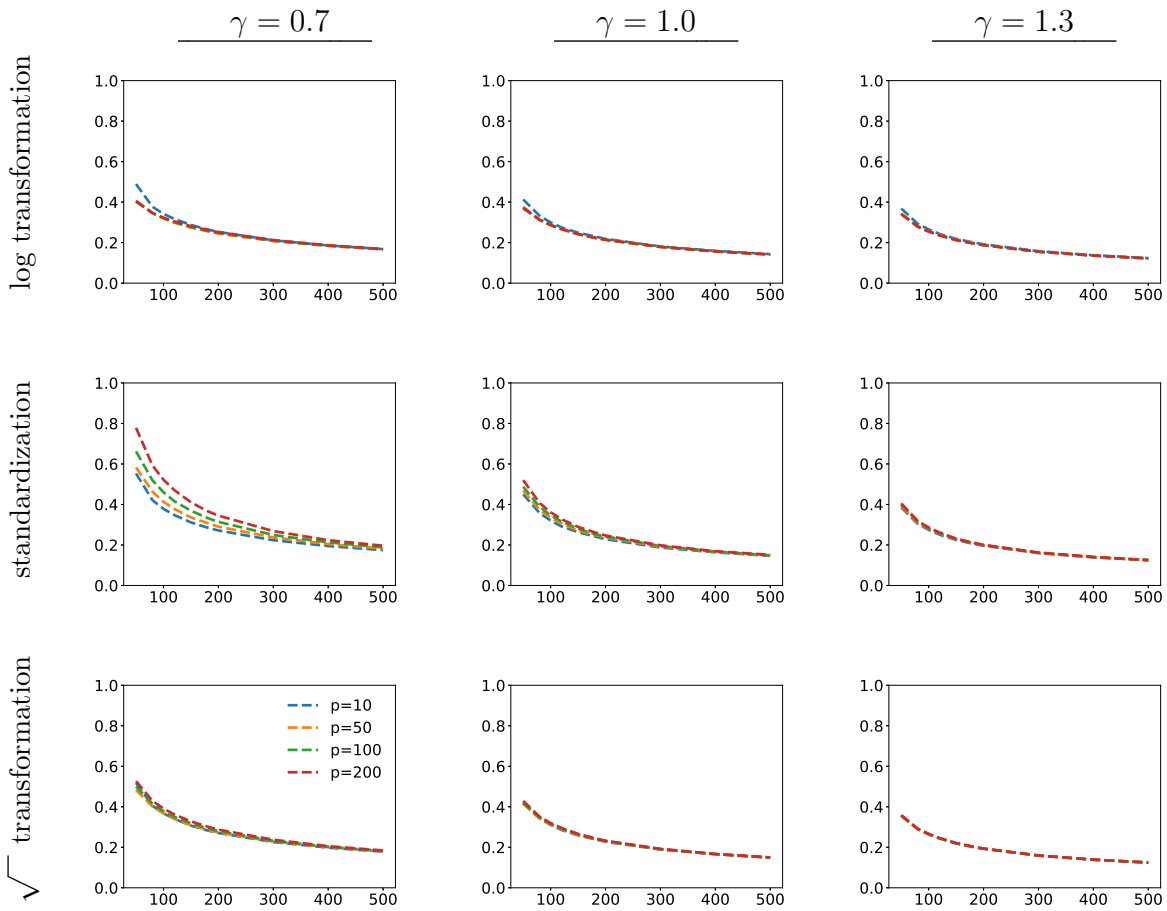


Figure C.3: (Average width versus n in simulation model (ii) with a polynomial decay profile). The plotting scheme is the same as described in the caption of Figure 4.3 in the main text.

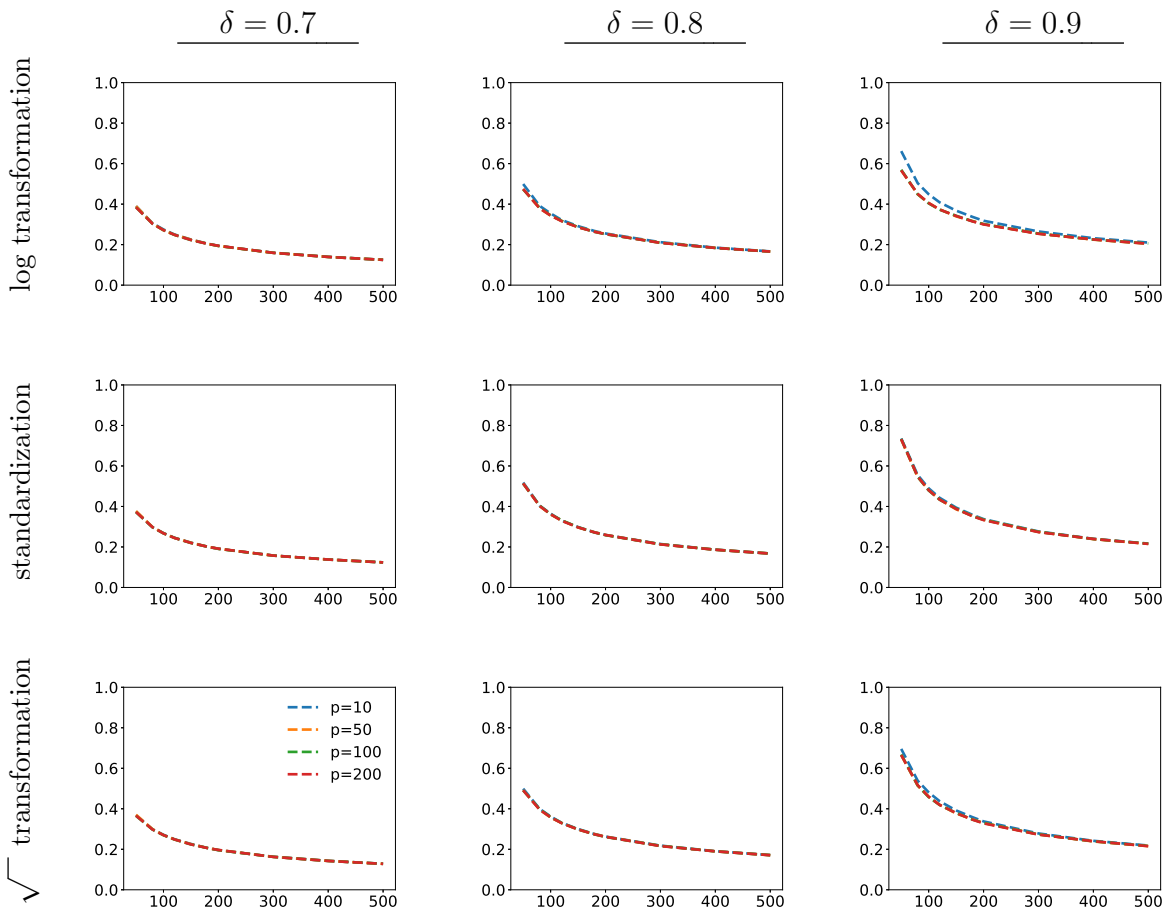


Figure C.4: (Average width versus n in simulation model (ii) with an exponential decay profile). The plotting scheme is the same as described in the caption of Figure 4.4 in the main text.

Nominal value of 90% in models (i) and (ii). The following eight figures are presented in the same manner as in the main text, except that they use a nominal value of 90%, and are based on both models (i) and (ii).

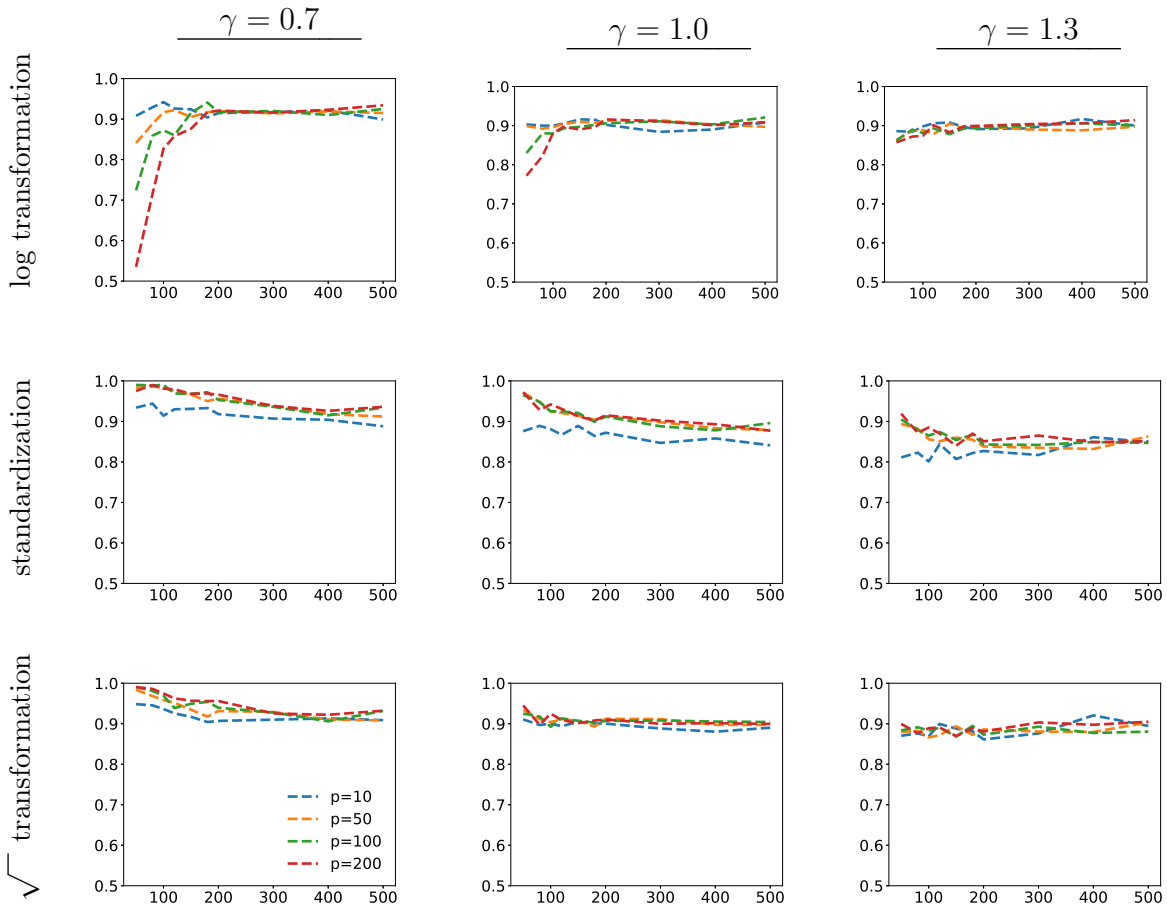


Figure C.5: (Simultaneous coverage probability versus n in simulation model (i) with a polynomial decay profile). In each panel, the y-axis measures $\mathbb{P}(\cap_{j=1}^{5}\{\lambda_j(\Sigma) \in \widehat{\mathcal{I}}_j\})$ based on a nominal value of 90%, and the x-axis measures n. The colored curves correspond to the different values of p, indicated in the legend. The three rows and three columns correspond to labeled choices of transformations and values of the eigenvalue decay parameter γ.

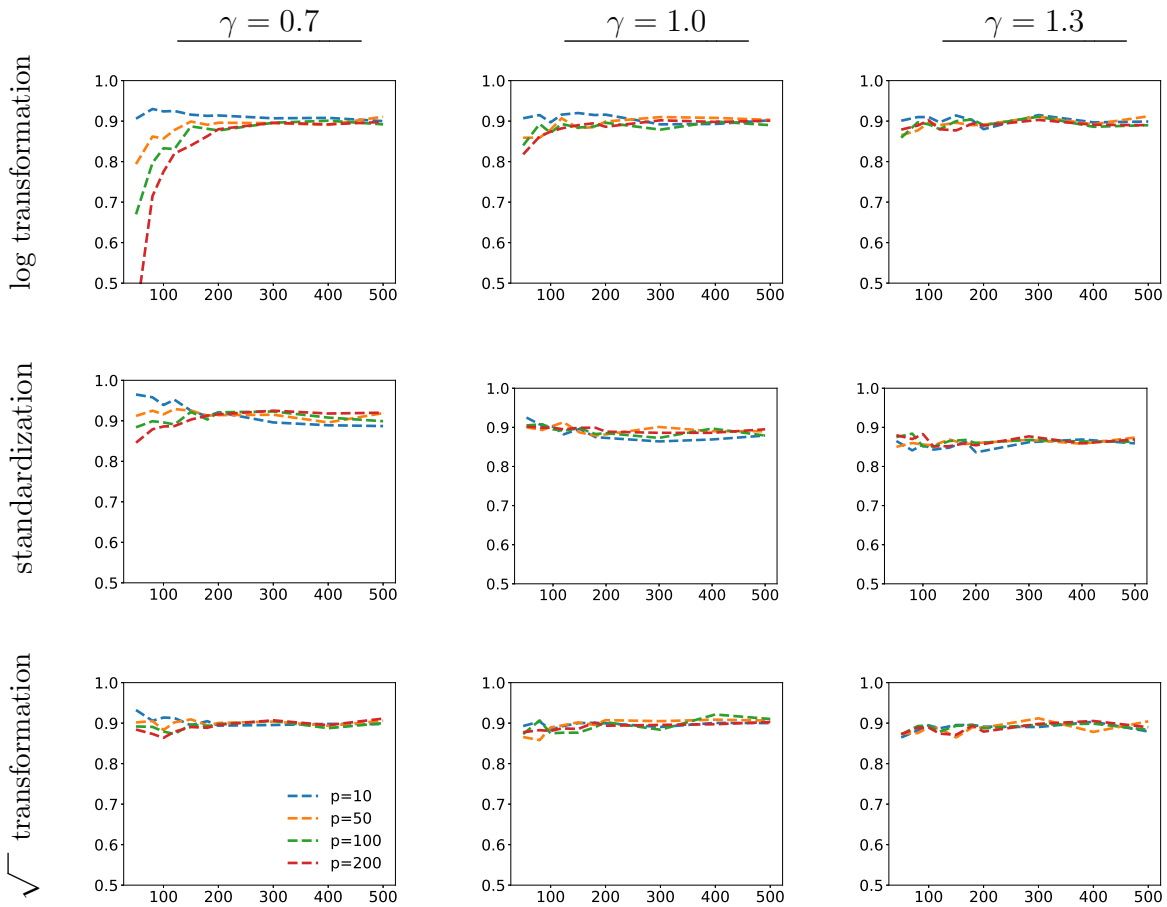


Figure C.6: (Simultaneous coverage probability versus n in simulation model (ii) with a polynomial decay profile). The plotting scheme is the same as described in the caption of Figure C.5 above.

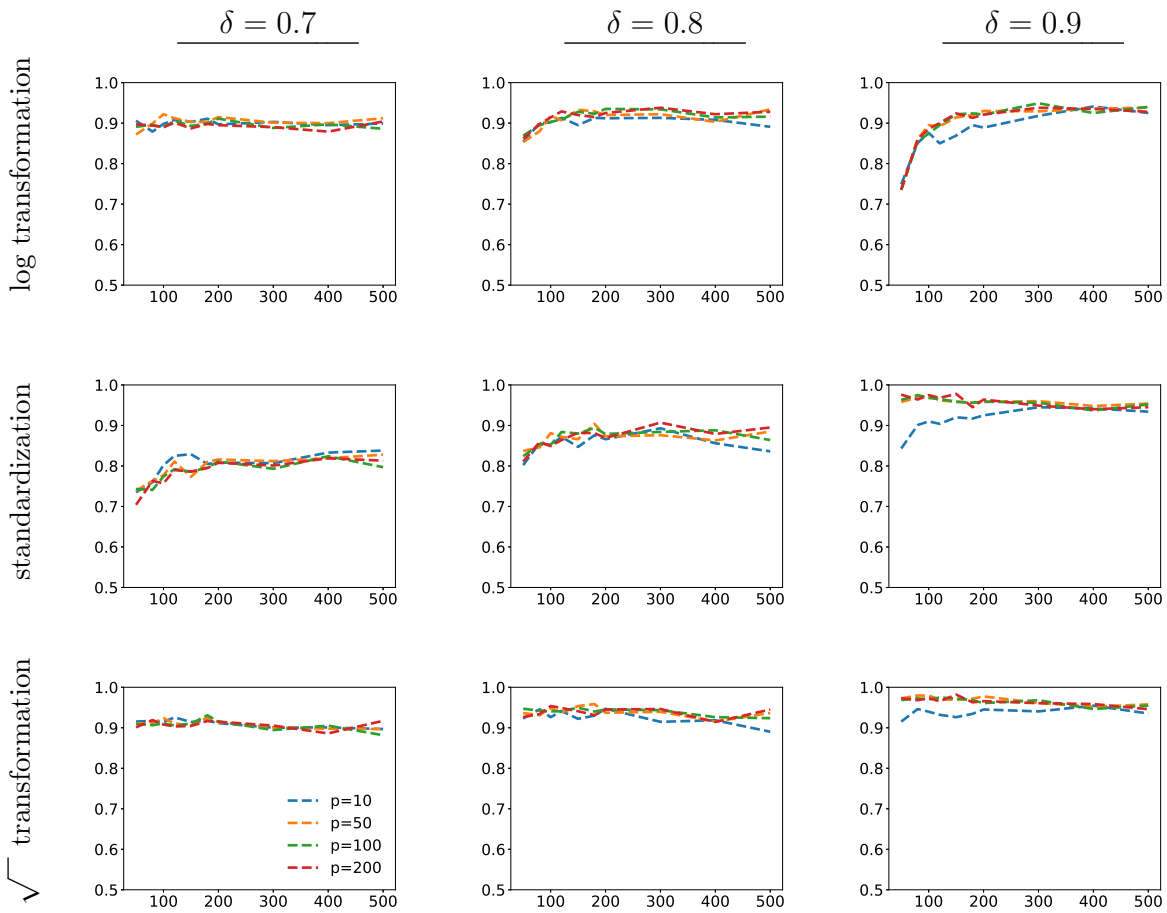


Figure C.7: (Simultaneous coverage probability versus n in simulation model (i) with an exponential decay profile). The plotting scheme is the same as described in the caption of Figure C.5 above, except that the three columns correspond to values of the eigenvalue decay parameter δ.

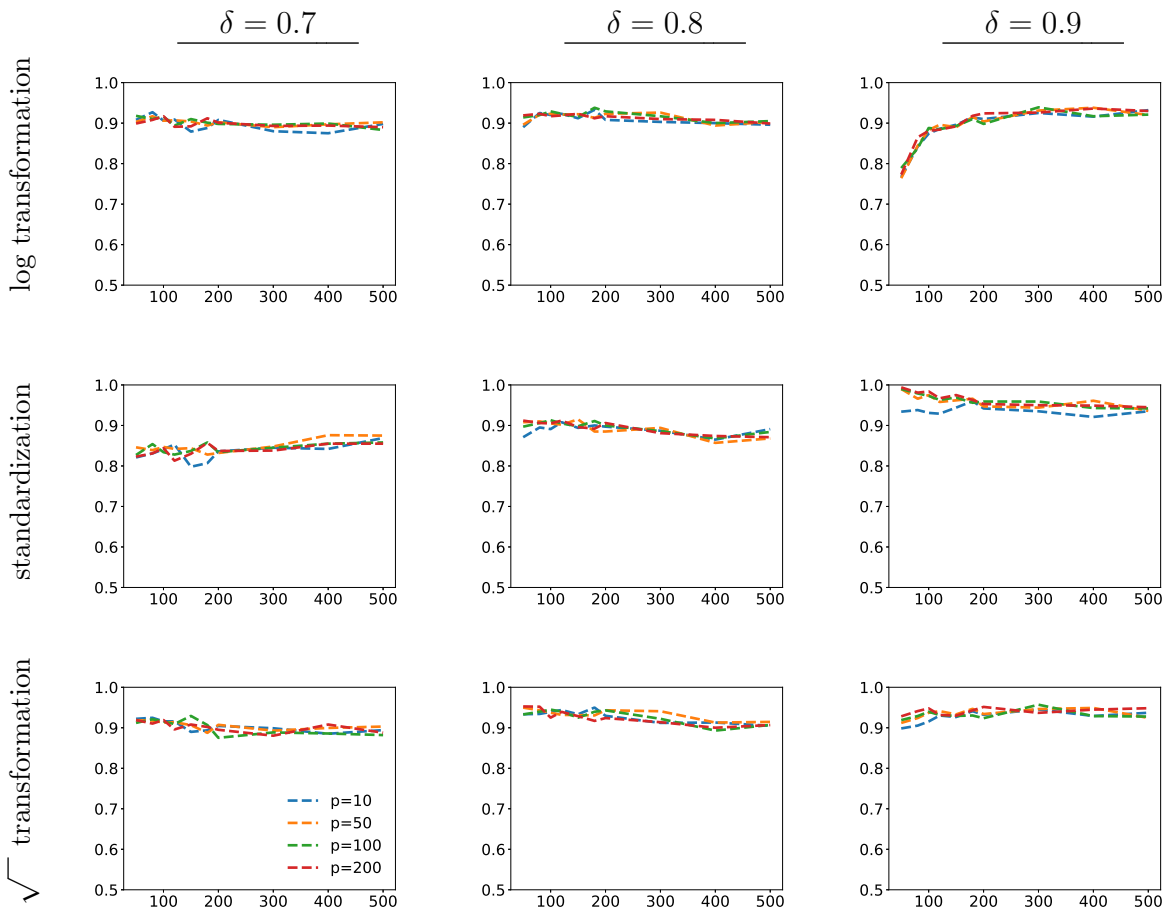


Figure C.8: (Simultaneous coverage probability versus n in simulation model (ii) with an exponential decay profile). The plotting scheme is the same as described in the caption of Figure C.5 above.

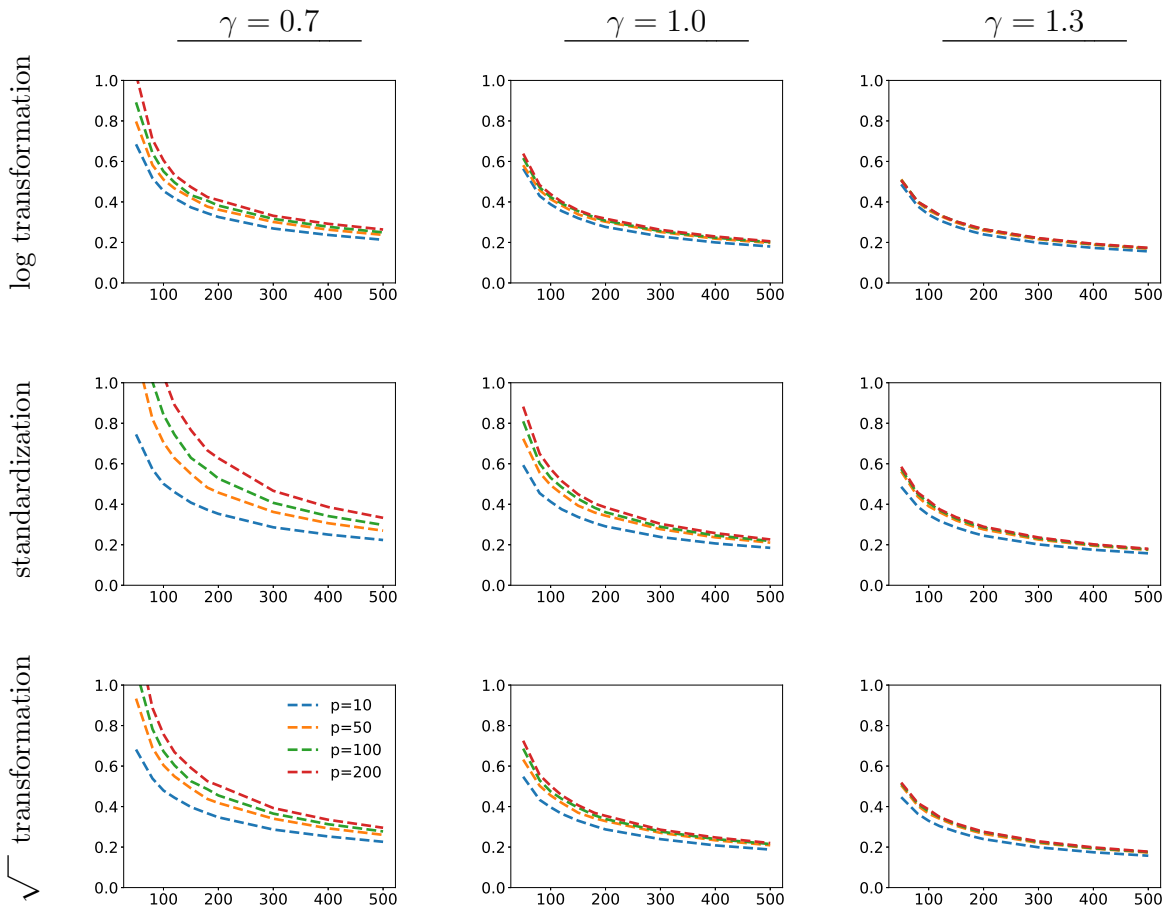


Figure C.9: (Average width versus n in simulation model (i) with a polynomial decay profile). In each of the nine panels, the y-axis measures the average width $\mathbb{E}[|\widehat{\mathcal{I}}_1| + \cdots + |\widehat{\mathcal{I}}_5|]/5$, and the x-axis measures n. The colored curves correspond to the different values of $p = 10, 50, 100, 200$, indicated in the legend. The three rows and three columns correspond to labeled choices of transformations and values of the eigenvalue decay parameter γ.

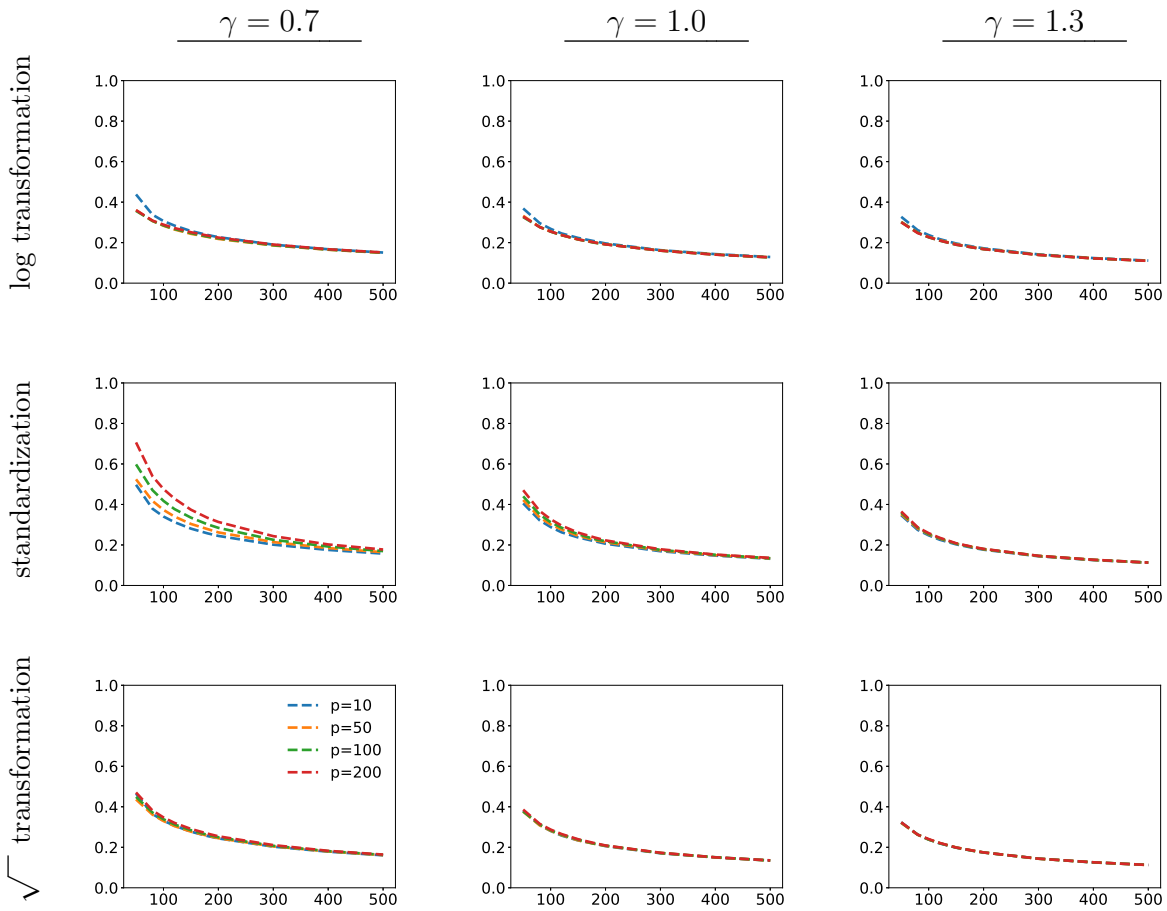


Figure C.10: (Average width versus n in simulation model (ii) with a polynomial decay profile). The plotting scheme is the same as described in the caption of Figure C.9 above.

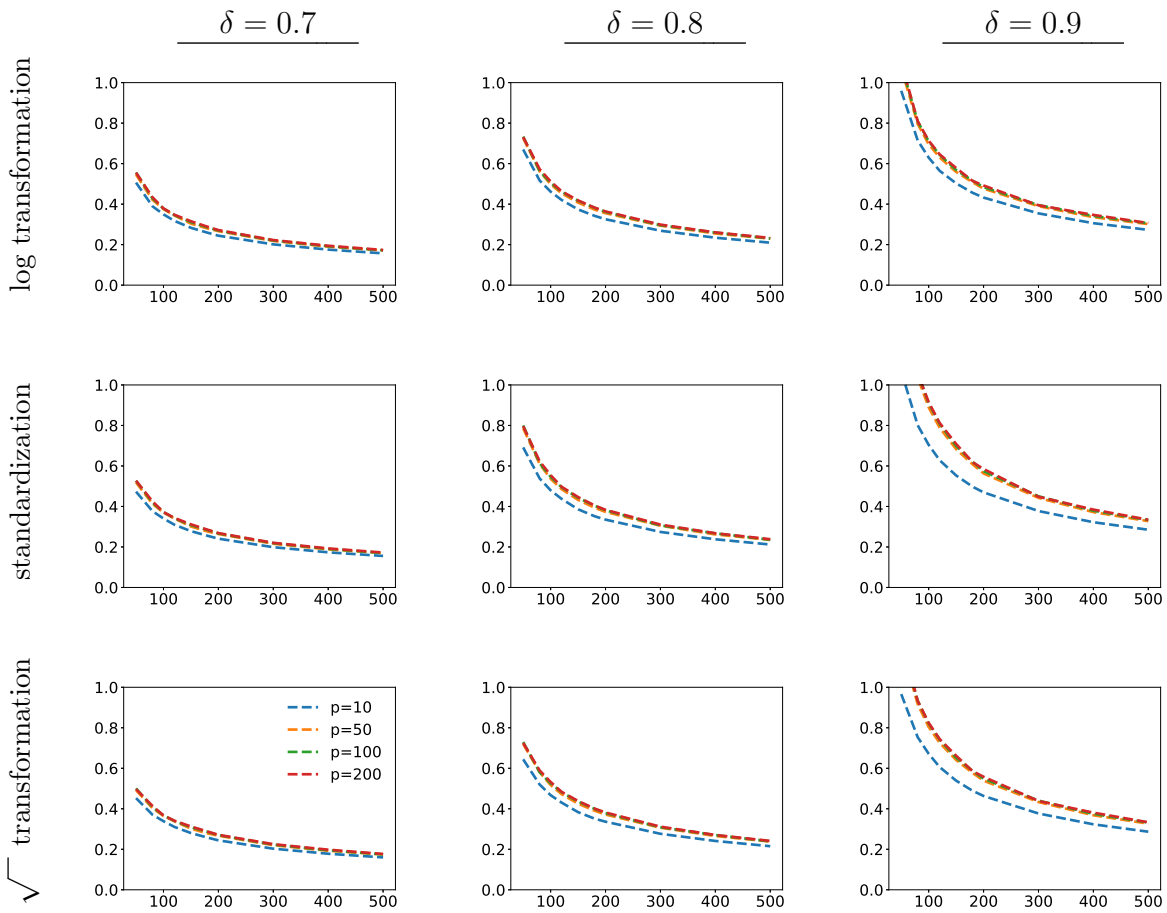


Figure C.11: (Average width versus n in simulation model (i) with an exponential decay profile). The plotting scheme is the same as described in the caption of Figure C.9 above, except that the three columns correspond to values of the eigenvalue decay parameter δ.

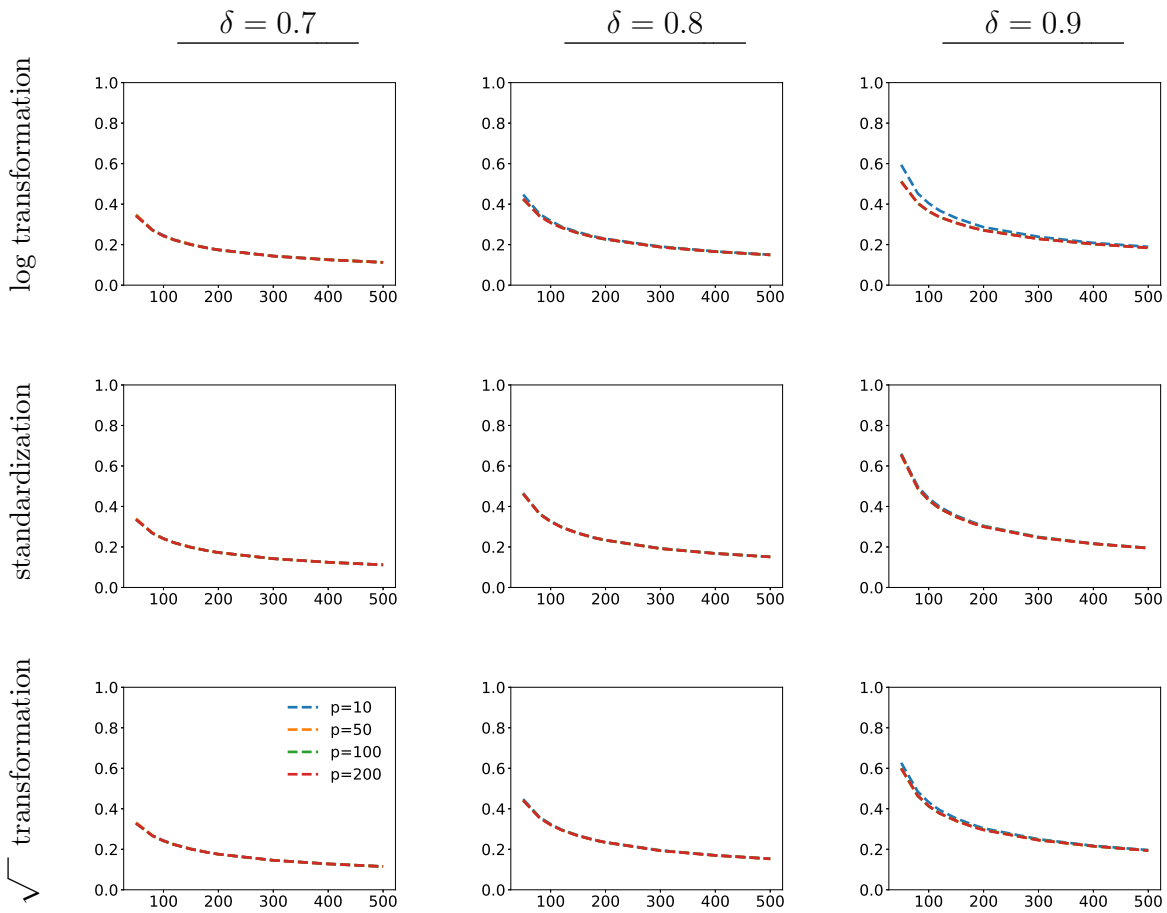


Figure C.12: (Average width versus n in simulation model (ii) with an exponential decay profile). The plotting scheme is the same as described in the caption of Figure C.11 above.

S9 Computational cost

The cost to compute a single bootstrap sample of $\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star) - \boldsymbol{\lambda}_k(\widehat{\Sigma})$ can be broken into two steps. The first step is to sample n points with replacement from the original set of n observations, which has a cost of $O(n\log(n))$. The second step consists of computing the largest k eigenvalues of $\widehat{\Sigma}^\star$. This can be done by computing the largest k singular values of the $n \times p$ matrix of resampled observations, which has a cost of $\mathcal{O}(npk)$ (Halko et al., 2011). (Note that the largest k eigenvalues of $\widehat{\Sigma}$ only need to be computed once, before any resampling is done.) So, the cost to compute B bootstrap samples of $\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star) - \boldsymbol{\lambda}_k(\widehat{\Sigma})$ on a single processor is $\mathcal{O}(B(n\log(n) + npk))$. However, it is common to compute bootstrap samples in a parallel manner, across say m processors, and in this case the cost per processor becomes $\mathcal{O}(\frac{B}{m}(n\log(n) + npk)$. In order to simplify this expression, it is natural to consider a scenario where $B/m = \mathcal{O}(1)$ and $\log(n) = \mathcal{O}(p)$, which leads to a cost per processor that is $\mathcal{O}(npk)$.

S9.1 Empirical computational cost

Table C.1 displays the time, in seconds, to compute a single bootstrap sample of $\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star) - \boldsymbol{\lambda}_k(\widehat{\Sigma})$ for different choices of n and p. Each entry reflects an average over 1000 trials with data generated under simulation model (i). The computations were done on a single Intel Xeon E5-2699v3 processor. Notably, even when $(n, p) = (500, 200)$, the computing time for each bootstrap sample is on the order of just 10^{-3} seconds. Hence, it is possible to generate hundreds bootstrap samples within about 1 second.

	p			
n	10	50	100	200
10	1.9e-04	4.2e-04	8.7e-04	2.6e-03
50	2.0e-04	4.4e-04	9.0e-04	2.8e-03
100	2.1e-04	4.6e-04	9.6e-04	2.9e-03
200	2.2e-04	5.1e-04	1.1e-03	3.0e-03
500	2.4e-04	6.5e-04	1.5e-03	4.0e-03

Table C.1: (Average time, in seconds, to compute one bootstrap sample from simulated data.) The rows correspond to the values of $n = 10, 50, 100, 200, 500$, and the columns correspond to the values of $p = 10, 50, 100, 200$.

Table C.2 displays analogous computing times for $\boldsymbol{\lambda}_k(\widehat{\Sigma}^\star) - \boldsymbol{\lambda}_k(\widehat{\Sigma})$, in seconds, based on the stock market data. Here, the computations were done on a single 3.5 GHz Dual-Core Intel Core i7 processor. The results show that there is little difference in computing time compared to the setting of synthetic data in Table C.1.

	p			
n	50	150	200	300
118	3.8e-04	1.4e-03	2.0e-03	4.0e-03

Table C.2: (Average time, in seconds, to compute one bootstrap sample from the stock market data.) The single row corresponds to the value of $n = 118$, and the columns correspond to the values of $p = 50, 150, 200, 300$.

S10 Additional discussion on Assumption 1(b)

Recall that Assumption 1(b) requires the condition $\min_{1 \le j \le k}(\lambda_j(\Sigma) - \lambda_{j+1}(\Sigma)) \gtrsim \lambda_1(\Sigma)$, which ensures that there are gaps between the leading eigenvalues $\lambda_1(\Sigma), \ldots, \lambda_{k+1}(\Sigma)$. To inspect whether or not this type of condition holds in practice, the paper (Hall et al., 2009) proposes a diagnostic method that constructs a preliminary set of *conservative* simultaneous confidence intervals for $\lambda_1(\Sigma), \ldots, \lambda_{k+1}(\Sigma)$. In exchange for their conservatism, these intervals have the property that they are not sensitive to the existence of

gaps. (See also (Lopes et al., 2023) for theoretical analysis related to this technique.) If none of the intervals overlap, then the user may conclude that the eigenvalues are adequately separated. However, if some of the intervals do overlap, then the paper (Hall et al., 2009) recommends that an adjusted form of bootstrapping be used to approximate the distribution of the statistic $\boldsymbol{\lambda}_k(\widehat{\Sigma}) - \boldsymbol{\lambda}_k(\Sigma)$.

S10.1 Sensitivity analysis

To study the sensitivity of the bootstrap to Assumption 1(b), we now discuss some numerical experiments with varying gaps between the leading population eigenvalues. The first three eigenvalues were specified as $(\lambda_1(\Sigma), \lambda_2(\Sigma), \lambda_3(\Sigma)) = (1 + g, 1, 1 - g)$ for a gap parameter $g \in \{0, 0.1, 0.2\}$, and the remaining eigenvalues were chosen to follow the polynomial decay profile $\lambda_j(\Sigma) = j^{-1}$ for $j \geq 4$. Next, we applied the bootstrap (as in Section 4.4) to construct simultaneous confidence intervals for $(\lambda_1(\Sigma), \ldots, \lambda_5(\Sigma))$ based on a nominal level of 95%. Figure C.13 contains a grid of plots in which the columns correspond to increasing values of the gap parameter g, and the rows correspond to the three transformation rules considered in Section 4.4. These plots are based on data generated from simulation model (i), and an analogous set of plots based on simulation model (ii) are given in Figure C.14.

As expected, Figures C.13 and C.14 show that the coverage accuracy of the bootstrap confidence intervals improves as the gap parameter increases. In the case when $g = 0$, the coverage accuracy is poor for all three transformation rules, even at large sample sizes. Next, when $g = 0.1$, the coverage is mostly accurate for sample sizes $n \geq 400$, but at smaller sample sizes the coverage generally falls below the desired level. Lastly, when $g = 0.2$, the coverage becomes more accurate when $n < 400$, especially when the square-root transformation is used.

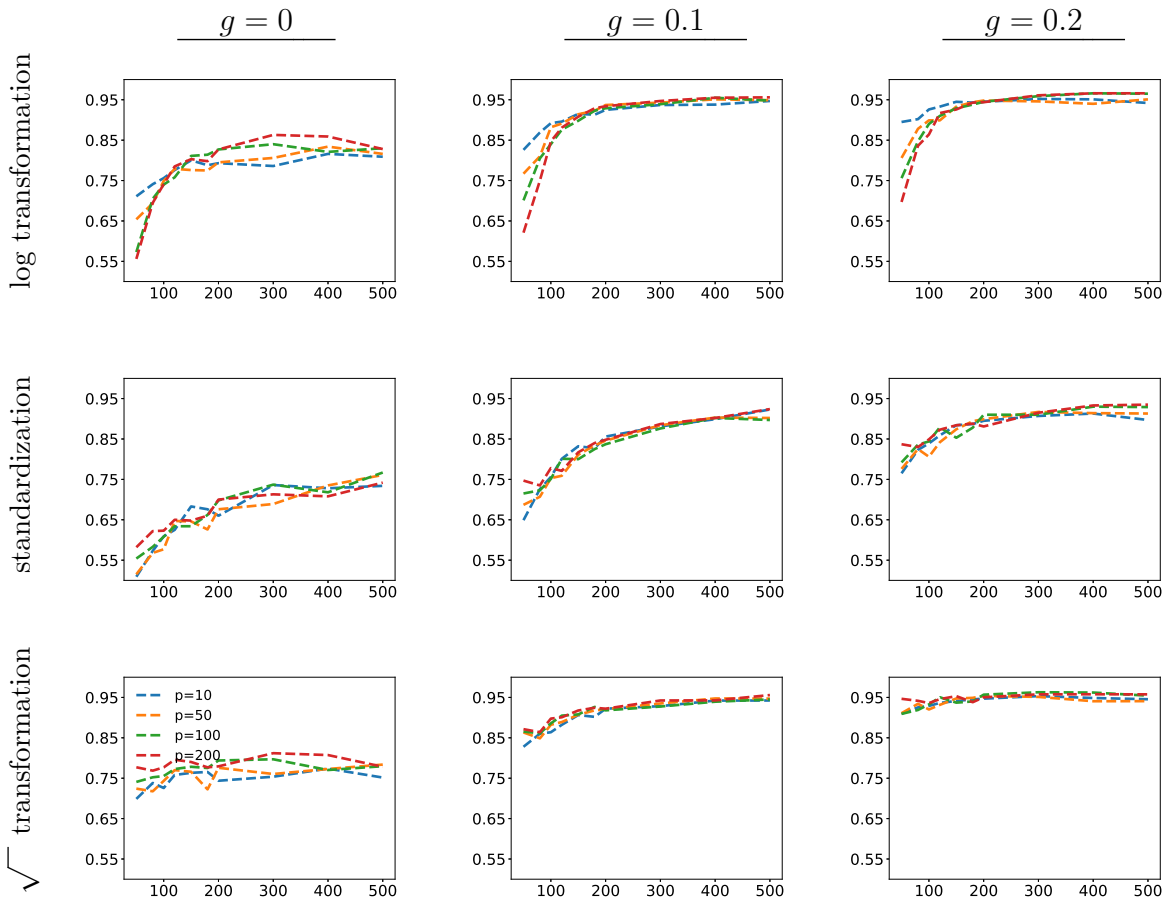


Figure C.13: (Simultaneous coverage probability versus n in simulation model (i) with the decay profile: $(\lambda_1)(\Sigma), \lambda_2(\Sigma), \lambda_3(\Sigma)) = (1+g, 1, 1-g)$ and $\lambda_j(\Sigma) = j^{-1}$ for $j \geq 4$). In each panel, the y-axis measures $\mathbb{P}(\cap_{j=1}^{5}\{\lambda_j(\Sigma) \in \widehat{\mathcal{I}}_j\})$ based on a nominal level of 95%, and the x-axis measures n. The colored curves correspond to the different values of p, indicated in the legend.

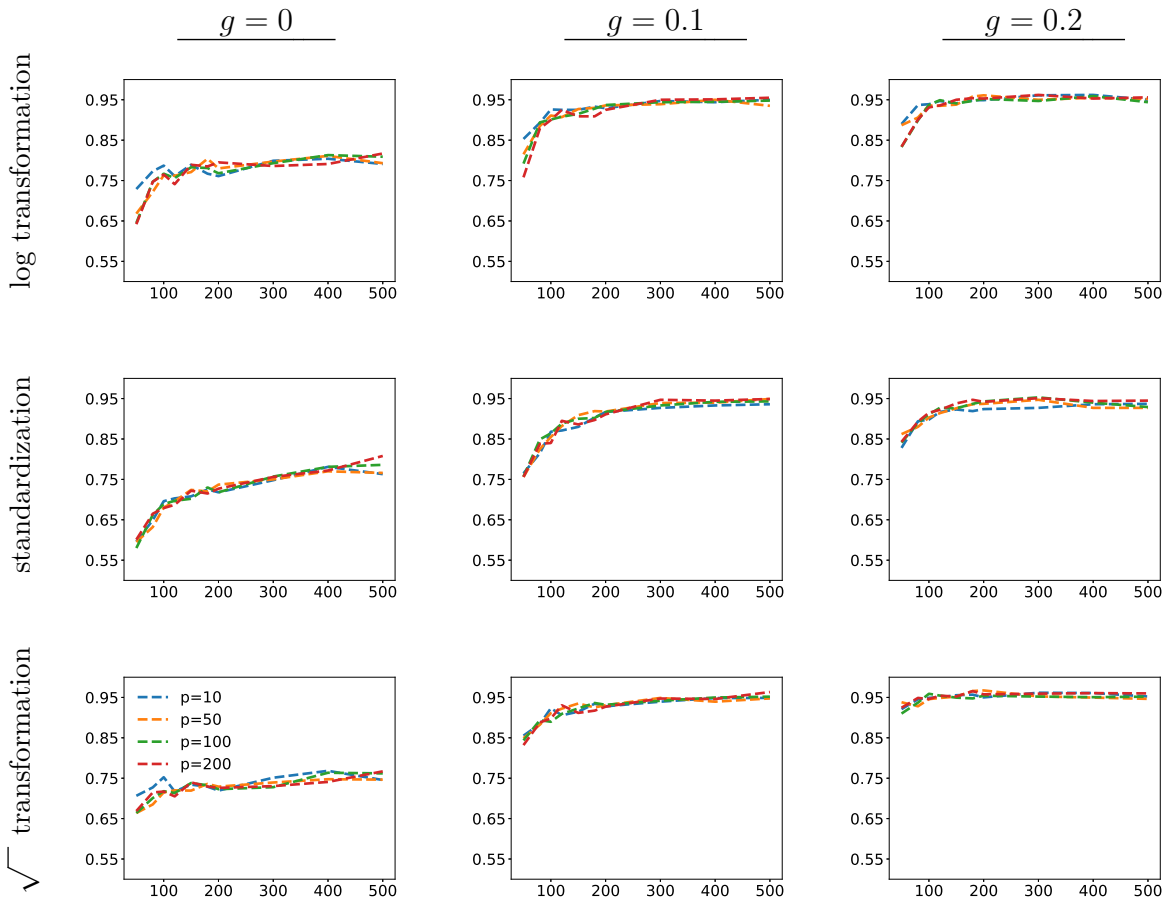


Figure C.14: (Simultaneous coverage probability versus n in simulation model (ii) with the decay profile: $(\lambda_1)(\Sigma), \lambda_2(\Sigma), \lambda_3(\Sigma)) = (1 + g, 1, 1 - g)$ and $\lambda_j(\Sigma) = j^{-1}$ for $j \geq 4$). The plotting scheme is the same as described in the caption of Figure C.13.

References

Abraham, C., P.-A. Cornillon, E. Matzner-Løber, and N. Molinari (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics 30*(3), 581–595.

Adams, R. A. and J. J. F. Fournier (2003). *Sobolev Spaces* (Second ed.), Volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley.

Araki, Y., S. Konishi, S. Kawano, and H. Matsui (2009). Functional logistic discrimination via regularized basis expansions. *Communications in Statistics—Theory and Methods 38*(16-17), 2944–2957.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*, 337–404.

Avery, M., Y. Wu, H. H. Zhang, and J. Zhang (2014). RKHS-based functional nonparametric regression for sparse and irregular longitudinal data. *Canadian Journal of Statistics 42*(2), 204–216.

Bai, Z. and H. Saranadasa (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 311–329.

Bai, Z. and J. W. Silverstein (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer.

Bali, J. L. and G. Boente (2009). Principal points and elliptical distributions from the multivariate setting to the functional case. *Statistics & Probability Letters 79*(17), 1858–1865.

Bali, J. L., G. Boente, D. E. Tyler, and J.-L. Wang (2011). Robust functional principal components: A projection-pursuit approach. *Annals of Statistics 39*(6), 2852–2882.

Bentkus, V. (2003). On the dependence of the Berry–Esseen bound on dimension. *Journal of Statistical Planning and Inference 113*(2), 385–402.

Beran, R. and M. S. Srivastava (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *Annals of Statistics 13*(1), 95–115.

Beran, R. and M. S. Srivastava (1987). Correction: Bootstrap tests and confidence regions for functions of a covariance matrix. *Annals of Statistics 15*(1), 470–471.

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association 45*(250), 164–180.

Berrendero, J. R., B. Bueno-Larraz, and A. Cuevas (2019). An RKHS model for variable selection in functional linear regression. *Journal of Multivariate Analysis 170*, 25–45.

Berrendero, J. R., A. Cuevas, and J. Torrecilla (2018). On the use of reproducing kernel Hilbert spaces in functional classification. *Journal of the American Statistical Association 113*(523), 1210–1218.

Besse, P. and J. O. Ramsay (1986). Principal components analysis of sampled functions. *Psychometrika 51*(2), 285–311.

Bickel, P. J. and D. A. Freedman (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics 9*(6), 1196–1217.

Boente, G., M. Salibián Barrera, and D. E. Tyler (2014). A characterization of elliptical distributions and some optimality properties of principal components for functional data. *Journal of Multivariate Analysis 131*, 254–264.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*, Volume 149. Springer Science & Business Media.

Brumback, B. A. and J. A. Rice (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association 93*(443), 961–976.

Bunea, F. and L. Xiao (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli 21*(2), 1200–1230.

Cai, T. T. and P. Hall (2006). Prediction in functional linear regression. *Annals of Statistics 34*(5), 2159–2179.

Cai, T. T. and M. Yuan (2010). Nonparametric covariance function estimation for functional and longitudinal data. Technical report.

Cai, T. T. and M. Yuan (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Annals of Statistics 39*(5), 2330–2355.

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

Cambanis, S., S. Huang, and G. Simons (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis 11*(3), 368–385.

Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics & Probability Letters 45*(1), 11–22.

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Statistica Sinica 13*(3), 571–591.

Cardot, H. and P. Sarda (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis 92*(1), 24–41.

Carey, J. R., P. Liedo, H.-G. Müller, J.-L. Wang, D. Senturk, and L. Harshman (2005). Biodemography of a long-lived tephritid: Reproduction and longevity in a large cohort of female mexican fruit flies, anastrepha ludens. *Experimental Gerontology 40*(10), 793–800.

Castro, M. H. d. and A. C. Piantella (2015). Eigenvalue decay of positive integral operators on compact two-point homogeneous spaces. Technical report.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*(5), 1281–1304.

Chang, C., Y. Chen, and R. T. Ogden (2014). Functional data classification: a wavelet approach. *Computational Statistics 29*(6), 1497–1513.

Chen, D., P. Hall, and H.-G. Müller (2011). Single and multiple index functional regression models with nonparametric link. *Annals of Statistics 39*(3), 1720–1747.

Chen, J. X. and M. Lopes (2020). Estimating the error of randomized Newton methods: A bootstrap approach. In *International Conference on Machine Learning*, pp. 1649–1659.

Chen, S., E. B. O'Dea, J. M. Drake, and B. I. Epureanu (2019). Eigenvalues of the covariance matrix as early warning signals for critical transitions in ecological systems. *Scientific Reports 9*(1), 1–14.

Chernick, M. R. (2011). *Bootstrap Methods: A Guide for Practitioners and Researchers*. Wiley.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Gaussian approximation of suprema of empirical processes. *Annals of Statistics 42*(4), 1564–1597.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017a). Central limit theorems and bootstrap in high dimensions. *Annals of Probability 45*(4), 2309–2352.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017b). Detailed proof of Nazarov's inequality. *arXiv:1711.10696*.

Chernozhukov, V., D. Chetverikov, K. Kato, and Y. Koike (2022). Improved central limit theorem and bootstrap approximations in high dimensions. *Annals of Statistics*.

Chiou, J.-M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction. *Annals of Applied Statistics 6*(4), 1588–1614.

Chiou, J.-M. and P.-L. Li (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 679–699.

Chiou, J.-M., H.-G. Müller, J.-L. Wang, and J. R. Carey (2003). A functional multiplicative effects model for longitudinal data, with application to reproductive histories of female medflies. *Statistica Sinica 13*(4), 1119.

Coffey, N., J. Hinde, and E. Holian (2014). Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics & Data Analysis 71*, 14–29.

Connor, G. and R. A. Korajczyk (1993). A test for the number of factors in an approximate factor model. *Journal of Finance 48*(4), 1263–1291.

Couillet, R. and M. Debbah (2011). *Random Matrix Methods for Wireless Communications*. Cambridge.

Crambes, C., A. Kneip, and P. Sarda (2009). Smoothing splines estimators for functional linear regression. *Annals of Statistics 37*(1), 35–72.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems 2*(4), 303–314.

D'Agostino, Ralph B., S., S. Grundy, L. M. Sullivan, P. Wilson, and for the CHD Risk Prediction Group (2001). Validation of the Framingham Coronary Heart Disease Prediction Scores: Results of a Multiple Ethnic Groups Investigation. *Journal of the American Medical Association 286*(2), 180–187.

Dai, X., H.-G. Müller, and W. Tao (2018). Derivative principal component analysis for representing the time dynamics of longitudinal and functional data. *Statistica Sinica 28*(3), 1583–1609.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge.

Delaigle, A. and P. Hall (2013). Classification using censored functional data. *Journal of the American Statistical Association 108*(504), 1269–1283.

Demmel, J., I. Dumitriu, and O. Holtz (2007). Fast linear algebra is stable. *Numerische Mathematik 108*(1), 59–91.

Diaconis, P. and B. Efron (1983). Computer-intensive methods in statistics. *Scientific American 248*(5), 116–131.

DiCiccio, T. J. (1984). On parameter transformations and interval estimation. *Biometrika 71*(3), 477–485.

DiCiccio, T. J. and B. Efron (1996). Bootstrap confidence intervals. *Statistical Science 11*(3), 189–228.

Dobriban, E. and S. Wager (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Annals of Statistics 46*(1), 247–279.

Dou, W. W., D. Pollard, and H. H. Zhou (2012). Estimation in functional regression for general exponential families. *Annals of Statistics 40*(5), 2421–2451.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields 95*(1), 125–140.

Eaton, M. L. and D. E. Tyler (1991). On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Annals of Statistics*, 260–271.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics 7*, 1–26.

El Karoui, N. and E. Purdom (2019). The non-parametric bootstrap and spectral analysis in moderate and high-dimension. In *International Conference on Artificial Intelligence and Statistics*, pp. 2115–2124.

Fabozzi, F. J., P. N. Kolm, D. A. Pachamanova, and S. M. Focardi (2007). *Robust Portfolio Optimization and Management*. Wiley.

Facer, M. R. and H.-G. Müller (2003). Nonparametric estimation of the location of a maximum in a response surface. *Journal of Multivariate Analysis 87*(1), 191–217.

Fan, J. and I. Gijbels (2018). *Local Polynomial Modelling and Its Applications*. Routledge.

Fan, J., Q. Sun, W.-X. Zhou, and Z. Zhu (2014). Principal component analysis for big data. *Wiley StatsRef: Statistics Reference Online*, 1–13.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *Annals of Statistics 27*(5), 1491–1518.

Ferraty, F. and P. Vieu (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis 44*(1-2), 161–173.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media.

Fisher, A., B. Caffo, B. Schwartz, and V. Zipunnikov (2016). Fast, exact bootstrap principal component analysis for $p > 1$ million. *Journal of the American Statistical Association 111*(514), 846–860.

Frahm, G. (2004). Generalized elliptical distributions: Theory and applications. *Ph.D. Thesis from the University of Köln, Germany*.

Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics 9*(6), 1218–1228.

Funahashi, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks 2*(3), 183–192.

Gajardo, Á., X. Dai, and H.-G. Müller (2021). Predictive distributions and the transition from sparse to dense functional data. *arXiv:2109.02236*.

Garcia-Escudero, L. A. and A. Gordaliza (2005). A proposal for robust curve clustering. *Journal of Classification 22*(2), 185–201.

Giacofci, M., S. Lambert-Lacroix, G. Marot, and F. Picard (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics 69*(1), 31–40.

Gu, M. and S. C. Eisenstat (1996). Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing 17*(4), 848–869.

Gunter, T., M. A. Osborne, R. Garnett, P. Hennig, and S. J. Roberts (2014). Sampling for inference in probabilistic models with fast bayesian quadrature. In *Advances in Neural Information Processing Systems*, Volume 27.

Guss, W. H. (2016). Deep function machines: Generalized neural networks for topological layer expression. *arXiv:1612.04799*.

Guss, W. H. and R. Salakhutdinov (2019). On universal approximation by neural networks with uniform guarantees on approximation of infinite dimensional maps. *arXiv:1910.01545*.

Ha, C. W. (1986). Eigenvalues of differentiable positive definite kernels. *SIAM Journal on Mathematical Analysis 17*(2), 415–419.

Halko, N., P.-G. Martinsson, and J. A. Tropp (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review 53*(2), 217–288.

Hall, P. (2013). *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media.

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics 35*(1), 70–91.

Hall, P., Y. K. Lee, B. U. Park, and D. Paul (2009). Tie-respecting bootstrap methods for estimating distributions of sets and functions of eigenvalues. *Bernoulli 15*(2), 380–401.

Hall, P., H.-G. Müller, and J.-L. Wang (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics 34*(3), 1493–1517.

Hall, P., D. S. Poskitt, and B. Presnell (2001). A functional data—analytic approach to signal discrimination. *Technometrics 43*(1), 1–9.

Han, F., S. Xu, and W.-X. Zhou (2018). On Gaussian comparison inequality and its application to spectral analysis of large random matrices. *Bernoulli 24*(3), 1787–1833.

Han, K., H.-G. Müller, and B. U. Park (2018). Smooth backfitting for additive modeling with small errors-in-variables, with an application to additive functional regression for multiple predictor functions. *Bernoulli 24*(2), 1233–1265.

Hardt, M., B. Recht, and Y. Singer (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on International Conference on Machine Learning*, Volume 48, pp. 1225–1234.

Hastie, T. and C. Mallows (1993). A statistical view of some chemometrics regression tools: Discussion. *Technometrics 35*(2), 140–143.

Heinzl, F. and G. Tutz (2014). Clustering in linear-mixed models with a group fused lasso penalty. *Biometrical Journal 56*(1), 44–68.

Hilgert, N., A. Mas, and N. Verzelen (2013). Minimax adaptive tests for the functional linear model. *Annals of Statistics 41*(2), 838–869.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks 4*(2), 251–257.

Hsing, T. and R. Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons.

Hsu, D., S. M. Kakade, and T. Zhang (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics 14*(3), 569–600.

Hu, J., W. Li, Z. Liu, and W. Zhou (2019). High-dimensional covariance matrices in elliptical distributions with application to spherical test. *Annals of Statistics 47*(1), 527–555.

Hu, Z., N. Wang, and R. J. Carroll (2004). Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika 91*(2), 251–262.

Ingersoll Jr., J. E. (1984). Some results in the theory of arbitrage pricing. *Journal of Finance 39*(4), 1021–1039.

Jacques, J. and C. Preda (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing 112*, 164–171.

Jacques, J. and C. Preda (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis 71*, 92–106.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 411–432.

James, G. M. and T. J. Hastie (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(3), 533–550.

James, G. M., J. Wang, J. Zhu, et al. (2009). Functional linear regression that's interpretable. *Annals of Statistics 37*(5A), 2083–2108.

Johnstone, I. M. and D. Paul (2018). PCA in high dimensions: An orientation. *Proceedings of the IEEE 106*(8), 1277–1292.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Jung, S., M. H. Lee, and J. Ahn (2018). On the number of principal components in high dimensions. *Biometrika 105*(2), 389–402.

Kayano, M., K. Dozono, and S. Konishi (2010). Functional cluster analysis via orthonormalized gaussian basis expansions and its application. *Journal of Classification 27*(2), 211–230.

Kokoszka, P. and M. Reimherr (2017). *Introduction to Functional Data Analysis*. CRC Press.

Koltchinskii, V., M. Löffler, and R. Nickl (2020). Efficient estimation of linear functionals of principal components. *Annals of Statistics 48*(1), 464–490.

Koltchinskii, V. and K. Lounici (2017). Normal approximation and concentration of spectral projectors of sample covariance. *Annals of Statistics 45*(1), 121–157.

Konishi, S. (1991). Normalizing transformations and bootstrap confidence intervals. *Annals of Statistics 19*(4), 2209 – 2225.

Laloux, L., P. Cizeau, M. Potters, and J.-P. Bouchaud (2000). Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance 3*(03), 391–397.

Ledoit, O. and M. Wolf (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics 40*(2), 1024–1060.

Ledoit, O. and M. Wolf (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis 139*, 360–384.

Lee, E. R. and B. U. Park (2012). Sparse estimation in functional linear regression. *Journal of Multivariate Analysis 105*, 1–17.

Lee, J. O. and K. Schnelli (2016). Tracy–Widom distribution for the largest eigenvalue of real sample covariance matrices with general population. *Annals of Applied Probability 26*(6), 3786–3839.

Lei, J. (2014). Adaptive global testing for functional linear models. *Journal of the American Statistical Association 109*(506), 624–634.

Leng, X. and H.-G. Müller (2006a). Classification using functional data analysis for temporal gene expression data. *Bioinformatics 22*(1), 68–76.

Leng, X. and H.-G. Müller (2006b). Time ordering of gene coexpression. *Biostatistics 7*(4), 569–584.

Li, H. and P. Ralph (2019). Local PCA shows how the effect of population structure differs along the genome. *Genetics 211*(1), 289–304.

Li, P.-L. and J.-M. Chiou (2011). Identifying cluster number for subspace projected functional data clustering. *Computational Statistics & Data Analysis 55*(6), 2090–2103.

Li, Y. and T. Hsing (2007). On rates of convergence in functional linear regression. *Journal of Multivariate Analysis 98*(9), 1782–1804.

Li, Y. and T. Hsing (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics 38*(6), 3321–3351.

Lin, M., Q. Chen, and S. Yan (2013). Network in network. *arXiv:1312.4400*.

Lin, Z., M. E. Lopes, and H.-G. Müller (2021). High-dimensional MANOVA via bootstrapping and its application to functional and sparse count data. *Journal of the American Statistical Association*, 1–15.

Lopes, M., N. B. Erichson, and M. Mahoney (2020). Error estimation for sketched SVD via the bootstrap. In *International Conference on Machine Learning*, pp. 6382–6392.

Lopes, M., S. Wang, and M. Mahoney (2018). Error estimation for randomized least-squares algorithms via the bootstrap. In *International Conference on Machine Learning*, pp. 3217–3226.

Lopes, M. E. (2022). Central limit theorem and bootstrap approximation in high dimensions with near $1/\sqrt{n}$ rates. *Annals of Statistics 50*(5), 2492–2513.

Lopes, M. E., A. Blandino, and A. Aue (2019). Bootstrapping spectral statistics in high dimensions. *Biometrika 106*(4), 781–801.

Lopes, M. E., N. B. Erichson, and M. W. Mahoney (2023). Bootstrapping the operator norm in high dimensions: Error estimation for covariance matrices and sketching. *29*(1), 428–450.

Lopes, M. E., Z. Lin, and H.-G. Müller (2020). Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional and multinomial data. *Annals of Statistics 48*(2), 1214–1229.

Lopes, M. E., S. Wang, and M. W. Mahoney (2019). A bootstrap method for error estimation in randomized matrix multiplication. *Journal of Machine Learning Research 20*(1), 1434–1473.

Lopes, M. E., S. Wu, and T. C. Lee (2020). Measuring the algorithmic convergence of randomized ensembles: The regression setting. *SIAM Journal on Mathematics of Data Science 2*(4), 921–943.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli 20*(3), 1029–1058.

Lunde, R., P. Sarkar, and R. Ward (2021). Bootstrapping the error of Oja's algorithm. In *Advances in Neural Information Processing Systems*, Volume 34, pp. 6240–6252.

Malfait, N. and J. O. Ramsay (2003). The historical functional linear model. *Canadian Journal of Statistics 31*(2), 115–128.

Matsui, H., T. Araki, and S. Konishi (2011). Multiclass functional discriminant analysis and its application to gesture recognition. *Journal of Classification 28*(2), 227–243.

Müller, H.-G. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics 32*(2), 223–240.

Müller, H.-G. (2016). Peter Hall, functional data analysis and random objects. *Annals of Statistics 44*(5), 1867–1887.

Müller, H.-G. and U. Stadtmüller (2005). Generalized functional linear models. *Annals of Statistics 33*(2), 774–805.

National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC) (2020). National health and nutrition examination survey data.

Naumov, A., V. Spokoiny, and V. Ulyanov (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probability Theory and Related Fields 174*(3-4), 1091–1132.

Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis*, pp. 169–187. Springer.

Nguyen, L. H. and S. Holmes (2019). Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology 15*(6), e1006907.

Nugent, C. (2017). S&P 500 Stock Data. Retrieved from Kaggle Datasets on 8/30/21.

Olive, D. J. (2017). *Robust Multivariate Analysis*. Springer.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 1617–1642.

Peng, J. and H.-G. Müller (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Annals of Applied Statistics 2*(3), 1056–1077.

Radchenko, P., X. Qiao, and G. M. James (2015). Index models for sparsely sampled functional data. *Journal of the American Statistical Association 110*(510), 824–836.

Ramsay, J. O. and C. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological) 53*(3), 539–561.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (Second ed.). Springer Series in Statistics. Springer, New York.

Ramsay, J. O. and B. W. Silverman (2007). *Applied Functional Data Analysis: Methods and Case Studies*. Springer.

Reade, J. B. (1983). Eigenvalues of positive definite kernels. *SIAM Journal on Mathematical Analysis 14*(1), 152–157.

Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological) 53*(1), 233–243.

Rice, J. A. and C. O. Wu (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics 57*(1), 253–259.

Rincón, M. and M. D. Ruiz-Medina (2012). Wavelet-RKHS-based functional statistical classification. *Advances in Data Analysis and Classification 6*(3), 201–217.

Roll, R. and S. A. Ross (1980). An empirical investigation of the arbitrage pricing theory. *Journal of Finance 35*(5), 1073–1103.

Rossi, F. and B. Conan-Guez (2005). Functional multi-layer perceptron: a non-linear tool for functional data analysis. *Neural Networks 18*(1), 45–60.

Rossi, F., B. Conan-Guez, and F. Fleuret (2002). Functional data analysis with multi layer perceptrons. In *International Joint Conference on Neural Networks*, Volume 3, pp. 2843–2848.

Rossi, F., N. Delannay, B. Conan-Guez, and M. Verleysen (2005). Representation of functional data in neural networks. *Neurocomputing 64*, 183–210.

Ruppert, D. (2014). *Statistics and Finance: An Introduction*. Springer.

Ruppert, D. and D. S. Matteson (2015). *Statistics and Data Analysis for Financial Engineering*. Springer.

Serban, N. and L. Wasserman (2005). CATS: Clustering after transformation and smoothing. *Journal of the American Statistical Association 100*(471), 990–999.

Shin, H. (2009). Partial functional linear regression. *Journal of Statistical Planning and Inference 139*(10), 3405–3418.

Shin, H. and S. Lee (2016). An RKHS approach to robust functional linear regression. *Statistica Sinica 26*(1), 255–272.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Annals of Statistics 24*(1), 1–24.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics*, 1187–1195.

Song, J. J., W. Deng, H.-J. Lee, and D. Kwon (2008). Optimal classification for time-course gene expression data using functional data analysis. *Computational Biology and Chemistry 32*(6), 426–432.

Spokoiny, V. and M. Zhilova (2015). Bootstrap confidence sets under model misspecification. *Annals of Statistics 43*(6), 2653–2675.

Stewart, T. A., C. Liang, J. L. Cotney, J. P. Noonan, T. J. Sanger, and G. P. Wagner (2019). Evidence against tetrapod-wide digit identities and for a limited frame shift in bird wings. *Nature Communications 10*(1), 1–13.

Stinchcombe, M. B. (1999). Neural network approximation of continuous functionals and continuous functions on compactifications. *Neural Networks 12*(3), 467–477.

Sun, X., P. Du, X. Wang, and P. Ma (2018). Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework. *Journal of the American Statistical Association 113*(524), 1601–1611.

Talagrand, M. (1989). Isoperimetry and integrability of the sum of independent Banach-space valued random variables. *Annals of Probability*, 1546–1570.

Terry, E. E., X. Zhang, C. Hoffmann, L. D. Hughes, S. A. Lewis, J. Li, M. J. Wallace, L. A. Riley, C. M. Douglas, and M. A. Gutierrez-Monreal (2018). Transcriptional profiling reveals extraordinary diversity among skeletal muscle tissues. *eLife 7*, e34613.

Tibshirani, R. (1988). Variance stabilization and the bootstrap. *Biometrika 75*(3), 433–444.

Tishby, N. and N. Zaslavsky (2015). Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*, pp. 1–5.

UK Power Networks (2015). Smartmeter energy consumption data in london households.

Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge.

Wagner, F. (2015). GO-PCA: An unsupervised method to explore gene expression data using prior knowledge. *PLOS One 10*(11), e0143196.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics (SIAM).

Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016). Functional data analysis. *Annual Review of Statistics and Its Application 3*, 257–295.

Wang, Q., Y. Lu, X. Zhang, and J. Hahn (2021). Region of interest selection for functional features. *Neurocomputing 422*, 235–244.

Wang, S., L. Qian, and R. J. Carroll (2010). Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika 97*(1), 79–93.

Wang, X., S. Ray, and B. K. Mallick (2007). Bayesian curve classification using wavelets. *Journal of the American Statistical Association 102*(479), 962–973.

Webb-Vargas, Y., S. Chen, A. Fisher, A. Mejia, Y. Xu, C. Crainiceanu, B. Caffo, and M. A. Lindquist (2017). Big data and neuroimaging. *Statistics in Biosciences 9*(2), 543–558.

Weyl, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen 71*(4), 441–479.

Wielandt, H. and R. R. Meyer (1967). *Topics in the Analytic Theory of Matrices*. Department of Mathematics, University of Wisconsin.

Yao, F., H.-G. Müller, and J.-L. Wang (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association 100*(470), 577–590.

Yao, F., H.-G. Müller, and J.-L. Wang (2005b). Functional linear regression analysis for longitudinal data. *Annals of Statistics 33*(6), 2873–2903.

Yao, J., J. Mueller, and J.-L. Wang (2021). Deep learning for functional data analysis with adaptive basis layers. In *International Conference on Machine Learning*, pp. 11898–11908.

Yao, J., S. Zheng, and Z. D. Bai (2015). *Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press.

Yuan, M. and T. T. Cai (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Annals of Statistics 38*(6), 3412–3444.

Zhang, X. and J.-L. Wang (2016). From sparse to dense functional data and beyond. *Annals of Statistics 44*(5), 2281–2321.

Zhou, H., D. Wei, and F. Yao (2022). Theory of functional principal components analysis for discretely observed data. *arXiv:2209.08768*.

Zhou, J., N.-Y. Wang, and N. Wang (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica 23*(1), 25.

Zhou, Y., S. Bhattacharjee, C. Carroll, Y. Chen, X. Dai, J. Fan, A. Gajardo, P. Z. Hadjipantelis, K. Han, H. Ji, C. Zhu, S.-C. Lin, P. Dubey, H.-G. Müller, and J.-L.

Wang (2022). *fdapace: Functional Data Analysis and Empirical Dynamics*. R package version 0.5.9.

Zhu, H., P. J. Brown, and J. S. Morris (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics 68*(4), 1260–1268.

Zhu, H., J. Fan, and L. Kong (2014). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association 109*(507), 1084–1098.

Zhu, H., M. Vannucci, and D. D. Cox (2010). A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics 66*(2), 463–473.